UNIVERSIDADE DE LISBOA
FACULDADE DE CIÊNCIAS
DEPARTAMENTO DE INFORMÁTICA

Identification of Subtypes of Autism Spectrum Disorder Patients using Machine Learning Methods

Ana Sofia Sousa Teixeira de Almeida

Mestrado em Bioinformática e Biologia Computacional

Dissertação orientada por:
Dr. Astrid Vicente
Dr. Hugo Martiniano

2023

Acknowledgements

As I finally reach the end of this chapter, I cannot help but express my gratitude to the people who, in some way, accompanied me during this journey. I knew from the beginning that it would not be an easy task, and indeed, what a rollercoaster it has been. It would not have been possible without your support.

First and foremost, I would like to thank my internal advisor, Dr. Astrid Vicente, who welcomed me and never let me forget how important and challenging research work is.

To my external advisor, Dr. Hugo Martiniano, and Joana Vilela, I am grateful for always being available and willing to discuss the existing problems, their different perspectives, and insights were crucial during the discussions. Thank you for your advice, suggestions, and encouragement.

I would also like to extend my gratitude to the entire team at the Department of Health Promotion and Non-Communicable Disease Prevention at INSA, for all their comments and suggestions on the work I developed and presented.

To my friends, I am thankful every day for the many lessons they have taught me and for always uplifting my mood.

Finally, a huge thank you to my family, especially my mother, for allowing me to navigate on this journey and constantly reminding me of my commitments.

Resumo

A Perturbação do Espectro do Autismo (PEA) é uma perturbação do neurodesenvolvimento caracterizada por alterações persistentes na comunicação e interação social e pela presença de padrões repetitivos e restritos de comportamentos, interesses ou atividades. O espectro clínico varia entre pacientes com um nível cognitivo normal ou acima da média, caracterizado por linguagem bem desenvolvida e, por vezes, apresentam um nível cognitivo acima da média, mas com dificuldades na interação social; e pacientes com quadros clínicos muito graves, que requerem assistência em tarefas básicas, apresentando um défice intelectual e pouca ou nenhuma linguagem.

A vasta heterogeneidade de sintomas nesta perturbação é reconhecida como uma característica importante da PEA. A investigação, o diagnóstico e o tratamento da PEA é dificultado devido à variedade e complexidade da sintomatologia clínica. A origem desta variabilidade clínica não é conhecida, sendo consistente com a ausência de biomarcadores diagnósticos. A PEA é diagnosticada através da sintomatologia sendo um processo complexo e demorado. O sistema de diagnóstico atual carece de uma abordagem baseada em evidências, sendo também necessária uma abordagem científica de modo a determinar quais as intervenções com maior probabilidade de serem eficazes para cada paciente diagnosticado com PEA.

O principal objetivo desta dissertação foi a identificação e caracterização de subgrupos de pacientes diagnosticados com PEA. O conjunto de dados analisado foi obtido no Hospital Pediátrico – Centro Hospitalar e Universitário de Coimbra e inclui informações clínicas de 949 crianças diagnosticadas com PEA entre 1999 e 2021. O diagnóstico foi baseado nos critérios do *Diagnostic and Statistical Manual of Mental Disorders* (DSM-IV e DSM-V), sendo a avaliação comportamental feita utilizando dois instrumentos de diagnóstico, o *Autism Diagnostic Interview - Revised* (ADI-R) e o *Autism Diagnostic Observation Schedule*(ADOS).

Foram examinadas um total de 307 variáveis contendo diferentes informações, tendo sido selecionadas apenas 11 variáveis para a construção de clusters, que evidenciam alterações no desenvolvimento das crianças, como a comunicação e capacidade de interação social. Deste modo, dos quatro domínios da ferramenta de diagnóstico ADI-R, três foram incluídos: Interações sociais recíprocas, Comportamento Restrito e Repetitivo, e Anormalias do Desenvolvimento Evidente Antes ou Depois dos 36 meses. Foram também selecionadas medidas dimensionais da sintomatologia da PEA, como os domínios Socialização, Comunicação e Vida Quotidiana da Vineland Adaptive Behavior Scales (VABS) (1ª e 2ª Edição) e as subescalas de Locomoção, Pessoal/Social, Audição e Fala, Coordenação Visual e Manual e Desempenho da Griffiths Mental Development Scales (GMDS) (Edição de 1984). Adicionalmente, de modo a investigar a natureza dos clusters formados, foram incluídas na análise dezoito variáveis: os Níveis de Gravidade segundo o ADOS, a cotação do ADI-R, Dimorfismos, Regressão de Linguagem, Atraso e Regressão do Desenvolvimento Psicomotor, Sexo, Audição, Visão, Condição Verbal, Perímetro Cefálico,

Apgar 3º e 5º, bem como o Diagnóstico e as idades em que a criança deu os Primeiros Passos, e disse as Primeiras Palavras e construiu as Primeiras Frases.

Todas as variáveis foram sujeitas a limpeza e processamento, através de um processo conduzido unicamente em Python (Version 3.8.12), enquanto a análise estatística foi implementada em Python e no software estatístico R (Versão 4.2.1). No final, foram selecionadas um total de 661 amostras clínicas, as quais foram examinadas para valores em falta. De modo a minimizar o viés de imputação dos valores em falta, as amostras com falta de dados em dois grupos inteiros de variáveis de cluster (ou seja, ADI-R, VABS e GMDS) foram excluídas da análise posterior. As variáveis descritivas dos clusters também sofreram alterações de modo a dicotomizar os valores das variáveis categóricas. Em relação às variáveis numéricas, foram investigados casos de *outliers*, e sempre que encontrados foram convertidos em valores em falta.

De forma a termos uma visão geral das características dos indivíduos na amostra, foram feitas estatísticas descritivas para todas as medidas incluídas. 586 crianças eram do sexo masculino (88,65%) e 75 do sexo feminino (11,35%). Os participantes exibiram uma ampla gama de sintomas e características, com variações significativas em todas as variáveis. Compararam-se também as características dos indivíduos incluídos no estudo com os indivíduos excluídos. Os indivíduos incluídos tinham características semelhantes em termos de domínio de Comunicação da VABS, o domínio de Anomalia do Desenvolvimento Evidente Antes ou Depois dos 36 meses do ADI-R e a maioria das subescalas da GMDS. No entanto, os indivíduos excluídos foram mais vezes classificados com sendo "Não Autistas" e obtiveram pontuações mais baixas nos domínios Social e Comportamento Restrito e Repetitivo do ADI-R e também pontuações mais baixas na subescala de Locomoção da GMDS. Por outro lado, obtiveram pontuações mais altas nas subescalas de Socialização e Competências da Vida Quotidiana da VABS.

A análise de correlações revelou que a maioria das variáveis não se encontram significativamente correlacionadas, exceto as fortes correlações encontrada entre os valores das subescalas/domínios dentro de cada grupo de variáveis. O domínio Socialização do ADI-R apresentou correlações inversas com as subescalas Audição e Fala da Griffiths, bem como com as subescalas Pessoal/Social.

Um total de 8 algoritmos de clusters foram avaliados usando dois métodos diferentes, de modo a aferir a validade interna e estabilidade dos clusters formados. No entanto, os resultados foram inconsistentes, com diferentes medidas de avaliação indicando diferentes métodos de cluster e diferentes números ótimos de partição dos mesmos. Adicionalmente, estes métodos de cluster não têm em consideração as correlações entre os diversos domínios/subescalas, o que poderia impactar os resultados.

Deste modo, a Análise de Componentes Principais (PCA) foi usada para reduzir as dimensões das variáveis e identificar os componentes que explicam a maior parte da variância dos dados. Foram identificados cinco Componentes Principais (CP), que explicam 89,41% da variação da amostra. A primeira CP explicou 60,37% da variância e está altamente correlacionado com os domínios/subescalas da VABS e GMDS, assim como o domínio da Socialização do ADI-R. A segunda CP foi responsável por 10,87% da variância e capturou os domínios do ADI-R. A terceira, quarta e quinta CP explicaram 7,87%, 5,63% e 4,66% da variância, respetivamente, e correlacionaram-se com domínios e subescalas específicos.

Acknowledgements

O Clustering Hierárquico solucionou um total de três clusters, com o Cluster 1 contendo 29% das amostras, o Cluster 2 contendo 32% e o Cluster 3, 39%. A solução de três clusters foi validada através de bootstrap. Os clusters foram comparados em termos de características de desenvolvimento e comportamento. Foram encontradas diferenças entre os três clusters para a maioria das variáveis. O Cluster 1 apresentou níveis mais baixos de capacidades intelectuais e adaptativas, juntamente com uma maior gravidade de sintomas a nível social, repetição de comportamentos e anormalidades de desenvolvimento. O Cluster 2 apresentou níveis semelhantes ao Cluster 1 relativamente às anomalias no desenvolvi- mento, mas níveis de gravidade mais altos no comportamento social, comunicação e adaptação do que o Cluster 3. Por fim, o Cluster 3 apresentou os valores mais altos nas capacidades de linguagem e adaptação e apresentou também a menor gravidade nos sintomas sociais e de desenvolvimento comparativamente aos restantes grupos, indicando ser deste modo o grupo com os menos défices.

O presente estudo fornece informações importantes sobre a compreensão da heterogeneidade da PEA e realça a necessidade de intervenções personalizadas com base em perfis individuais. Os resultados sugerem que ter em consideração diversos domínios comportamentais e de desenvolvimento, bem como sintomas básicos de autismo, pode ajudar a diferenciar subgrupos de crianças com PEA e fornecer perfis de desenvolvimento mais abrangentes.

Palavras chave: Perturbação do espectro do autismo, diagnostico, analise estatística, aprendizagem automática, clusters

Abstract

Autism Spectrum Disorder (ASD) is a neurodevelopmental disorder affecting brain structure and neuronal connectivity, characterized by a shortage of social interaction and communication, as well as repetitive patterns of behavior and restricted interests. Highly heterogeneous clinical features pose great challenges for ASD diagnosis, such that children who receive a diagnosis of ASD have a range of vastly different presentations, trajectories, and outcomes.

Identifying ASD subgroups can be helpful for researchers and clinicians to gain insights into distinct characteristics and patterns within these groups, as well as identify specific factors that may influence long-term outcomes. Moreover, it can also contribute to advancing scientific knowledge about ASD. It allows researchers to explore the underlying mechanisms, genetic factors, environmental influences, and brain processes that may be specific to each subgroup. This deeper understanding can lead to more targeted research efforts, improved diagnostic tools, and the development of innovative therapies for different subgroups.

This study used Hierarchical Clustering on Principal Component (HCPC) to analyze a sample of 661 children, aged 1-8 years, diagnosed with ASD following DSM-IV or DSM-V criteria and reaching the thresholds for ASD from the Autism Diagnostic Interview–Revised and the Autism Diagnostic Observation Schedule. In total, 11 variables were selected for cluster analysis, which included, apart from the diagnostic/screening ones, measures that can capture variations in children's development, such as communication and social abilities assessed through Vineland Adaptive Behavior Scales and The Griffiths Mental Development Scales.

Our analysis identified three distinct subgroups based on multiple developmental and behavioral domains. Cluster 1 exhibited lower levels of intellectual and adaptive abilities, accompanied by more severe social symptoms, repetitive behaviors, and developmental abnormalities. In comparison, Cluster 2 displayed similar levels of developmental abnormalities as Cluster 1, but demonstrated higher severity in social interactions, communication, and adaptive behavior than Cluster 3. On the other hand, Cluster 3 showcased the highest scores in language and adaptive abilities, and presented the lowest severity across social and developmental symptoms among all three clusters, indicating the least impairments. These findings emphasize the importance of considering multiple developmental and behavioral domains, as well as core symptoms of autism, in order to distinguish subgroups of young children with ASD and provide more comprehensive developmental profiles.

Keywords: Autism spectrum disorders, diagnosis, statistic analysis, machine learning, clusters

Index

List of Figures

List of Tables

List of Abbreviations

AbDev Abnormality of Development Evident At or Before 36 months

ADHD Attention Deficit Hyperactivity Disorder

ADI-R Autism Diagnostic Interview-Revised

ADOS Autism Diagnostic Observation Schedule

ASD Autism Spectrum Disorder

BAP Broader Autism Phenotype

CNVs Copy Number Variants

DSM Diagnostic and Statistical Manual of Mental Disorders

DSM-II Diagnostic and Statistical Manual of Mental Disorders 2^{nd} Edition

DSM-III Diagnostic and Statistical Manual of Mental Disorders 3^{rd} Edition

DSM-IV Diagnostic and Statistical Manual of Mental Disorders 4^{th} Edition

DSM-V Diagnostic and Statistical Manual of Mental Disorders 5^{th} Edition

FA Factor Analysis

FE Feature Extraction

FRAXA Fragile X syndrome

FRAXE Fragile XE syndrome

FS Feature Selection

FMM Factor Mixture Modeling

GAI General Ability Index

GWAS Genome-Wide Association Studies

GMDS Griffiths Mental Development Scales

GMDS-ER Griffiths Mental Development Scales-Extended Revised

GMDS-III Griffiths Mental Development Scales 3rd Edition

GQ General Developmental Quotient

HCPC Hierarchical Clustering on Principal Component

HP-CHUC Centro Hospitalar e Universitário de Coimbra

IQ Intelligence Quotient

LCA Latent Class Analysis

LDA Linear Discriminant Analysis

ML Machine Learning

NGS Next-generation sequencing

PC Principal Component

PCA Principal Component Analysis

PDD-NOS Pervasive Developmental Disorder Not Otherwise Specified

RRB Restricted and Repetitive Behavior

SNVs Single Nucleotide Variants

SVD Singular Value Decomposition

VABS Vineland Adaptive Behavior Scale

VABS-II Vineland Adaptive Behavior Scale 2^{nd} Edition

VABS-III Vineland Adaptive Behavior Scale 3^{rd} Edition

WAIS Wechsler Adult Intelligence Scale

WAIS-R Wechsler Adult Intelligence Scale Revised

WAIS-III Wechsler Adult Intelligence Scale 3^{rd} Edition

WAIS-IV Wechsler Adult Intelligence Scale 4^{th} Edition

WES Whole-Exome Sequencing

WGS Whole-Genome Sequencing

Chapter 1

Introduction

1.1 Historical Background

Autism spectrum disorder ASD is a common neurodevelopmental disorder affecting brain structure and neuronal connectivity, characterized by a shortage of social interaction and communication, as well as repetitive patterns of behavior and restricted interests. There have been numerous changes in our understanding of autism over the last 60 years since it was first discovered. In 1943, Leo Kanner defined autism as a syndrome of childhood onset in which he reported repetitive and ritualistic activities, as well as language delay and social isolation in children with intellectual disabilities [1]. Well before that time, during the early eighteenth century, notable individuals such as John Haslam in the United Kingdom and Jean Itard in France documented cases of classic autism [2, 3]. Eugen Bleuler coined the term "autism" by combining the Greek word "autos," meaning "self," with the intention of describing the self-focused cognition associated with schizophrenia [4, 5]. The distinction between autism and schizophrenia was controversial until the 1960s. However, Kanner considered early infantile autism a distinct disease entity and thought it to be a very rare disorder that would be easy to identify and diagnose. The concept of ASD was not introduced until the early 1980s by Wing and Gillberg.[6, 7, 8, 9]. Wing introduced the concept of a "triad of impairments," which encompassed deficits in social communication, both verbal and nonverbal, as well as a lack of symbolic play in imagination, forming the foundation of all ASD. She also designed the term Asperger's syndrome to refer to the form of "high functioning" autism spectrum disorder that Hans Asperger first identified, around the same time that Kanner identified a "low-functioning" variant of autism [10]. The Diagnostic and Statistical Manual of Mental Disorders Diagnostic and Statistical Manual of Mental Disorders (DSM) has undergone numerous revisions throughout the years, changing both the definition and the criteria for diagnosing autism. The second edition of the DSM, Diagnostic and Statistical Manual of Mental Disorders 2^{nd} Edition (DSM-II), was released in 1968 and classified autism as a mental illness, a form of childhood schizophrenia characterized by a detachment from reality [11]. It was not until 1980, in the Diagnostic and Statistical Manual of Mental Disorders 3^{rd} Edition (DSM-III), that autism was established as a distinct diagnosis and referred to it as a "pervasive developmental illness" apart from schizophrenia. Unlike earlier versions of the manual, which relied heavily on clinician observations and interpretations, the DSM-III introduced specific criteria that were essential for diagnosing autism spectrum disorder. It established three fundamental characteristics of autism, all of which emerged within the first 30 months of life: a notable absence of interest in social interaction, significant speech impairments, and peculiar reactions to the environment. [12]. However, in 1987, the DSM-III underwent a revision that significantly altered the criteria for autism. The concept of autism was broadened with the addition of pervasive developmental disorder not otherwise

specified Pervasive Developmental Disorder Not Otherwise Specified (PDD-NOS) for "mild autism". Furthermore, the 30-month requirement was also dropped [13]. Although the term "spectrum" was not used in the manual, the definition of autism started to become one more and more. This revision reflected that autism researchers were beginning to realize that autism is a spectrum of conditions rather than a single condition. In the revised manual, 16 specific criteria were outlined, encompassing the three previously established domains. To receive a diagnosis, a minimum of 8 criteria needed to be met. The inclusion of PDD-NOS provided an opportunity for physicians to identify children who didn't fully meet the criteria for autism but still required developmental or behavioral support and intervention. The first edition to officially classify autism as a spectrum disorder was the Diagnostic and Statistical Manual of Mental Disorders 4^{th} Edition (DSM-IV), published in 1994 and updated in 2000. Within this particular edition, the manual delineated five distinct conditions, each with its specified features. In addition to autism and PDD-NOS, it encompassed Asperger's disorder, which falls within the milder range of the spectrum. It also accounted for "childhood disintegrative disorder," characterized by significant developmental reversals and regressions, as well as Rett syndrome, primarily affecting movement and communication, and commonly observed in girls. [14]. The categorization supported the then-current theory of genetic underpinning of autism, and each category would ultimately be connected to a distinct set of problems and treatments. Researchers sought to identify genes linked to autism during the 1990s. Numerous researchers attempted to identify a list of "autism genes" after the Human Genome Project's completion in 2003. Many were discovered however, none were specifically linked to autism. It became apparent that it would be very hard to find a genetic ground and a corresponding treatment for the five conditions included in the DSM-IV. For this reason, medical experts decided to classify autism as an umbrella diagnosis with a spectrum ranging from mild to severe. Around this time, the autism diagnosis was not approached consistently by clinicians, being conceivable to have two distinct diagnoses of Asperger syndrome or PDD-NOS for the same condition. To prevent the lack of consistency, the term 'autism spectrum disorder' was introduced in Diagnostic and Statistical Manual of Mental Disorders 5^{th} Edition (DSM-V), replacing the different subcategories that were previously used – autistic disorder, Asperger's disorder, and pervasive developmental disorder – not otherwise specified PDD-NOS [15]. The DSM-V defines autism spectrum disorder as "persistent difficulties with social communication and social interaction" and "restricted and repetitive patterns of behaviors, activities or interests" (this includes sensory behavior), present since early childhood, to the extent that these "limit and impair everyday functioning". Each group consists of particular behaviors, some of which clinicians are required to recognize, as DSM-V now includes a condition called 'social communication disorder', separate from 'autism spectrum disorder'. This diagnosis is given when someone exhibits social interaction and social communication difficulties but does not show restricted, repetitive patterns of behavior, interests, or activities.

1.2 Clinical Aspects

Even though individuals with ASD vary greatly from one another, currently this single-spectrum disorder is based on two domains: social communication and restricted, repetitive, or unusual sensory-motor behaviors [16, 17]. Additionally, DSM-V explicitly acknowledges that ASD can coexist with other diseases, such as mental conditions (eg, attention-deficit hyperactivity disorder Attention Deficit Hyperactivity Disorder (ADHD)) and genetic disorders like fragile X syndrome. Although these symptoms can appear in several psychiatric diseases, it is the clustering of these symptoms in the same person that makes the disorder so intriguing. The broad range of features in individual children is one of the most

difficult challenges of detecting ASD. The autism spectrum comprises an extremely variable phenotype with uncertain endpoints, particularly at the mild end of the spectrum. Different symptoms appear at different times throughout the life span and the degree of each of the basic impairments varies greatly among children with ASD [15]. ASD signs arise in children as early as the age of three, although they may not fully manifest until they are of school age or even later [18, 19, 20]. Although social deficiencies may appear sooner and be more specific, they might be subtle and are less frequently identified by parents [21]. According to more recent studies, the average age a child receives an ASD diagnosis is between four and five years [22, 23]. An ASD diagnosis is important to access early intervention. In ASD, early intervention is associated with reduced cognitive, behavioral, and functional deficits [24]. According to research, early intervention improves adaptive functioning and Intelligence Quotient (IQ) in ASD patients while reducing challenging behaviors [25, 26, 27]. A delayed diagnosis may lead to inadequate treatments [28], ineffective school placement, and restricted access to services and activities that are beneficial for children with ASD [22], whereas early ASD diagnosis can speed early intervention and effective outcomes, emphasizing the need for early ASD diagnosis [29]. There are several factors that may influence the age at which a child obtains a diagnosis. First, individuals who have a diagnosed sibling are more likely to get a diagnosis sooner than those who are the first in the family to seek one. Secondly, there are systematic delays in diagnosis for children from minority racial/ethnic groups, with poor socioeconomic status, and/or with milder severity of ASD symptoms [30, 31, 32]. Additionally, it has been noted that people in rural areas are more likely to get a diagnosis later than those in urban areas [19]. One of the early indications of ASD is inadequate joint attention, which includes an absence of pointing, eye gaze shifting, and facial expression exhibition [33]. Due to difficulties combining nonverbal and verbal communication, body language is also impacted, which can make it appear odd or excessive. Children with ASDs have universal deficiencies in social relatedness, which is defined as the intrinsic need to connect with others and exchange similar emotional experiences, they are content to be alone and disregard their parent's pleas for attention, and rarely make eye contact or use gestures or vocalizations to attract others attention [34]. In later years they may fail to develop, maintain, and understand peer relationships because they have difficulties sharing the emotional state of others in cooperative and group situations as well as a lack of exhibiting and sharing of interests [15]. Social communication impairments are marked by delays in language development without nonverbal compensation efforts such as gestures. However, children with milder symptoms, particularly those with normal cognitive abilities, may be able to communicate. Their speech may not be effective or fluent, and they may lack communicative intent [15]. Moreover, they may present difficulties in starting and maintaining conversations, repetitive and stereotyped language, and a lack of creativity and imitative play. One of ASD's core symptoms is an insistence on sameness and inflexible adherence [35]. A restricted repertoire of interests, behaviors, and activities are manifested by abnormal over-focus on certain themes, adherence to non-functional routines and rituals, repetitive, stereotyped motor mannerisms, and concern with object parts rather than the whole [15]. Apart from the pattern of symptoms and the age at which they first appear, another factor to be considered is whether the child has an intellectual disability or other psychiatric illnesses.

1.3 Comorbidities

Comorbidity is defined as the co-existence of one or more secondary diseases or disorders with a primary disease or disorder [36]. A comorbid condition is a 2^{nd} order diagnosis that has core symptoms that are distinct from the primary disorder. It is important to be aware of potential co-occurring conditions in children with ASD in order to correctly diagnose one disorder as primary and another as secondary. A

study of over 2000 children with ASD found that 83% had an additional developmental diagnosis, 10% had at least one psychiatric diagnosis, and 16% had at least one neurologic diagnosis [37]. Previously, rates of comorbid intellectual disability (ID) in ASD patients ranged from 50% to 70%, with most current estimate by The Centers for Disease Control and Prevention being around 33%, with ID defined as an IQ $\leqslant$ 70 [38, 39]. Other common co-occurring medical issues include gastrointestinal diseases, such as dietary limitations and food selectivity, sleep disorders, obesity, and seizures [40, 41, 42, 43]. Anxiety, attention deficit, hyperactivity disorder, obsessive-compulsive disorder, and mood disorders or other disruptive behavior disorders are some of the additional behavioral or psychiatric co-occurring conditions in ASD [40]. Comorbidity symptoms may overlap with and conceal ASD symptoms, leading to a considerable delay in an ASD diagnosis [44]. Healthcare professionals can greatly benefit from an accurate diagnosis of comorbidities with ASD by giving children the best treatment regimens possible while reducing potential side effects [45, 46]. Additionally, the awareness of disorders co-occurring with ASD is extremely important in order to prevent the occurrence of diagnostic overshadowing biases [47]. This phenomenon happens when a clinician misdiagnosis a patient's co-occurring disorder by attributing their symptoms to the primary disorder instead of recognizing the co-occurring disorder [48]. However, due to several reasons, including communication disorders, the ambiguity of symptoms, how they differ from those in the general population, or how they evolve over time, it is not always simple to recognize comorbid conditions in children with ASD. The assumption that abnormal behaviors and symptoms are "simply part of ASD" is another aspect that exacerbates these issues. Another crucial limitation is the absence of screening diagnostic methods for these disorders [49]. Many of the behaviors and symptoms that are frequently linked to ASD may really be the result of other biological disorders. For example, aggression and self-injurious behavior may also be related to the presence of pain and the child's inability to communicate about his/her condition [50]. Food refusal may be related to the high food selectivity seen in autistic children, but it may also indicate the presence of a food allergy or intolerance, or it may have a more localized cause, such as dental problems.

1.4 Prevalence

ASD is one of the most common neurodevelopmental disorders, affecting people of all races, ethnicities, and socioeconomic backgrounds. In Western countries, the diagnosis rate of ASD has increased from 4.5 per 10,000 children in 1943 [1] to 110 per 10,000 children in 2009 [51]. Over the last three decades, continuous epidemiological surveys have been conducted around the world to determine the prevalence of ASD. ASD cases have been increasing across the globe, according to regular surveillance [52, 53, 54, 55, 56, 57];. More recently, Zeidan et al. reported an ASD median prevalence of 65 per 10,000 [58]. Studies conducted in Europe have reported different estimates for the prevalence of ASD. Based on register-based studies, the estimated pooled prevalence in Europe is 0.8%. Meanwhile, population-based studies have reported a higher pooled prevalence of 1.4%.[59]. According to several reports, the incidence of ASD is substantially higher in males as compared to females. However, the male-to-female ratios in ASD are quite variable [60, 61]. This heterogeneity is currently understudied, making it difficult to understand. When compared to girls, males are four times more likely to be diagnosed with ASD [38, 62, 63]. Because ASDs might go unnoticed, misdiagnosed, or diagnosed at a later stage in females, there is insufficient evidence of female patients suffering from ASD [64, 65]. According to studies, there may be a discrepancy in prevalence because women disguise their clinical signs [63]. Recent research supports a 'female protective effect' as an explanation for the higher frequency of ASD in males than in females [66]. According to a 2000 survey of the Portuguese population, the preva-

lence of ASD is 1:1000 children in mainland Portugal and 1.6:1000 children in the Azores. A higher male-to-female ratio (about 2.9:1) was also noted in the Portuguese population. There were significant geographical disparities in prevalence at the time, with greater estimates in the Center, Lisbon and Vale do Tejo, and Azores [67]. More recently, as part of the Autism Spectrum Disorders in the European Union (ASDEU) project, the prevalence of ASD was analyzed in children from the center region of Portugal. The researchers reported a prevalence of ASD of 0.5%, which is higher than the 0.125% observed in the central region in the year 2000 [68]. Correct estimates of the prevalence of ASD are essential for every country to assess the economic and healthcare burdens and devote appropriate finances and resources to children with ASD and their families. Even though researchers have discovered a significant increase in the incidence of ASD worldwide, there is still a lot of variability in the prevalence rate across developed and developing countries. Most of the time, the rate of ASD is substantially increasing in developed countries [69]. There is a great disparity in service delivery throughout the world with low-resource countries having more problems than high-resource countries [70]. As a result, low- and middle-income regions of developed countries have not seen as much of a rise in the rate of ASD as in high-resource communities and countries [69]. Most people with ASD living in low- and moderate-income countries are not correctly identified due to a lack of knowledge, awareness, and socioeconomic differences. As a result, global prevalence studies did not include these unidentifiable people. Because of this inconsistency, experts have been debating whether this is a true increase. Researchers have proposed several explanations for the rise in prevalence over time, including changed diagnostic criteria [71, 72], improved diagnostic equipment [73], a broader definition of ASD, the use of various study methodologies, cultural variations, and increased knowledge and recognition of ASD [72, 74].

1.5 Diagnostic tools

1.5.1 Autism Diagnostic Interview-Revised (ADI-R)

Autism Diagnostic Interview-Revised (ADI-R) comprises a standardized, semi-structured clinical interview for caregivers. The interview consists of 93 questions and focuses on three content areas or domains: social interaction quality; communication and language; and repetitive, restricted, and stereotyped interests and behavior. The test is carried out by an experienced clinician in the presence of a parent or caregiver who is familiar with the developmental history and behavior of the individual being evaluated. The clinician scores the responses based on the caregiver's description of the individual under evaluation. After that, the interview generates scores in each of the three areas, as previously mentioned. The clinician assigns a number between 0 and 3 to each item. When "behavior of the type specified in the coding is not present," a score of 0 is assigned, and a score of 3 is assigned when the specified behavior is of "extreme severity." When scores in all three content areas of communication, social interaction, and behavioral patterns meet or surpass the stipulated cut-offs, a diagnosis of ASD is given. The entire test administration and scoring process normally take 2 to 3 hours. Individuals with a mental age of more than 18 months can use the ADIR instrument [75].

1.5.2 Autism Diagnostic Observation Schedule (ADOS)

The Autism Diagnostic Observation Schedule (ADOS) is a semi-structured evaluation that monitors communication, social interaction, and imaginative play-related behavioral aspects in individuals suspected of having ASD [76]. ADOS is administered by qualified clinicians and involves a series of organized and semi-structured tasks with the referred person in a controlled environment. Since the 1980s,

when it was first introduced, ADOS has increased the range of ages that can be assessed. The ADOS was initially designed to assess children between the ages of 5 and 12 who were experiencing delays in their expressive language development. The ADOS was modified in 2000, extending its application to both adults and younger children [77]. In 2012 was released the most updated version, the ADOS-II, includes a toddler module for assessing children as young as 12 months [78]. The ADOS test consists of four modules, as well as a fifth toddler module, based on the children's development and capacity to communicate verbally. Each module is designed to assess different age and language levels, from those who have no expressive or receptive language to those who are verbally competent. Module 1, which contains 29 items, is designed for children who have limited or no phrase ability. Children who use phrases but lack fluency are given Module 2 (28 items). Module 3 has 28 items and is appropriate for children that have good communication abilities and play with age-appropriate toys, while Module 4 has 31 items and is appropriate for older children, mostly adolescents, with complete verbal fluency. The examiner selects the module that is most appropriate for a given child or adult based on expressive language level and chronological age, and then uses structured activities and materials, as well as less structured interactions, to create 'standard' contexts in which social, communicative, and other behaviors can be observed. The diagnostician records the participant's answer to each activity within each module, and overall ratings are given at the end. After that, the ratings are used to create a diagnostic. An algorithm, which sums the scores of particular items from the measure, yields a classification indicative of autism, ASD or non-spectrum conditions, based on empirically derived cutoffs (Note that ADOS-II classification alone is not sufficient to make a diagnosis of an ASASD). The ADOS essentially consists of a 30- to 45-minute observation period during which the examiner offers the person being evaluated pre-planned social circumstances in which a specific type of behavior is expected to occur.

1.6 Assessment tools for Adaptive, social, and cognitive abilities

Several other instruments, in addition to ASD diagnostic tools, have been developed to assess the phenotype of ASD in patients. The Vineland Adaptive Behavior Scale Vineland Adaptive Behavior Scale (VABS) [79] and the Wechsler Adult Intelligence Scale Wechsler Adult Intelligence Scale (WAIS) [80] are two commonly used tools. In addition, the Griffiths Mental Development Scales provide Griffiths Mental Development Scales (GMDS) an overall measure of a child's development [81].

1.6.1 Vineland Adaptive Behavior Scale (VABS)

The Vineland Adaptive Behavior Scale (VABS) assesses children's and adults' adaptive and social behavior which are characterized as the ability of a person to manage changes in their immediate surroundings and everyday activities. Similarly to ADOS and ADI-R, the VABS uses a semi-structured interview format to collect information from parents about their child's behavior in a range of circumstances. The Vineland Adaptive Behavior Scale 2^{nd} Edition (VABS-II) is the Second Revision of the early Vineland Social Maturity Scale [82]. VABS' first revision was published in 1984 [79] and in 2005, VABS-II was designed to measure five major aspects of adaptive behavior in daily settings for children and adults from birth to 90 years of age [83]. The VABS-II items are divided into 11 simple to more complex behavior subcategories spanning four main domains: communication which measures the receptive, expressive, and written abilities of children; daily living skills cover personal, domestic, and community interactions skills of a child; socialization focuses on measuring the interpersonal interactions, play and leisure time and coping abilities; and motor abilities (fine and gross). An optional Maladaptive Behavior

Index is also provided. For all VABS domains, a numerical value is generated throughout an individual's observations, with higher scores equating to higher adaptive skills. Two of the domains (i.e., motor skills and maladaptive behaviors) do not specifically address the basic symptoms of ASD. The VABS-II Adaptive Behavior Composite, a standardized score indicating the individual's general degree of adaptive functioning, is composed of the communication, daily living skills, and socialization scales. The Motor skills domain is also taken into account when calculating the Composite Score for children from birth to age six. The rationale for not including the Motor skills in subjects older than 6 years in the original scale is that Motor skills should be fully developed by this age. VABS has been widely used in clinics and across trials. However, a further revision of the scale is now available. The Vineland Adaptive Behavior Scale 3^{rd} Edition (VABS-III) is the latest iteration of the Vineland scale [84]. The domain and subdomain structure of the VABS-III is identical to that of the VABS-II. In comparison to the VABS-II, it includes a number of updates, such as updated normative reference data and revised item content that reflect changes in everyday life (such as advances in electronic technology). The correlation between the VABS-II and the previous version of the Vineland scale varied from 0.69 to 0.96 across domains/subdomains and between ages. How the VABS-II measures up against the VABS-III scale and how the link will be made are both unknown.

1.6.2 Griffiths Mental Development Scales (GDMS)

Originally published in 1954, the Griffiths Mental Development Scales GMDS were the first published scales designed to assess mental development in children under two years of age [85]. The GMDS 1984 Edition consists of six functional domains, each of which is assessed on a different subscale. These subscales are Locomotor which evaluates gross motor skills, including balance, coordination, and control of movements; Personal/Social which measures the acquisition of skills necessary for social and independent development; Hearing and Speech allows the evaluation of hearing (in the sense of active listening), receptive language, and expressive language; Eye and Hand-coordination focus on manual dexterity, visual monitoring skills, and fine motor skills; Performance concerning visual perception awareness, including working speed and accuracy, and Practical Reasoning which consists of questions that evaluate a child's ability to use information gained from their surroundings to solve problems and to understand mathematical concepts and moral problems, starting at the age of two [86]. This version has been updated several times since its original publication. Re-standardization for children aged 2 to 8 was completed in 2006, and the scale was renamed Griffiths Mental Development Scales-Extended Revised Griffiths Mental Development Scales-Extended Revised (GMDS-ER), and the content and scoring procedures have been revised and updated to reflect current knowledge and practice in child development assessment [87]. GMDS-ER provides a General Developmental Quotient (GQ) in addition to assessments of six functional domains. The Griffiths Scales are composed of a variety of items that cover the main aspects of a child's development and are arranged in order of gradually increasing difficulty. Each subscale is scored based on the child's performance on a series of tasks, with scores ranging from 0 to 15. The GQ score is calculated by averaging the subscale scores and multiplying by 10, with a possible range of 50 to 150. The classification of the child's development based on the GQ score is as follows:

Table 1.1: Griffiths' General Developmental Quotient and Classification

General Quotient	Classification
130 and above	Very Superior
120-129	Superior
110-119	High Average
90-109	Average
80-89	Low Average
70-79	Borderline
69 and below	Extremely low

The Griffiths Scales are widely used as a developmental assessment tool for children throughout the world. The third edition of the GMDS was published in 2016 [88]. Griffiths Mental Development Scales 3rd Edition (GMDS-III) is the product of a thorough re-standardization of the GMDS – described as the "gold standard" in child development testing. It assesses the rate of development in infants from birth to 6 years.

1.6.3 Intelligence Quotient (IQ)

The Wechsler Adult Intelligence Scale WAIS is the most widely used cognitive measure of adolescent and adult intelligence [89]. It is currently in its fourth edition Wechsler Adult Intelligence Scale 4^{th} Edition (WAIS-IV) released in 2008 [80]. The WAIS was initially introduced in 1955 and consists of both verbal and non-verbal (performance IQ) scales. The Wechsler Adult Intelligence Scale Revised (WAIS-R), a revised form of the WAIS, was introduced in 1981 and consisted of six verbal and five performance subtests. The Wechsler Adult Intelligence Scale 3^{rd} Edition (WAIS-III), a subsequent revision of the WAIS and the WAIS-R, was released in 1997 [90]. It provided scores for Verbal IQ, Performance IQ, and Full-Scale IQ. The WAIS-IV is the most updated edition of the WAIS test and includes 10 core subtests and five supplemental subtests, using the scaled scores from the 10 core subtests to calculate the Full-Scale IQ. The verbal/performance IQ scores from earlier editions were eliminated and replaced with the index scores. The General Ability Index General Ability Index (GAI) was included, which consists of the Similarities, Vocabulary, and Information subtests from the Verbal Comprehension Index and the Block Design, Matrix Reasoning, and Visual Puzzles subtests from the Perceptual Reasoning Index. The GAI is clinically useful because it can be used to assess cognitive abilities that are less susceptible to processing speed and working memory impairments. The WAIS-IV test is suitable for use with individuals ages 16 to 90. The Wechsler Intelligence Scale for Children (WISC, 6-16 years) and the Wechsler Preschool and Primary Scale of Intelligence (WPPSI, 212-7 years, 7 months) are used to assess intelligence in people younger than 16 [91].

1.7 The genetic architecture of ASD

Over the last two decades, significant progress in understanding the genetic architecture of ASD has been made due to technological genomics advancements. Researchers have identified large structural variations within the genomes of each individual, and it is now possible to investigate the genetic variants linked to ASD and determine their underlying genomic architecture [92, 93]. The transmission type of a

genetic risk factor can be via inherited or de novo variants [94]. De novo variations are regarded as major genetic risk factors and can be of three types: copy number variants Copy Number Variants (CNVs), single nucleotide variants Single Nucleotide Variants (SNVs), and indels [95]. Microarray techniques have allowed for genome-wide investigation of CNVs in large-scale lineage cohorts, successfully identifying recurrent locations of de novo CNVs. Genome-Wide Association Studies (GWAS) have investigated the polygenic risk factors associated with common variants in case-control testing. Next-generation sequencing (NGS) technology has made it possible to identify genetic variants and assess rare or low-frequency genetic changes that were previously undetectable using array-based methods. Whole-Exome Sequencing (WES) has allowed for the detection of deleterious variants that modify the protein-coding sequence and the identification of several genes strongly linked to ASD [96, 97]. Whole-genome sequencing Whole-Genome Sequencing (WGS) has allowed the discovery of rare or low-frequency noncoding variants that were previously unknown in WES research, as well as a comparison of the risk associated with coding and noncoding mutations in terms of human diseases and features [98, 99]. Despite this progress, the majority of people with ASD do not have a single identifiable etiology [100]. Nonetheless, the combined efforts in analyses and methodological development in the past decade have provided an opportunity to explore and assess the contribution of genetic variants in disease development, allowing for a more in-depth look at the genetic landscape of ASD in a time- and cost-effective manner.

1.8 Etiology

The research on ASD genetics has been critical in the last decade, not only to understand and explain its phenotypic heterogeneity but also to develop new diagnostic tools and therapies. To date, hundreds of genes are thought to be involved with ASD, resulting in a broad spectrum of distinct phenotypes including multiple linguistic and social deficits with numerous associated sub-phenotypes. It is uncertain that a single condition or event plays a significant role in the etiology of ASD; rather, none of the risk factors identified so far is a necessary and sufficient condition for ASD. Even in the case of syndromic or secondary ASD, which refers to ASD caused by a single underlying condition, such as fragile X syndrome or tuberous sclerosis, none of these etiologies are specific to ASD because they each include a varied proportion of individuals with and without ASD[101]. Currently, ASD appears to have a complex, multifactorial etiology with genetic, environmental, and developmental factors playing a key role in the onset of ASD all contributing in unknown and varied ways, as many epidemiological studies have shown [102].

1.8.1 Genetic risk factors for ASD

ASD is a complex neurodevelopmental disorder with high heritability. Around 75% of ASD patients are diagnosed with ASD, and about 20% have a positive family history of ASD. Siblings of ASD children are more likely to be diagnosed with ASD, with a sibling recurrence risk of 32.2% for families with multiple ASD probands [103]. Syndromic ASD affects about 25% of patients and is characterized by the co-occurrence of autistic traits with dysmorphic features or congenital defects. Furthermore, the probability of sibling recurrence is smaller (4–6%) and family history is less common (9%) [70].

Studies have shown that the genetic component of ASD is significant, with genetic heritability accounting for 45-56% of cases [104]. Concordance rates for ASD are higher in monozygotic twins (60-90%) than in dizygotic twins (5-40%), confirming the genetic fingerprint [105, 106].

Based on all available information, the genetic heritability of ASD is thought to play a significant role

in ASD onset, alongside environmental and epigenetic variables. Even though the majority of studies aimed at identifying the etiological foundation of ASD have focused on the genetic component, genetic variations have only been linked to a small percentage of ASD patients [107, 108]. According to Colvert et al. and Hallmayer et al. environmental factors can account for up to 30% and 55% of the ASD liability, respectively [106, 109]. In line with these two studies, Frazier et al. found a low heritability rate for categorically defined patients, indicating the environmental contribution [110]. They did, however, find higher heritability rates in severely affected patients, confirming prior concepts of genetic influence on ASD development. Common genetic variants with minor effects are thought to contribute to the development of complex traits in ASD [111].

1.8.2 Environmental factors

Despite the strong genetic etiology of ASD, several researchers have also looked into the significance of environmental factors, as previously mentioned. According to recent studies, environmental risk factors such as pharmaceutical drugs, exposure to ubiquitous xenobiotics, advanced parental age, nutritional deficits, uterine environment, and many more may account for up to 40%–50% of the variance in ASD liability [112, 113, 114, 115]. Environmental factors are anticipated to take effect in the early developmental stages, when the developing brain is particularly vulnerable to external factors that might have negative effects, potentially modifying the neuropathological events that lead to ASD onset. However, while there is significant evidence for some putative risk factors, backed up by association studies as well as in vitro and in vivo investigations, many others have described relatively weak associations [115].

Role of parental age and perinatal risk factors in ASD etiology

Parental age is a well-established environmental risk factor for ASD. Advanced paternal age has been linked to the emergence of bipolar disorder, schizophrenia, ADHD, and ASD [116]. A meta-analysis of 27 studies found that a 10-year rise in maternal and paternal age is associated with a 20% higher incidence of ASD in children [117]. Moreover, age-related methylation alterations in sperm may also increase the incidence of ASD in offspring [118]. Perinatal risk factors are among the most investigated causes of ASD, with umbilical cord complications, birth trauma, multiple births, maternal hemorrhage, low birth weight, neonatal anemia, genital malformation, ABO or Rh blood group incompatibility, and hyperbilirubinemia all showing statistically significant associations with ASD risk [119, 120]. Fetal distress, induced labor, cesarean birth, and management age of less than 36 weeks are also associated with an increased risk of ASD [121]. ASD risk is significantly influenced by the mother's health, given that the nutrients available to promote fetal growth are determined by the mother's diet throughout pregnancy [122]. Deficiencies or excesses in micronutrients such as folic acid, zinc, iron, vitamin D, and omega-3 fatty acids can hamper neurodevelopment. Additionally, a possible risk factor could be the mother's use of drugs, alcohol, or tobacco during pregnancy. However, there is not much evidence linking drinking alcohol and smoking to ASD [123, 124].

Presence of autistic-like traits in unaffected relatives of ASD probands

The presence of subclinical ASD characteristics in families with autistic children is commonly referred to as the Broader Autism Phenotype (BAP) [125]. The BAP are milder manifestations of heritable autistic-like traits that are frequently observed in unaffected relatives of ASD patients. When compared to families of typically developing children or families of children with Down's syndrome, this congregation of ASD-like traits is significantly more prevalent among parents and relatives of probands

with ASD[125, 126, 127]. BAP is also heritable [128] and according to research by Robinson et al. (2011), sub-diagnostic autism features are just as heritable as ASD [129]. Therefore, features of BAP may be an endophenotype, a measurable phenotypic trait that can be heritable, independent of the state, co-segregates in families, and is more prevalent in the relatives of probands than in the general population [130, 131]. Numerous research has revealed links between parental BAP features and the severity of a child's ASD, making parental BAP traits an unquestionable risk factor for ASD in children [132, 133, 134]. It is still unclear how the autism spectrum disorders' personality features are passed on from parents to children.

1.9 ASD heterogeneity and research challenges

The vast heterogeneity of symptom presentation has long been acknowledged as a defining characteristic of ASD. Research, diagnosis, and treatment of ASD are complicated by the wide variety and complexity of autistic symptomatology [135, 136, 137]. Perhaps the most outstanding reason why understanding heterogeneity is of high importance is the fact that individuals with ASDrespond to treatment in vastly different ways. Existing literature suggests that the majority of treatment approaches, such as early intensive behavioral intervention and naturalistic developmental behavioral intervention have variable levels of effectiveness and occasionally may not significantly affect core characteristics of ASD, such as social-communication difficulties [138, 139, 140, 141]. Additionally, there are currently no medical interventions that significantly affect the core features of autism [142, 143]. Most recent best practice recommendations specifically emphasize the critical need for future research to identify factors that explain heterogeneity in response to treatment, in order to better individualize treatment and intervention approaches and to better target changes in core or functionally impairing symptomatology [138, 139, 140, 141].

Heterogeneity also limits basic scientific progress toward understanding ASD. Studies that use case-control designs to find "biomarkers" implicitly assume that, if such a factor existed, it would completely distinguish all cases from all controls. However, we have not yet found any biomarkers for ASD that reliably and consistently meet this demanding standard [144, 145]. One possible explanation is that high-impact biomarkers are likely exclusive to certain subsets of autistic individuals, that is may provide information about one subtype of ASD but not another. In other words, a high-impact biomarker may provide information about a subset of autistic individuals rather than the entire patient population. Some researchers have "given up on a single explanation for ASD". In contrast, others have suggested that ASD should not be considered a single disorder [146, 147]. Instead, they suggest that within ASD, there could be groups of distinct disorders with many etiologies [147].

Heterogeneity research can be challenged due to a shortage of datasets with sufficient size to adequately address such problems. Big, open data is particularly valuable in research relating to complex heterogeneous disorders such as ASD. The inconsistent results and significant heterogeneity of ASD make it necessary to use big and open data in order to address important challenges, such as defining the heterogeneity and potential subtypes of ASD. This research is of high importance since finding more homogeneous subtypes would help to better understand the diversity in etiologies. Different subtypes may require different treatment interventions and different types of care, as behaviourally distinct diagnostic subtypes may be associated with specific responses to specific interventions. Moreover, they may also be influenced by many biological and environmental factors. Several studies have looked for empirically derived clusters of participants, within reasonably large datasets, in an effort to define the subtype structure of ASD.

Unfortunately, because the numbers and characteristics of identified subtypes vary across studies, the identification of replicable subtypes remains elusive. These inconsistencies may result from methodological variations in the age ranges covered, measurements included, and statistical techniques used. Previous research has stressed the importance of evaluating multiple sources of adaptive behavior and intellectual functioning when attempting to identify subtypes within ASD in order to construct more comprehensive profiles [136, 148, 149, 150]. When only one measure is used to derive subtypes, the findings are restricted to the characteristics of that specific measure making it less likely to account for the complete spectrum of children's developmental abilities and behaviors. Between two and four phenotypically distinct ASD subtypes have been determined using these data in the past. For instance, Kim et al. clustered young children referred for suspected ASD at 22 months, into four subtypes based on symptom severity, developmental skills, and adaptive functioning [149]. The four clusters were characterized by distinct clinical profiles evidenced by verbal and nonverbal, adaptive, and sociability abilities measured by the Mullen Scales of Early Learning and Vineland Adaptive Behavior Scales (VABS), in addition to the ASD symptoms. Therefore, including measures on multiple developmental and behavioral domains as subgtyping features could provide additional information to identify clinically meaningful subtypes. Prior research in this field identified symptom structures and subtypes by using statistical methods, such as Factor Analysis (FA), Latent Class Analysis (LCA) [151, 152], Factor Mixture Modeling (FMM) [153, 136], and Cluster Analysis [148, 154, 155]. The advantages of these methodologies vary depending on the statistical approach used. FA, for instance, focuses on variable-level grouping in order to identify the relationships between variables and the structures of developmental and behavioral symptoms. LCA and FMM, on the other hand, stratify individuals into a predetermined number of classes according to the factors and pre-identified models. Cluster analysis is an exploratory method that enables subtypes to be derived at the subject level based on individual characteristics. Exploratory methods (e.g., FA and cluster analysis), are more data-driven and suitable when prior findings remain inconclusive, in contrast to modeling techniques (e.g., LCA and FMM) which adopt a theory-driven approach.

1.10 Machine Learning

The term Machine Learning (ML) typically refers to the process of applying prediction models to data or of finding meaningful groupings within data. ML is especially helpful when the dataset being analyzed is too large (contains many individual data points) or too complex (contains a large number of features) for human analysis and/or when it is desired to automate the process of data analysis to establish a reproducible and time-effective pipeline. These characteristics are frequently present in biological datasets, which have grown significantly in size and complexity over the past few decades. When applying ML in biology, there are two objectives. First, in the absence of experimental data, develop precise predictions and utilize these to direct future studies. The second is to employ machine learning to deepen our understanding of biology.

1.10.1 Types of machine learning

Machine Learning algorithms can be broadly divided into two groups, supervised and unsupervised learning. The two main classes are distinguished by the presence of labels in the training data subset. In the former, labels, such as clinical diagnoses, are known and used to identify the best decision rule. In the latter, the algorithm entirely relies on the innate structure of the unlabeled input data to derive a

judgment regarding class membership.

1.10.1.1 Supervised machine learning methods

Supervised machine learning involves predetermined output attributes in addition to the use of input attributes [156]. The algorithms attempt to forecast and classify the preset attribute, and their accuracy and misclassification, along with other performance measures, depending on the numbers of the predetermined attribute successfully predicted or categorized or otherwise. The supervised learning algorithms are further divided into classification and regression algorithms [156, 157].

1.10.1.2 Unsupervised machine learning methods

Unsupervised data learning, on the other hand, analyzes data that has not been labeled (i.e. there is no separation between input and output) while making assumptions about the underlying structural properties of the data (such as algebraic, combinatorial, or probabilistic). Since no training samples are used in this process, the learning algorithm looks for patterns and correlations in the supplied data. These algorithms' main applications include clustering and dimensionality reduction.

Clustering

Clustering methods are used to predict groupings of similar data points in a dataset and are usually based on some measure of similarity between data points in such a way that objects in the same cluster share or contain more similar characteristics than the objects in other clusters. Clustering has been widely used in many fields such as pattern recognition, molecular biology, and bioinformatics. More specifically, in biomedical research, it has also been used to study many diseases, including ASD, by identifying co-regulated gene modules and disease subtypes in the context of precision medicine [148, 155, 158]. Clustering has become a crucial technique for data analysis research due to its simplicity of use and interpretation, especially in the study of complex datasets. Various clustering models have been developed over time. Han and Kamber divided the many clustering methods created for handling static data into five main groupings: partitioning, hierarchical, density-based, grid-based, and model-based methods [159, 160]. The most significant and widely applied in medical sciences are Hierarchical methods, which create clusters based on a linking criterion and a distance function; Partitioning methods in which the cluster is represented by a single mean vector and Density-based methods where clusters are defined by the region where observations are denser and more similar.

Partitioning methods

Partitioning methods create k partitions of the data from a collection of n unlabeled data tuples, where each partition represents a cluster with at least one object and $k \leq n$. If each object belongs to only one cluster, the partition is crisp; otherwise, it is fuzzy since each object may belong to more than one cluster to a different degree. The k-means algorithm and the k-medoids algorithm are two well-known heuristic techniques for crisp partitions [161]. With the k-means algorithm, each cluster is represented by the average value of the objects in the cluster, and with the k-medoids algorithm, each cluster is represented by the object that is closest to the center of the cluster [162, 163, 164].

Hierarchical Methods

Data objects are arranged into groups in a tree structure using the hierarchical clustering approach. This clustering method does not require a predetermined number of clusters and typically employs a distance matrix calculated using a predefined distance function. The most common distance functions are Euclidean distance and the Gower method [165]. A dendrogram is constructed using the distances between the objects, merging several groups at specific distances. The dendrogram's y-axis indicates the distance between clusters whereas the x-axis displays the objects. Hierarchical clustering is primarily determined by the distance function and linkage criterion used to form clusters. The linkage criterion includes single, complete, average, and ward linkage clustering. These linkage methods are primarily used to assess the degree of dissimilarity between two groups of observations. Hierarchical clustering methods are broadly classified into two types: agglomerative and divisive. Agglomerative methods use a bottom-down approach, beginning by grouping every object into its own cluster and then combining those clusters into ever-larger clusters until every object is in a single cluster, or until specific termination requirements, like the required number of clusters, are reached. In contrast, the Divisive method is a top-down approach that does just the opposite. In this method, each observation is given its own cluster after which splits are performed successively until each observation has its own cluster. A pure hierarchical clustering method is limited in its ability to adjust after a merge or split decision has been made.

Density-based Methods

The main idea behind density-based algorithms is to keep forming a cluster as long as the density (number of objects or data points) in the "neighborhood" is higher than a predetermined limit [166].

Currently, conducted research and applications combine supervised and unsupervised machine learning algorithms. The goals of the analysis and the data at hand influence the choice of ML technique. The number of predictor variables/features available in the data, as well as the quality of the data, are important data considerations. In general, a small but informative feature space increases the model's generalizability and prevents overfitting while improving data quality greatly improves the analysis [167]. As a result, the first steps in a machine learning analysis are often data preprocessing and dimension reduction.

Dimensionality Reduction

Dimensionality reduction is the preprocessing step to identify and remove redundant features and noisy and irrelevant data. Reducing the number of variables in a data set naturally reduces accuracy, but the idea of dimensionality reduction is to trade a little accuracy for simplicity. Smaller data sets are easier to visualize and examine, and since there are fewer unnecessary variables to handle, machine learning algorithms can analyze data much more quickly and easily. Dimensionality reduction methods can be categorized mainly into Feature Selection Feature Selection (FS) and Feature Extraction Feature Extraction (FE). In the FE method, features are reprojected into a new, lower-dimensional space. A few examples of feature extraction techniques are Singular Value Decomposition Singular Value Decomposition (SVD), Linear Discriminant Analysis Linear Discriminant Analysis (LDA), and Principal Component Analysis (PCA). In contrast, the FS method aims to select a limited subset of features that maximize relevance to the target (i.e. class label) while minimizing redundancy. To name a few popular feature selection techniques, there are Information Gain, Relief, Chi Squares, Fisher Score, and Lasso.

Principal Component Analysis

Principal Component Analysis is a statistical procedure that transforms a set of observations of potentially correlated variables into a set of values of linearly uncorrelated variables called principal components. The data dimension is reduced through the following steps:

1. Data Standardization

The goal of this step is to standardize the range of the continuous initial variables so that each one contributes equally to the analysis. Standardization must be performed before PCA, in part because the latter is quite sensitive to variations in the initial variables. It is best practice to make the data unit-free and center it at mean zero to improve process efficiency.

2. Covariance Matrix Computation

PCA attempts to understand how the variables in the input data set differ from the mean in relation to each other, or to see if there is any relationship between them. Because variables are sometimes so highly correlated that they contain redundant information. Therefore, the covariance matrix is computed to find these relationships. The covariance matrix is a symmetric matrix with the same number of rows and columns as the number of dimensions in the data. Calculating the covariance between the pairwise means tells us how the features or variables diverge from each other.

3. Eigenvectors and Eigenvalues of the Covariance Matrix Computation

Eigenvectors and eigenvalues are linearly independent vectors and scalars, respectively, that must be computed from the covariance matrix in order to determine the principal components of the data. Principal components are new variables that are created by linear combinations or mixtures of the initial variables. These new variables (i.e., principal components) are uncorrelated and the majority of the information within the initial variables is compressed into the first components. By eliminating the components with little information, PCA allows for minimizing dimensionality without sacrificing much information. Geometrically, Principal Components represent the data directions that explain the most variance. The axes' directions with the largest variance (most information) are the eigenvectors of the covariance matrix. Additionally, eigenvalues are simply the coefficients attached to eigenvectors and indicate the amount of variance held by each Principal Component. By ranking eigenvectors in order of their eigenvalues, highest to lowest, we get the Principal Components in order of significance.

4. Feature Vector

In this step, it must be decided how many principal components are required and how much information loss may be tolerated in order to create a matrix of vectors, called a feature vector. Therefore, the feature vector is just a matrix with the eigenvectors of the components that we choose to maintain as its columns. As a result, it can be considered the initial stage in the process of dimensionality reduction since, if just p of the eigenvectors (components) out of n is retained, the resulting data set will only have p dimensions.

5. Data recast along the principal component axis

Here, the goal is to reorient the data from the original axes to those represented by the principal components by using the feature vector produced using the eigenvectors of the covariance matrix.

1.11 Unsupervised machine learning methods evaluation

1.11.1 Approaches to clustering stability

The overarching goal of stability analysis is to characterize the reproducibility of a clustering. A natural solution is to validate a clustering by obtaining an independent sample, or many independent samples, from the underlying population. However, due to constraints like time and money, this is rarely feasible. Cluster stability approaches rely on perturbations to the original dataset as an alternative to collecting new data for validation. Stability measures capture how well partitions and clusters are preserved under perturbations to the original dataset. The underlying assumption is that a good clustering of the data will be replicated throughout an ensemble of perturbed datasets that are nearly identical to the original data. Stability is popular in a variety of fields and applications. Clinical studies, for example, have used stability to identify stable disease clusters based on phenotype, to characterize disease subgroups in longitudinal studies, and to identify stable clusters of symptoms to promote better patient care [168, 169, 170, 171]. Over the past years, various cluster stability approaches have been developed and are being used to address some of the issues and constraints associated with clustering. These stability approaches differ significantly in the way similarity between clustering is calculated and, in the way, small perturbations to the original dataset are generated. Stability has addressed some foundational issues such as an estimate of confidence in an item's membership in a cluster, an estimate of confidence to cluster, and an overall estimate of confidence for a dataset clustering.

Resampling for stability estimation

Bootstrapping is an easy-to-implement method that makes it possible to create replicated datasets of the same size. For a dataset X0$\in$RN$\times$p with N observations, the data are resampled with replacement to generate bootstrap replications X1, X2, ... XB, that are the same size as the data [172]. A given bootstrap replication has the inherent property that an observation xi may happen just once, several times, or not at all. The bootstrap approach depends on the distance between clusters as well as their dispersion. In brief, within-cluster dispersion is defined as the sum of distances from each point in the cluster to the cluster center divided by the number of points in the cluster while between-cluster dispersion is defined as the distance between two cluster centers [173]. Jain and Moreau (1987) bootstrap this stability measure and average overall k clusters. The focus of their study was on choosing the best model for the optimal value of k (number of clusters) that produced the most stable data partitions. The ideal k was found to be the least variable measure that minimizes the criterion, representing the most stable clusterings. Although demonstrations were limited to k-means and hierarchical clustering with different linkage functions, this application is adaptable to more broad clustering algorithms. Additionally, it was demonstrated that statistics were useful for comparing clustering techniques [173]

Chapter 2

Thesis Aims

While existing intervention approaches do not assist all patients equally, it is now very difficult to forecast the course of the disease and what may work better for each individual. The work underlying this thesis aimed to improve our current knowledge of ASD biological etiology by analyzing clinical data from the Portuguese population sample of ASD patients. By characterizing the different subtypes of individuals with ASD, we can better understand the underlying mechanisms that contribute to the development and progression of ASD as well as the variability in symptoms and outcomes associated with ASD which may be helpful in the development of more personalized interventions for each subtype. Therefore the main goal of this thesis is to identify subtypes of ASD patients using machine learning methods. The specific objectives are listed as follows:

1) To define phenotypic clusters using clinical variables of large datasets of ASD patients;

2) To define the characteristics of individuals within each subtype.

Chapter 3

Methods

3.1 Participants

The ASD original dataset was obtained from the Hospital Pediátrico – Centro Hospitalar e Universitário de Coimbra Centro Hospitalar e Universitário de Coimbra (HP-CHUC). This cohort was set up to study ASD developmental trajectories and comprises clinical information from ASD patients. Between 1999 and 2021, 949 children aged 1–18 years were recruited at the time of diagnosis in the HP-CHUC covering most of the central regions of Portugal. ASD diagnosis, ascertained by a trained multidisciplinary team based on DSM-IV and DSM-V criteria [14, 15], relied on the Second Edition of the Autism Diagnostic Observation Schedule [77], the Revised Version of the Autism Diagnostic - Interview [75], Vineland Adaptive Behavior Scales Second Edition [83] and The Griffiths Mental Development Scales Second Edition [86]. The parents of the 949 children originally included, were asked to answer a questionnaire covering sociodemographic characteristics and the mother's lifestyle during the pregnancy. Approval was obtained from the ethical committees of the institutions participating, and informed consent was granted by all subjects or their parents/legal guardians. Following data protection laws, study participants' identities remained confidential.

3.2 ASD diagnosis, clinical assessment instruments

Individuals meeting DSM-IV or DSM-V criteria and reaching the thresholds for ASD from the Autism Diagnostic Interview - Revised (ADI-R) and the Diagnostic Observation Schedule (ADOS) were classified as ASD cases. Furthermore, we decided to include individuals having a diagnosis of ASD based solely on one instrument and who did not have information or did not meet the thresholds for ASD from the other. Individuals who did not meet DSM-IV criteria for an ASD diagnosis on both the ADOS and the ADI-R were excluded. Karyotype results, as well as Fragile X syndrome Fragile X syndrome (FRAXA) and Fragile XE syndrome Fragile XE syndrome (FRAXE) results, were analyzed in order to identify possible cases of non-idiopathic autism (e.g., known neurogenetic disorders, known chromosomal abnormalities) and consequently removed from the study. Moreover, samples identified with severe intellectual disabilities and global developmental delay in the clinical diagnosis were also excluded. In total, 878 individuals met the criteria for inclusion. The ASD dataset comprises a total of 307 variables containing different information, such as sociodemographic, genetic and comorbidities,

family history, assessment tools for adaptive, social, and cognitive abilities as well as other relevant information (ie. Apgar, mother's behavior and complications during pregnancy). Diagnostic evaluations of ASD involve the assessment of multiple domains, including cognitive, language, restricted and repetitive behaviors, social deficits, and adaptive skills.

3.3 Variables Selection

In order to uncover significant subgroups within the autism spectrum, we included measures that can capture variations in children's development, such as communication and social abilities. Therefore, apart from the diagnostic/screening measures designed to classify children as having ASD or not (ie. ADI-R or ADOS), dimensional measures of ASD symptoms (ie. VABS and GMDS) were also included in this analysis. Both VABS and GMDS assess adaptive function, which is an important distinguishing factor in ASD. For the purpose of the current study, we only included subscale scores in order to capture different aspects of the developmental domains that are measured. Composite scores combine these subscales and may miss some important linkages to predictor variables. Therefore, we believe they do not bring new information to the analysis and were not included.

In total, 11 variables were selected for cluster analysis (Table 3.1). From ADI-R's four domains, only three were included; Reciprocal Social Interactions, Restricted and Repetitive Behavior (RRB), and the Abnormality of Development Evident At or Before 36 months (AbDev). The ADI-R communication domain is divided into verbal and non-verbal scores. However, this domain was not included because the verbal and non-verbal are measured on different, finite scales. Consequently, it would be inappropriate to directly compare the severity of communication deficiencies between non-verbal and verbal individuals. From the Vineland Adaptive Behavior Scales (1^{st} and 2^{nd} Edition), standard scores of Socialization, Communication, and Daily Living Skills domains were included. The Motor Skill domain was not included because it had 77% of missing values. Five subscales of Griffiths Mental Development Scales (1984 Edition), Locomotor, Personal/Social, Hearing and Speech, Eye and Hand-coordination, and Performance were also included. The Practical Reasoning subscale was not included because it had 15% zero values and 23% missing values. ADOS domains were not selected for clustering as the domain values were not standardized and presented more than 20% of missing data. On the other hand, ADOS is a diagnostic tool like ADI-R and we decided to make use of it to confirm our cluster results in the characterization process. Therefore, to investigate the nature of the clusters formed, we included eighteen variables in our analysis, such as Clinical severity (ADOS), ADI-R quotation, Dysmorphisms, Family Psychiatric History, Language Regression, Psychomotor Developmental Delay and Regression, Gender, Audition, Vision, Verbal Status, Head Circumference, Apgar 3^{rd}, and 5^{th} as well as Diagnose, First Walk, First Words and First Phrases Ages (Table 3.1).

Table 3.1: List of Variable Names and Full Decriptions

Measures	Variable Name	Full Description
	ADIR_Soc	Social
Autism Diagnostic Interview-Revised	ADIR_RRB	Restricted Repetitive Behaviour
(Standard Scoress)	ADIR_AbDev	Abnormality of development
		evident at or before 36 months
	VABS_Com	Communication
Vineland Adaptive Behaviour Scales-II	VABS_Aut	Daily Living Skills
(Standard Scores)	VABS_Soc	Socialization
	QD_M	Locomotion
Griffiths Mental Development Scales	QD_PS	Personal/Social
(Standard Scoress)	QD_L	Hearing /Language
	QD_EH	Eye-Hand Coordination
	QD_R	Performance
Head Circumference	HC	Head Circumference
		(centimeters)
Apgar 1º	Apgar_1	Apgar Test Score
		(1 minute after birth)
Apgar 5º	Apgar_5	Apgar Test Score
		(5 minute after birth)
Diagnosis Age	Age_Diag	Age at inclusion
Walking Age	Walk_Age	Age when started walking (months)
First Words Agw	First_Words_Age	Age when said the first words (months)
First Phrases Age	First_Phrases_Age	Age when said the first sentences (months)
Gender	Gender	Gender
Clinical Severity (ADOS) score	ADOS_sev	ADOS Severity Score
ADI-R Quotation	ADIR_quot	Result of ADI-R diagnoses
Dysmorphysms	Dysmorphysms	Presence of Dysmorphysms
Language Regression	Language_Reg	Regression in Language
Audition	Audition	Audition Problems
Vision	Vision	Vision Problems
Verbal	Verbal	Verbal Status
Family Psychiatric History	Psyc_Family_Hist	Presence of Psychological
		History in Family
PMD Regression	PMD_Regression	Regression of Psychomotor
		Development
PMD Delay Always	PMD_Delay	Delay of Psychomotor
		Development Since Always

3.4 Data Analysis

Data cleaning and preprocessing treatment were conducted solely using Python version 3.8.12 [174], whereas statistical analysis was implemented in both Python and statistical software R version 4.2.1 [175]. We first generated a descriptive analysis profile of all variables to provide an overview of the database. This step allowed us to identify which variables were suitable to be used in the cluster analysis. Unfortunately, a significant number of variables with relevant clinical interest had to be left out because they had more than 30% of missing data. We then proceed to clean and process our data. A new categorical variable (Verbal) was created through the combination of verbal and non-verbal scores of the ADI-R communication domain. GMDS subscales were carefully revised as some inconsistencies were found. Some individuals had scored in several GMDS subscales but were older than the age at which the Griffiths test is performed. Therefore, all samples that were older than 8 years old at the time of Griffith's evaluation were removed. We also verified that some samples that were classified as non-verbal by ADI-R did not present values for the Hearing and Speech domain of the GMDS or for its reevaluation. Therefore, the missing values of this domain, for these samples, were altered to zero, since they did not score because they are not verbal. Moreover, there were samples classified as non-verbal by the ADI-R, however, they had very high values in the Hearing/Language domain and did not present values for its reevaluation. Within these selected samples, we then removed the ones with Hearing/Language domain values greater than 79 following the Griffiths' General Developmental Quotient Classification Table previously mention in Section 1.6.2 (Table 1.1). Whenever available, GMDS subscales were updated with the scores obtained during the child reassessment process. Clinical samples were examined for missing values. To minimize missing value imputation bias, samples with missing data in two entire cluster variable groups (i.e ADI-R, VABS, and GMDS) were excluded from further analysis. Clusters' descriptive variables also underwent a cleaning process in order to dichotomize the values of the categorical variables. Regarding numerical variables, cases of *outliers* were investigated, and whenever found were converted into missing values. In order to give a general overview of the individual's characteristics in the sample, descriptive statistics were computed for all measurements included (Tables 3.2, 3.3 and 3.4 and Figures 7.1, 7.2 and 7.3).

Table 3.2: Descriptive Statistics of Cluster Variables

Measures	Developmental Domain	Subdomains	N	Mean	SD	Range
Autism Diagnostic Interview-Revised (Standard Scores)	Social Ability	Social	638	18.97	6.52	0-30
	Repetitive Behaviours	Restricted Repetitive Behaviour	637	4.49	2.00	0-23
	Developmental Abnormalities	Abnormality of development evident at or before 36 months	629	4.03	1.56	0-25
Vineland Adaptive Behaviour Scales-II (Standard Scores)	Communication	Comunication	524	63.80	22.35	20-121
	Adaptive Behaviour	Daily Living Skills	524	60.47	19.40	20-111
	Social Ability	Socialization	526	66.18	17.54	20-106
Griffiths Mental Development Scales (Standard Scores)	Gross Motor Ability	Locomotion	628	87.17	20.6	28-157
	Adaptive Behaviour	Personal/Social	628	70.35	22.23	19-161
	Receptive and Expressive Language	Hearing and Speech	640	63.01	32.09	0-158
	Fine Motor Ability	Eye-Hand Coordination	628	74.62	25.98	15-150
	Visuospatial Ability	Performance	628	81.78	27.32	14-166

Table 3.3: Descriptive Statistics of Numeric Categorization Variables

Numerical Measures	N	Mean	SD	Range
Head Circumference (cm)	580	34.70	3.24	10-75
Apgar 1º	482	8.56	1.25	3-10
Apgar 5º	619	9.82	0.52	7-10
Diagnosis Age (years)	480	3.67	1.56	1-8
Walking Age (months)	626	14.06	2.60	8-22
First Words Age (months)	596	23.25	9.99	9-50
First Phrases Age (months)	456	41.36	10.91	18-66

Table 3.4: Descriptive Statistics of Categorical Variables

Categorical Measures	N	Values
Gender	663	Male / Female
Clinical Severity (ADOS)	557	Autism / ASD / Non-Spectrum
ADIR Quotation	643	Positive / Negative
Dysmorphysms	639	Yes / No
Language Regression	625	Yes / No
Audition	638	Normal / Not Normal
Vision	602	Normal / Not Normal
Verbal	638	Yes / No
Family Psychiatric History	648	Yes / No
PMD Regression	619	Yes / No
PMD Delay Always	642	Yes / No

We also compared the main characteristics between the included and excluded individuals to assess any bias introduced by the variable and sample selection process in the dataset (Table 4.1). In order to compare the two groups of samples, the *Mann-Whitney* (MW) nonparametric test and the Chi-Square test of Independence were applied to numeric and categorical variables, respectively. The measures included in our analysis reflect different sorts of statistical variables, including ordinal, finite variables (e.g. ADI-R, Apgar), and continuous, finite variables (e.g. VABS, GMDS, Walk Age). Therefore, variables were transformed into percentile ranks, using a built-in R function, to allow for more comparable measurements. The relationships among the domains measured were obtained by computing pairwise Spearman's rank correlation coefficients [176]. VIM R package's aggregation plot was used to visualize and analyze the proportion of missing data by variable [177]. The histogram showed that VABS subscales had the highest proportion of missing values, with more than 20%, followed by the majority of GMDS subscales, around 5% and lastly, ADIR subgroups and GMDS Hearing and Speech subscale having less than 5% (Figure 3.1). Moreover, the missing pattern plot shows that 70% of the samples do not have missing data for any cluster variable and around 20% of samples have missing data in all VABS subscales (Figure 3.1). To handle missing data, we used the missForest R package [178] that implements the Random Forest algorithm, a decision tree-based supervised machine learning method: (a) missing values are filled in using median/mode imputation; (b) missing values are marked as 'Predict' and the others as training rows, which are fed into a Random Forest model; (c) generated prediction is then filled in to produce a transformed dataset. Imputation error was assessed using the Normalized Root Mean Square Error [178], and the missing values imputation error was 0.288.

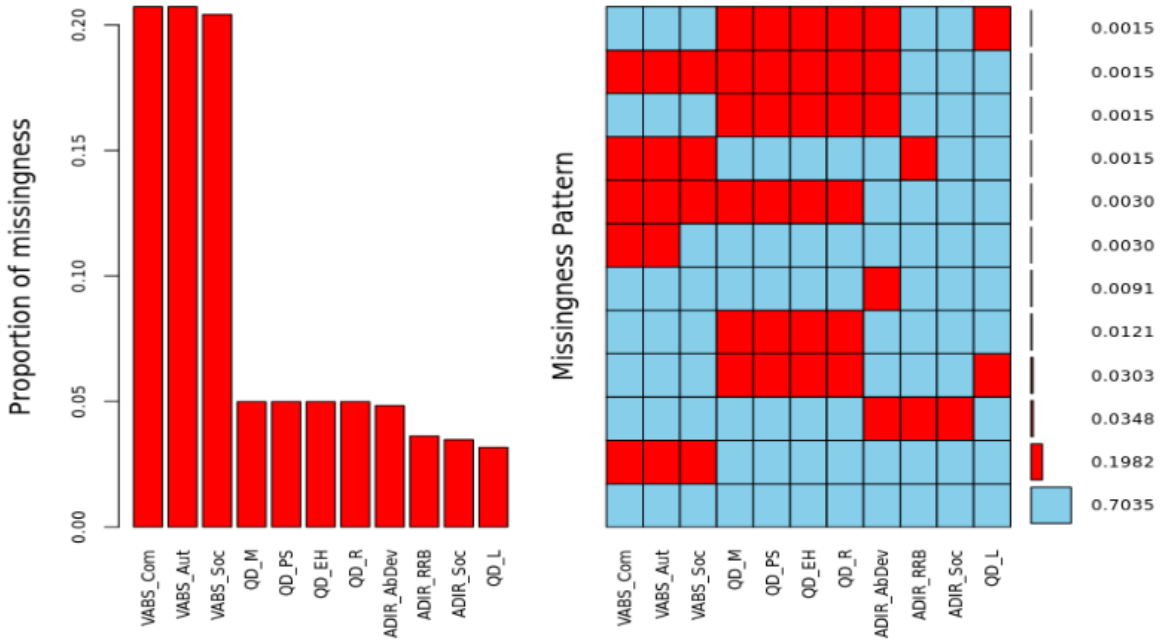


Figure 3.1: Proportion of Missing Data and Missingness Pattern. Histogram shows the proportions (percentage) of Missing Values. Pattern Chart shows the proportion of Non-Missing (Blue) and Missinng (Red) Values.

3.5 Clustering analysis of ASD clinical data

Before applying any clustering methods, evaluating if the data contains meaningful (i.e non random-structures) clusters is important. The Hopkins statistics [179] allowed us to assess the spatial randomness of the data by measuring the probability of our data being generated by a uniform data distribution. The obtained value was 0.642 which is above the threshold (0.5), therefore we can reject the null hypothesis and conclude that the data is not uniformly distributed (i.e., contains meaningful clusters.). Prior to clustering, a Euclidean distance matrix was calculated using percentile-ranked data. We started by evaluating different clustering methods using both 'CLValid'[180] and 'NbClust' [179] packages in R. CLValid used several internal validation indices to assess the quality of clustering results for the different numbers of clusters (range of 2 to 6). It then selected the number of clusters that maximized the average validation index across seven methods selected (i.e. kmeans, hierarchical, agglomerative nesting and divisive hierarchical, clustering large applications, model-based and self-organizing tree algorithm). On the other hand, NbClust compared the performance of 8 clustering algorithms (kmeans, ward.D2, single, complete, average, mcquitty, median and centroid) by computing 30 clustering quality indices for the range of possible numbers of clusters (0 to 6). It then provided a consensus clustering solution by selecting the number of clusters that were most frequently suggested by the indices across all the clustering methods being evaluated. Since the results were not consistent, we decided to approach a different method, the Hierarchical Clustering on Principal Component HCPC because it combines three standard methods used in multivariate data analysis: 1) principal component methods as a denoising step which can lead to a more stable clustering; 2) hierarchical clustering and 3) partitioning clustering for a consolidation to increase robustness. Implementation of the HCPC method was done using the

'FactoMineR' package [181]. First, we conducted a PCA on the percentile-ranked data and identified the components that account for the majority of the variance in order to reduce data dimensions. The first set of Principal Components Principal Component (PC)s that accounts for 89% of the variation were selected for clustering and provided insights into inherent structures among the variables that drove the clustering. Next, a Hierarchical Cluster was performed on the PCs by computing the Euclidean distance using Ward's minimum variance method. To visualize the processes of the hierarchical clustering step, a dendrogram was produced showing the stepwise process of case groupings. Then, for a range of potential cluster numbers, it calculates the within- and between-group sum of squares (also termed inertia) and chooses the number of clusters for which the change in between-group variance is minimized. The optimal number of clusters was determined based on the elbow method by plotting the within-cluster variances for different numbers of clusters, the inertia gains, and the shape of the dendrogram tree. After cutting the tree to the desired number of clusters, a k-means consolidation was performed to increase the clusters' robustness, which is included in the HCPC method by default. The robustness of the selected clusters was assessed with the Jaccard similarity index [182] produced by the bootstrap function (clusterboot) in the 'fpc' package [183]. Lastly, descriptive statistics were computed to assess the distinctiveness of clusters. To determine where the significant differences lay between each pair of the clusters found, post hoc comparisons (*Kruskal-Wallis* and Fisher's exact tests) across each of the clusters and descriptive variables were conducted [184, 185]. Considering the increase in the Type I error owing to multiple comparisons, univariate tests were also performed for each of the variables with a Bonferroni correction to the significance level (at 0.05), considering a total of 18 variables (7 numeric and 11 categorical) to corrected p-value of 0.003. [186].

Chapter 4

Results

4.1 Descriptive Statistics

The current sample included 661 individuals, 586 (88.65%) were males and 75 (11.35%) were females, which is significantly higher (8:1) than the ASD general population gender ratio of 4:1 (Table 4.1). Children's ages at enrollment ranged from 1 to 8 years old, with a mean age of 3.67 (SD = 1.56) years (Table 4.1). Participants present varying symptoms and characteristics, as seen by the wide ranges and large variances in all variables. Characteristics of the individuals in the study were similar to those excluded in terms of VABS Communication domain, ADI-R AbDev domain, and almost all GMDS subescales. However, children excluded from the study presented more cases of Non-Autism based on the ADOS classification, lower scores on ADI-R Social and RRB domains as well as in GMDS' Locomotion subscale domain than those in the study individuals (Table 4.1). Moreover, excluded individuals also presented higher scores for VABS' Socialization and Daily Living Skills subscales (Table 4.1). Therefore, it's important to highlight that the study might not be capturing the full spectrum of autism, as individuals with certain characteristics were excluded.

Most of the variables were not significantly correlated (Figure 4.1). However, in addition to the strong correlations between pairs of subscale scores within each scale (e.g., the correlation coefficient of VABS communication and VABS socialization was 0.93), the ADI-R Socialization domain is inversely correlated with Griffiths' Hearing and Speech (r = - 0.5) and Personal/Social (r = -0.47) subscales (Figures 4.1 and 7.4).

Table 4.1: Comparison of main characteristics between the study population (n = 661) and the Excluded one (n = 354) in children with ASD.

Characteristics	Study Population (N = 661)	Excluded Population (N = 354)	P value[a]
Gender	N (%)	N (%)	
Male	586 (88.65)	295 (82.87)	0.013
Female	75 (11.35)	61 (17.13)	
Clinical Severity (ADOS)			
Autism	443 (79.53)	126 (54.08)	< 0.001
Autism Spectrum Disorder	108 (19.39)	66 (17.60)	
Non-Autism	6 (1.08)	41 (28.32)	
	N = 480	N = 188	
Age at inclusion (years)	3.67 (1.56)	6.61 (3.62)	< 0.001
	Mean (SD)	Mean (SD)	
ADIR Subdomains	N = 638	N = 287	
Social	18.97 (6.52)	16.20 (9.31)	< 0.001
	N = 637	N = 287	
Restricted Repetitive Behaviour	4.49 (2.00)	4.11 (2.77)	0.040
	N = 629	N = 276	
Abnormality of development evident at or before 36 months	4.03 (1.56)	3.79 (1.44)	0.073
VABS Subdomains	N = 524	N = 191	
Comunication	63.8 (22.35)	67.25 (19.25)	0.102
	N = 524	N = 191	
Daily Living Skills	60.47 (19.4)	66.89 (18.39)	< 0.001
	N = 526	N = 191	
Socialization	66.18 (17.54)	71.43 (16.15)	0.001
Griffiths Subdomains	N = 628	N = 256	
Locomotion	87.17 (20.60)	81.34 (28.09)	0.027
	N = 628	N = 256	
Personal/Social	70.35 (22.23)	68.41 (27.79)	0.325
	N = 640	N = 255	
Hearing and Speech	63.01 (32.09)	61.24 (34.01)	0.230
	N = 628	N = 256	
Eye-Hand Coordination	74.62 (25.98)	71.27 (28.77)	0.119
	N = 628	N = 256	
Performance	81.78 (27.32)	78.20 (32.90)	0.142

[a] *Mann–Whitney* test for quantitative variables and Chi^2 test for qualitative variables

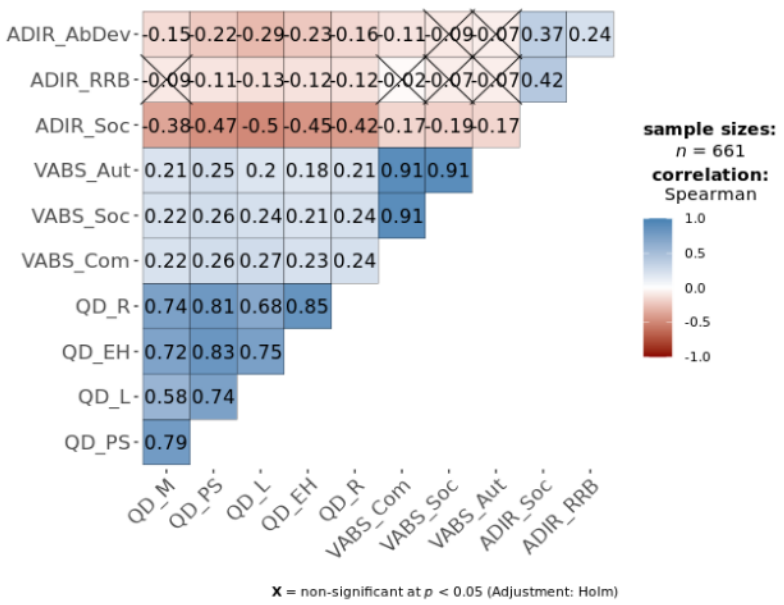


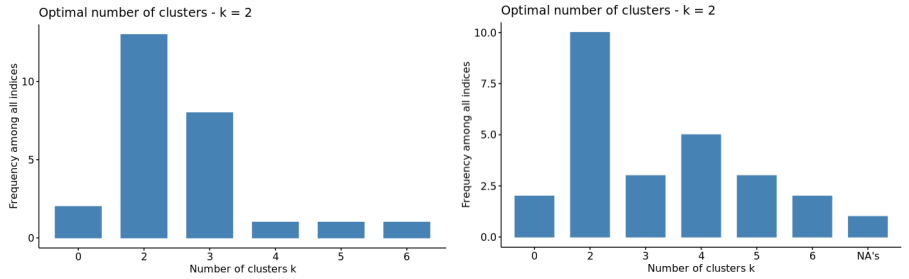
Figure 4.1: Correlogram of Spearman Correlations among Cluster Variables

Note: Each block in the chart represents the correlation between two variables that share the same row and column as the block. The numerical values enclosed within the blocks denote the correlation coefficients for the corresponding variable pairs. The color of the blocks conveys the strength of the correlation, based on the spectrum on the right-hand side of the chart. The darker shades of color signify larger correlation coefficients, with blue indicating a positive correlation and red representing a negative correlation. When a block is white, it signifies that the variables in the corresponding row and column are not significantly correlated

As previously mentioned in Section3.5, we evaluated different clustering methods for internal validity and stability using two different functions. The results obtained were not consistent since different validation measures point to different clustering methods and different optimal numbers of clusters (Table 4.2 and Figure 4.2).

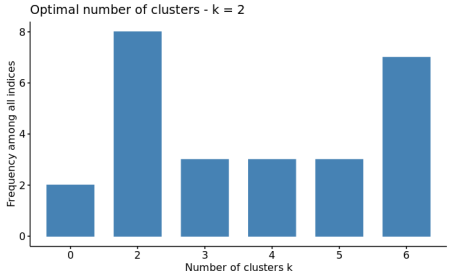
Table 4.2: Cluster Validation Measures

Validation Measures		Score	Method	Clusters
Internal	Connectivity	91.771	Hierarchical	2
	Dunn	0.187	CLARA	2
	Silhouette	0.372	Kmeans	2
Stability	APN	0.029	DIANA	2
	AD	0.837	SOTA	6
	ADM	0.033	DIANA	2
	FOM	0.196	Kmeans	6

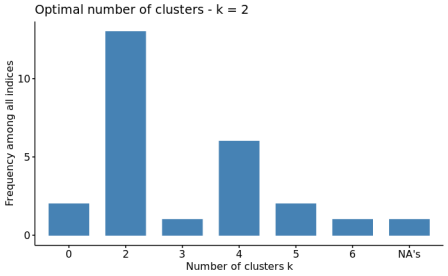


(a) Kmeans

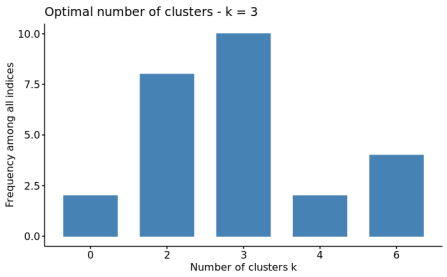
(b) Centroid



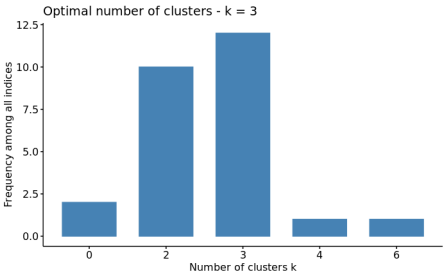
(c) Median



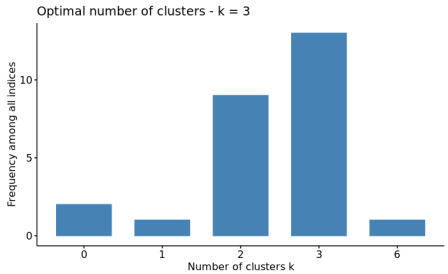
(d) Mcquitty



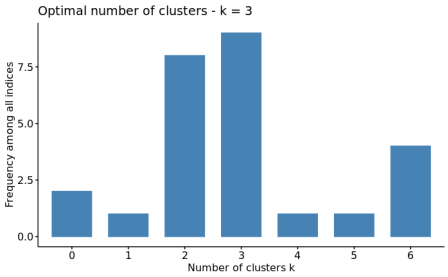
(e) Ward.D2



(f) Average



(g) Complete



(h) Single

Figure 4.2: Plot of Optimal Number of Clusters for the Selected Clustering Methods using NbClust

Moreover, these clustering methods did not have into account the inter-correlated domains which could overly affect the results. The significant correlations found across variables (Figure 4.1) suggested that certain variables might account for overlapping variances within the sample. As a result, it is possible to use PCA to reduce the dimensions of the feature variables for cluster analysis by identifying PCs that explain the majority of the variance.

4.2 Principal Component Analysis

PCA results identified 5 PCs that accounted for 89.41% of the sample variation (Figure 4.3 and Table 7.1). The identification of the five PCs enabled the data dimension to be reduced from 11 variables to five components. Among the five PCs, the first two PCs had absolute Eigenvalues larger than one (Table

7.1). As the criterion of absolute Eigenvalues larger than one is arguably overly conservative, in the current analysis, more PCs were retained in order to maintain the majority of the information (89% of the variance) for the cluster analysis (Figure 4.3 and Table 7.1).

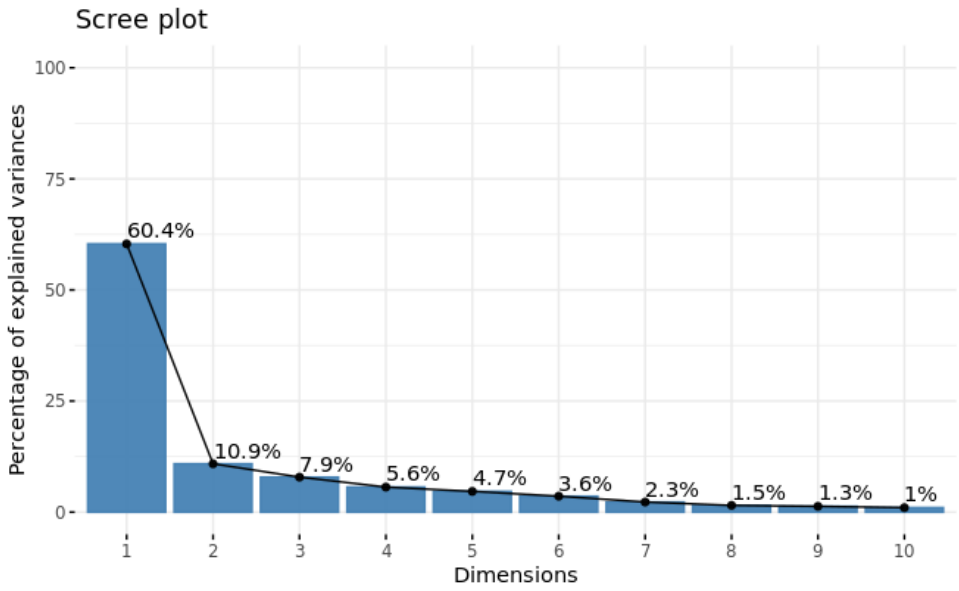


Figure 4.3: PCA Scree Plot

The factor structures of the five PCs were dissected to show how variables contribute to the PCs: PC 1 represented 60.37% of the variance and was highly correlated with the VABS and GMDS subscales as well as the ADI-R Socialization domain (Figures 4.3 and 7.5 and Table 4.3). PC 2 accounted for 10.87% of the variance and captured ADI-R domains (Figures 4.3 and 7.6 and Table 4.3). PC 3 took on 7.87% of the variance and primarily correlated with RRB and Abnormality of Development Evident At or Before 36 months domains (Figure 4.3 and Table 4.3). PC 4 presented 5.63% of the variance and was mainly correlated with VABS Daily Living Skills and Socialization subscales as well as the Griffiths Locomotion subscale (Figure 4.3 and Table 4.3). Finally, PC 5 explained 4.66% of the variance and was strongly correlated with the ADI-R Socialization domain (Figure 4.3 and Table 4.3).

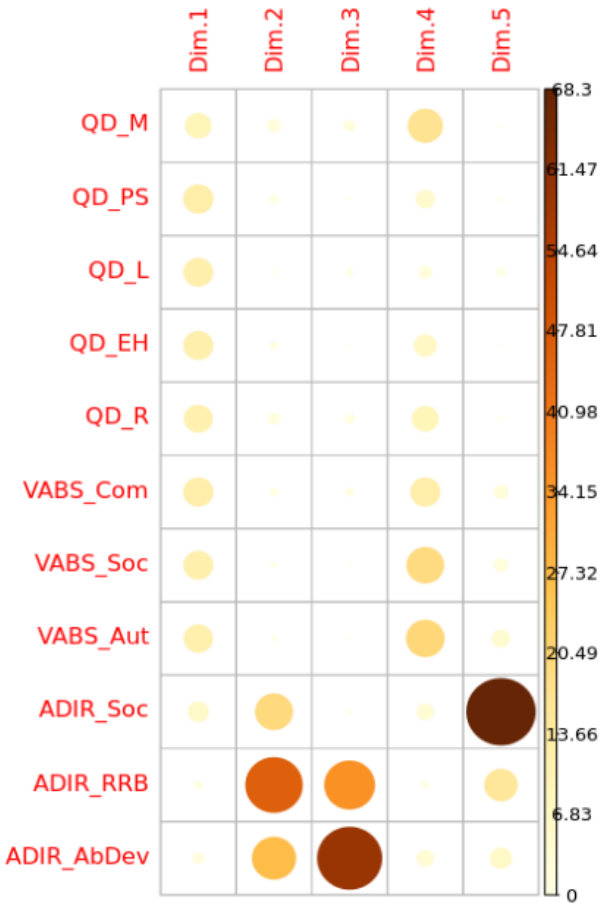


Figure 4.4: Principal Components and Cluster Variables Correlation Plot

Note: Each dot in the chart highlights the most contributing variables for each Principal Component. Based on the spectrum on the right-hand side of the chart, the numbers indicate the percentage of these contributions, with darker shades and lighter shades of color indicating higher and lower contributions, respectively.

Table 4.3: Variables' Correlation Coefficient and Contributions to Principal Components

Principal Component 1			Principal Component 2			Principal Component 3		
Percentage of variance explained		60.37%	Percentage of variance explained		10.87%	Percentage of variance explained		7.87%
Subdomain Variables	r[a]	Contribution[b] (%)	Subdomain Variables	r[a]	Contribution[b] (%)	Subdomain Variables	r[a]	Contribution[b] (%)
VABS Comunication	0.90	12.27	ADIR RRB	0.74	46.34	ADIR DevAb	0.72	59.59
Griffiths Personal/Social	0.90	12.22	ADIR DevAb	0.57	27.01	ADIR RRB	-0.56	35.84
Griffiths Eye-Hand Coordination	0.89	11.94	ADIR Socialization	0.49	19.76	Griffiths Locomotion	0.12	1.62
VABS Socialization	0.89	11.89	Griffiths Locomotion	0.16	2.12	Griffiths Performance	0.10	1.23
Griffiths Hearing and Speech	0.89	11.83	Griffiths Performance	0.14	1.61	VABS Comunication	-0.09	0.90
VABS Daily Living Skills	0.87	11.45	Griffiths Personal/Social	0.10	0.90	Griffiths Hearing and Speech	-0.07	0.53
Griffiths Performance	0.86	11.14	Griffiths Eye-Hand Coordination	0.10	0.78	ADIR Socialization	-0.04	0.19
Griffiths Locomotion	0.78	9.24	VABS Comunication	0.10	0.77	Griffiths Personal/Social	0.02	0.06
ADIR Socialization	-0.62	5.76	VABS Socialization	0.06	0.33	VABS Daily Living Skills	0.02	0.03
ADIR DevAb	-0.32	1.50	VABS Daily Living Skills	0.06	0.28	Griffiths Eye-Hand Coordination	0.01	0.01
ADIR RRB	-0.22	0.76	Griffiths Hearing and Speech	0.03	0.1	VABS Socialization	-0.01	0.01

Principal Component 4			Principal Component 5		
Percentage of variance explained		5.63%	Percentage of variance explained		4.66%
Subdomain Variables	r[a]	Contribution[b] (%)	Subdomain Variables	r[a]	Contribution[b] (%)
VABS Daily Living Skills	-0.36	20.45	ADIR Socialization	0.59	68.30
VABS Socialization	-0.35	19.45	ADIR RRB	-0.28	15.24
Griffiths Locomotion	0.32	17.02	ADIR AbDev	-0.18	6.26
VABS Comunication	-0.28	12.25	VABS Daily Living Skills	0.15	4.34
Griffiths Performance	0.24	9.3	VABS Comunication	0.11	2.44
Griffiths Eye-Hand Coordination	0.21	6.82	VABS Socialization	0.11	2.34
Griffiths Personal/Social	0.17	4.77	Griffiths Hearing and Speech	-0.07	0.85
ADIR AbDev	-0.15	3.75	Griffiths Personal/Social	-0.03	0.19
ADIR Socialization	0.15	3.58	Griffiths Locomotion	0.01	0.04
Griffiths Hearing and Speech	0.11	2.03	Griffiths Eye-Hand Coordination	0	0.00
ADIR RRB	-0.06	0.55	Griffiths Performance	0.00	0.00

[a] Correlation coefficients between the subdomain scores and principal components.

[b] Contribution = r^2/(total r^2 of the component) × 100%

4.3 Hierarchical Clustering

The Hierarchical Clustering produced a three-cluster solution based on the findings of the dendrogram and inertia graph (Figures 4.5, 4.5a and 7.7). The dendrogram showed that in the first to last hierarchy, the cases merged into three clusters (Figure 4.5). According to the elbow method graph, the three-cluster solution gained notably larger within-cluster variance than the two-cluster solution, which clearly highlighted the elbow point (Figure 4.5b).

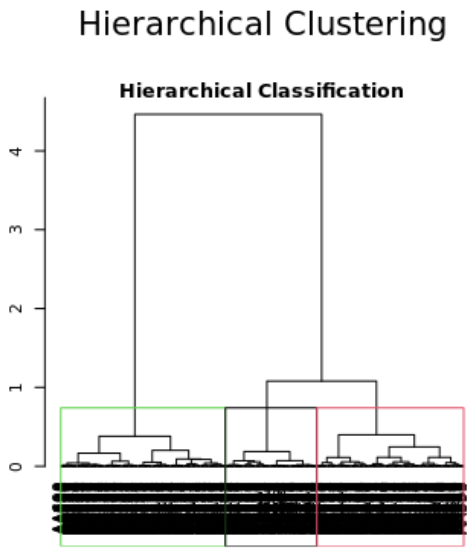
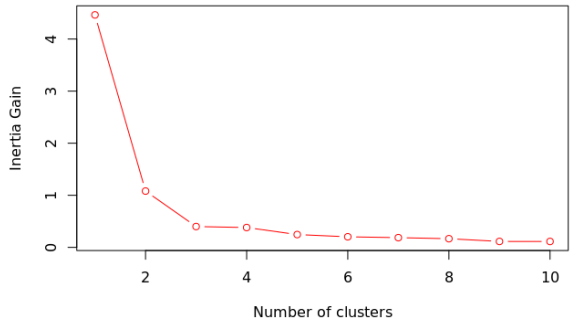
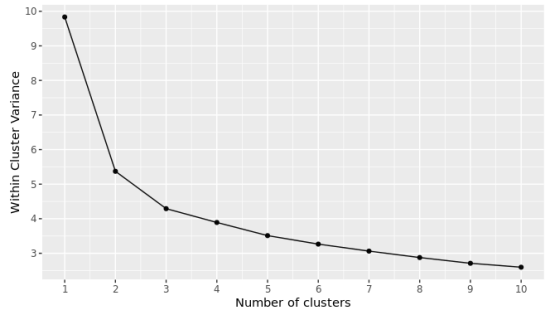


Figure 4.5: Hierarchical Clustering Plot. The x-axis represents individual cases, and the hierarchical brackets above them indicate the hierarchical clustering at each level. At the third to last hierarchy, the current dendrogram revealed that the cases were grouped into three clusters



(a) Inertia Gain Scree Plot



(b) Elbow Method Plot. Note: The elbow method graph demonstrates that the within-cluster variance of the three-cluster solution was significantly higher than that of the two-cluster solution. Moreover, the ratio of inertia gain between the three-cluster and four-cluster solutions was lower (ratio = 0.40) compared to the ratios of the two-cluster to three-cluster (ratio = 1.08).

The three-cluster solution was therefore selected for the current sample, with 194 (29%) children in Cluster 1, 209 (32%) in Clusters 2, and 258 (39%) in Cluster 3 (Table 4.5). The scatterplot was constructed on the first two PCs to visualize the distribution of each cluster and revealed that the three clusters were well separated on the first two PCs (Image 4.6). Specifically, Cluster 1 and Cluster 3 were separated by both PC 1 and PC 2 while Cluster 2 was majorly separated from Cluster 1 and 3 on PC2. Overall, the cluster's validation through bootstrapping showed that the three clusters were true and consistent (Table 4.2).

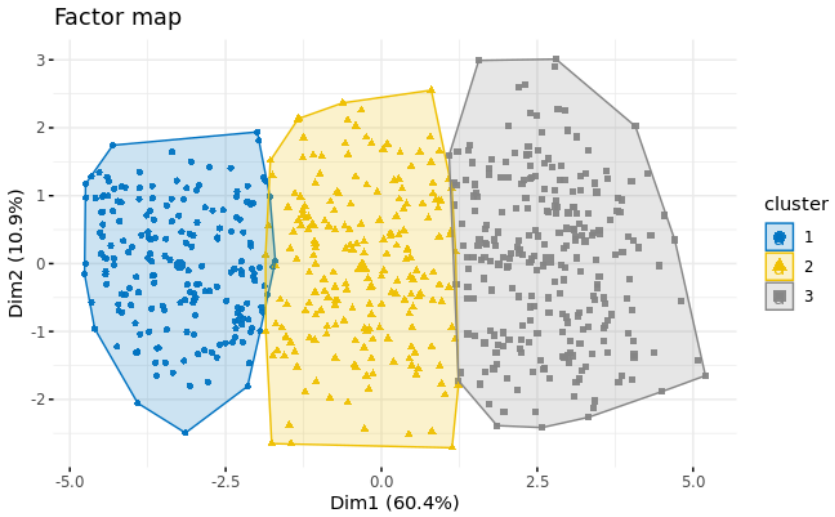


Figure 4.6: Scatter Plot of the Three Cluster on the First Two Principal Components

4.4 Subtypes Phenotypic Description

The descriptive statistics and post hoc comparison results were detailed in Table 4.4. Pairwise Wilcox tests indicated that almost all of the input variable distributions were significantly different between the three subtypes (clusters). As can be observed, all of the eleven variables showed a significant difference between Clusters 1 and 3. Apart from ADI-R RRB and AbDev domains, all of the comparisons between Clusters 1 and 2 were significant at the nominated level. Similarly, all comparisons between Clusters 2 and 3 except RRB were at the nominated significance level. Furthermore, these results show that Cluster 3 scored significantly higher than Cluster 2 on VABS and GMDS subscales, but not on the ADI-R domains. On the other hand, Cluster 1 scored significantly higher than Cluster 3 only in ADI-R domains. Cluster 1 is significantly different from Cluster 2 for all variables except for RRB and AbDev domains. To further describe the nature of the three subtypes obtained in this study, we used the non-imputed data and conducted *Kruskal-Wallis* and Fisher exact tests on the cluster variables as well as the seven numerical variables and eleven categorical variables, that were selected for the descriptive analysis (Table 4.5, 4.7 and 4.6). Univariate tests also showed significant effects, as presented in Tables 4.5 and 4.6. Again, it was necessary to perform post hoc comparisons across the three subtypes to determine where the actual between-subtypes differences lay and the results of this analysis. Cluster 1 comprised the samples that had the most delays, with above-sample average scores for all ADI-R domains and below-sample average scores for all VABS and GMDS subscales (Table 4.5 Cluster 1). Children in Cluster 1 also had impaired language skills, children were much older when they started to talk when compared to individuals in Cluster 3 and with the highest percentage of non-verbal individuals when compared to the other two clusters. Regarding walking, children also started later than those in Cluster 3 (Table 4.6). Cluster 1 also has the highest number of female patients and the most likely to have a DSM-IV diagnosis of Autism Disorder. Moreover, a substantial number of children has always shown a psychomotor development delay when compared to numbers found in the other two clusters (Table 4.7). In short, Cluster 1 generally had the highest symptom-presence scores with the most delays/impairments across developmental areas among the three clusters. Cluster 2 included 32% of the current sample and was characterized by a milder developmental and behavioral profile (Table 4.5 Cluster 2). Children in Cluster 2 did not present significantly different levels to Cluster 1 in ADI-R RRB and AbDev domains

but presented lower scores in the Social domain. Cluster 2 was significantly different from Cluster 1 having higher scores on Social and AbDev domains. On the other hand, Cluster 2 had significantly lower scores in all subscales of VABS and GMDS than Cluster 3 but higher than the scores of Cluster 1 (Table 4.5). Moreover, with no significant difference from Cluster 1, children in Cluster 2 also presented high levels of severity in language ability, as they were older than children of Cluster 3 when they first started to say words and construct phrases (Table 4.6). Additionally, the majority of children in Cluster 2 have been identified as non-verbal but still, have more verbal cases than Cluster 1. The majority of children in Cluster 2 are identified as having Autism however, this cluster has more cases of patients identified with ASD than Cluster 1 (Table 4.7). Also, more than 50% have always presented a delay in psychomotor development. Thus, children in Cluster 2 showed salient impairments in the core ASD symptom area, despite relatively high development and intellectual abilities compared to Cluster 1 (Table 4.6). Cluster 3 was the largest cluster with above-sample average abilities across all VABS and GMDS subscales (Table 4.5 Cluster 3). On the other hand, children in Cluster 3 had lower ADI-R scores within the sample. Among the three clusters, children in Cluster 3 had the lowest severity scores across all subdomains. Children in Cluster 3 started speaking relatively sooner than those in Cluster 1 and Cluster 2. Moreover, they also started walking sooner than those in Cluster 1 4.6). Cluster 3 had the lowest number of female patients and higher cases of patients being identified as ASD and Non-Spectrum than the other two clusters combined, as well as more cases classified as negative by ADI-R. Relatively to verbal status, Cluster 3 had the highest number of verbal children when compared to Cluster 1 and 2. Additionally, the majority of these patients did not always have a delay in psychomotor development, unlike the other two clusters (Table 4.7). In summary, children in Cluster 3 tended to have relatively higher language skills and adaptive behaviors and fewer ASD-related deficits compared to the other two clusters.

Table 4.4: Cluster Comparisons of Scores for Developmental and Behavioral Characteristics (SD)

			Cluster 1 N = 194 (29%)		Cluster 2 N = 209 (32%)		Cluster 3 N = 258 (39%)		KW results		Post hoc test
Measures	Developmental Domain	Subdomains	Mean	SD	Mean	SD	Mean	SD	Chi^2 (df=2)	P-value	
ADI-R (Standard Scores)	Social Ability	Social	24.49	4.85	18.40	5.45	15.26	5.23	230.05	< 0.001	1 > 2 > 3
	Repetitive Behaviours	Restricted Repetitive Behaviour	4.89	1.77	4.42	1.78	4.20	2.20	20.73	< 0.001	1 > 2 ; 1 > 3
	Developemental Abnormalities	Abnormality of development evident at or before 36 months	4.49	1.75	4.16	1.19	3.52	1.47	55.65	< 0.001	1 > 3 ; 2 > 3
VABS-II (Standard Scores)	Communication	Comunication	38.13	11.99	59.46	11.24	82.24	11.58	500.34	< 0.001	3 > 2 > 1
	Adaptive Behaviour	Daily Living Skills	38.30	13.41	59.34	10.18	74.97	9.33	457.70	< 0.001	3 > 2 > 1
	Social Ability	Socialization	47.51	11.65	64.52	9.05	79.15	9.80	467.53	< 0.001	3 > 2 > 1
GMDS (Standard Scores)	Gross Motor Skills Ability	Locomotion	67.73	16.40	86.35	12.64	102.45	13.98	348.71	< 0.001	3 > 2 > 1
	Adaptive Behaviour	Personal/Social	46.72	12.43	67.94	11.30	90.30	13.79	474.63	< 0.001	3 > 2 > 1
	Receptive and Expressive Language	Hearing and Speech	30.77	14.37	56.99	18.95	93.67	19.83	467.66	< 0.001	3 > 2 > 1
	Fine Motor Skills Ability	Eye-Hand Coordination	45.75	14.13	73.33	14.07	97.44	14.92	464.31	< 0.001	3 > 2 > 1
	Visuospatial Skills Ability	Performance	52.63	15.91	81.08	15.58	104.19	17.74	430.15	< 0.001	3 > 2 > 1

Table 4.5: Cluster Comparisons of Non-Imputed Scores for Developmental and Behavioral Characteristics (SD)

		Cluster 1 N = 194 (29%)		Cluster 2 N = 213 (32%)		Cluster 3 N = 256 (39%)		KW results		Post hoc test
Measures	Range	Mean	SD	Mean	SD	Mean	SD	Chi^2 (df=2)	P-value	
ADIR (Standard Scores)										
Social	0-30	24.58	4.90	18.38	5.56	15.25	5.29	222.63	< 0.001	1 > 2 > 3
Restricted Repetitive Behaviour	0-23	4.91	1.79	4.43	1.82	4.20	2.22	21.06	< 0.001	1 > 2 ; 1 > 3
Abnormality of development evident at or before 36 months	0-25	4.53	1.77	4.18	1.22	3.53	1.49	54.26	< 0.001	1 > 3 ; 2 > 3
VABS-II (Standard Scores)										
Comunication	20-121	36.00	12.33	59.54	12.39	82.34	12.07	373.85	< 0.001	3 > 2 > 1
Daily Living Skills	20-111	36.32	14.99	59.06	11.34	74.98	9.80	335.62	< 0.001	3 > 2 > 1
Socialization	20-106	45.36	13.46	64.50	10.08	79.21	10.29	344.31	< 0.001	3 > 2 > 1
GMDS (Standard Scores)										
Locomotion	28-157	67.61	16.63	86.36	12.94	103.08	14.24	329.92	< 0.001	3 > 2 > 1
Personal/Social	19-161	46.69	12.53	67.81	11.43	90.77	14.14	448.97	< 0.001	3 > 2 > 1
Hearing and Speech	0-158	30.78	14.36	56.88	18.77	94.41	20.24	449.39	< 0.001	3 > 2 > 1
Eye-Hand Coordination	15-150	45.76	14.26	73.39	14.33	98.11	15.23	440.18	< 0.001	3 > 2 > 1
Performance	14-166	52.56	16.09	81.22	15.95	105.08	18.00	409.73	< 0.001	3 > 2 > 1

Table 4.6: Statistic Analysis of Non-Imputed Numeric Variables across Clusters (SD)

		Cluster 1 N = 194 (29%)		Cluster 2 N = 209 (32%)		Cluster 3 N = 258 (39%)		KW results		Post hoc test
Measures	Range	Mean	SD	Mean	SD	Mean	SD	Chi^2 (df=2)	P-value	
Head Circumference (cm)	10-75	34.50	1.78	35.09	4.44	34.53	2.86	0.26	0.876	
Apgar 1º	3-10	8.37	1.51	8.59	1.21	8.66	1.08	2.41	0.300	
Apgar 5º	7-10	9.76	0.60	9.82	0.55	9.87	0.42	3.84	0.147	
Diagnosis Age (months)	1-8	3.61	1.58	3.63	1.45	3.75	1.63	0.58	0.749	
Walking Age (months)	8-22	14.44	2.72	14.16	2.59	13.72	2.48	8.72	0.013	1 > 3
First Words Age (months)	9-50	25.09	11.06	24.38	10.44	21.37	8.65	12.12	0.002	1 > 3 ; 2 > 3
First Phrases Age (months)	18-66	46.00	13.28	44.53	11.10	38.40	9.28	41.27	0.001	1 > 3 ; 2 > 3

Table 4.7: Statistic Analysis of Non-Imputed Categorical Variables across Clusters

Measures	Value	Cluster 1 N = 194 (29%) N	(%)	Cluster 2 N = 213 (32%) N	(%)	Cluster 3 N = 256 (39%) N	(%)	Fisher's Exact Test
Gender	Female	37	19.07	21	10.04	17	6.59	< 0.001
	Male	155	80.93	188	89.95	241	93.41	
Clinical Severity (ADOS)	Autism	124	96.12	159	87.36	160	65.04	< 0.001
	ASD	5	3.88	22	12.09	81	32.92	
	Non Spectrum	0	0.00	1	0.55	5	2.03	
ADIR Quotation	Positive	186	98.94	189	92.04	206	81.89	< 0.001
	Negative	2	1.06	16	7.96	46	18.11	
Dysmorphysms	Yes	19	10.11	18	8.91	17	6.83	0.483
	No	169	89.89	184	91.09	232	93.17	
Language Regression	Yes	29	15.76	26	12.87	20	8.37	0.055
	No	155	84.24	176	87.13	219	91.63	
Audition	Normal	184	97.35	199	98.03	241	97.97	0.889
	Not Normal	5	2.65	4	1.97	5	2.03	
Vision	Normal	156	88.64	176	91.19	207	88.84	0.672
	Not Normal	20	11.36	17	8.81	26	11.16	
Verbal	Yes	36	19.15	77	38.69	161	64.14	< 0.001
	No	152	80.85	122	61.31	90	35.86	
Family Psychiatric History	Yes	89	46.60	100	48.54	100	39.84	0.141
	No	102	53.40	106	51.46	151	60.16	
PMD Regression	Yes	13	7.22	10	5.03	7	2.92	0.121
	No	167	92.78	189	94.97	233	97.08	
PMD Delay Always	Yes	149	79.68	116	57.43	115	45.45	< 0.001
	No	38	20.32	86	42.57	138	54.55	

Chapter 5

Discussion

This analysis identified three distinct subtypes among 661 children with ASD based on five PCs derived from domains of three diagnostic tools (Table 4.4). The participants were divided into three distinct subtypes using data from standardized tests, and most of these variables showed a significant difference between Cluster 1 (the severe-functioning group), Clusters 2 (the moderate-functioning group), and Cluster 3 (the mild-functioning group), indicating that Cluster 1 has more severe ASD-related characteristics than Clusters 2 and 3. This finding is supported by the elevated ADI-R's Social and RRB scores for Cluster 1 which indicates that the severity of the identified ASD core symptoms is greater for this cluster than for Cluster 2 or 3 (Table 4.4). Cluster 1 also scored lower on the other standardized tests which measured the general level of adaptive and behavioral function. Therefore, it appears that the children with ASD in this subgtype have higher requirements or are less functional. It was possible to determine whether each of the three subtypes could be described by a set of impairments by analyzing the descriptive data for the three subtypes. Compared to the rest of the sample, Cluster 1 children had relatively lower language and adaptive abilities and higher severity for social symptoms, repetitive behaviors, and development abnormalities. Children in Cluster 2 had similar levels of developmental anomalies as Cluster 1 and similar RRB levels to both Cluster 1 and 3, but higher severity levels in social, communication, and adaptive behavior than Cluster 3. Children in Cluster 3 earned the highest scores in language and adaptive abilities and the lowest severity across social and developmental symptoms of all three subtypes, showing the least impairments, with its peak impairment in RRB (Table 4.4). The findings reported here show that diagnostic distinctions between Clusters 1, 2, and 3 can be made based on the hypothesis put forth by Waterhouse et al that less evidence of atypical performance in adaptive behavior is an important basis for making distinctions within the autism spectrum [187]. The first two PCs served as the main distinguishing factors between these three subtypes, with PC 1 representing adaptive and social behavior as well as mental development variables and PC 2 reciprocal social interactions, and repetitive behaviors/interests (Table 4.3). Compared to the conventional approach of using individual items or subdomains as clustering features, the HCPC method has some advantages. In order to explore heterogeneity, the PCA reduced dimensions to meaningful features and accounted for common variances while retaining the greatest variance in the sample. As a result, the PC-based clusters were more likely to be replicated, which could potentially lead to more consistent results across samples. Additionally, the PCA in the HCPC revealed insights into the relationships and constructs within the domains across several measurements. PC 1 was highly correlated with the social and adaptive behavior abilities measured on the VABS and GMDS, indicating that scores on both tools covaried and likely captured the same domain of adaptive behavior in the current sample. This result was in line with other research indicating relationships between these subdomains on Griffiths and Vineland [188]. PC 2, as well as PC 3 and PC 5, were

correlated with domains measured on the ADI-R with high contribution of RRB, Daily Living Skills, and Socialization, respectively. PC 4 primarily measured Daily Living Skills on VABS. The number of meaningful clusters derived from the current data is consistent with prior findings [189, 154, 150]. Concerning core ASD symptoms, we found that the severity of deficits in Social and AbDev domains varied across subgroups, regardless of children's functioning in other subscales: Cluster 3 was characterized the less severe deficits in both core symptoms, while more severe deficits characterized Cluster 1 and 2. On the contrary, previous studies discovered subgroups with heterogeneous levels of social and RRBs (e.g., high social severity and low RRBs, or low social severity and high RRBs) [154, 136]. This study mainly focused on younger children, thus it's possible that the symptom profiles would differ among children of this age with a different ascertainment source. According to previous research [190, 191, 135], we discovered two subtypes that could be primarily distinguished by their level of adaptive and behavior ability and overall severity of ASD: Cluster 1 had severe symptom severity as well as severe dysfunction in adaptive behavior and language skills and Cluster 3 showed the opposite profile. However, we also identified a subtype (Cluster 2) with moderated dysfunctions in adaptive behavior and language skills. Children in Cluster 2 showed comparably higher levels of mental and language ability and adaptive behavior functioning than in Cluster 1 but significantly lower than Cluster 3 (Table 4.4 and 4.5). The social deficits and unusual repetitive behaviors and interests of children in Cluster 2 may seem more noticeable to caregivers because of this group's higher adaptive behavior ability than those of children in Cluster 1 [192].

Researchers such as Prior et al. have suggested that rather than specific symptom patterns, the distinguishing factor amongst ASD-based subtypes is related to the severity of social and cognitive impairments [193]. This study backed the idea of a spectrum for autism, which places children on a continuum from severe functioning to mild functioning [194, 195, 152, 150]. The idea of a spectrum of autism-based disorders was also formerly supported by Sevin et al., who argued that the severity of impairment in the main characteristics of ASD (i.e. social communication and interaction and restricted and ritualistic repetitive behaviors) is the most significant basis for understanding the variety of behaviors that are identified as belonging to the autism spectrum. The current study accepts and confirms the existence of an "autism spectrum" and offers some support for characterizing the symptom profile at various positions along this spectrum [196, 194, 195, 152, 150]. These varied subtypes and the differences they show from some other research emphasize the importance of accurate evaluations and gathering information from a variety of sources (e.g., caregivers, service providers, and professionals) [197]. In order to comprehend children's strengths and weaknesses across domains, it is advised that researchers and practitioners administer numerous developmental and behavioral assessments reported by multiple informants beyond diagnostic tests. When collecting caregiver-reported information to capture child behaviors and needs across settings, caregivers' opinions and experiences should be valued. In the past, different diagnostic labels (such as Asperger's, (PDD-NOS, etc.) were used to define "subtypes" of individuals with ASD. Even though these categorical categories were abandoned in the DSM-V due to their poor reliability and validity, there is consensus among academics and clinicians that some other methods of identifying subtypes within ASD are urgently needed [198, 199]. The results of the current subtype study made clear how important it is to include all aspects of development and behavior as separate variables when subgtyping. The employment of more advanced methods, like HCPC, would benefit data-driven approaches designed to capture a continuum of several developmental domains. In the future, studies could validate subtype profiles with various samples by employing model-based subtyping techniques (such as latent class analysis). Furthermore, given the progress made in ASD genetics, it may be interesting to include genetic traits in subtyping attempts [148, 155]. For our understanding of prognosis, developmental tra-

jectories, and outcome prediction, it is also crucial to monitor the development of confirmed subgroups longitudinally [194, 200]. Even though some studies have looked at the outcomes of predetermined subtypes on the autism spectrum, the subtype's findings from the current sort of analysis could offer empirically derived subtypes to guide the exploration of differential outcomes and related influential factors [201, 202, 203].

Limitations

One notable constraint is the presence of missing data within the study population. The absence of complete information poses a challenge, as it can potentially lead to biased or incomplete cluster assignments, thereby compromising the accuracy and generalizability of the findings. Additionally, the small sample size (N = 661) employed in the cluster analysis warrants consideration as another significant limitation. A larger sample could potentially include more cases with different features and enhance the likelihood that there would be suitable sampling to identify all potential subgroups. More samples would also be advantageous for Cluster Analysis because they would enable cross-validation using both training and validation data [204]. The current study, however, did not incorporate a validation analysis to reproduce the subgroup findings with a different sample of ASD children. Additionally, because the majority of the current sample's autistic children are male, generalizability to other ASD subtypes may be constrained. Moreover, the study focused on individuals aged 1 to 8 years old. Once again, this indicates that the study's sample is not entirely representative of the entire population of individuals diagnosed with autism. Therefore, these results require replication in a separate sample of individuals on the spectrum. Future research should also take into account the heterogeneity caused by additional characteristics, such as age, gender, and the prevalence of comorbid conditions, which were not taken into account in this study. Another potential limitation is the fact that the Cluster Analysis was based on data gathered at the time the child enrolled in the original study, which means that the data and any inferences that come from it are restricted to this particular point in the child's life. Although the use of data from a one-time point helps control additional sources of variance, future research should look at how empirically derived subtypes change over time as a result of developmental evolution or intervention [202, 200]. Lastly, the PCA in the HCPC analytic model was employed as a data reduction strategy and only produced preliminary results about the relationships between the developmental aspects of ASD in children. In order to shed more light on the heterogeneity of ASD, additional analyses are required to discover latent factor structures and linkages collected across various measures and to reveal more information about correlations between frequently used developmental and behavioral measures in the ASD community.

Moreover, the study emphasized the importance of accurate evaluations and gathering information from multiple sources, such as caregivers, service providers, and professionals, to comprehensively understand children's strengths and weaknesses across various domains. Finally, the research contributes to the broader understanding of ASD as a spectrum disorder, confirming the existence of an autism spectrum and offering support for characterizing symptom profiles at different positions along this spectrum. This study highlights the need for continued research to validate subgroup profiles, include genetic traits in subtyping attempts, and monitor the development of confirmed subgroups longitudinally for a better understanding of prognosis and outcomes.

Chapter 6

Conclusions

The premise behind research that seeks to categorize certain behavioral variations or subtypes of autistic children is that doing so will enable us to better comprehend distinct etiologies. Genetic investigations of ASD have focused on locating major genes (or even genes of modest effects) or identifying chromosomal regions with high probabilities of harboring susceptibility genes. However, these investigations have not been very successful. With this work, we attempted to determine whether specific behavioral phenotypes in a sample of autistic children could be distinguished using autistic behavioral symptoms and developmental measures. Study results reveal that by concurrently taking into account several domains of development and behavior, the heterogeneity of ASD in children can be better described. Using a Hierarchical Cluster Analysis, three distinct subtypes were identified, with significant differences between the subtypes in terms of ASD-related characteristics, severity, and levels of functioning. Cluster 1 consisted of children who received a diagnosis of ASD and had lower levels of adaptive behavior abilities, as well as higher severity of social symptoms, repetitive behaviors, and developmental anomalies. Cluster 2 had similar levels of developmental anomalies as Cluster 1, but higher severity levels in social and adaptive behavior than Cluster 3. Cluster 3 earned the highest scores in language and adaptive abilities and had the lowest severity across social and developmental symptoms of all three clusters, indicating the least impairments.

The implications of this research extend beyond mere categorization. By incorporating numerous symptom profile characteristics in the subtyping process, we can gain valuable insights into differential treatment responses. This understanding is vital in developing personalized interventions that can maximize developmental outcomes for children with ASD. Furthermore, this research opens avenues for investigating the reasons behind optimal developmental outcomes in certain individuals, shedding light on the factors that contribute to positive outcomes.

To ensure the long-term impact and applicability of these findings, further research is needed. Longitudinal studies that monitor the stability and development of these empirically determined subgroups over time will provide crucial insights into the trajectory of ASD heterogeneity. Additionally, exploring potential genetic differences between the identified subtypes could uncover valuable genetic markers and deepen our understanding of the underlying mechanisms of ASD.

In summary, this study not only enhances our comprehension of ASD heterogeneity but also emphasizes the pressing need for tailored interventions based on individual profiles. By considering various

symptom profile characteristics and utilizing advanced analytical techniques, we can advance our understanding of treatment responses and uncover factors that contribute to optimal developmental outcomes. Ultimately, this research has the potential to positively impact the lives of individuals with ASD by informing targeted interventions and promoting better outcomes for all.

Bibliography

[1] Leo Kanner and others. Autistic disturbances of affective contact. *Nervous child*, 2(3):217–250, 1943. Publisher: New York.

[2] GEORGE E VAILLANT. Historical notes: John Haslam on early infantile autism. *American Journal of Psychiatry*, 119(4):376–376, 1962. Publisher: Am Psychiatric Assoc.

[3] Jean-Marc-Gaspard Itard. *The wild boy of Aveyron (Rapports et mémoires sur le sauvage de l'Aveyron).* The Century Co, 1932.

[4] Eugen Bleuler. *Lehrbuch der psychiatrie*. Springer-Verlag, 2013.

[5] Daniel Hell, Christian Scharfetter, and Arnulf Möller. *Eugen Bleuler-Leben und Werk*. Huber, 2001.

[6] L Waterhouse, L Wing, and D Fein. Re-evaluating the syndrome of autism in the light of empirical research. *Autism: Nature, diagnosis, and treatment*, pages 263–281, 1989. Publisher: Guilford New York.

[7] Lorna Wing. Language, social, and cognitive impairments in autism and severe mental retardation. *Journal of autism and developmental disorders*, 11(1):31–44, 1981. Publisher: Springer.

[8] Lorna Wing. The Continuum of Autistic Characteristics. In Eric Schopler and Gary B. Mesibov, editors, *Diagnosis and Assessment in Autism*, pages 91–110. Springer US, Boston, MA, 1988.

[9] Christopher Gillberg and Suzanne Steffenburg. Outcome and prognostic factors in infantile autism and similar conditions: A population-based study of 46 cases followed through puberty. *Journal of autism and developmental disorders*, 17:273–287, 1987. Publisher: Springer.

[10] Hans Asperger. Die „Autistischen psychopathen" im kindesalter. *Archiv für psychiatrie und nervenkrankheiten*, 117(1):76–136, 1944. Publisher: Springer-Verlag Berlin/Heidelberg.

[11] American Psychiatric Association. *Diagnostic and statistical manual of mental disorders*. Author, Washington, DC, 2nd ed. edition, 1968.

[12] American Psychiatric Association. *Diagnostic and statistical manual of mental disorders*. Author, Washington, DC, 3rd ed. edition, 1980.

[13] American Psychiatric Association. *Diagnostic and Statistical Manual of Mental Disorders*. Author, Washington, DC, 3rd ed., revised edition, 1987.

[14] American Psychiatric Association. *Diagnostic and statistical manual of mental disorders*. Author, Arlington, VA, 4th ed. edition, 2000.

BIBLIOGRAPHY

[15] American Psychiatric Association. *Diagnostic and statistical manual of mental disorders*. Author, Washington, DC, 5th ed. edition, 2013.

[16] Margaret L Bauman and Thomas L Kemper. Neuroanatomic observations of the brain in autism: a review and future directions. *International journal of developmental neuroscience*, 23(2-3):183–187, 2005. Publisher: Elsevier.

[17] Christian O'Reilly, John D Lewis, and Mayada Elsabbagh. Is functional brain connectivity atypical in autism? A systematic review of EEG and MEG studies. *PloS one*, 12(5):e0175870, 2017. Publisher: Public Library of Science San Francisco, CA USA.

[18] Catherine Lord and Rhiannon Luyster. Early diagnosis of children with autism spectrum disorders. *Clinical Neuroscience Research*, 6(3-4):189–194, 2006. Publisher: Elsevier.

[19] David S Mandell, Maytali M Novak, and Cynthia D Zubritsky. Factors associated with age of diagnosis among children with autism spectrum disorders. *Pediatrics*, 116(6):1480–1486, 2005. Publisher: American Academy of Pediatrics.

[20] Paul T Shattuck, Maureen Durkin, Matthew Maenner, Craig Newschaffer, David S Mandell, Lisa Wiggins, Li-Ching Lee, Catherine Rice, Ellen Giarelli, Russell Kirby, and others. Timing of identification among children with an autism spectrum disorder: findings from a population-based surveillance study. *Journal of the American Academy of Child & Adolescent Psychiatry*, 48(5):474–483, 2009. Publisher: Elsevier.

[21] Bengt Sivberg. Parents' detection of early signs in their children having an autistic spectrum disorder. *Journal of Pediatric Nursing*, 18(6):433–439, 2003. Publisher: Elsevier.

[22] Denise Brett, Frances Warnell, Helen McConachie, and Jeremy R Parr. Factors affecting age at ASD diagnosis in UK: no evidence that diagnosis age has decreased between 2004 and 2014. *Journal of autism and developmental disorders*, 46:1974–1984, 2016. Publisher: Springer.

[23] Lonnie Zwaigenbaum, Margaret L Bauman, Wendy L Stone, Nurit Yirmiya, Annette Estes, Robin L Hansen, James C McPartland, Marvin R Natowicz, Roula Choueiri, Deborah Fein, and others. Early identification of autism spectrum disorder: Recommendations for practice and research. *Pediatrics*, 136(Supplement_1):S10–S40, 2015. Publisher: American Academy of Pediatrics Elk Grove Village, IL, USA.

[24] Christine Fountain, Marissa D King, and Peter S Bearman. Age of diagnosis for autism: individual and community factors across 10 birth cohorts. *Journal of Epidemiology & Community Health*, 65(6):503–510, 2011. Publisher: BMJ Publishing Group Ltd.

[25] Iliana Magiati, Xiang Wei Tay, and Patricia Howlin. Cognitive, language, social and behavioural outcomes in adults with autism spectrum disorders: A systematic review of longitudinal follow-up studies in adulthood. *Clinical psychology review*, 34(1):73–86, 2014. Publisher: Elsevier.

[26] Maria McGarrell, Olive Healy, Geraldine Leader, Jennifer O'Connor, and Neil Kenny. Six reports of children with autism spectrum disorder following intensive behavioral intervention using the Preschool Inventory of Repertoires for Kindergarten (PIRK®). *Research in Autism Spectrum Disorders*, 3(3):767–782, 2009. Publisher: Elsevier.

[27] Inalegwu P Oono, Emma J Honey, and Helen McConachie. Parent-mediated early intervention for young children with autism spectrum disorders (ASD). *Evidence-Based Child Health: A Cochrane Review Journal*, 8(6):2380–2479, 2013. Publisher: Wiley Online Library.

[28] Sandra L Harris and Jan S Handleman. Age and IQ at intake as predictors of placement for young children with autism: A four-to six-year follow-up. *Journal of autism and developmental disorders*, 30:137–142, 2000. Publisher: Springer.

[29] Geraldine Dawson. Early behavioral intervention, brain plasticity, and the prevention of autism spectrum disorder. *Development and psychopathology*, 20(3):775–803, 2008. Publisher: Cambridge University Press.

[30] David S Mandell, Lisa D Wiggins, Laura Arnstein Carpenter, Julie Daniels, Carolyn DiGuiseppi, Maureen S Durkin, Ellen Giarelli, Michael J Morrier, Joyce S Nicholas, Jennifer A Pinto-Martin, and others. Racial/ethnic disparities in the identification of children with autism spectrum disorders. *American journal of public health*, 99(3):493–498, 2009. Publisher: American Public Health Association.

[31] David S Mandell, Knashawn H Morales, Ming Xie, Lindsay J Lawer, Aubyn C Stahmer, and Steven C Marcus. Age of diagnosis among Medicaid-enrolled children with autism, 2001–2004. *Psychiatric Services*, 61(8):822–829, 2010. Publisher: Am Psychiatric Assoc.

[32] Amy M Daniels and David S Mandell. Explaining differences in age at autism spectrum disorder diagnosis: A critical review. *Autism*, 18(5):583–597, 2014. Publisher: Sage Publications Sage UK: London, England.

[33] Lonnie Zwaigenbaum and Melanie Penner. Autism spectrum disorder: advances in diagnosis and evaluation. *Bmj*, 361, 2018. Publisher: British Medical Journal Publishing Group.

[34] Warren Jones and Ami Klin. Attention to eyes is present but in decline in 2–6-month-old infants later diagnosed with autism. *Nature*, 504(7480):427–431, 2013. Publisher: Nature Publishing Group UK London.

[35] Luc Lecavalier, James Bodfish, Clare Harrop, Allison Whitten, Desiree Jones, Jill Pritchett, Richard Faldowski, and Brian Boyd. Development of the behavioral inflexibility scale for children with autism spectrum disorder and other developmental disabilities. *Autism Research*, 13(3):489–499, 2020. Publisher: Wiley Online Library.

[36] Johnny L Matson and Marie S Nebel-Schwalm. Comorbid psychopathology with autism spectrum disorder in children: An overview. *Research in developmental disabilities*, 28(4):341–352, 2007. Publisher: Elsevier.

[37] Susan E Levy, Ellen Giarelli, Li-Ching Lee, Laura A Schieve, Russell S Kirby, Christopher Cunniff, Joyce Nicholas, Judy Reaven, and Catherine E Rice. Autism spectrum disorder and co-occurring developmental, psychiatric, and medical conditions among children in multiple populations of the United States. *Journal of Developmental & Behavioral Pediatrics*, 31(4):267–275, 2010. Publisher: LWW.

[38] Matthew J Maenner, Kelly A Shaw, Jon Baio, Anita Washington, Mary Patrick, Monica DiRienzo, Deborah L Christensen, Lisa D Wiggins, Sydney Pettygrove, Jennifer G Andrews, and others.

Prevalence of autism spectrum disorder among children aged 8 years—autism and developmental disabilities monitoring network, 11 sites, United States, 2016. *MMWR Surveillance summaries*, 69(4):1, 2020. Publisher: Centers for Disease Control and Prevention.

[39] Marshalyn Yeargin-Allsopp, Catherine Rice, Tanya Karapurkar, Nancy Doernberg, Coleen Boyle, and Catherine Murphy. Prevalence of autism in a US metropolitan area. *Jama*, 289(1):49–55, 2003. Publisher: American Medical Association.

[40] GN Soke, MJ Maenner, D Christensen, M Kurzius-Spencer, and LA29524016 Schieve. Prevalence of co-occurring medical and behavioral conditions/symptoms among 4-and 8-year-old children with autism spectrum disorder in selected areas of the United States in 2010. *Journal of autism and developmental disorders*, 48:2663–2676, 2018. Publisher: Springer.

[41] Isaac S Kohane, Andrew McMurry, Griffin Weber, Douglas MacFadden, Leonard Rappaport, Louis Kunkel, Jonathan Bickel, Nich Wattanasin, Sarah Spence, Shawn Murphy, and others. The co-morbidity burden of children and young adults with autism spectrum disorders. *PloS one*, 7(4):e33224, 2012. Publisher: Public Library of Science San Francisco, USA.

[42] Marko Kielinen, Heikki Rantala, Eija Timonen, Sirkka-Liisa Linna, and Irma Moilanen. Associated medical disorders and disabilities in children with autistic disorder: a population-based study. *Autism*, 8(1):49–60, 2004. Publisher: Sage Publications.

[43] Carol Curtin, Sarah E Anderson, Aviva Must, and Linda Bandini. The prevalence of obesity in children with autism: a secondary data analysis using nationally representative data from the National Survey of Children's Health. *BMC pediatrics*, 10(1):1–5, 2010. Publisher: BioMed Central.

[44] Tara Stevens, Lei Peng, and Lucy Barnard-Brak. The comorbidity of ADHD in children diagnosed with autism spectrum disorder. *Research in Autism Spectrum Disorders*, 31:11–18, 2016. Publisher: Elsevier.

[45] Ryan T Thorson and Johnny L Matson. Cutoff scores for the autism spectrum disorder–comorbid for children (ASD-CC). *Research in Autism Spectrum Disorders*, 6(1):556–559, 2012. Publisher: Elsevier.

[46] Francisca JA Van Steensel, Susan M Bögels, and Esther I de Bruin. Psychiatric comorbidity in children with autism spectrum disorders: A comparison with children with ADHD. *Journal of child and family studies*, 22:368–376, 2013. Publisher: Springer.

[47] Emily Simonoff, Andrew Pickles, Tony Charman, Susie Chandler, Tom Loucas, and Gillian Baird. Psychiatric disorders in children with autism spectrum disorders: prevalence, comorbidity, and associated factors in a population-derived sample. *Journal of the American Academy of Child & Adolescent Psychiatry*, 47(8):921–929, 2008. Publisher: Elsevier.

[48] James K. Luiselli, editor. *Children and youth with autism spectrum disorder (ASD): recent advances and innovations in assessment, education, and intervention*. Oxford University Press, Oxford, 2014.

[49] Hadeel Faras, Nahed Al Ateeqi, and Lee Tidmarsh. Autism spectrum disorders. *Annals of Saudi medicine*, 30(4):295–300, 2010. Publisher: King Faisal Specialist Hospital & Research Centre.

[50] Jane Summers, Ali Shahrami, Stefanie Cali, Chantelle D'Mello, Milena Kako, Andjelka Palikucin-Reljin, Melissa Savage, Olivia Shaw, and Yona Lunsky. Self-injury in autism spectrum disorder and intellectual disability: Exploring the role of reactivity to pain and sensory input. *Brain sciences*, 7(11):140, 2017. Publisher: MDPI.

[51] Centers for Disease Control, Prevention, et al. Changes in prevalence of parent-reported autism spectrum disorder in school-aged us children: 2007 to 2011-2012. *National Health Statistics Reports*, 65:1–11, 2013.

[52] Deborah L Christensen, Matthew J Maenner, Deborah Bilder, John N Constantino, Julie Daniels, Maureen S Durkin, Robert T Fitzgerald, Margaret Kurzius-Spencer, Sydney D Pettygrove, Cordelia Robinson, and others. Prevalence and characteristics of autism spectrum disorder among children aged 4 years—early autism and developmental disabilities monitoring network, seven sites, United States, 2010, 2012, and 2014. *MMWR Surveillance Summaries*, 68(2):1, 2019. Publisher: Centers for Disease Control and Prevention.

[53] Minha Hong, Sang Min Lee, Saengryeol Park, Seok-Jun Yoon, Young-Eun Kim, and In-Hwan Oh. Prevalence and economic burden of autism spectrum disorder in South Korea using national health insurance data from 2008 to 2015. *Journal of Autism and Developmental Disorders*, 50:333–339, 2020. Publisher: Springer.

[54] Der-Chung Lai, Yen-Cheng Tseng, Yuh-Ming Hou, and How-Ran Guo. Gender and geographic differences in the prevalence of autism spectrum disorders in children: Analysis of data from the national disability registry of Taiwan. *Research in developmental disabilities*, 33(3):909–915, 2012. Publisher: Elsevier.

[55] Marit Maria Elisabeth van Bakel, Malika Delobel-Ayoub, Christine Cans, Brigitte Assouline, Pierre-Simon Jouk, Jean-Philippe Raynaud, and Catherine Arnaud. Low but increasing prevalence of autism spectrum disorders in a French area from register-based data. *Journal of Autism and Developmental Disorders*, 45:3255–3261, 2015. Publisher: Springer.

[56] Tamara May, Amanda Brignell, and Katrina Williams. Autism spectrum disorder prevalence in children aged 12–13 years from the longitudinal study of Australian children. *Autism Research*, 13(5):821–827, 2020. Publisher: Wiley Online Library.

[57] Melinda Randall, Emma Sciberras, Amanda Brignell, Elfriede Ihsen, Daryl Efron, Cheryl Dissanayake, and Katrina Williams. Autism spectrum disorder: Presentation and prevalence in a nationally representative Australian sample. *Australian & New Zealand Journal of Psychiatry*, 50(3):243–253, 2016. Publisher: Sage Publications Sage UK: London, England.

[58] Jinan Zeidan, Eric Fombonne, Julie Scorah, Alaa Ibrahim, Maureen S Durkin, Shekhar Saxena, Afiqah Yusuf, Andy Shih, and Mayada Elsabbagh. Global prevalence of autism: A systematic review update. *Autism Research*, 15(5):778–790, 2022.

[59] Marco Solmi, Minjin Song, Dong Keon Yon, Seung Won Lee, Eric Fombonne, Min Seo Kim, Seoyeon Park, Min Ho Lee, Jimin Hwang, Roberto Keller, et al. Incidence, prevalence, and global burden of autism spectrum disorder from 1990 to 2019 across 204 countries. *Molecular Psychiatry*, pages 1–9, 2022.

[60] F Icasiano, P Hewson, P Machet, C Cooper, and A Marshall. Childhood autism spectrum disorder in the Barwon region: a community based study. *Journal of paediatrics and child health*, 40(12):696–701, 2004. Publisher: Wiley Online Library.

[61] Eric Fombonne, Christiane Du Mazaubrun, Christine Cans, and Hélène Grandjean. Autism and associated medical disorders in a French epidemiological survey. *Journal of the American Academy of Child & Adolescent Psychiatry*, 36(11):1561–1569, 1997. Publisher: Elsevier.

[62] Patricia JM Van Wijngaarden-Cremers, Evelien van Eeten, Wouter B Groen, Patricia A Van Deurzen, Iris J Oosterling, and Rutger Jan Van der Gaag. Gender and age differences in the core triad of impairments in autism spectrum disorders: a systematic review and meta-analysis. *Journal of autism and developmental disorders*, 44:627–635, 2014. Publisher: Springer.

[63] María Tubío-Fungueiriño, Sara Cruz, Adriana Sampaio, Angel Carracedo, and Montse Fernández-Prieto. Social camouflaging in females with autism spectrum disorder: A systematic review. *Journal of Autism and Developmental Disorders*, 51:2190–2199, 2021. Publisher: Springer.

[64] Rachel Loomes, Laura Hull, and William Polmear Locke Mandy. What is the male-to-female ratio in autism spectrum disorder? A systematic review and meta-analysis. *Journal of the American Academy of Child & Adolescent Psychiatry*, 56(6):466–474, 2017. Publisher: Elsevier.

[65] Lindsay A Olson, Lisa E Mash, Annika Linke, Christopher H Fong, Ralph-Axel Müller, and Inna Fishman. Sex-related patterns of intrinsic functional connectivity in children and adolescents with autism spectrum disorders. *Autism*, 24(8):2190–2201, 2020. Publisher: SAGE Publications Sage UK: London, England.

[66] Meghan Styles, Dalal Alsharshani, Muthanna Samara, Mohammed Alsharshani, Azhar Khattab, M Walid Qoronfleh, and Nader Izzeddin Al-Dewik. Risk factors, diagnosis, prognosis and treatment of autism. *Frontiers in Bioscience*, 25(9):1682–1717, 2020. Publisher: Frontiers in Bioscience.

[67] Guiomar Oliveira, Assunção Ataíde, Carla Marques, Teresa S Miguel, Ana Margarida Coutinho, Luísa Mota-Vieira, Esmeralda Gonçalves, Nazaré Mendes Lopes, Vitor Rodrigues, Henrique Carmona da Mota, and others. Epidemiology of autism spectrum disorder in Portugal: prevalence, clinical characterization, and medical conditions. *Developmental Medicine & Child Neurology*, 49(10):726–733, 2007. Publisher: Wiley Online Library.

[68] Célia Rasga, João Xavier Santos, Cátia Café, Alexandra Oliveira, Frederico Duque, Ana Nunes, Guiomar Oliveira, and Astrid Moura Vicente. Prevalência da perturbação do espectro do autismo na região Centro de Portugal: um estudo no âmbito do projeto ASDEU. *Boletim Epidemiológico Observações*, 9(27):47–51, 2020. Publisher: Instituto Nacional de Saúde Doutor Ricardo Jorge, IP.

[69] Natasha Nassar, Glenys Dixon, Jenny Bourke, Carol Bower, Emma Glasson, Nick De Klerk, and Helen Leonard. Autism spectrum disorders in young children: effect of changes in diagnostic practices. *International journal of epidemiology*, 38(5):1245–1254, 2009. Publisher: Oxford University Press.

[70] Mayada Elsabbagh, Gauri Divan, Yun-Joo Koh, Young Shin Kim, Shuaib Kauchali, Carlos Marcín, Cecilia Montiel-Nava, Vikram Patel, Cristiane S Paula, Chongying Wang, and others. Global

prevalence of autism and other pervasive developmental disorders. *Autism research*, 5(3):160–179, 2012. Publisher: Wiley Online Library.

[71] Suniti Chakrabarti and Eric Fombonne. Pervasive developmental disorders in preschool children: confirmation of high prevalence. *American Journal of Psychiatry*, 162(6):1133–1141, 2005. Publisher: Am Psychiatric Assoc.

[72] Katrina Williams, Sarah MacDermott, Greta Ridley, Emma J Glasson, and John A Wray. The prevalence of autism in Australia. Can it be established from existing data? *Journal of paediatrics and child health*, 44(9):504–510, 2008. Publisher: Wiley Online Library.

[73] Simon Baron-Cohen, Fiona J Scott, Carrie Allison, Joanna Williams, Patrick Bolton, Fiona E Matthews, and Carol Brayne. Prevalence of autism-spectrum conditions: UK school-based population study. *The British journal of psychiatry*, 194(6):500–509, 2009. Publisher: Cambridge University Press.

[74] Claudia L Hilton, Robert T Fitzgerald, Kelley M Jackson, Rolanda A Maxim, Christopher C Bosworth, Paul T Shattuck, Daniel H Geschwind, and John N Constantino. Brief report: Underrepresentation of African Americans in autism genetic research: A rationale for inclusion of subjects representing diverse family structures. *Journal of autism and developmental disorders*, 40:633–639, 2010. Publisher: Springer.

[75] Catherine Lord, Michael Rutter, and Ann Le Couteur. Autism Diagnostic Interview-Revised: a revised version of a diagnostic interview for caregivers of individuals with possible pervasive developmental disorders. *Journal of autism and developmental disorders*, 24(5):659–685, 1994. Publisher: Springer.

[76] Catherine Lord, Michael Rutter, Susan Goode, Jacquelyn Heemsbergen, Heather Jordan, Lynn Mawhood, and Eric Schopler. Austism diagnostic observation schedule: A standardized observation of communicative and social behavior. *Journal of autism and developmental disorders*, 19(2):185–212, 1989. Publisher: Springer.

[77] Catherine Lord, Susan Risi, Linda Lambrecht, Edwin H Cook, Bennett L Leventhal, Pamela C DiLavore, Andrew Pickles, and Michael Rutter. The autism diagnostic observation schedule—generic: A standard measure of social and communication deficits associated with the spectrum of autism. *Journal of autism and developmental disorders*, 30:205–223, 2000.

[78] Catherine Lord, M Rutter, P DiLavore, S Risi, K Gotham, S Bishop, et al. Autism diagnostic observation schedule–2nd edition (ados-2). *Los Angeles, CA: Western Psychological Corporation*, 284, 2012.

[79] Sara S Sparrow, David A Balla, and Domenic V Cicchetti. *Vineland Adaptive Behavior Scales VABS: Expanded Form Manual*. American Guidance Service, 1984.

[80] David Wechsler. Wechsler adult intelligence scale–Fourth Edition (WAIS–IV). *San Antonio, TX: NCS Pearson*, 22(498):1, 2008.

[81] Ruth Griffiths. The Griffiths mental development scales from birth to 2 years. *Manual. The 1996 revision Huntley: Association for research in infant and child development*, 1996. Publisher: Test agency.

[82] Edgar Arnold Doll. *Vineland social maturity scale: Condensed manual of directions*. American Guidance Service, 1965.

[83] Sara S. Sparrow, Domenic V. Cicchetti, and David A. Balla. *Vineland Adaptive Behavior Scales, Second Edition*. AGS Publishing, Circle Pines, MN, 2005.

[84] Sara S Sparrow, Domenic V Cicchetti, and Celine A Saulnier. *Vineland-3: Vineland adaptive behavior scales*. PsychCorp, 2016.

[85] R. Griffiths. *The abilities of babies: a study in mental measurement*. McGraw-Hill, 1954.

[86] Ruth Griffiths. The abilities of young children: A comprehensive system of mental measurement for the first eight years of life. *(No Title)*, 1984.

[87] R Griffiths. The griffiths mental developmental scales, extended revised. *UK: Association for Research in Infant and Child Development, The Test Agency*, 2006.

[88] E Green, L Stroud, S Bloomfield, J Cronje, C Foxcroft, K Hurter, H Lane, M Candice, P McAlinden, R Paradice, and others. Griffiths scales of child development. *Hogrefe Ltd.: Oxford, UK*, 2016.

[89] David Wechsler. Wechsler adult intelligence scale–. *Archives of Clinical Neuropsychology*, 1955.

[90] Psychological Corporation. *WAIS-III and WMS-III Technical Manual*. Psychological Corporation, 1997.

[91] David Wechsler. *WISC-V: Weschler Intelligence Scale for Children*. Pearson, 2014.

[92] Jeffrey M Kidd, Gregory M Cooper, William F Donahue, Hillary S Hayden, Nick Sampas, Tina Graves, Nancy Hansen, Brian Teague, Can Alkan, Francesca Antonacci, and others. Mapping and sequencing of structural variation from eight human genomes. *Nature*, 453(7191):56–64, 2008. Publisher: Nature Publishing Group UK London.

[93] Jonathan Sebat, B Lakshmi, Jennifer Troge, Joan Alexander, Janet Young, Par Lundin, Susanne Manér, Hillary Massa, Megan Walker, Maoyen Chi, et al. Large-scale copy number polymorphism in the human genome. *Science*, 305(5683):525–528, 2004.

[94] Donald F Conrad et al. Variation in genome-wide mutation rates within and between human families. *Nature genetics*, 43(7):712–714, 2011.

[95] Pauline Chaste, Kathryn Roeder, and Bernie Devlin. The Yin and Yang of autism genetics: how rare de novo and common variations affect liability. *Annual review of genomics and human genetics*, 18:167–187, 2017. Publisher: Annual Reviews.

[96] Silvia De Rubeis, Xin He, Arthur P Goldberg, Christopher S Poultney, Kaitlin Samocha, A Ercument Cicek, Yan Kou, Li Liu, Menachem Fromer, Susan Walker, and others. Synaptic, transcriptional and chromatin genes disrupted in autism. *Nature*, 515(7526):209–215, 2014. Publisher: Nature Publishing Group UK London.

[97] Ryan N Doan, Elaine T Lim, Silvia De Rubeis, Catalina Betancur, David J Cutler, Andreas G Chiocchetti, Lynne M Overman, Aubrie Soucy, Susanne Goetze, Autism Sequencing Consortium, and others. Recessive gene disruptions in autism spectrum disorder. *Nature genetics*, 51(7):1092–1098, 2019. Publisher: Nature Publishing Group US New York.

[98] Ryan K C Yuen, Daniele Merico, Matt Bookman, Jennifer L Howe, Bhooma Thiruvahindrapuram, Rohan V Patel, Joe Whitney, Nicole Deflaux, Jonathan Bingham, Zhuozhi Wang, and others. Whole genome sequencing resource identifies 18 new candidate genes for autism spectrum disorder. *Nature neuroscience*, 20(4):602–611, 2017. Publisher: Nature Publishing Group US New York.

[99] Elizabeth K Ruzzo, Laura Pérez-Cano, Jae-Yoon Jung, Lee-kai Wang, Dorna Kashef-Haghighi, Chris Hartl, Chanpreet Singh, Jin Xu, Jackson N Hoekstra, Olivia Leventhal, and others. Inherited and de novo genetic risk for autism impacts shared networks. *Cell*, 178(4):850–866, 2019. Publisher: Elsevier.

[100] Caroline M Dias and Christopher A Walsh. Recent advances in understanding the genetic architecture of autism. *Annual Review of Genomics and Human Genetics*, 21:289–304, 2020. Publisher: Annual Reviews.

[101] Arianna Benvenuto, Romina Moavero, Riccardo Alessandrelli, Barbara Manzi, and Paolo Curatolo. Syndromic autism: causes and pathogenetic pathways. *World journal of pediatrics*, 5:169–176, 2009. Publisher: Springer.

[102] Craig J Newschaffer, Lisa A Croen, Julie Daniels, Ellen Giarelli, Judith K Grether, Susan E Levy, David S Mandell, Lisa A Miller, Jennifer Pinto-Martin, Judy Reaven, and others. The epidemiology of autism spectrum disorders. *Annu. Rev. Public Health*, 28:235–258, 2007. Publisher: Annual Reviews.

[103] Sally Ozonoff, Gregory S Young, Alice Carter, Daniel Messinger, Nurit Yirmiya, Lonnie Zwaigenbaum, Susan Bryson, Leslie J Carver, John N Constantino, Karen Dobkins, and others. Recurrence risk for autism spectrum disorders: a Baby Siblings Research Consortium study. *Pediatrics*, 128(3):e488–e495, 2011. Publisher: American Academy of Pediatrics Elk Grove Village, IL, USA.

[104] Sven Sandin, Paul Lichtenstein, Ralf Kuja-Halkola, Henrik Larsson, Christina M. Hultman, and Abraham Reichenberg. The Familial Risk of Autism. *JAMA*, 311(17):1770, May 2014.

[105] Hu, Devlin, and Debski. ASD Phenotype—Genotype Associations in Concordant and Discordant Monozygotic and Dizygotic Twins Stratified by Severity of Autistic Traits. *International Journal of Molecular Sciences*, 20(15):3804, August 2019.

[106] Emma Colvert, Beata Tick, Fiona McEwen, Catherine Stewart, Sarah R. Curran, Emma Woodhouse, Nicola Gillan, Victoria Hallett, Stephanie Lietz, Tracy Garnett, Angelica Ronald, Robert Plomin, Frühling Rijsdijk, Francesca Happé, and Patrick Bolton. Heritability of Autism Spectrum Disorder in a UK Population-Based Twin Sample. *JAMA Psychiatry*, 72(5):415, May 2015.

[107] Evan E. Eichler, Jonathan Flint, Greg Gibson, Augustine Kong, Suzanne M. Leal, Jason H. Moore, and Joseph H. Nadeau. Missing heritability and strategies for finding the underlying causes of complex disease. *Nature Reviews Genetics*, 11(6):446–450, June 2010.

[108] Barbara Wiśniowiecka-Kowalnik and Beata Anna Nowakowska. Genetics and epigenetics of autism spectrum disorder—current evidence in the field. *Journal of Applied Genetics*, 60(1):37–47, February 2019.

[109] Joachim Hallmayer. Genetic Heritability and Shared Environmental Factors Among Twin Pairs With Autism. *Archives of General Psychiatry*, 68(11):1095, November 2011.

[110] Thomas W. Frazier, Lee Thompson, Eric A. Youngstrom, Paul Law, Antonio Y. Hardan, Charis Eng, and Nathan Morris. A Twin Study of Heritable and Shared Environmental Contributions to Autism. *Journal of Autism and Developmental Disorders*, 44(8):2013–2025, August 2014.

[111] Lambertus Klei, Stephan J Sanders, Michael T Murtha, Vanessa Hus, Jennifer K Lowe, A Jeremy Willsey, Daniel Moreno-De-Luca, Timothy W Yu, Eric Fombonne, Daniel Geschwind, Dorothy E Grice, David H Ledbetter, Catherine Lord, Shrikant M Mane, Christa Lese Martin, Donna M Martin, Eric M Morrow, Christopher A Walsh, Nadine M Melhem, Pauline Chaste, James S Sutcliffe, Matthew W State, Edwin H Cook, Kathryn Roeder, and Bernie Devlin. Common genetic variants, acting additively, are a major source of risk for autism. *Molecular Autism*, 3(1):9, December 2012.

[112] Trent Gaugler, Lambertus Klei, Stephan J Sanders, Corneliu A Bodea, Arthur P Goldberg, Ann B Lee, Milind Mahajan, Dina Manaa, Yudi Pawitan, Jennifer Reichert, Stephan Ripke, Sven Sandin, Pamela Sklar, Oscar Svantesson, Abraham Reichenberg, Christina M Hultman, Bernie Devlin, Kathryn Roeder, and Joseph D Buxbaum. Most genetic risk for autism resides with common variation. *Nature Genetics*, 46(8):881–885, August 2014.

[113] Wenlin Deng, Xiaobing Zou, Hongzhu Deng, Jianying Li, Chun Tang, Xueqin Wang, and Xiaobo Guo. The Relationship Among Genetic Heritability, Environmental Effects, and Autism Spectrum Disorders: 37 Pairs of Ascertained Twin Study. *Journal of Child Neurology*, 30(13):1794–1799, November 2015.

[114] Lisa R. Edelson and Kimberly J. Saudino. Genetic and Environmental Influences on Autistic-Like Behaviors in 2-Year-Old Twins. *Behavior Genetics*, 39(3):255–264, May 2009.

[115] Sven Bölte, Sonya Girdler, and Peter B. Marschik. The contribution of environmental exposure to the etiology of autism spectrum disorder. *Cellular and Molecular Life Sciences*, 76(7):1275–1297, April 2019.

[116] M Janecka, J Mill, M A Basson, A Goriely, H Spiers, A Reichenberg, L Schalkwyk, and C Fernandes. Advanced paternal age effects in neurodevelopmental disorders—review of potential underlying mechanisms. *Translational Psychiatry*, 7(1):e1019–e1019, January 2017.

[117] S. Wu, F. Wu, Y. Ding, J. Hou, J. Bi, and Z. Zhang. Advanced parental age and autism risk in children: a systematic review and meta-analysis. *Acta Psychiatrica Scandinavica*, 135(1):29–41, January 2017.

[118] Stefanie Atsem, Juliane Reichenbach, Ramya Potabattula, Marcus Dittrich, Caroline Nava, Christel Depienne, Lena Böhm, Simone Rost, Thomas Hahn, Martin Schorsch, Thomas Haaf, and Nady El Hajj. Paternal age effects on sperm *FOXK1* and *KCNA7* methylation and transmission into the next generation. *Human Molecular Genetics*, page ddw328, September 2016.

[119] Hannah Gardener, Donna Spiegelman, and Stephen L. Buka. Prenatal risk factors for autism: comprehensive meta-analysis. *British Journal of Psychiatry*, 195(1):7–14, July 2009.

[120] Hannah Gardener, Donna Spiegelman, and Stephen L. Buka. Perinatal and Neonatal Risk Factors for Autism: A Comprehensive Meta-analysis. *Pediatrics*, 128(2):344–355, August 2011.

[121] Chengzhong Wang, Hua Geng, Weidong Liu, and Guiqin Zhang. Prenatal, perinatal, and postnatal factors associated with autism: A meta-analysis. *Medicine*, 96(18):e6696, May 2017.

[122] Aa Jackson and Sm Robinson. Dietary guidelines for pregnancy: a review of current evidence. *Public Health Nutrition*, 4(2b):625–630, April 2001.

[123] Shiming Tang, Ying Wang, Xuan Gong, and Gaohua Wang. A Meta-Analysis of Maternal Smoking during Pregnancy and Autism Spectrum Disorder Risk in Offspring. *International Journal of Environmental Research and Public Health*, 12(9):10418–10431, August 2015.

[124] Brittany N. Rosen, Brian K. Lee, Nora L. Lee, Yunwen Yang, and Igor Burstyn. Maternal Smoking and Autism Spectrum Disorder: A Meta-analysis. *Journal of Autism and Developmental Disorders*, 45(6):1689–1698, June 2015.

[125] P. Bolton, H. Macdonald, A. Pickles, P. Rios, S. Goode, M. Crowson, A. Bailey, and M. Rutter. A Case-Control Family History Study of Autism. *Journal of Child Psychology and Psychiatry*, 35(5):877–900, July 1994.

[126] Emre Bora, Aydan Aydın, Tuğba Saraç, Muhammed Tayyib Kadak, and Sezen Köse. Heterogeneity of subclinical autistic traits among parents of children with autism spectrum disorder: Identifying the broader autism phenotype with a data-driven method: BAP and LCA in parents. *Autism Research*, 10(2):321–326, February 2017.

[127] Wouter De la Marche, Ilse Noens, Jan Luts, Evert Scholte, Sabine Van Huffel, and Jean Steyaert. Quantitative autism traits in first degree relatives: evidence for the broader autism phenotype in fathers, but not in mothers and siblings. *Autism*, 16(3):247–260, May 2012.

[128] Ed Sucksmith, Ilona Roth, and Rosa Anna Hoekstra. Autistic traits below the clinical threshold: re-examining the broader autism phenotype in the 21st century. *Neuropsychology review*, 21:360–389, 2011. Publisher: Springer.

[129] Elise B. Robinson. Evidence That Autistic Traits Show the Same Etiology in the General Population and at the Quantitative Extremes (5%, 2.5%, and 1%). *Archives of General Psychiatry*, 68(11):1113, November 2011.

[130] Irving I. Gottesman and Todd D. Gould. The Endophenotype Concept in Psychiatry: Etymology and Strategic Intentions. *American Journal of Psychiatry*, 160(4):636–645, April 2003.

[131] Todd D Gould and Irving I Gottesman. Psychiatric endophenotypes and the development of valid animal models. *Genes, brain and behavior*, 5(2):113–119, 2006.

[132] John N Constantino and Richard D Todd. Intergenerational transmission of subthreshold autistic traits in the general population. *Biological psychiatry*, 57(6):655–660, 2005.

[133] Christina R Maxwell, Julia Parish-Morris, Olivia Hsin, Jennifer C Bush, and Robert T Schultz. The broad autism phenotype predicts child functioning in autism spectrum disorders. *Journal of neurodevelopmental disorders*, 5(1):1–7, 2013.

[134] Noah J Sasson, Kristen SL Lam, Morgan Parlier, Julie L Daniels, and Joseph Piven. Autism and the broad autism phenotype: familial patterns and intergenerational transmission. *Journal of neurodevelopmental disorders*, 5(1):1–7, 2013.

[135] Vicki Bitsika, CF Sharpley, and S Orapeleng. An exploratory analysis of the use of cognitive, adaptive and behavioural indices for cluster analysis of asd subgroups. *Journal of Intellectual Disability Research*, 52(11):973–985, 2008.

[136] Stelios Georgiades, Peter Szatmari, Michael Boyle, Steven Hanna, Eric Duku, Lonnie Zwaigenbaum, Susan Bryson, Eric Fombonne, Joanne Volden, Pat Mirenda, and others. Investigating phenotypic heterogeneity in children with autism spectrum disorder: a factor mixture modeling approach. *Journal of Child Psychology and Psychiatry*, 54(2):206–215, 2013. Publisher: Wiley Online Library.

[137] Anne Masi, Marilena M DeMayo, Nicholas Glozier, and Adam J Guastella. An overview of autism spectrum disorder, heterogeneity and treatment options. *Neuroscience bulletin*, 33:183–193, 2017.

[138] Lonnie Zwaigenbaum, Margaret L Bauman, Roula Choueiri, Connie Kasari, Alice Carter, Doreen Granpeesheh, Zoe Mailloux, Susanne Smith Roley, Sheldon Wagner, Deborah Fein, and others. Early intervention for children with autism spectrum disorder under 3 years of age: recommendations for practice and research. *Pediatrics*, 136(Supplement_1):S60–S81, 2015. Publisher: American Academy of Pediatrics Elk Grove Village, IL, USA.

[139] Connie Wong, Samuel L Odom, Kara A Hume, Ann W Cox, Angel Fettig, Suzanne Kucharczyk, Matthew E Brock, Joshua B Plavnick, Veronica P Fleury, and Tia R Schultz. Evidence-based practices for children, youth, and young adults with autism spectrum disorder: A comprehensive review. *Journal of autism and developmental disorders*, 45:1951–1966, 2015.

[140] Lorna French and Eilis MM Kennedy. Annual Research Review: Early intervention for infants and young children with, or at-risk of, autism spectrum disorder: a systematic review. *Journal of Child Psychology and psychiatry*, 59(4):444–456, 2018. Publisher: Wiley Online Library.

[141] Laura Schreibman, Geraldine Dawson, Aubyn C Stahmer, Rebecca Landa, Sally J Rogers, Gail G McGee, Connie Kasari, Brooke Ingersoll, Ann P Kaiser, Yvonne Bruinsma, and others. Naturalistic developmental behavioral interventions: Empirically validated treatments for autism spectrum disorder. *Journal of autism and developmental disorders*, 45:2411–2428, 2015. Publisher: Springer.

[142] Melissa L McPheeters, Zachary Warren, Nila Sathe, Jennifer L Bruzek, Shanthi Krishnaswami, Rebecca N Jerome, and Jeremy Veenstra-VanderWeele. A systematic review of medical treatments for children with autism spectrum disorders. *Pediatrics*, 127(5):e1312–e1321, 2011.

[143] Oliver D Howes, Maria Rogdaki, James L Findon, Robert H Wichers, Tony Charman, Bryan H King, Eva Loth, Gráinne M McAlonan, James T McCracken, Jeremy R Parr, et al. Autism spectrum disorder: Consensus guidelines on assessment, treatment and research from the british association for psychopharmacology. *Journal of Psychopharmacology*, 32(1):3–29, 2018.

[144] Lynn Waterhouse, Eric London, and Christopher Gillberg. ASD validity. *Review Journal of Autism and Developmental Disorders*, 3:302–329, 2016. Publisher: Springer.

[145] Ralph-Axel Müller and David G Amaral. Time to give up on Autism Spectrum Disorder?, 2017. Issue: 1 Pages: 10–14 Publication Title: Autism Research Volume: 10.

[146] Francesca Happé, Angelica Ronald, and Robert Plomin. Time to give up on a single explanation for autism. *Nature neuroscience*, 9(10):1218–1220, 2006. Publisher: Nature Publishing Group US New York.

[147] Daniel H Geschwind and Pat Levitt. Autism spectrum disorders: developmental disconnection syndromes. *Current opinion in neurobiology*, 17(1):103–111, 2007. Publisher: Elsevier.

[148] OJ Veatch, Jeremy Veenstra-VanderWeele, Melissa Potter, MA Pericak-Vance, and JL Haines. Genetically meaningful phenotypic subgroups in autism spectrum disorders. *Genes, Brain and Behavior*, 13(3):276–285, 2014. Publisher: Wiley Online Library.

[149] So Hyun Kim, Suzanne Macari, Judah Koller, and Katarzyna Chawarska. Examining the phenotypic heterogeneity of early autism spectrum disorder: subtypes and short-term outcomes. *Journal of Child Psychology and Psychiatry*, 57(1):93–102, 2016. Publisher: Wiley Online Library.

[150] Shuting Zheng, Kara A Hume, Harriet Able, Somer L Bishop, and Brian A Boyd. Exploring developmental and behavioral heterogeneity among preschoolers with ASD: A cluster analysis on principal components. *Autism Research*, 13(5):796–809, 2020. Publisher: Wiley Online Library.

[151] Scott D Tomchek, Lauren M Little, John Myers, and Winnie Dunn. Sensory subtypes in preschool aged children with autism spectrum disorder. *Journal of autism and developmental disorders*, 48:2139–2147, 2018. Publisher: Springer.

[152] Lidan Zheng, Rachel Grove, and Valsamma Eapen. Spectrum or subtypes? A latent profile analysis of restricted and repetitive behaviours in autism. *Research in Autism Spectrum Disorders*, 57:46–54, 2019. Publisher: Elsevier.

[153] Rose F Eagle, Raymond G Romanczyk, and Mark F Lenzenweger. Classification of children with autism spectrum disorders: A finite mixture modeling approach to heterogeneity. *Research in Autism Spectrum Disorders*, 4(4):772–781, 2010. Publisher: Elsevier.

[154] Hannah Cholemkery, Juliane Medda, Thomas Lempp, and Christine M Freitag. Classifying autism spectrum disorders by adi-r: subtypes or severity gradient? *Journal of autism and developmental disorders*, 46:2327–2339, 2016.

[155] Muhammad Asif, Hugo FMC Martiniano, Ana Rita Marques, João Xavier Santos, Joana Vilela, Celia Rasga, Guiomar Oliveira, Francisco M Couto, and Astrid M Vicente. Identification of biological mechanisms underlying a multidimensional ASD phenotype using machine learning. *Translational psychiatry*, 10(1):43, 2020. Publisher: Nature Publishing Group UK London.

[156] Sotiris B Kotsiantis, Ioannis Zaharakis, P Pintelas, and others. Supervised machine learning: A review of classification techniques. *Emerging artificial intelligence applications in computer engineering*, 160(1):3–24, 2007. Publisher: Amsterdam.

[157] Ethem Alpaydin. *Introduction to machine learning*. MIT press, 2020.

[158] Vinay Prasad, Tito Fojo, and Michael Brada. Precision oncology: origins, optimism, and potential. *The Lancet Oncology*, 17(2):e81–e86, 2016.

[159] Jiawei Han and Micheline Kamber. *Data mining: concepts and techniques*. The Morgan Kaufmann series in data management systems. Elsevier ; Morgan Kaufmann, Amsterdam ; Boston : San Francisco, CA, 2nd ed edition, 2006. OCLC: ocm63401845.

[160] Leonard Kaufman and Peter J Rousseeuw. *Finding groups in data: an introduction to cluster analysis*. John Wiley & Sons, 2009.

[161] J MacQueen. Classification and analysis of multivariate observations. In *5th Berkeley Symp. Math. Statist. Probability*, pages 281–297. University of California Los Angeles LA USA, 1967.

[162] K. Krishna and M. Narasimha Murty. Genetic K-means algorithm. *IEEE Transactions on Systems, Man and Cybernetics, Part B (Cybernetics)*, 29(3):433–439, June 1999.

[163] L Meng, QH Wu, and ZZ Yong. A genetic hard c-means clustering algorithm. *DYNAMICS OF CONTINUOUS DISCRETE AND IMPULSIVE SYSTEMS SERIES B*, 9:421–438, 2002. Publisher: WATAM PRESS.

[164] Vladimir Estivill-Castro and Alan T Murray. *Spatial clustering for data mining with genetic algorithms*. Citeseer, 1997.

[165] John C Gower. A general coefficient of similarity and some of its properties. *Biometrics*, pages 857–871, 1971. Publisher: JSTOR.

[166] Martin Ester, Hans-Peter Kriegel, Jörg Sander, Xiaowei Xu, and others. A density-based algorithm for discovering clusters in large spatial databases with noise. In *kdd*, volume 96, pages 226–231, 1996. Issue: 34.

[167] Pang-Ning Tan, Michael Steinbach, and Vipin Kumar. Data mining introduction. *People's Posts and Telecommunications Publishing House, Beijing*, 2006.

[168] Chris Newby, Liam G. Heaney, Andrew Menzies-Gow, Rob M. Niven, Adel Mansur, Christine Bucknall, Rekha Chaudhuri, John Thompson, Paul Burton, Chris Brightling, and on behalf of the British Thoracic Society Severe Refractory Asthma Network. Statistical Cluster Analysis of the British Thoracic Society Severe Refractory Asthma Registry: Clinical Outcomes and Phenotype Stability. *PLoS ONE*, 9(7):e102987, July 2014.

[169] Matthew J Loza, Ian Adcock, Charles Auffray, Kiang F Chung, Ratko Djukanovic, Peter J Sterk, Vedrana S Susulic, Elliot S Barnathan, Frederik Baribaud, and Philip E Silkoff. Longitudinally stable, clinically defined clusters of patients with asthma independently identified in the ADEPT and U-BIOPRED asthma studies. *Annals of the American Thoracic Society*, 13(Supplement 1):S102–S103, 2016. Publisher: American Thoracic Society.

[170] Matthew J Loza, Ratko Djukanovic, Kian Fan Chung, Daniel Horowitz, Keying Ma, Patrick Branigan, Elliot S Barnathan, Vedrana S Susulic, Philip E Silkoff, Peter J Sterk, and others. Validated and longitudinally stable asthma phenotypes based on cluster analysis of the ADEPT study. *Respiratory research*, 17(1):1–21, 2016. Publisher: BioMed Central.

[171] Hee-Ju Kim, Deborah B McGuire, Lorraine Tulman, and Andrea M Barsevick. Symptom clusters: concept analysis and clinical implications for cancer nursing. *Cancer nursing*, 28(4):270–282, 2005. Publisher: LWW.

[172] Bradley Efron and R.J. Tibshirani. *An Introduction to the Bootstrap*. Chapman and Hall/CRC, 0 edition, May 1994.

[173] Anil K Jain and JV Moreau. Bootstrap technique in cluster analysis. *Pattern Recognition*, 20(5):547–568, 1987. Publisher: Elsevier.

[174] Python Software Foundation. Python 3.8.12 documentation. `https://docs.python.org/3.8/index.html`, 2021. Accessed on [insert date of access].

[175] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2021. Version 4.2.1.

[176] Charles Spearman. The proof and measurement of association between two things. *The American Journal of Psychology*, 15(1):72–101, 1904.

[177] Samuel Berger, Johannes Groß, Frank Konietschke, Robert Prill, Daniel Römer, Matthias Templ, Claus Weihs, and Andreas Ziegler. *VIM: Visualization and Imputation of Missing Values*, 2021. R package version 6.1.1.

[178] Daniel J. Stekhoven and Peter Bühlmann. *missForest: Nonparametric missing value imputation using random forest*, 2012. R package version 1.4.

[179] Malika Charrad, Nizar Ghazzali, Vincent Boiteau, and Ali Niknafs. Nbclust: An r package for determining the relevant number of clusters in a data set. *Journal of Statistical Software*, 61(6):1–36, 2014.

[180] Gina M. L. C. Brock, Irit Aharony, Ryan P. Browne, Leon Eyrich Jessen, and Rasmus Waagepetersen. *CLValid: Validation of Clustering Results*, 2019. R package version 0.7.2.

[181] J. Philippe A. Lever and Hervé Abdi. *FactoMineR: Multivariate Exploratory Data Analysis and Data Mining*, 2017. R package version 1.41.

[182] Anil K Jain and Richard C Dubes. *Algorithms for Clustering Data*, volume 345. Prentice-Hall, Inc., 1988.

[183] Christian Hennig. *fpc: Flexible Procedures for Clustering*, 2021. R package version 2.2-12.

[184] William H Kruskal and W Allen Wallis. Use of ranks in one-criterion variance analysis. *Journal of the American Statistical Association*, 47(260):583–621, 1952.

[185] Ronald Aylmer Fisher. The arrangement of field experiments. *Journal of the Ministry of Agriculture*, 33(2):503–513, 1926.

[186] Carlo Emilio Bonferroni. Teoria statistica delle classi e calcolo delle probabilità. *Pubblicazioni del R Istituto Superiore di Scienze Economiche e Commerciali di Firenze*, 8:3–62, 1935.

[187] Lynn Waterhouse, Deborah Fein, and Charlotte Modahl. Neurofunctional mechanisms in autism. *Psychological review*, 103(3):457, 1996.

[188] Ira L. Cohen and Michael J. Flory. Autism Spectrum Disorder Decision Tree Subgroups Predict Adaptive Behavior and Autism Severity Trajectories in Children with ASD. *Journal of Autism and Developmental Disorders*, 49(4):1423–1437, April 2019.

[189] Donna Spiker, Linda J Lotspeich, Sue Dimiceli, Richard M Myers, and Neil Risch. Behavioral phenotypic variation in autism multiplex families: evidence for a continuous severity gradient. *American Journal of Medical Genetics*, 114(2):129–136, 2002.

[190] Deborah Fein, Michael Stevens, Michelle Dunn, Lynn Waterhouse, Doris Allen, Isabelle Rapin, and Carl Feinstein. Subtypes of pervasive developmental disorder: Clinical characteristics. *Child Neuropsychology*, 5(1):1–23, 1999.

[191] Leigh J Beglinger and Tristram H Smith. A review of subtyping in autism and proposed dimensional classification model. *Journal of Autism and Developmental Disorders*, 31:411–422, 2001.

[192] Jennifer Richler, Marisela Huerta, Somer L Bishop, and Catherine Lord. Developmental trajectories of restricted and repetitive behaviors and interests in children with autism spectrum disorders. *Development and psychopathology*, 22(1):55–69, 2010.

[193] Margot Prior, Richard Eisenmajer, Susan Leekam, Lorna Wing, Judith Gould, Ben Ong, and David Dowe. Are there subgroups within the autistic spectrum? a cluster analysis of a group of children with autistic spectrum disorders. *The Journal of Child Psychology and Psychiatry and Allied Disciplines*, 39(6):893–902, 1998.

[194] Michael C Stevens, Deborah A Fein, Michelle Dunn, Doris Allen, Lynn H Waterhouse, Carl Feinstein, and Isabelle Rapin. Subgroups of children with autism by cluster analysis: A longitudinal examination. *Journal of the American Academy of Child & Adolescent Psychiatry*, 39(3):346–352, 2000.

[195] Sylvie Verté, Hilde M Geurts, Herbert Roeyers, Yves Rosseel, Jaap Oosterlaan, and Joseph A Sergeant. Can the children's communication checklist differentiate autism spectrum subtypes? *Autism*, 10(3):266–287, 2006.

[196] Jay A Sevin, Johnny L Matson, David Coe, Steven R Love, Mary J Matese, and Debra A Benavidez. Empirically derived subtypes of pervasive developmental disorders: A cluster analytic study. *Journal of Autism and Developmental Disorders*, 25:561–578, 1995.

[197] Marisela Huerta and Catherine Lord. Diagnostic evaluation of autism spectrum disorders. *Pediatric Clinics*, 59(1):103–111, 2012.

[198] Rebecca Grzadzinski, Marisela Huerta, and Catherine Lord. Dsm-5 and autism spectrum disorders (asds): an opportunity for identifying asd subtypes. *Molecular autism*, 4(1):1–6, 2013.

[199] Meng-Chuan Lai, Michael V Lombardo, Bhismadev Chakrabarti, and Simon Baron-Cohen. Subgrouping the autism "spectrum": reflections on dsm-5. *PLoS biology*, 11(4):e1001544, 2013.

[200] Stelios Georgiades, Somer L Bishop, and Thomas Frazier. Editorial perspective: Longitudinal research in autism–introducing the concept of 'chronogeneity'. *Journal of Child Psychology and Psychiatry*, 58(5):634–636, 2017.

[201] Deborah K Anderson, Jessie W Liang, and Catherine Lord. Predicting young adult outcome among more and less cognitively able individuals with autism spectrum disorders. *Journal of child psychology and psychiatry*, 55(5):485–494, 2014.

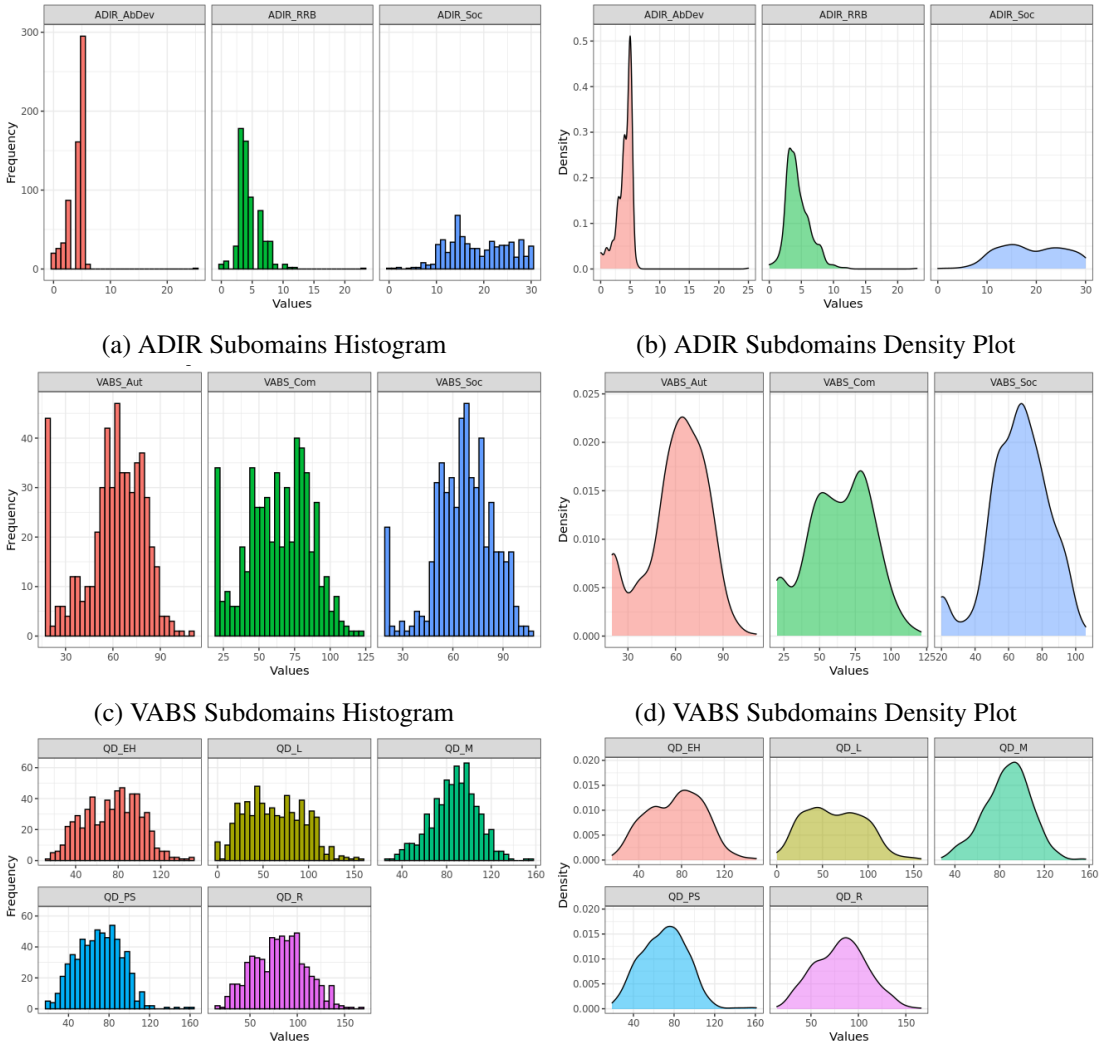
[202] Catherine Lord, Somer Bishop, and Deborah Anderson. Developmental trajectories as autism phenotypes. In *American Journal of Medical Genetics Part C: Seminars in Medical Genetics*, volume 169, pages 198–208. Wiley Online Library, 2015.

[203] Åsa Hedvall, Joakim Westerlund, Elisabeth Fernell, Fritjof Norrelgen, Liselotte Kjellmer, Martina Barnevik Olsson, Lotta Höglund Carlsson, Mats A Eriksson, Eva Billstedt, and Christopher Gillberg. Preschoolers with autism spectrum disorder followed for 2 years: those who gained and those who lost the most in terms of adaptive functioning outcome. *Journal of Autism and Developmental Disorders*, 45:3624–3633, 2015.

[204] Ovsanna Tadevosyan-Leyfer, Michael Dowd, Raymond Mankoski, BRIAN Winklosky, SARA Putnam, Lauren McGrath, Helen Tager-Flusberg, and Susan E Folstein. A principal components analysis of the autism diagnostic interview-revised. *Journal of the American Academy of Child & Adolescent Psychiatry*, 42(7):864–872, 2003.

Chapter 7

Annex



(a) ADIR Subomains Histogram

(b) ADIR Subdomains Density Plot

(c) VABS Subdomains Histogram

(d) VABS Subdomains Density Plot

(e) GDMS Subscales Histogram

(f) GDMS Subscales Density Plot

Figure 7.1: Histogram and Density Plot of Cluster Variables

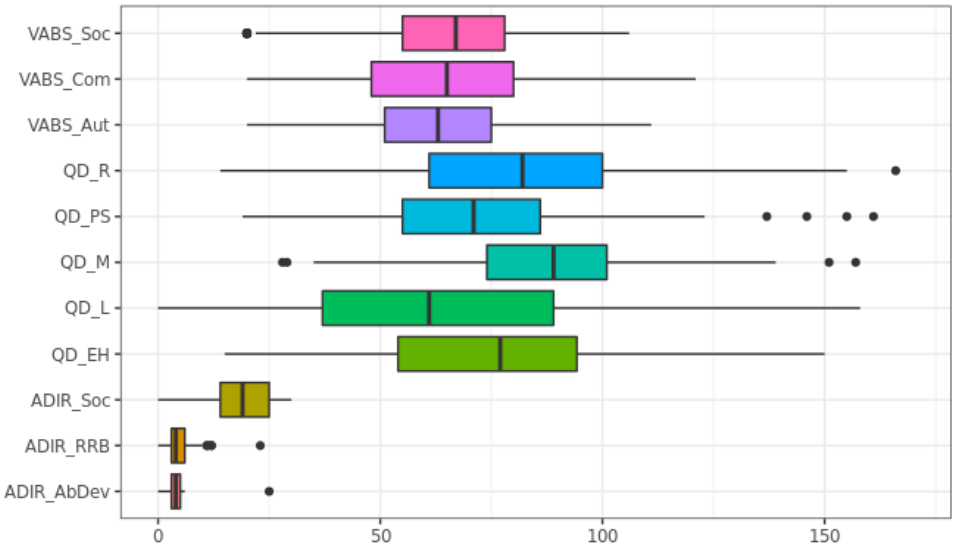
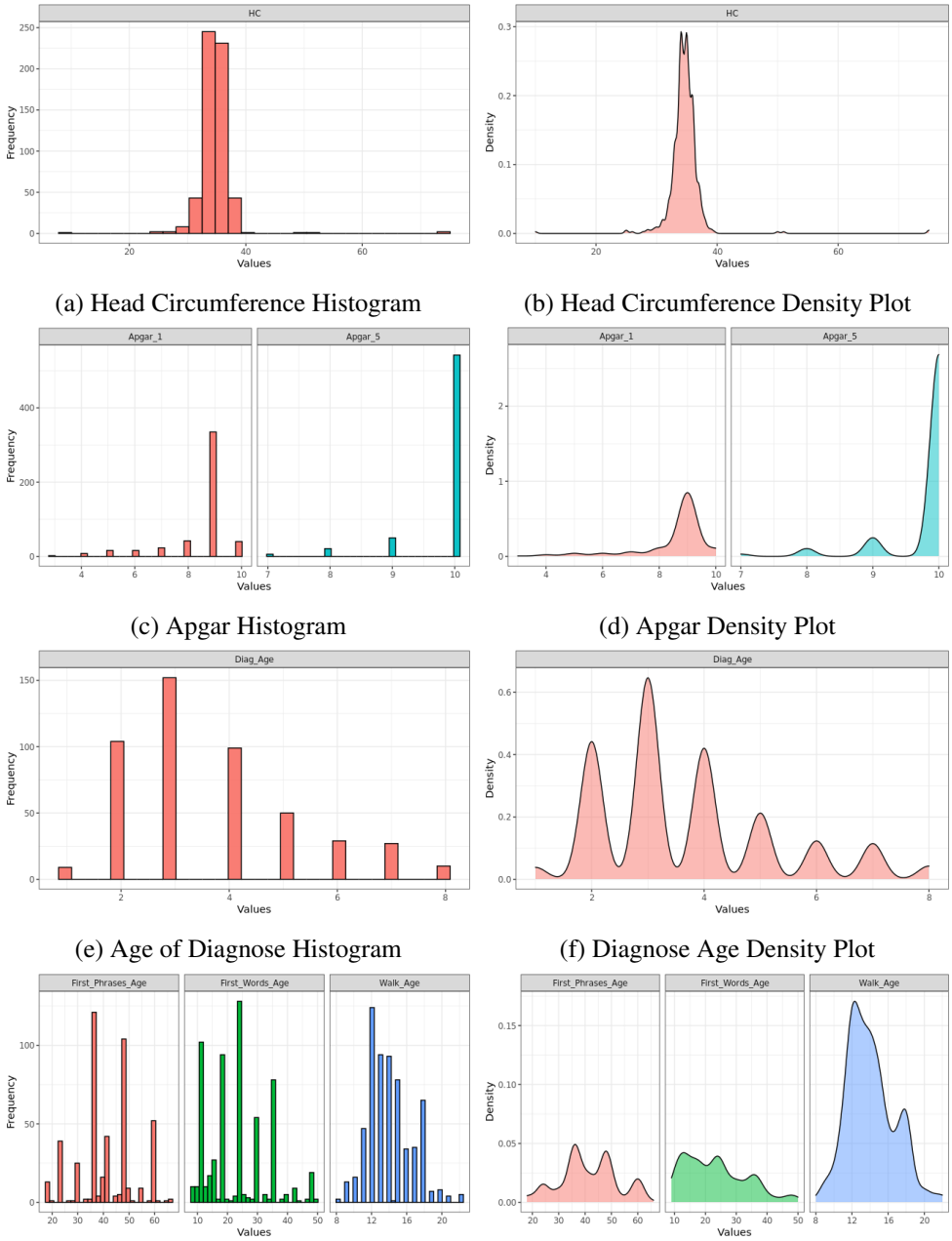


Figure 7.2: Boxplot of Cluster Variables

7. ANNEX



(a) Head Circumference Histogram

(b) Head Circumference Density Plot

(c) Apgar Histogram

(d) Apgar Density Plot

(e) Age of Diagnose Histogram

(f) Diagnose Age Density Plot

(g) Age of first Phrases, Words and Walk Histogram

(h) Age of first Phrases, Words and Walk Density Plot

Figure 7.3: Histogram and Density Plot of Numeric Categorization Variables

Table 7.1: Principal Components Eigenvalues and Variance Percentage

Principal Components (PC)	Eigenvalue Variance	Variance Percent	Percent Cumulative
PC 1	6.64	60.37	60.37
PC 2	1.20	10.87	71.24
PC 3	0.87	7.87	79.11
PC 4	0.62	5.63	84.74
PC 5	0.51	4.66	89.41
PC 6	0.39	3.58	92.99
PC 7	0.25	2.27	95.26
PC 8	0.17	1.52	96.77
PC 9	0.14	1.30	98.07
PC 10	0.11	1.01	99.08
PC 11	0.10	0.92	100

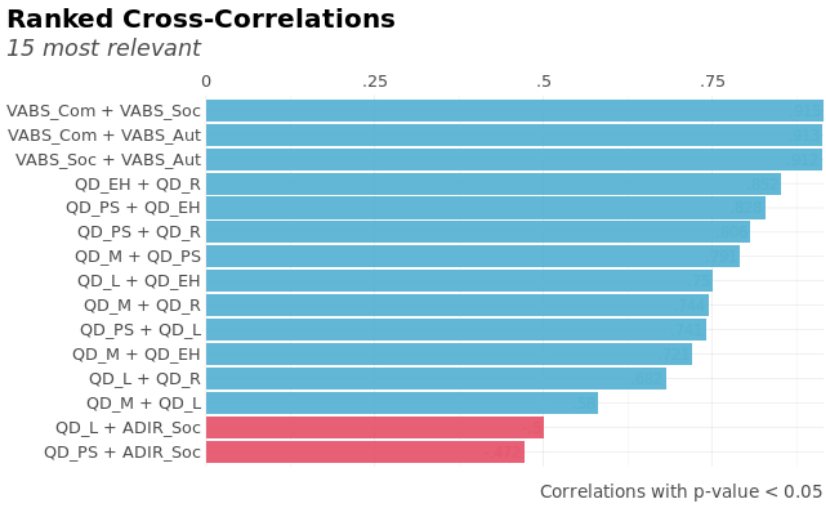


Figure 7.4: Cluster Variables Cross-Correlations Plot. Displayed the top 15 couples of variables (by correlation coefficient). Blue bars indicate a positive correlation while red bars indicate a negative correlation.

7. ANNEX

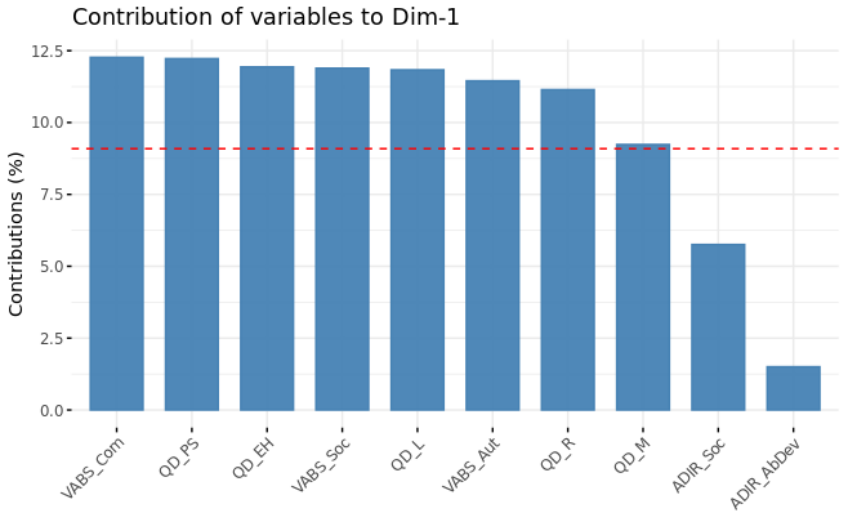


Figure 7.5: Contribution of Variables to the 1st Principal Component

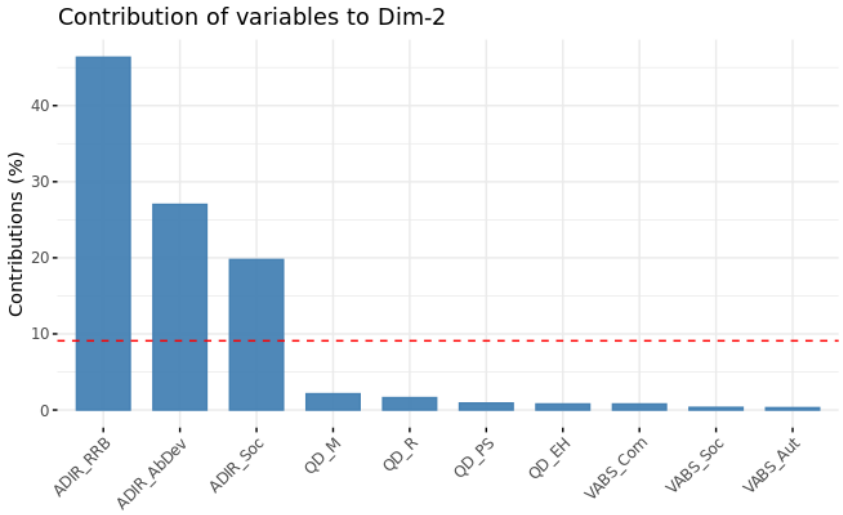
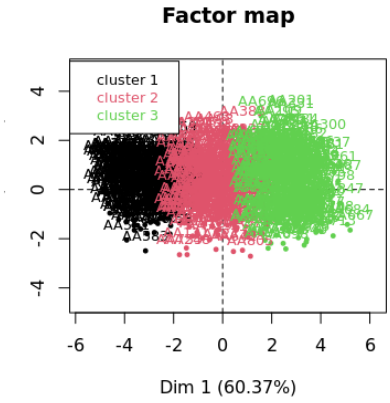
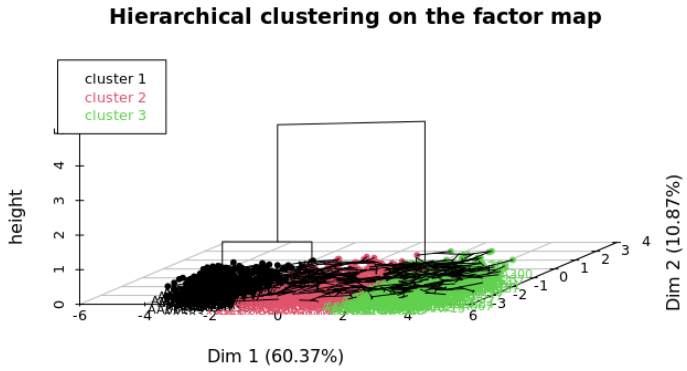


Figure 7.6: Contribution of Cluster Variables to the 2nd Principal Component



(a) Factor Map



(b) Hierarchical Clustering on factor map

Figure 7.7: Factor Map and Hierarchical Clustering on Factor Map

Note: Each dot in the chart highlights the most contributing variables for each Principal Component PC. The numerical values on the right-hand side of the chart denote the percentage of variance that each variable contributes to explain the different PC. The color of the dots conveys the range, based on the spectrum on the right-hand side. The darker shades of color signify a higher contribution.