UNIVERSITY OF TRENTO
BRUNO KESSLER FOUNDATION

Department of Psychology and Cognitive Science

PhD in Psychology and Education

XXXII cycle

Stratification of autism spectrum conditions by deep encodings

Supervisors:

Prof. Paola Venuti

Prof. Cesare Furlanello

PhD candidate:

Isotta Landi

Academic Year 2018/2019

Contents

Summary

This work aims at developing a novel machine learning method to investigate heterogeneity in neurodevelopmental disorders, with a focus on Autism Spectrum Conditions (ASCs). In ASCs, heterogeneity is shown at several levels of analysis, e.g., genetic, behavioral, throughout developmental trajectories, which hinders the development of effective treatments and the identification of biological pathways involved in gene-cognition-behavior links.

ASC diagnosis comes from behavioral observations, which determine the cohort composition of studies in every scientific field (e.g., psychology, neuroscience, genetics). Thus, uncovering behavioral subtypes can provide stratified ASC cohorts that are more representative of the true population. Ideally, behavioral stratification can (1) help to revise and shorten the diagnostic process highlighting the characteristics that best identify heterogeneity; (2) help to develop personalized treatments based on their effectiveness for subgroups of subjects; (3) investigate how the longitudinal course of the condition might differ (e.g., divergent/convergent developmental trajectories); (4) contribute to the identification of genetic variants that may be overlooked in case-control studies; and (5) identify possible disrupted neuronal activity in the brain (e.g., excitatory/inhibitory mechanisms).

The characterization of the temporal aspects of heterogeneous manifestations based on their multi-dimensional features is thus the key to identify the etiology of such disorders and establish personalized treatments. Features include trajectories described by a multi-modal combination of Electronic Health Records (EHRs), cognitive functioning and adaptive behavior indicators. This thesis contributes in particular to a data-driven discovery of clinical and behavioral trajectories of individuals with complex disorders and ASCs. Machine learning techniques, such as deep learning and word embedding, that proved successful for e.g., natural language processing and image classification, are gaining ground in healthcare research for precision medicine. Here, we leverage these methods to investigate the feasibility of learning data-driven pathways that have been difficult to identify in the clinical practice to help disentangle the complexity of conditions whose etiology is still unknown.

In Chapter 1 we present a new computational method, based on deep learning, to stratify patients with complex disorders; we demonstrate the method on Multiple Myeloma, Alzheimer's disease, and Parkinson's disease, among others. We use clinical records from a heterogeneous patient cohort (i.e., multiple disease dataset) of $\sim 1.6M$ temporally-ordered EHR sequences from the Mount Sinai Health system's data warehouse to learn unsupervised patient representations. These representations are then leveraged to identify subgroups within complex condition cohorts via hierarchical clustering. We investigate the enrichment of terms that code for comorbidities, medications, laboratory tests and procedures, to clinically validate our results.

A data analysis protocol is developed in Chapter 2 that produces behavioral embeddings from observational measurements to represent subjects with ASCs in a latent space able to capture multiple levels of assessment (i.e., multiple tests) and the temporal pattern of behavioral-cognitive profiles. The computational framework includes clustering algorithms and state-of-the-art word and text representation methods originally developed for natural language processing. The aim is to detect subgroups within ASC cohorts towards the identification of possible subtypes based on behavioral, cognitive, and functioning aspects. The protocol is applied to ASC behavioral data of 204 children and adolescents referred to the Laboratory of Observation Diagnosis and Education (ODFLab) at the University of Trento.

In Chapter 3 we develop a case study for ASCs. From the learned representations of Chapter 1, we select $1,439$ individuals with ASCs and investigate whether such representations generalize well to any disorder. Specifically, we identify three subgroups within individuals with ASCs that are further clinically validated to detect clinical profiles based on different term enrichment that can inform comorbidities, therapeutic treatments, medication side effects, and screening policies.

This work has been developed in partnership with ODFLab (University of Trento) and the Predictive Models for Biomedicine and Environment unit at FBK. The study reported in Chapter 1 has been conducted at the Institute for Next Generation Healthcare, Department of Genetics and Genomic Sciences, Icahn School of Medicine at Mount Sinai (NY).

List of abbreviations

AD Alzheimer's Disease

ADHD Attention Deficit Hyperactivity Disorder

ADI-R Autism Diagnostic Interview - Revised

ADOS Autism Diagnostic Observation Schedule

ADT Androgen Deprivation Therapy

AE Autoencoder

ASCs Autism Spectrum Conditions

ASDEU Autism Spectrum Disorder in the European Union

BC Breast Cancer

BIQ Brief Intelligence Quotient

CD Crohn's disease

CE Cross Entropy

CNN Convolutional Neural Network

COPD Chronic Obstructive Pulmonary Disease

CPTs Current Procedural Terminologies

CSS Calibrated Severity Score

DP Deep Patient

DSM-5 Diagnostic and Statistical Manual of Mental Disorders – Fifth Edition

ECG Electrocardiogram

EHRs Electronic Health Records

FMI Fowlkes-Mallows Index

FSIQ Full Scale Intelligence Quotient

GForms Google Form questionnaires

GI Gastrointestinal

GloVe Global vectors

GMDS Griffiths Mental Development Scales

GQ General Quotient

GWAS Genome Wide Association Studies

HCT Hematopoietic Cell Transplantation

IDs Intellectual Disabilities

IQ Intelligence Quotient

LSTM Long Short-Term Memory

ML Machine Learning

MM Multiple Myeloma

MSDW Mount Sinai Health System's data warehouse

NLP Natural Language Processing

ODFLab Laboratory of Observation Diagnosis and Education

PAD Peripheral Artery Disease

PC Prostate Cancer

PD Parkinson's disease

PDD-NOS Pervasive Developmental Disorders – Not Otherwise Specified

PPMI Progression Markers Initiative

PSA Prostate-specific Antigen

PSI-SF Parental Stress Index - Short Form

RF Random Forest

RRB Restricted and Repetitive Behaviors

RT Radiation Therapy

SA Social Affect

SNPs Single-Nucleotide Polymorphisms

SRS Social Responsiveness Scale

SVD Singular Value Decomposition

SVM Support Vector Machine

TD Typical Development

T2D Type 2 Diabetes

TFIDF Term Frequency - Inverse Document Frequency

UMAP Uniform Manifold Approximation and Projection

Chapter 1

Complex disorder stratification via deep unsupervised learning of Electronic Health Records

Abstract

Hospital-based Electronic Health Records (EHRs) provide a heterogeneous but structured information that can be used for the discovery of patterns associated to patient's health status. In particular, we aim at using EHRs to identify health (or disease) trajectories and possibly cluster them into groups of patients that may have common patterns of disease progression or response to treatment. However, there are multiple challenges in analyzing EHR data and obtaining a coherent representation usable to train machine learning methods for early diagnosis, discovery of most effective treatments, and prediction of disease progression.

In this chapter, we introduce a machine learning approach for the pre-processing of such representation. In particular, we apply deep learning to learn individual patient representations from heterogeneous EHR sequences. Such representations can be leveraged for disease subtyping, considering EHRs from large scale hospital cohorts and comprising of different clinical concepts (e.g., ICD-9 diagnosis, medications). In particular, to transform patient sequences into low-dimensional latent vectors, we propose the ConvAE deep learning architecture, which provides an *end-to-end*, unsupervised representation learning model. The ConvAE representations were compared to different baselines in a disease-specific detection task aimed at disease stratification analysis at scale.

The ConvAE architecture obtained the best results in identifying eight different complex conditions, including neurodevelopmental (e.g. Attention Deficit Hyperactivity Disorder – ADHD) and cancer (e.g. Multiple Myeloma). For disease subtyping, we applied hierarchical clustering to complex disorder cohorts and we identified more homogeneous subtypes tied to different disease progressions, symptoms, and response to treatments.

We discuss how the stratification of informative patient subgroups tied to characteristic patterns within the possible heterogeneous conditions can guide personalized medicine.

1.1. Introduction

Electronic Health Records (EHRs) have revolutionized healthcare and biomedical research over the past three decades. In recent years, with the advent of new hardware and storage capabilities, EHRs have evolved into a complex framework, housing massive amounts of digital data. These digital records can be *structured*, typically including disease diagnoses, medication prescriptions, performed procedures, and laboratory test results, or *unstructured*, storing progress reports, and free-text clinical notes. In providing a snapshot of a patient's state of health, EHRs have created new opportunities to investigate the predictors and properties of health-related events across large and heterogeneous populations, which can provide clarity and insights into the future effects of medical decisions. At the individual level, these trajectories are becoming the basis for personalized medicine. Across patient cohorts, EHRs and derived trajectories provide a vital resource to understand population health management and make better decisions for healthcare operational policies.

However, EHRs are challenging to represent and utilize in computational models because of their high dimensionality, heterogeneity, sparseness, random errors, and systematic biases. Namely, patient clinical annotations are not uniform across healthcare facilities, where the same clinical phenotype can be expressed using different codes and terminologies. For example, a patient diagnosed with Type 2 Diabetes Mellitus (T2D) can be identified by a variety of EHR-related inputs, such as 1) laboratory values of hemoglobin A1C greater than 7.0%; 2) the presence of the `250.00` ICD-9 code; or 3) "Type 2 Diabetes Mellitus" mentioned in free-text (i.e., unstructured) clinical notes. These challenges have motivated recent developments of machine learning methods that attempt to accurately identify diseases from EHRs for modeling and risk prediction (see, e.g., [1, 2, 3]).

Given a specific disease, heterogeneity among patients usually leads to different progression patterns and may require different types of interventions, despite equivalence of these patients at the diagnostic level. This is particularly evident for *complex disorders*, whose disease etiology is still mostly unknown, possibly due to multiple genetic, environmental, and lifestyle factors. Patients with complex disorders may differ on multiple levels of analyses (e.g., different clinical measures such as laboratory tests or medications, or different comorbidities) and in response to treatments throughout the disease trajectory, making these conditions difficult to evaluate. For example, heterogeneity of Autism Spectrum Conditions (ASCs) has been detected by comorbidity stratification studies. Indeed, coexisting conditions are registered in 70% of individuals with ASC [4]. Genetic studies report hundreds of ASC-linked risk genes and a heritability at around $50 - 80\%$ [5, 6] as well. Moreover, the investigation of developmental trajectories [7] differentiates behavioral growth patterns according to their progression and severity level of core symptoms along with differentiations of cognitive and language impairments relative to the norm. This variability presumably reflects the existence of multiple subgroups tied to diverse genetic and environmental etiologies and that, although convergent to the behavioral autistic profile, show a range of clinical phenotypic characteristics under behavioral, linguistic and cognitive levels.

Several different conditions have been referred to as *complex*, such as Parkinson's disease (PD), Multiple Myeloma (MM) and T2D. Clinical heterogeneity is common among PD patients and includes both motor and nonmotor symptoms (e.g., cognitive impairment) along with varying rates of disease progression [8]. Stratification studies have addressed PD heterogeneity at the clinical, pathological, and genetic level. Moreover, longitudinal approaches aim at informing prognostic models that can be used at the individual level [9]. MM haematologic malignancy presents heterogeneity at the molecular, clinical, and clonal level and only part of the prognostic variability can be explained by underlying biology. In particular, age and the extent of disease involvement (e.g., renal function and number of bone lesions) are considered important outcome determinants [10]. T2D is a metabolic disorder strongly influenced by genetics and environment. Among lifestyle factors, obesity is considered a strong indicator of T2D risk. The multiple phenotypic manifestations of this condition are thought to stem from the disruption of one or more etiological pathways, such as beta cell function, beta cell mass, insulin action, glucagon secretion/action, incretin secretion/action, and fat distribution [11]. While patients for these conditions all report with the same diagnosis, a range of etiologies and manifestations are clearly present.

Personalized medicine focuses on the use of patient-specific data to tailor treatment to an individual's unique health status. However, even seemingly simple diseases can show different degrees of complexity that can create challenges for identification, treatment, and prognosis. For instance, despite being Mendelian disorders, both cystic fibrosis and Huntington's disease exhibit a broad array of symptoms and varying degrees of phenotypic manifestations [12, 13]. Multiple data types in patient EHR histories offer a way to examine disease complexity, and present an opportunity to refine individually ICD-coded diseases into subtypes and tailor personalized treatments. Stratifying patients according to their clinical phenotypes into informative subgroups can also help identify high-impact biomarkers that can be overlooked. In Genome Wide Association Studies (GWAS), it has been reported that phenotypic heterogeneity substantially reduces the magnitude of association of genetic variants [14]. This suggests that the generalizability and subsequent validity of case-control studies may suffer due to a failure to identify genetic heterogeneity. From a computational perspective, patient stratification (or patient subtyping) can be considered as an unsupervised learning task to group patients according to their historical records [15] towards data-driven discoveries without an a priori knowledge.

The method proposed in this chapter is a novel *end-to-end*, unsupervised deep learning architecture to derive patient representations from heterogeneous cohorts of EHR sequences. The model learns patient representations from a heterogeneous EHR dataset of $\sim 1.6M$ patients from the Mount Sinai Health System's data warehouse (MSDW) located in New York. To demonstrate the feasibility of deriving clinically meaningful associations (e.g., diseases progression and clinical encounters) from learned representations of diseases, we further cluster patients into subgroups within diagnoses using hierarchical clustering to clinically evaluate them. We successfully identify informative subgroups within different complex disorders. For example, patients with MM divide in 5 subgroups according to comorbid conditions, treatments, and therapy side effects.

1.1.1. Background: searching for heterogeneity with deep learning embeddings

Studies on disease stratification in EHRs are often based on disease-specific cohorts of patients and manually selected features. Generally, EHR clinical terms are aggregated at the patient level, and each patient is represented by a vector. This vector can be used to derive disease subtypes in a population via clustering or topological analysis. For example, Li and colleagues [16] identify three distinct subgroups of T2D from topology-based patient-to-patient networks. Subgroup I is characterized by T2D complications, diabetic nephropathy, and diabetic retinopathy; subgroup II is enriched for cancer and cardiovascular diseases; and subgroup III is strongly associated with cardiovascular and neurological diseases. These findings were then validated through genotyping, and demonstrate that the enriched phenotypes and biological functions at gene level match the clinical differences and disease comorbidities within each subgroup. These distinctions suggest that tailored treatments rather than blanket approaches for T2D may be more effective for disease outcomes. In other bodies of work, statistical modeling and machine learning are applied to 1) stratify patients with hypertension and determine their responsiveness to therapy [17]; 2) identify different subgroups of patients at risk of heart failure in order to derive different types of treatments and considerations [18]; 3) separate progression stages of chronic kidney disease [19]; 4) subgroup EHRs of children with ASCs via hierarchical clustering [20]; and 5) distinguish subgroups of patients with PD according to longitudinal data of motor progression and cognitive functioning [21].

Natural Language Processing algorithms, such as Word2vec continuous *Skip-gram* and *Bag-of-words* methods [22], have been tailored to EHRs to perform medical concept contextual embeddings via neural networks. These contextual embeddings represent structured electronic records in a metric space, and identify semantic similarities between terms. Utilizing these word encodings, patients can be represented as a multidimensional vector that captures term dependencies throughout their health trajectories.

By leveraging medical term encodings learned from over a million patients, Glicksberg et al. [23] attempt to generate automated disease representations. Each patient's clinical history is summarized by weighted averages of the corresponding medical embeddings over time windows. Next, the distance between the patient's embedded history to a disease key-word representation is computed to generate disease cohorts. Medical concept embeddings have also been used to extract supervised patient representations for similarity studies. In particular, Zhu and colleagues [24] propose a patient similarity evaluation framework that preserves the temporal properties of EHRs. Each patient is represented by an embedding matrix of EHRs that include four years of data (i.e., 40 terms). Each matrix is then input into a convolutional neural network trained to learn pairwise similarities of patients from Chronic Obstructive Pulmonary Disease (COPD), diabetes, heart failure, and obesity cohorts. Similarly, Suo et al. [25] develop a time fusion convolutional neural network trained on fixed-sized sub-frames of the embedded patient matricies. After computing the weighted average of the output vectors, the authors capture both the local and the global contributions of these at different time intervals. Patient similarity scores are then used to predict COPD, diabetes, and obesity.

Recently, unsupervised deep learning has been applied to healthcare data for patient stratification. Such approaches are generally tested on quantitative synthetic data or small cohorts of patients with a specified disease, and explicitly model the timeline of care events. For example, Baytas et al. [15] develop a time-aware Long Short-Term Memory (LSTM) that applies an autoencoder framework to learn representations of patients for disease subtyping. The proposed model is initially tested on synthetic EHRs. Afterwards, the model is used to stratify patient data from the Parkinson's Progression Markers Initiative (PPMI) clinical and longitudinal study. The authors are able to identify two distinct subgroups of patients, one of which shows imaging assessment measures consistent with

PD patients. Similarly, Zhang and colleagues [26] apply LSTM to 466 patients with idiopathic PPMI markers to perform patient subtyping, and obtain three different subclusters that differ in disease progression patterns and symptom severity.

From clinical features to unsupervised representations

From this body of work it emerges that existing studies have commonly identified patient subgroups from disease specific EHR datasets by utilizing manually selected features [15, 16, 26]. Moreover, unsupervised representations have also been derived without considering temporal aspects of data [27]. However, because EHRs are longitudinal, capturing the dependencies between elements of the patient record in sequence allows models to learn more robust representations. These ordered representations can then be used to obtain patient subgroups. Moreover, since EHRs tend to be incomplete, using diverse cohorts to derive disease-specific subgroups allow us to adequately capture the features of heterogeneity within the population of interest.

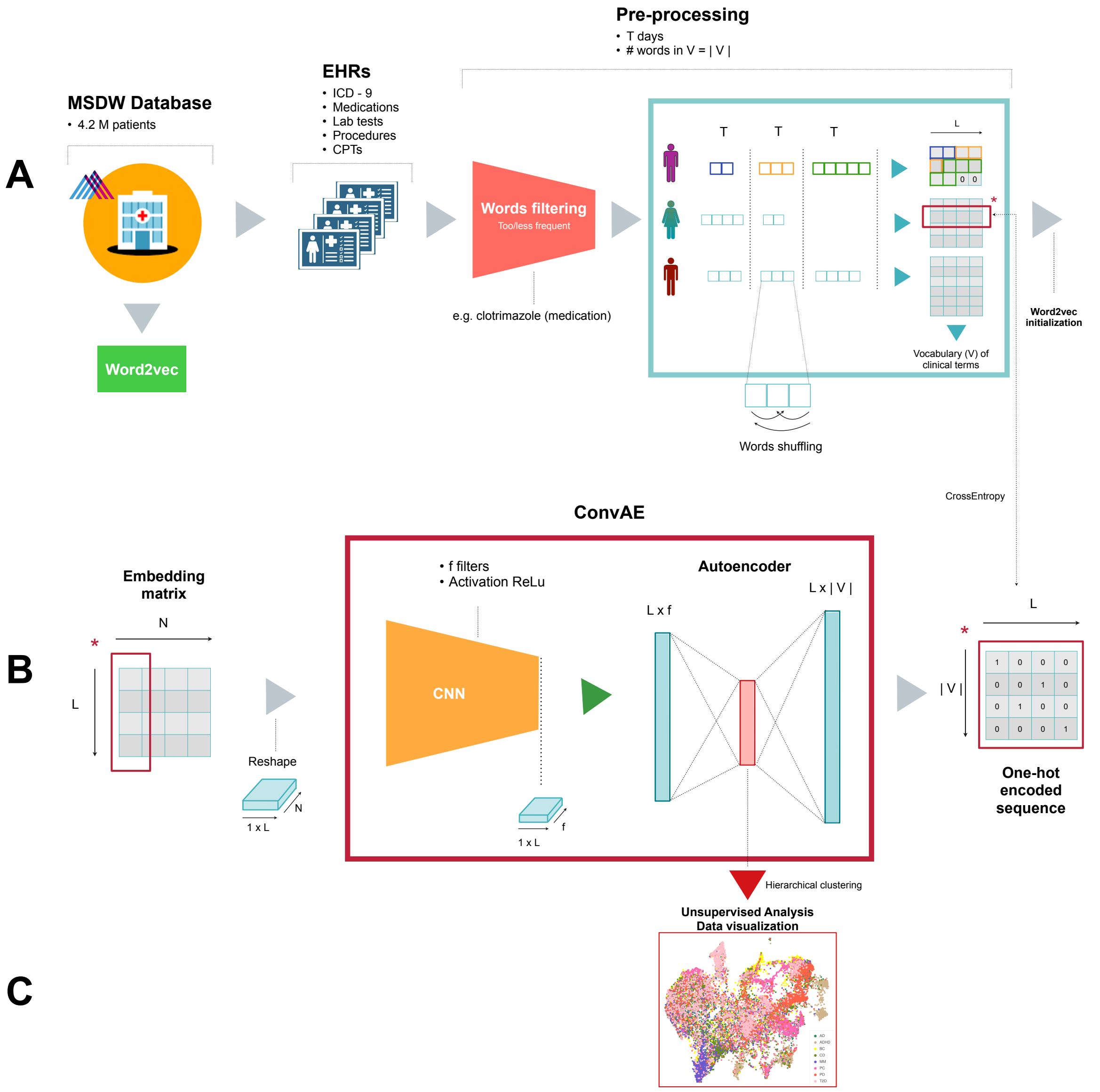
The proposed method is a novel *end-to-end*, unsupervised deep learning architecture to derive patient representations from a large domain-free EHR dataset. This approach favors scalability and generalizability: first, it includes embeddings of medical terms (i.e., ICDs, medications, procedure codes, laboratory tests) to clinically contextualize concepts into clear subgroups of disorders, and, second, a combination of convolutional neural networks and autoencoders (ConvAE), to model the temporal aspects of patient data and to enable the application of an unsupervised architecture. This architecture eliminates the need for manual *ad hoc* feature engineering. In addition, it processes the whole EHR sequence, regardless of the length of patient history, and does not require explicit modeling of events within patient care timelines. By generating empirically-based disease subgroups, this architecture can help identify high-impact biomarkers within complex disorders, whose effect may be masked in case-control studies. As such, we are able to isolate properties of these patient subgroups that can then be used to advance precision medicine through the development of novel and more personalized treatments. Ideally, when a new patient enters the medical system, their health status progression can be tied to a specific subgroup, thereby informing prognosis and effective treatments.

1.2. Material and methods

The proposed pipeline to derive patient representation subtypes from EHRs is based on three steps: A) data pre-processing; B) *end-to-end* unsupervised modeling, which includes medical concept embeddings, and ConvAE (see Section 1.2.3); and C) clustering analysis (see Figure 1.1).

1.2.1. Patient medical records

We use an EHR dataset from the Mount Sinai Health System's data warehouse, located in New York City. This dataset is generated from Mount Sinai Health System's healthcare and clinical operations,



[Courtesy of Nicole Bussola, PhD candidate]

Figure 1.1: Unsupervised individual representation pipeline. (A) Data pre-processing; (B) *End-to-end* unsupervised deep representation modeling; (C) Clustering and visualization.

which includes a high volume of structured, semi-structured and unstructured data. These data include inpatient, outpatient, and emergency room visits. Patients in the system can have up to 12 years of follow up data unless they are transferred or move their residence away from the hospital system. The hospital dataset contains approximately 4.2 million patients, spanning the years from 1980 to 2015. For this study, we utilize pseudo de-identified patient histories, which include only structured data including: diagnoses, medications, laboratory tests, procedures, current procedural terminologies (CPTs), and patient demographics. EHR data for patients are first extracted from the clinical data warehouse and every patient is represented as a longitudinal sequence s_p of length L_p of aggregated temporally-ordered medical terms, i.e., $s_p = (w_1,\ w_2,\ \ldots, w_{L_p})$. Each term w_i is a medical concept (i.e., diagnosis, medication, procedure, CPT code, or laboratory test) that the patient p has represented in their EHR history. As an example, a term w_i can be ICD-9 diagnostic code `362.11`, which corresponds to *Hypertensive retinopathy*. The entire collection of terms makes up a vocabulary V.

1.2.2. Data pre-processing

Data pre-processing comprises: 1) filtering of the least frequent medical terms; 2) dropping of redundant terms within fixed time frames; 3) subsequencing of long sequences to include the complete patient history while leveraging a convolutional framework.

Filtering Taking into account patient sequences, we evaluate medical concept occurrences to filter the least frequent terms. To select *clinical stop words*, we consider the whole EHR corpus as a document D and each patient sequence s_p as a sentence. For each term w of the vocabulary V we compute the probability of finding w in corpus D. We multiply this probability by the sum of the probabilities to find w in a sentence s_p for all sentences in the corpus. In particular, let P be the set of all patients and V the vocabulary of terms, then $\forall w \in V$:

$$\mathbb{P}(w \in D) \sum_{p \in P} \mathbb{P}(w \in s_p) = \frac{\#\{s \in D;\ w \in s\}}{|D|} \sum_{p \in P} \frac{\#\{w_i \in s_p,\ w_i = w\}}{|s_p|} \tag{1.1}$$

where $|D|$ is the number of sentences in document D and $|s_p|$ is the length of a patient sequence. The filtering score from Eq. (1.1) includes document frequency, i.e., the number of patients with at least one occurrence of a term w, and term frequency, which is the total number of occurrences of term w in a patient sequence. To eliminate terms that constitute noise (i.e., occur multiple times in few patients, or are too general and not informative) we drop terms with filtering scores inferior to a cut-off.

Drop duplicates Every encounter of a patient with the healthcare system can span multiple days and include repeated concepts. To avoid noise from overlapping medical concepts that are artifacts of an EHR system with repeated codes, we further drop duplicate medical concepts in non-overlapping intervals of length T days, as done previously by [23, 28]. Next, within the same time window, we randomly shuffle patient concepts after deletion, given that multiple events within the same encounter occur randomly.

Subsequencing All sequences with less than 3 concepts are removed, and the longest sequences are truncated to a maximum of $5,000$ terms. Patient sequences are then chopped into subsequences of fixed length L that will be used to train ConvAE model. A typical value for L is 32, see Section 1.2.5. Each patient sequence becomes

$$s_p = [(w_1, \ldots, w_L), (w_{L+1}, \ldots, w_{2L}), \ldots]$$

and subsequences shorter than L are padded with 0 up to length L. This is done to account for the complete patient care history in future processing, which requires uniform dimensions. For the sake of clarity, in the following we present the architecture as applied to a subsequence of terms $(w_1, \ldots, w_L)$.

1.2.3. ConvAE architecture

Here, we describe the ConvAE architecture that models vector representations of patients from the pre-processed raw patient subsequences obtained previously. The architecture consists of three stacked modules that take patient term subsequences as inputs, and reconstruct them within an autoencoder (AE). The hidden layer of the AE module becomes the latent patient history representation (see Figure 1.1 B).

Medical concept embedding See Figure 1.1 (A). Given a subsequence $s = (w_1, w_2, \ldots, w_L)$, the first module assigns each medical concept w to an N-dimensional embedding vector v_w. Specifically, a subsequence can alternatively be represented as an $(L \times N)$ matrix $E = (v_{w_1},\ v_{w_2},\ \ldots, v_{w_L})^T$, where L is the patient subsequence length, and N is the embedding dimension. The embedding matrix is learned during training and is initialized with pre-trained embeddings on the complete Mount Sinai data warehouse ($\sim 4M$ patients). This step is conceived to simultaneously represent the semantic relationships between medical concepts from contextual terms within an EHR, and derive a progressive patient representation as well. This structure retains temporal information because the rows of matrix E are temporally ordered according to patient visits, and also enables a fine-grained representation of an individual medical history via vectors that represent a single medical term from a specific visit.

ConvAE See Figure 1.1 (B). The second module applies a convolutional neural network (CNN) to extract local temporal patterns in patient subsequences, while the third module leverages an AE to learn a representation of the encoded temporal patterns. The CNN applies temporal filters to each patient EHR embedding matrix and returns the filtered response. CNN filters applied to an EHR embedding matrix usually perform a one-side convolution operation across time. This means that a filter can be defined as $k \in \mathbb{R}^{h \times N}$, where h is the variable window size and N is the embedding dimension [24, 25]. Differently from images, where the convolutional filter generates relevant spatial information in 2 dimensions, the second dimension of a patient embedding matrix has no spatial meaning. In fact, only the first dimension carries temporal information that can be captured via filter sliding.

Our approach differs in that it processes embedding matrices as they were RGB images carrying a third "depth" dimension. We reshape the $(L \times N)$ embedding matrix into $\tilde{E} \in \mathbb{R}^{1 \times L \times N}$ and we consider the embedding dimensions as channels. We then apply f filters $\mathbf{k} \in \mathbb{R}^{1 \times h \times N}$ to padded input to keep the same output dimension and learn more sophisticated features that may grasp temporally-further sequence characteristics. In particular, we obtain, for each filter j

$$(R)_j = \text{ReLU}(\sum_{i=0}^{N-1} \mathbf{k}_i \star \tilde{\mathbf{e}}_i + \mathbf{b}_j), \; j = 1, \ldots, f \tag{1.2}$$

with the output matrix $R \in \mathbb{R}^{1 \times L \times f}$, $\mathbf{k}_i$ the h-dimensional weight matrix at depth i, $\tilde{\mathbf{e}}_i \in \mathbb{R}^{1 \times L}$ the input matrix i-th embedding dimension, and $\mathbf{b}$ the bias vector. In Eq. (1.2), $\text{ReLU}(x) = \max(0, x)$ is the Rectified Linear Unit function, and $(\star)$ the convolution function. Max Pooling is then applied, and the maximum filter activations are retained. The output is then reshaped into a concatenated vector of dimension Lf. With images, we expect different neurons along the depth dimension to activate in presence of edges or blobs of color. As a result, the algorithm would learn different weights for each embedding dimension to highlight relevant interdependencies of medical terms, and tune vector representations of patient histories to identify the most relevant characteristics of their semantic space. These filtered responses are then applied to an AE which outputs a vector that estimates the given input subsequence. Specifically, a one-layer AE takes a dense vector $\mathbf{x} \in \mathbb{R}^{Lf}$ as an input and transforms it to a hidden representation $\mathbf{y} = a(\mathbf{W}\mathbf{x} + \mathbf{b})$, where $a(\cdot)$ is a non-linear activation function, $\mathbf{W}$ is the weight coefficient matrix, and $\mathbf{b}$ is the bias vector. The latent representation is then mapped back to a reconstructed vector $\mathbf{z} \in \mathbb{R}^{|V|L}$ that aims at recreating the initial one-hot encoded sequence s with $\mathbf{z} = a(\mathbf{W}'\mathbf{y} + \mathbf{b}')$, and $|V|$ the dimension of the vocabulary of medical terms. We extract the hidden representation $\mathbf{y}$, a H-dimensional vector, as the encoded representation of the initial patient subsequence input. For each patient, we then compute the component-wise average of all subsequence encodings to obtain a unique latent vector representation of dimension H.

To train the ConvAE architecture, we set up a multi-class multi-label classification task for ConvAE to reconstruct each initial input subsequence of medical terms, from their one-hot encoded representations. These medical terms have been previously indexed in the medical concepts vocabulary from our EHR data. Given a subsequence of medical concepts s, the ConvAE is trained by minimizing the Cross Entropy (CE) loss

$$\text{CE}(\text{Softmax}(O), \; s) = -\frac{1}{L} \sum_{j=1}^{L} \log(\text{Softmax}(O^j)_{w_j}),$$

where O is the output of the ConvAE module reshaped into a matrix of dimension $|V| \times L$, w_j is the

j-th element of sequence s that correspond to a term indexed in vocabulary V and

$$\text{Softmax}(O^j)_i = \frac{\exp O_i^j}{\sum_{i=1}^{|V|} \exp O_i^j} \quad i = 1, ..., |V|. \tag{1.3}$$

Since the objective function consists of only self-reconstruction errors, the model can be trained without any supervised training samples.

1.2.4. Hierarchical clustering

Agglomerative hierarchical clustering is a widely used technique to detect possible subgroups in a dataset, identifying most similar elements according to a distance function. Each individual element starts out as its own cluster, then clusters are progressively merged together minimizing a loss function. Different choices for loss functions are available, e.g., Ward's method recursively merges the pair of clusters for which the within cluster variance increase is minimum. More specifically, the distance between two points $\mathbf{x}_i$, $\mathbf{x}_j$ is the Euclidean distance, i.e., $d(\mathbf{x}_i, \mathbf{x}_j) = \sqrt{\sum_k (x_{ik} - x_{jk})^2}$, and the distance that we want to minimize at subsequent iterations between clusters C_1 and C_2 is:

$$\tilde{d}(C_1, C_2) = \sqrt{\frac{|C_2| + |s|}{T} d(C_2, s)^2 + \frac{|C_2| + |t|}{T} d(C_2, t)^2 - \frac{1}{T} d(s, t)^2}$$

where $T = |C_2| + |s| + |t|$, $|\cdot|$ denotes the set cardinality, and $s,\ t$ are the elements of cluster C_1.

Hierarchical clustering is an unsupervised clustering technique that aims at identifying patterns in the absence of predetermined labels. We apply hierarchical clustering to our hidden patient representations (see Figure 1.1 C) in order to validate the ConvAE architecture and identify informative subgroups within complex conditions. A shortcoming of the hierarchical clustering method is that it requires the number of clusters to be defined in advance. A way to automate this decision process is via the Elbow Method, which empirically selects the smallest number of clusters that minimize the increase in explained variance. With a top-down approach, first, it lists the minimum distances $\tilde{d}$, which correspond to the within cluster variance increase at a new iteration. This variance tends towards zero as the number of clusters increases, until we observe single-point clusters. Then, the Elbow Method returns the minimum number of clusters for which the variance increase is relatively small. This value corresponds to the "elbow" of the curve $(n_c, \tilde{d}(\cdot))$, where n_c is the number of clusters, and $\tilde{d}(\cdot)$ is the merging score from $n_c - 1$ to n_c clusters.

1.2.5. Implementation details

The implementation details reported here are empirically derived by trying to balance efficiency and effectiveness of the model. We validate the model on smaller subsets of data (i.e., using disease cohorts and random patients) with time interval T equal to $\{15,\ 30\}$; subsequence length L equal to $\{32,\ 64\}$; and embedding dimension N spanning $\{100,\ 128,\ 200\}$.

All modules are implemented in Python, version 3.5.2, and `scikit-learn` and `pytorch` libraries

	Fold-1		**Fold-2**	
	Training	**Test**	**Training**	**Test**
Patients	$741,177$	$751,979$	$740,922$	$751,900$
Median sequence length	41	40	41	40
Subsequences	$3,636,014$	$3,656,238$	$3,644,596$	$3,647,426$
Mean N of subseq per patient (sd)	4.91 (12.13)	4.86 (12.06)	4.92 (12.14)	4.85 (12.06)
Vocabulary size	$32,799$	$32,156$	$32,875$	$32,210$

Table 1.1: Description of pre-processed training and test sets' characteristics

[29, 30]. Computations were run on a server with an NVIDIA TITAN V GPU. Code can be found at https://github.com/landiisotta/ehr_stratification.

Dataset For this study, we utilize pseudo de-identified health histories of $1,608,741$ patients from MSDW, which include $57,464$ clinical concepts in the form of structured data (see Section 1.2.1). The medical concept dataset for this study consists of ICD-9 diagnosis codes (all ICD-10 codes are backwards mapped to corresponding ICD-9 codes), medications (normalized with RxNorm), CPTs and procedures, and laboratory tests (normalized with LOINC). The cohort includes $900,932$ females, $691,321$ males, and $16,488$ not declared; mean age of the population as of 2015 is 47.29 years ($sd = 23.79$). Data is randomly partitioned into two-fold train and test sets at 50% resulting in two mirror dataset splits, Fold-1 and Fold-2, which results in $804,370$ and $804,371$ subjects in test sets, respectively.

During the filtering process, we use Eq. (1.1) to discard terms that return a score less than 10^{-6}. That is, we discard terms with a document frequency that range from 1 to 10. Examples of least scoring terms (e.g., filter score $< 10^{-10}$, and document frequency equal to 1) for Fold-1 and Fold-2 training sets are *clotrimazole*, an antifungal medication, and *torsemide*, a medication to reduce extra fluid in the body, respectively. The upper bound of the term frequency list is composed of diagnostic codes which are frequent in document and with multiple occurrences. These are central to patient representations, and are retained. The *clinical stop words* are removed from both test sets.

After the filtering process, we remove duplicates in intervals of $T = 15$ days and shuffle the terms in both training and test sets. Subsequently, patients with less than 3 medical terms are removed and the longest sequences are cut at $5,000$ medical terms. Finally, sequences are split in subsequences of length $L = 32$. In total, $24,665$ medical terms are filtered out from Fold-1, decreasing its vocabulary size to $32,799$. On the other hand, $24,589$ terms are dropped from Fold-2, and the vocabulary size becomes $32,875$. Description of folds' characteristics are reported in Table 1.1.

Architecture implementation In total, we train the architecture on $\sim 3M$ subsequences that represent the co-occurrence of medical concepts. To acquire an initialization baseline for the embedding module of the ConvAE architecture, we use the Word2vec algorithm [22] to generate the initial 100-dimensional medical concept embeddings. The semantic word embeddings are estimated using the continuous Skip-gram model. We consider all collected subsequences as sentences and medical concepts as words. In total, we obtain $31,659$ medical concept embeddings in Fold-1 and $31,715$ in Fold-2, after applying the Word2vec embedding algorithm. The remaining concepts are randomly initialized, and subsequence padding is initialized at $\mathbf{0}$ (i.e. the null vector). These embedding vectors are then used as input to the ConvAE module, and are trained within the architecture.

The convolutional neural network applies 50 filters of length 5 with stride size 1 followed by a ReLU activation function. Next, we apply 1-dimensional Max Pooling, using the same convolution parameters, to obtain the final filtered responses. The autoencoder is a fully-connected neural network with 4 hidden layers with 200, 100, 200 and $|V| \times 32$ hidden nodes in each layer, where $|V|$ is the vocabulary size of the selected training set. We use the ReLU activation function in the first three layers and Softplus activation (i.e., $a(x) = \log(1 + \exp(x))$) in the output layer. Dropout layers with $p = 0.5$ are applied in the first two layers for regularization. The model is trained using Cross Entropy loss with the Adam optimizer (learning rate $= 10^{-5}$ and weight decay $= 10^{-5}$) [31]. The model is trained for 5 epochs on all training data with batch size equal to 128.

To determine the optimal architecture configuration, we run alternative 1-layer, 2-layer, and multikernel CNN module configurations and compare performance on test sets. In the 2-layer CNN, all parameters and structures remain unchanged, except for the number of filters set to 25 in the second layer. Similarly, the multikernel CNN performs parallel training of distinct filters with different dimensions, while all other aspects remain unchanged. Then, the CNN concatenates the outputs and feeds them to the next layer. In our study, we set parallel kernel dimensions to 3, 5, and 7.

1.2.6. Patient representation evaluation

To evaluate our model trained on the complete heterogeneous cohort of patients, we quantify the extent to which patient representations from test sets reflect their clinical phenotypes, and determine whether these representations can be leveraged for within-disorder stratification. First, we test model performance against baseline models for the task of subgrouping specific cohorts of patients. Next, we investigate if clinically relevant subtypes can be extracted from individual complex disorder cohorts. In particular, patient representations are first learned from a heterogeneous dataset, and then these representations are used to subtype different cohorts of patients via hierarchical clustering. This replicates a common approach in clinical research, where attempts to identify latent patterns within a cohort of patients can contribute to the development of improved personalized therapeutics [32].

To this aim, the clustering validation is based on two steps, and applies hierarchical clustering with Ward's method and Euclidean distance. First, we select a subset of patients with different complex disorders (i.e., neoplasm, metabolic disorder, neurodevelopmental disorder, neurological disorder, inflammatory disease) and run an external validation to investigate if patients with the same condition cluster together. Second, we perform an internal validation to identify the optimal number of subclusters that best disentangles heterogeneity on the independent disorder dataset.

External clustering validation External validation measures the extent to which cluster labels match externally supplied labels. After selecting specific diseases, we perform hierarchical clustering on the ConvAE representations of the heterogeneous cohort test sets with the number of clusters equal to the number of selected diseases. We then measure the extent that externally supplied labels match cluster labels using the cluster *Entropy* and *Purity* indicators.

In particular, for each cluster j, we compute p_{ij}, i.e., the probability that a member of cluster j belongs to class i. Hence, $p_{ij} = \frac{m_{ij}}{m_j}$, where m_j is the number of elements in cluster j and m_{ij} is the number of element in cluster j of class i. The Entropy for each cluster becomes: $E_j = -\sum_i p_{ij} \log_2 p_{ij}$. The conditional Entropy $H(\text{disease}|\text{cluster})$ is then computed as

$$H(\text{disease}|\text{cluster}) = \sum_j \frac{m_j}{m} E_j,$$

where m is the total number of elements in the complex disease dataset.

We also use a purity measure to determine the most represented disease class in each cluster. For a cluster j, the Purity P_j is defined as $P_j = \max_i p_{ij}$, where p_{ij} is computed as before. The overall purity of a clustering result is the weighted average of P_j:

$$P = \sum_j \frac{m_j}{m} P_j$$

The best clustering configuration has low entropy and high purity.

Evaluation Baselines We compare the results of the ConvAE architecture to a set of baseline algorithms on external clustering validation. Each approach outputs patient representations on test sets that are externally validated via hierarchical clustering for a subset of patients with complex disorders.

- **Raw Count**: a sparse patient representation, where each patient is encoded into a count vector that has the length of the vocabulary. More specifically, each individual health history s_p is represented as an integer vector $\mathbf{x} \in \mathbb{Z}^{|V|}$, where each element is the count of the corresponding clinical word in the patient sequence, i.e., $x_i = \#\{w_i;\ w_i \in s_p\}$.

- **Singular Value Decomposition (SVD) of raw counts**: SVD is a linear dimensionality reduction technique that reduces a $(m \times n)$ real matrix X into its singular values (i.e., square roots of non-negative eigenvalues), based on the decomposition $X = U\Sigma V^T$, were U and V are and $(m \times m)$ and $(n \times n)$ real or complex unitary matrices, respectively. In this study, we apply a truncated version of SVD transformation that only computes the k largest singular values of the raw count encodings. Assuming fixed $k = 100$, the training matrix with m samples and n features is decomposed into $X = U_k \Sigma_k V_k^T$, with $U_k \Sigma^T$ of dimension $(m \times k)$. The transformation of the test matrix X' with m' samples and n features provides the matrix $X'V_k$ of dimension $(m' \times k)$.

- **SVD of the Term Frequency - Inverse Document Frequency (TFIDF) matrix**: The

TFIDF weighting scheme combines term frequency and inverse document frequency scores to produce a composite weight for each term in a corpus D. Given a document d comprised of terms t, the normalized term frequency of a particular $\tilde{t}$ is defined as

$$\mathrm{tf}_{\tilde{t},d} = \frac{\#\{t \in d;\ t = \tilde{t}\}}{\#\{t \in d\}}$$

i.e., the number of occurrences of $\tilde{t}$ over the total number of terms in a document d. The inverse document frequency of the term $\tilde{t}$, on the other hand, is defined as

$$\mathrm{idf}_{\tilde{t}} = \log\left(\frac{N}{df_{\tilde{t}}}\right)$$

where N is the total number of documents in the corpus and document frequency $\mathrm{df}_{\tilde{t}} = \#\{d \in D;\ \tilde{t} \in d\}$. The TFIDF score for the term $\tilde{t}$ in document d is defined as:

$$\text{tf-idf}_{\tilde{t},d} = \mathrm{tf}_{\tilde{t},d} \cdot \mathrm{idf}_{\tilde{t}}$$

When truncated SVD is applied to term-document matrices, such matrices are transformed to a semantic space of low dimensionality. In this context, we apply TFIDF computation to patient sequences, which utilizes the entire dataset as a corpus and patient EHR sequences as documents. Hence, we downscale the weights of terms that occur too often, and attribute more relevance to terms that occur with less frequency. Each patient EHR is then represented as a vector of length $|V|$, storing the corresponding weight for each term. In this application, the TFIDF matrix is reduced via truncated SVD with feature dimension 100.

- **Deep Patient [27]**: each sequence of raw counts is input to a stack of denoising autoencoders to produce latent representations of patient histories within the EHR. The architecture is implemented as described in [27], with a deep feature dimension equal to 100, and trained with batch size 32 until model convergence (< 5 epochs).

Recurrent Neural Networks

EHR sequences carry an intrinsic sequential pattern that encourages the use of Recurrent Neural Networks, commonly utilized for time series data and text processing, as well as in a wide range of other applications [33]. In particular, Long Short-Term Memory Units (LSTM) has been applied to medical sequences to detect short- and long-range temporal dependencies for feature representations. These feature representations are particularly useful in storing local and global information that can be applied in both supervised and unsupervised settings [3]. LSTM core components are the cell state and the forget, input, and output gates. Each cell state at time t carries relevant information generated from input x_t and hidden state h_{t-1}, determined by the input and forget gates, along with the previous cell state c_{t-1}. The new cell state runs through the output gate, and combining the previous hidden state along with the current input activations, form the hidden state at time t.

A drawback of such an approach to EHR data is the assumption of uniformly distributed elapsed time between elements of a sequence. EHRs lack a rigorous longitudinal structure of term sequences, where subsequent visits can span from several days to several years. Such irregularity can affect the representations of temporal changes throughout disease trajectories. To overcome this limitation, a Time-aware LSTM (T-LSTM) architecture proposed by Baytas and colleagues [15] attempts to capture the temporal dynamics of sequential data with time irregularities. In particular, the T-LSTM architecture discounts the short-term contributions by the amount of time spanning between two subsequent records still retaining the global profile of patients. T-LSTM is trained on both a prediction and an unsupervised task, where representations are learned within an autoencoder structure. Experiments use both synthetic and real-world data (PPMI). Data structures are sequences of one hot encoded vectors each one representing an admission event. While T-LSTM approach addresses the temporal dimensions of disease progression representation, it still relies on highly engineered data (need for an *a priori* definition of admission event) and only relies on term co-occurrences within admission events. These shortcomings may result in representations with substantially lower resolution. In discovering latent relationships and hierarchical concepts from real-world data we do not significantly intervene on the temporal structure of the data. Hence, in this work, we do not include Recurrent Neural Networks models as a baseline because, with their need for structured admission events, they would not be directly comparable to our architecture.

Internal clustering validation To detect subclusters within each disorder, we apply hierarchical clustering to patient representations from separate disorder cohorts with the number of clusters ranging from 2 to 15. The best number of clusters is selected via the Elbow Method.

Frequent term inspection To investigate which clinical features best describe different subclusters, we rank diagnosis codes, laboratory tests, medications, CPT codes, respect to the counts of terms occurring in the concept histories of patients in the same cluster. We compute percentages of patients whose sequence includes a specific term both with respect to the counts of patients within a subcluster and among patients in the complete disease dataset. Ranking first maximizes the relative percentage within the subcluster, and second, the percentage over the whole dataset. Most frequent concepts are then analyzed and we use a pairwise chi-squared test to determine whether the distributions of present/absent counts with respect to the detected subgroups are significantly different.

1.3. Results and discussion

To visualize the distribution of the encoded representations of patients and the labels of interest (i.e., disease classes, external clustering labels, subgroup labels) we apply Uniform Manifold Approximation and Projection for dimensionality reduction (UMAP; [34]). UMAP *number of neighbors* and *minimum distance* parameters are set at 20 and 0.0, respectively, for external validation, and at 10 and 0.0, respectively, for internal validation. Recently, UMAP has been proposed as operative space for Deep Clustering [35].

Complex disorder	Full dataset	Test Fold-1	Test Fold-2
Type 2 diabetes	$100,580$	$50,253$	$50,327$
Parkinson's disease	$6,274$	$3,124$	$3,150$
Alzheimer's disease	$6,685$	$3,374$	$3,311$
Multiple myeloma	$3,882$	$1,947$	$1,935$
Prostate cancer	$28,909$	$14,401$	$14,508$
Breast cancer	$16,486$	$8,330$	$8,156$
Crohn's disease	$13,409$	$6,668$	$6,741$
ADHD	$12,822$	$6,510$	$6,312$

ADHD = Attention Deficit Hyperactivity Disorder

Table 1.2: Disease cohort counts.

	Entropy*	Purity*	N diseases†
ConvAE 1-layer CNN	**2.61** (0.04, $[2.58, 2.67]$)	**0.31** (0.02, $[0.31, 0.35]$)	**6.50** (0.62)
ConvAE 2-layer CNN	2.75 (0.02, $[2.74, 2.78]$)	0.26 (0.01, $[0.26, 0.29]$)	5.93 (0.50)
ConvAE multikernel CNN	2.66 (0.03, $[2.64, 2.70]$)	0.30 (0.02, $[0.29, 0.33]$)	5.94 (0.47)
Raw count	2.90 (0.02, $[2.88, 2.92]$)	0.18 (0.01, $[0.18, 0.20]$)	4.76 (0.70)
SVD raw count	2.90 (0.01, $[2.90, 2.92]$)	0.19 (0.01, $[0.18, 0.20]$)	5.13 (0.79)
SVD TFIDF	2.85 (0.02, $[2.84, 2.87]$)	0.21 (0.01, $[0.21, 0.23]$)	5.83 (0.76)
DP	2.81 (0.02, $[2.80, 2.84]$)	0.24 (0.01, $[0.23, 0.25]$)	5.96 (0.74)

* Mean (sd, CI); † Mean (standard deviation)

CNN = Convolutional Neural Network; SVD = Singular Value Decomposition; TFIDF = Term Frequency - Inverse Document Frequency; DP = Deep Patient

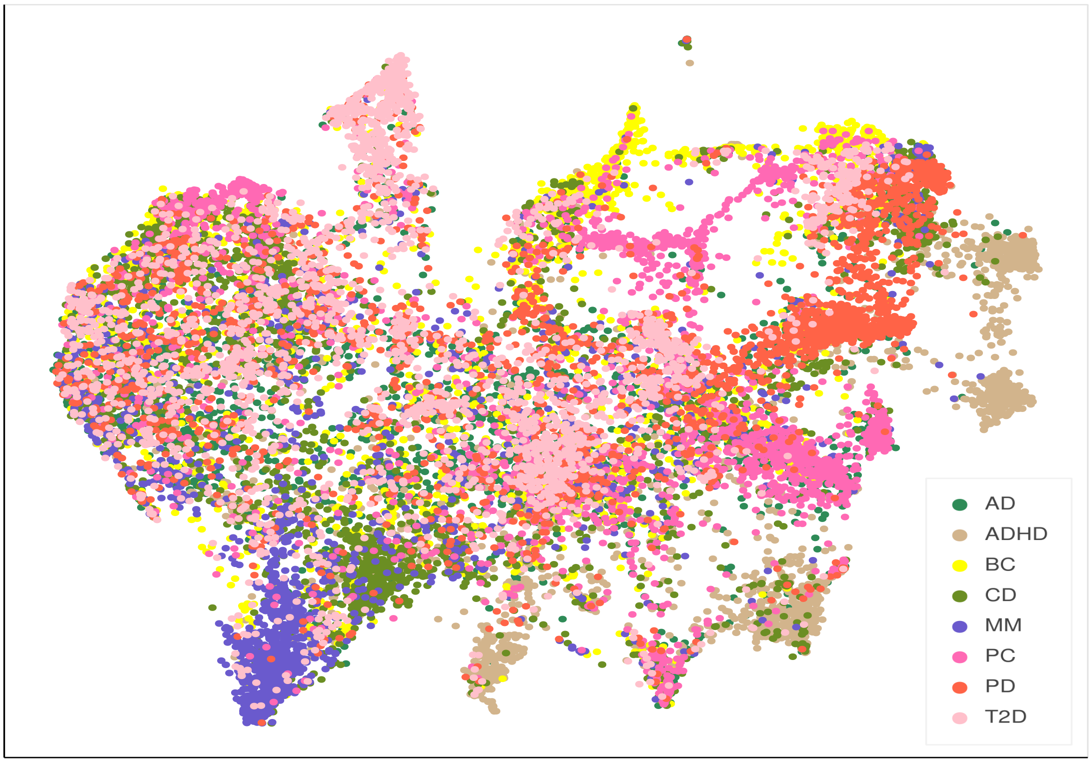
Table 1.3: External validation performances for ConvAE models and baselines.

1.3.1. External validation

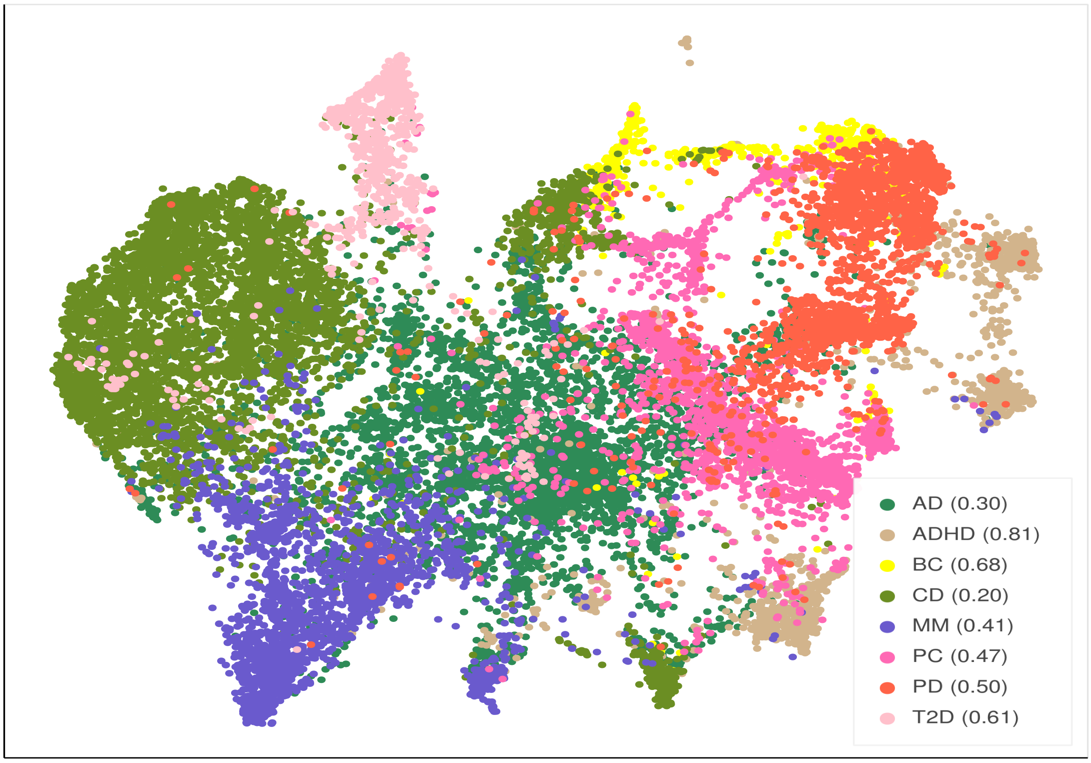
We select 8 complex disorders to validate normalized patient encodings. In particular, we extract patients with T2D, MM, PD, Alzheimer's disease (AD), Crohn's disease (CD), Malignant neoplasm of female breast (BC), Malignant neoplasm of prostate (PC), and Attention Deficit Hyperactivity Disorder (ADHD). For patient selection, we rely on SNOMED Clinical Term identifiers and we control for a corresponding ICD-9 code in patient EHRs. We exclude patients with comorbidities. Disease numerosities for the complete dataset and the two-fold test sets are reported in Table 1.2.

In an ideal validation setting, all patients with the same complex disorder are clustered together yielding an entropy score of 0, and a purity score equal to 1 (i.e., probability that members clustered together belong to the same cohort).

As can be observed in Table 1.2, disease cohorts in test folds are unbalanced, varying from $50K$ to $1.9K$, so we randomly sample $5,000$ patients from each disease cohort in the full dataset and obtain subsampled test sets from Fold-1 and Fold-2 of roughly $2,500$ subjects per disease cohort, respectively. We iterate the subsampling process 100 times and average out the performance results. We report



(a)



(b)

Figure 1.2: Uniform Manifold Approximation and Projection (UMAP) visualization of Fold-2 encodings. (a) complex disorder classes; (b) hierarchical clustering labels.

average Entropy and Purity scores, and the mean number of correctly detected diseases from the purity score for the three ConvAE architectures (i.e., 1-layer CNN, 2-layer CNN, multikernel CNN) and baselines in Table 1.3.

We observe that all three ConvAE model configurations perform better than the baselines in entropy and purity scores, whereas DP encodings perform better than 2-layer and multikernel CNN models in the average number of detected diseases. The 1-layer-CNN configuration has the best overall performances in all three scores, closely followed by the architecture with multikernel CNN. UMAP representation of $19,409$ encodings of patients from Fold-2 test set with their disease cohort labels is shown in Figure 1.2 (a). The clustering labels and the purity percentage scores of each cluster dominating disease are listed in the legend of Figure 1.2 (b).

The most visible and well separated points are from patients with ADHD detected with 80% purity by hierarchical clustering. Two main clusters of patients with PC, and one cluster of patients with PD are also visible. Most individuals with BC are scattered throughout the plot, except for a small group in the top area of the figure that is detected by hierarchical clustering with 68% BC purity. Close to a large assembly of individuals with MM, correctly identified by a cluster, a small sized subgroup of subjects with CD is not detected by the clustering algorithm and it is included in the MM cluster. This can be caused by shared phenotypic profiles (e.g., late-onset groups with similar comorbidities) that lead to latent representations that end up close to each other. Furthermore, although patients with T2D are mainly scattered across the center of the point cloud, a cluster with 61% purity is identified. The encodings of the AD cohort are of particular interest. There are no visible clusters among these patients, and corresponding clusters includes patients with other conditions. AD, CD, and T2D classes are highly entangled, and the lack of patterning could be due to confounders, such as sequences lengths, sex, age, or noise. Such entanglement may also reflect a shared phenotypic characterization that drives the learning process into displaying similar EHR progressions closely together. An interactive plot, implemented with Bokeh visualization library, that allows the visualization of individual diseases is available in the Supplementary Material folder.

1.3.2. Internal clinical validation

The latent semantic spaces for the two-fold datasets are different by construction, and this influences patient representations to an extent. We run internal validation on Fold-1 and Fold-2 disease cohorts, considering separately their UMAP plots.

Next, we focus on T2D, PD, AD, MM, PC, and BC for internal validation. For each cohort, we select patients with the corresponding ICD-9 code, in particular, *Diabetes mellitus (250.00)* for T2D; *Paralysis agitans (332.0)* for PD; *Alzheimer's disease (331.0)* for AD; *Multiple Myeloma without mention of having achieved remission (203.00)* for MM; *Malignant neoplasm of prostate (185)* for PC; and *Malignant neoplasm of female breast (174.9)* for BC. To reduce noise in the sequence encodings, patient ConvAE subsequence representations are considered from the first instance of diagnosis forward, and are averaged; sequences shorter than 3 terms are dropped.

	Fold-1		**Fold-2**	
	Numerosity	**N clusters**	**Numerosity**	**N clusters**
T2D	48,688	3	48,759	3
PD	3,052	2	3,071	2
AD	3,201	3	3,150	3
MM	1,884	5	1,883	4
PC	8,522	2	8,645	3
BC	7,964	2	7,838	2

T2D = Type 2 Diabetes; PD = Parkinson's Disease; AD = Alzheimer's Disease; MM = Multiple Myeloma; PC = Prostate Cancer; BC = Breast Cancer

Table 1.4: Disease cohorts and number of subclusters.

First, we find the best number of subclusters in each cohort via the Elbow Method applied to hierarchical clustering, then we perform medical term enrichment analysis and clinical phenotyping. In order to detect whether or not sex count distributions differ by group, we use multiple pairwise chi-squared tests. The existence of differences in patient age and sequence length in the EHR is determined using multiple pairwise two-tailed t-tests. To assess the significantly enriched terms in subgroups we run multiple pairwise chi-squared tests. Significance level for the Bonferroni-corrected p-values is set at 0.05. We expect cohorts from both folds to be similarly enriched and, when possible, we aim to compare the subgroups that we find to subtypes from the literature.

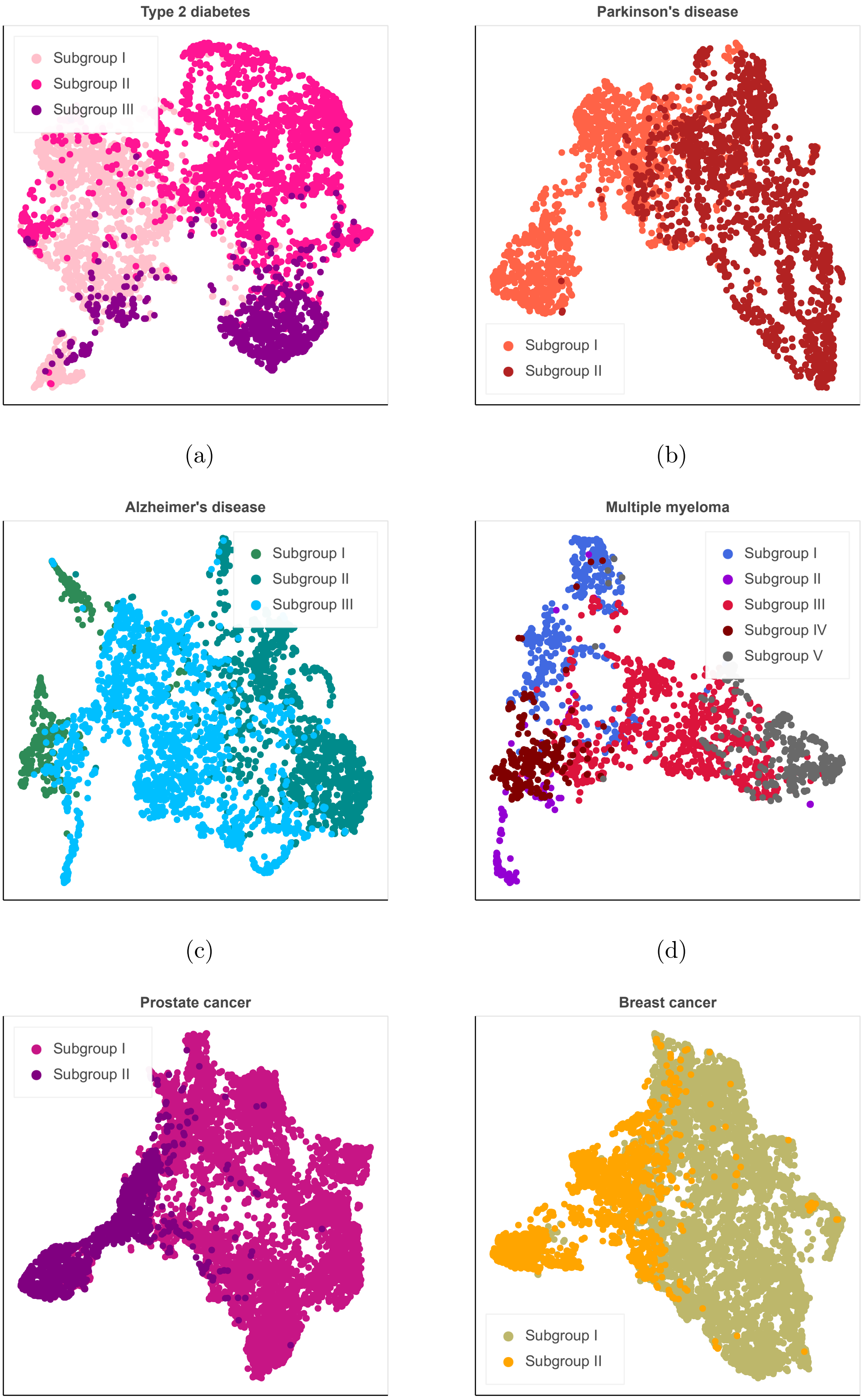
In Table 1.4, the disease cohorts and the number of subgroups found are reported for both folds. For each disease cohort, the top five most frequent terms from ICD-9, medications, laboratory tests, and CPTs are listed in Tables 1.5 - 1.10 from Fold-1 and Tables 1.11 - 1.16 from Fold-2. We do not report procedure terms because of their redundancy and overlapping with CPT codes. As an example, *Electrocardiogram complete, Electrocardiogram tracing* are often found among procedures, and they correspond to the CPT code 93000, *Electrocardiogram, routine ECG with at least 12 leads; with interpretation and report*, which indicates that the clinical interpretation of the Electrocardiogram (ECG) tracing always follows an ECG procedure. In the following, clinical subgroups for both folds are described. Results from the two folds are mostly similar.

In the following, we report the clinical validation of disorder subgroups identified by hierarchical clustering for six different complex disorders including neoplasms, metabolic disorder, and neurological disorders to test the validity of our representations.

Fold-1 disease cohorts

Type 2 Diabetes T2D is characterized by hyperglycemia, insulin resistance, and relative impairment in insulin secretion. An exact pathogenesis of T2D is still unknown because each clinical feature can arise through genetic and/or environmental influences. The global prevalence of T2D among adults over 18 years of age has risen to 8.5% in 2014[1]. T2D is often accompanied by other conditions

[1] https://www.who.int/news-room/fact-sheets/detail/diabetes (Accessed on September 17, 2019)



Type 2 diabetes

(a)

Parkinson's disease

(b)

Alzheimer's disease

(c)

Multiple myeloma

(d)

Prostate cancer

(e)

Breast cancer

(f)

Figure 1.3: Fold-1 complex disorder subgroups. A subsample of 5, 000 patients with T2D is displayed in Figure (a).

that also increase cardiovascular risk. These clinical conditions are referred to as metabolic syndrome [36].

Patients with T2D are clustered into three different subgroups (see Figure 1.3 a) by our modeling architecture. Subgroup I consists of 18,325 patients, subgroup II of 22,659, and subgroup III of 7,704 patients. The most relevant terms among ICD codes, medications, laboratory tests, and CPTs for each subgroup are reported in Table 1.5. Mean age is > 60 years and it increases progressively across the three subgroups, with subgroup I consisting of the youngest cohort of patients ($M = 64.77$, $sd = 14.32$), and subgroup III consisting of the oldest cohort ($M = 69.42$ $sd = 11.29$). Sex counts significantly differ between groups, with more females than males in subgroup I and more males than females in subgroups II and III. Mean sequence lengths are significantly different between groups, although they are similarly distributed. Older patients (subgroup III) have substantially shorter sequences probably due to a greater probability of death and subsequently dropping out of the EHR system.

In the stratification process of patients with T2D, we are able to identify a subgroup of patients characterized by cardiovascular disorders (subgroup III). Patients in this group often have patient records with terms for coronary artery diseases, i.e., *Coronary atherosclerosis (414.00/01), Angina pectoris (413.9)*, which are serious risk factors for heart failure. These subjects are also treated with antiplatelet therapy (i.e., *Acetylsalicylic acid, Clopidrogel*) to prevent cardiovascular events (e.g., stroke) and are likely to receive invasive procedures to treat severe arteriopathy. For instance, among the 10 most frequent terms, 30% of patients in subgroup III undergo *Percutaneous Transluminal Coronary Angioplasty* ($p < 0.001$), a procedure to open up blocked coronary arteries. Subgroup II appears to cluster patients that are more inclined towards or have developed diabetic nephropaty and peripheral artery disease (PAD), defined as atherosclerotic occlusive disease of lower extremities. In particular, *Creatinine* and *Urea nitrogen* laboratory tests, which measure creatinine and urea metabolic byproduct levels to estimate renal function, suggest monitoring of kidney disease. Diabetic nephropathy may lead to diabetic kidney disease, a complex and phenotypically heterogeneous disease with overlapping etiological pathways [37]. Moreover, signs of metabolic syndrome, i.e., *Hypertension (401.9), Hyperlipidemia (272.4)*, combined with analgesic drugs (i.e., *Paracetamol, Oxycodone*), may indicate the presence of vascular lesions at the peripheral level, manifested as ischemic rest pain or ulceration. Previous estimates have shown that individuals with T2D older than 50 years have a 29% prevalence of PAD, with age, disease duration, and peripheral neuropathy that are associated risk factors [38]. Subgroup I includes patients with signs of metabolic syndrome and are treated with *Metformin*, an oral hypoglycemic medication.

Based on clinical profiles, we seem to identify subgroups tied to different disease progressions. Subgroup I is the initial mild symptom severity group, characterized by common T2D symptoms (i.e., metabolic syndrome). Subgroups II/III show concomitant conditions associated to T2D progression and worsening symptoms. Specifically, subgroup II is characterized by microvascular problems (i.e., diabetic nephropathy, neuropathy) and/or PAD, whereas subgroup III shows severe cardiovascular problems (i.e., coronary atherosclerosis, angina pectoris).

Our results confirm, in part, what observed by Li et al. [16], which report three distinct subgroups of patients characterized by 1) microvascular diabetic complications (i.e., diabetic nephropathy, diabetic retinopathy); 2) cancer of bronchus and lungs; and 3) cardiovascular diseases and mental illness. In particular, we detect the same microvascular and cardiovascular disease groups, which are consequences of T2D. In contrast, we are unable to detect Li's group 2 characterized by cancer, an epiphenomenon that can be caused by secondary immunodeficiency in patients with T2D [39, 40]. Li et al. utilize the same cohort of patient histories as in this study, and of the $2,472$ patients from their study, we identified $1,050$ patients in Fold-1 data. We compare the similarity of the clusters we have obtained to those found by Li and colleagues via the Fowlkes-Mallows index (FMI), which is an external validation similarity measure of two cluster analyses [41, 42].

FMI is theorized for hierarchical clustering [41] and it assumes two hierachical clusterings of the same number of objects n, named A_1 and A_2, with k clusters. Clusters are then labeled from 1 to k and a matrix $M = (m_{ij}),\ i, j = 1, ..., k$ is created, with m_{ij} the number of objects in common between the i-th cluster of A_1 and the j-th cluster of A_2. The FMI then becomes:

$$FMI = \frac{T_k}{\sqrt{P_k Q_k}}$$

where

$$T_k = \sum_{i=1}^{k}\sum_{j=1}^{k} m_{ij}^2 - n$$
$$P_k = \sum_{i=1}^{k} m_{i\cdot}^2 - n$$
$$Q_k = \sum_{j=1}^{k} m_{\cdot j}^2 - n$$

with $m_{i\cdot} = \sum_j m_{ij}$, and $m_{\cdot j} = \sum_i m_{ij}$.

FMI scores range from 0 to 1, where 1 represents identical clustering and 0 for purely independent label assignments. More in general, it has been shown [42] that FMI can be defined as an information-based index:

$$FMI = \sqrt{\frac{TP}{TP + FP} \cdot \frac{TP}{TP + FN}}$$

where TP is the number of true positives, i.e. the number of pairs of points that are assigned to the same cluster in both A_1 and A_2, FP is the number of false positives, i.e., the number of pairs of points that are in the same cluster for A_1, but not for A_2, and FN (false negatives) the opposite. We obtain FMI $= 0.40$, which suggests that only a portion of patients in groups from [43] are identified by our clustering as sharing the same characteristics. This may entails that associated clinical characteristics overlap to a greater extent than hypothesized by Li and colleagues, which may have been overlooked because they collected shorter EHR sequences (i.e., 60 day intervals).

Parkinson's Disease PD is a progressive neurodegenerative disease now recognized to be a complex condition with diverse clinical features that include neuropsychiatric, motor, and nonmotor manifestations [8]. PD prevalence increases with age and is estimated at 1% in people over 60 years old [44]. Pathophysiology of PD ascribes parkinsonian signs, such as bradykinesia, to dopamine depletion from the basal ganglia, which results in major disruptions in the connections to the thalamus and motor cortex.

From a PD cohort of $3,052$ patients, we identify two subgroups of $1,368$ and $1,684$ patients, respectively, as shown in Figure 1.3 (b). Each subgroup's most frequent clinical terms and statistics are displayed in Table 1.6. We observe that there is no difference in the frequency distribution of female/male counts between the two subgroups, whereas mean sequence length and mean age significantly differ between the subgroups. In particular, PD-I has shorter sequences and lower mean age in years ($M = 70.76,\ sd = 13.52$) than PD-II.

One group seems to be characterized by nonmotor/independent features and longer course of the disease, whereas the other group is dominated by motor symptoms. In particular, subgroup I presents as a tremor-dominant subgroup for the presence of *Essential tremor (333.1)* code, with *Anxiety state (300.00)*. Subgroup II is described by nonmotor and independent symptoms, i.e., *Constipation (564.00)* and *Fatigue (780.79)*, respectively. In subgroup I, motor clinical features have led, with all probabilities, to a misdiagnosis of essential tremor, which is an action tremor that typically involves the hands. Parkinsonian tremor, on the contrary, although can be present during postural maneuvers and action, is much more severe at rest and decreases with purposeful activities. However, when the tremor is severe, it may be difficult to distinguish action tremor from resting tremor. Moreover, some patients with PD may have a re-emergent tremor, which is a postural tremor that manifests after a latency of few seconds. These individual characteristics may lead to the misdiagnosis of essential tremor [45]. Moreover, anxiety states, emotional excitement, and stressful situations can exacerbate the tremor, and the presence of anxiety comorbidity may have delayed PD diagnosis. Brain MRI, usually nondiagnostic in PD, is prescribed for few patients from subgroup I (13%) suggesting its use for differential diagnosis, i.e., to investigate the presence of chronic/vascular encephalopathy.

Considering subgroup II characterization, constipation is one of the most common nonmotor problems related to autonomic dysfunction and slowed intestinal transit time in PD. Fatigue is a common problem in patients with PD, with prevalence rates ranging from $33 - 58\%$, that manifests as diminished activity level, and increased perception of effort, disproportionate to activities [46]. Fatigue is considered a feature that does not depend on disease progression [47].

The PD stratification study on PPMI data by Zhang et al. [26] has identified three distinct subgroups of patients with PD based on severity of both motor and nonmotor symptoms. In particular, one subgroup includes patients with moderate functional decay in motor ability and stable cognitive ability. Another subgroup presents mild functional decay in both motor and nonmotor symptoms, and the third subgroup is characterized by rapid progression of both motor and nonmotor symptoms. The representations of patient from EHRs for stratification purposes exclude the analysis of PD symptom severity, in that EHRs lack quantitative measures of these. However, differently from Zhang and colleagues, our approach is able to discriminate between specific motor and nonmotor symptoms, and also suggests a longer, but not necessarily more severe, disease course for subgroup II.

Alzheimer's Disease AD is a progressive neurodegenerative disorder characteristic of older age (> 65 years) and its causes and pathogenesis remain unknown. AD is the most common cause of dementia, a decline in cognition involving one or more cognitive domains (e.g., learning and memory, language, executive functioning) [48]. Globally, an estimated 43.8 million people are affected by dementia [49] and there is very little difference between men and women in incidence and prevalence of dementia or AD. Due to life expectancy, the raw numbers indicate that more women than men present with AD, particularly over the age of 85 years, due to differences in life expectancy by sex. The hallmark neuropathologic changes in AD are diffuse and neuritic plaques, marked by extracellular amyloid beta deposition, and neurofibrillary tangles, comprised of the intracellular accumulation of hyperphosphorylated tau protein.

We identify three subgroups within patients with AD (see Figure 1.3 c). In particular, AD-I ($N = 399$) is characterized by younger age ($M = 54.45$ $sd = 22.48$) and shorter sequences ($M = 32.24$, $min/max = [3; 448]$), subgroup II includes 1,170 patients that are significantly older than subgroup III patients ($N = 1,632$), and sequence lengths are not significantly different between subgroups II and III. Moreover, significantly more females than males are affected by AD in all subgroups. For term frequencies and statistics, refer to Table 1.7.

Subgroup I appears to be characterized by patients with early-onset AD, i.e., patients whose dementia symptoms have typically developed between the age of 30 and 60 years, and initial neurocognitive disorder. Subgroup II includes patients with late-onset AD, mild neuropsychiatric symptoms and cerebrovascular disease. Subgroup III is characterized by individuals with typical onset and mild-to-moderate dementia symptoms. Early-onset AD affects 5% of the individuals with AD in the US[2] and, because clinicians do not usually look for AD in younger patients, the diagnostic process includes extensive evaluations of patient symptoms. In particular, given that a certain AD diagnosis can only be provided post-mortem through brain plaques examination and it is unlikely at a young age, clinicians first rule out other causes that can lead to early-onset dementia (i.e., differential diagnosis). We find evidence of this practice in subgroup I, which includes postmenopausal women, identifiable by mean age greater than 50, *Osteoporosis (733.00)* diagnosis with calcium supplement therapy, and menopausal hormone treatment (i.e., *Estradiol*). Patients in this group are tested for infectious diseases (i.e., HIV, Syphilis, Hepatitis C, Chlamydia/Gonorrhoea), that are possible causes of early-onset dementia [50], and screened via structural neuroimaging, e.g., *MRI/PET brain* procedures ($p < 0.01$ I vs II). As cognitive dysfunctions that may be mistaken for dementia can also be caused by depression and other psychiatric conditions, the presence of CPT term *Psychiatric service/procedure* suggests psychiatric evaluations to exclude depressive pseudodementia. After the differential diagnosis process and the exclusion of other possible causes, eventually these patients have received a diagnosis of AD.

Subgroup II is identified by *Dementia without behavioral disturbance (294.10)* ICD-9 code. Neu-

[2] https://www.alz.org/alzheimers-dementia/what-is-alzheimers/younger-early-onset (Accessed on October 14, 2019)

ropsychiatric symptoms are common in AD and other types of dementia and include behavioral disturbances such as agitation, aggression, and delusions. One or more behavioral symptoms are observed in $60-90\%$ of patients with dementia, and the prevalence increases with the onset of cognitive impairment [51]. The absence of behavioral disturbances in 39% of patients, and their high average age ($M = 84.96,\ sd = 9.61$), suggest a late AD onset, with a progression characterized by a slower rate of cognitive ability decline. Moreover, the presence of *Acetylsalicylic acid* antiplatelet medication possibly indicates the co-occurrence of cerebrovascular disease, which affects blood vessels and blood supply to the brain. Cerebrovascular diseases are common in aging, and can be associated with AD [52]. Among CPTs and procedures, *Head CT* ($p < 0.001$) may have been performed to prevent or identify structural abnormalities related to cerebrovascular disease.

Subgroup III includes 408 patients treated with *Donepezil*, a cholinesterase inhibitor, that is a primary treatment for the cognitive symptoms of AD and it is usually administered to patients with mild-to-moderate AD producing small improvement in cognition, neuropsychiatric symptoms, and activities of daily living [53].

Overall, we are able to distinguish patients with AD according to AD onset, disease progression, and severity of cognitive impairment. Subgroups I and II, where Alzheimer's associated dementia can be mistaken for other forms of dementia such as infectious disease and vascular dementia, suggest the need for extensive screening for comorbid diseases as a diagnostic tool for patients showing signs of idiopathic dementia.

Multiple Myeloma MM is characterized by the clonal proliferation of plasma cells in the bone marrow, producing a monoclonal immunoglobulin. Signs of MM can be: 1) bone pain with lytic lesions; 2) increased total serum protein concentration and/or presence of a monoclonal protein in the urine or serum; 3) anemia; 4) hypercalcemia; 5) acute kidney injury with or without proteinuria (i.e., light chain deposition disease, amyloidosis, myeloma kidney). Worldwide in 2016, there were $138,509$ cases of MM with an age-standardized incidence rate of 2.1 per $100,000$ persons [54]. MM is a disease of older adults, with median age at diagnosis equal to 66 years [55].

Considering Figure 1.3 (d), we obtain 5 subgroups of patients with MM with mean age > 63 years, and no differences across types in terms of count by sex, except for between subgroups IV and V (see Table 1.8). Mean sequence lengths are significantly different across subgroups, except for those in subgroups I and V, with subgroup III reporting the greatest mean sequence length ($M = 1204.13$, $min/max = [15; 4,849]$). Subgroup I includes older patients than all other groups ($M = 69.97, sd = 14.00$), whereas mean age for the other subgroups are not significantly different from one another.

Inspecting the most frequent ICD-9 terms for subgroup I (see Table 1.8), we observe that patients are characterized by pulmonary manifestations. Subgroup II shows bone-related signs of MM, and subgroup III shows signs of gastrointestinal problems. Among MM signs, renal function is estimated for patients in subgroup IV, whereas subgroup V shows signs of peripheral neuropathy.

Pulmunary manifestations in subgroup I include *Pleura effusion (`511.9`)*, which is a rare pulmonary manifestation of amyloidosis [56] and MM coexists with primary amyloidosis in $10-15\%$ of cases (i.e., superimposed amyloidosis). Moreover, when amyloidosis is accompanied by proteinuria (i.e., excess proteins in urine), this determines protein concentration in blood to decrease causing fluids to accumulate in the pleural cavity. Subgroup I can thus include patients with amyloidosis and proteinuria. More rarely, amyloidosis can develop into MM in older patients [57].

The presence of *Disorders of bone and cartilage (733.99)* ICD-9 code characterizes group II and, moreover, *Disease of salivary glands (527.9)* can be a sign of superimposed amyloidosis in patients with MM [58]. Gastrointestinal problems in subgroup III are associated with the ICD-9 code *Diarrhea (787.91)*, whereas subgroup V shows *Inflammatory/toxic neuropathy (357.9/89)* diagnosis. We expect these diagnoses to indicate side effects from anti-cancer medications. In particular, we find *Bortezomib* in combination to *Dexamethasone* in both subgroups III and V. Bortezomib is administered to 47% of patients from subgroup III, although not shown in Table 1.8 because ranked at 7th position ($p < 0.001$), and to 26% of patients in group V. Bortezomib is used 1) for patients ineligible for hematopoietic cell transplantation (HCT); 2) as a maintenance therapy; or 3) in conjunction with HCT for newly-diagnosed patients [59]. Peripheral nerve damage is one of the most significant non-hematologic toxicities of bortezomib [60]. Clinical manifestations are gastrointestinal side effects (i.e., diarrhea, constipation) [61] that are thought to be caused by autonomic peripheral neuropathy, and motor neuropathy, consisting of mild to severe distal weakness in the lower extremities [62]. Although unlikely, given the close distribution of clusters III and V in Figure 1.3 (d), and the clear presence of side effects from anti-cancer drugs in subgroup III, neurologic complications can also be caused by MM itself. Such neurologic complications can be due to spinal cord compression from an extramedullary plasmacytoma, or by peripheral neuropathy, which is rare and usually caused by superimposed amyloidosis [63]. The ICD-9 term *Counseling (V65.40)* in subgroup V likely denotes an encounter to treat severe pain linked to neurologic diseases or psychological support.

Creatinine laboratory testing indicates renal function estimate for patients in subgroup IV. Moreover, 30% of patients undergo surgery, and are likely eligible for HCT. Depending on eligibility, age, and disease course, HCT can be autologous or allogenic, and nonmyeloablative or myeloablative (i.e., decreased bone marrow activity). Further analysis on the kind of HCT procedure performed would provide information related to the probability of relapse, mortality risk, and response to treatment for this group.

In conclusion, we have identified 5 subgroups of patients showing different typical signs of MM. In some cases, these signs are further characterized by rare pulmonary diseases in older patients, and possibly superimposed amyloidosis, therapy-related side effects, or non-specific HCT procedures.

Malignant neoplasm of prostate Prostate cancer is the most common cancer in men worldwide, with an estimated $1,600,000$ cases annually [64]. Clinical manifestations of prostate cancer are usually absent at the time of diagnosis, and over 90% of patients present with local or locoregional disease due to the widespread use of PC screening (i.e., use of prostate-specific antigen – PSA, or digital rectal examination). In the remainder of patients, the diagnosis is typically based on symptoms of advanced cancer (e.g., urinary tract obstruction, hematuria, bone pain). Specifically, at the time of diagnosis, it is estimated that 77% of patients have localized cancer, 13% show regional lymph node involvement, and 6% present metastasis. PC is most frequently diagnosed among men aged $65-74$, and median age at diagnosis is 66 years[3]. Clinical behavior of PC is heterogeneous and may range from a screen detected asymptomatic, microscopic, well differentiated tumor, that may never become clinically significant, to the rare screen detected or clinically symptomatic aggressive cancer, that causes metastases, morbidity, and death. Treatment approaches to PC include: active surveillance, radical prostatectomy, or radiation therapy (RT) for patients with low-risk PC; prostatectomy or RT in combination with Androgen Deprivation Therapy (ADT) for patients with higher-risk, but localized PC; RT and ADT for patients with clinical evidence of lymph node involvement.

We find 2 subgroups of patients with PC, that are displayed in Figure 1.3 (e). Mean age between the two groups is not significantly different, whereas subgroup I includes patients with significantly longer sequences ($M = 219.18$, $min/max = [9; 4,817]$). Patients in subgroup I (20%) report *Personal history of PC (V10.46)* and *Ondansetron* medication, to prevent RT side effect (i.e., nausea). The latter suggests that there exists patients with recurrent prostate cancer that have either received prostatectomy in the past, and hence RT and ADT is required, or, have already received RT and thus require a radical approach. Patients in subgroup II show signs of effective PSA screening, indicating probable localized and asymptomatic PC. The diagnoses of *Nocturia (788.43)*, *Impotence of organic origin (607.84)*, *Urinary frequency (788.41)*, and treatments for male sexual dysfunctions, i.e., *Tadalafil, Sildenafil*, can all be signs of side effects from PC treatments [65]. Among them, at least 22% likely received a prostatectomy (*Surgery* CPT term). Dissimilarly to the second subgroup, where PSA screening is highly frequent (see Table 1.9), patients in the first subgroup do not have PSA terms among top-ranked terms. This suggests that subgroup I includes patients that already received prostatectomy, and thus PSA screening is less frequent. Patients in subgroup I appear to have been in the healthcare system for longer (i.e., similar age, but longer sequences than subgroup II), and also to have been diagnosed with PC earlier (i.e., at a younger age) relative to patients clustered in subgroup II.

According to our analysis, we are able to identify two same-age groups of patients along with diverging disease courses. The first group includes patient with recurrent PC, whereas the second includes patients showing prostatectomy treatment side effects, but for whom the screening policy via PSA proved to be effective.

Malignant neoplasm of breast (female) Malignant neoplasm of breast is the most frequent cancer in females worldwide, with more than $2,000,000$ new cases registered in 2018[4]. Female brest cancer is most fequently diagnosed among women aged $55-64$ and median age at diagnosis is 62 years[5]. BC shows both pathologic and biologic heterogeneity, with various types of breast carcinoma that differ in appearance and behavior [43]. Different molecular subtypes of BC have been identified based on gene expression profiling [66].

[3] https://seer.cancer.gov/statfacts/html/prost.html (Accessed on September 17, 2019)
[4] http://gco.iarc.fr/ (Accessed on September 17, 2019)
[5] https://seer.cancer.gov/statfacts/html/breast.html (Accessed on September 17, 2019)

In our cohort, we are able to identify two subgroups (see Figure 1.3 f) that significantly differ in average age and sequence length. In particular, subgroup I includes older patients ($M = 66.67$, $sd = 14.27$) and longer sequences ($M = 323.43$, $min/max = [6; 4,849]$) than subgroup II. In inspecting clinical terms from subgroup I (see Table 1.10), we observe that subjects in subgroup I do not display a clear disease pattern. Patients in subgroup II, on the other hand, are younger and present a high number of screening-related medical terms (e.g., *Mammography screening*). Moreover, the presence of *Abnormal mammogram (793.80)* and *Carcinoma in situ of breast (233.0)* suggests an early-stage diagnosis. However, the prescription of *Ondansetron* medication suggests the presence of patients with more severe BC (e.g., metastatic) that receive chemotherapy treatment.

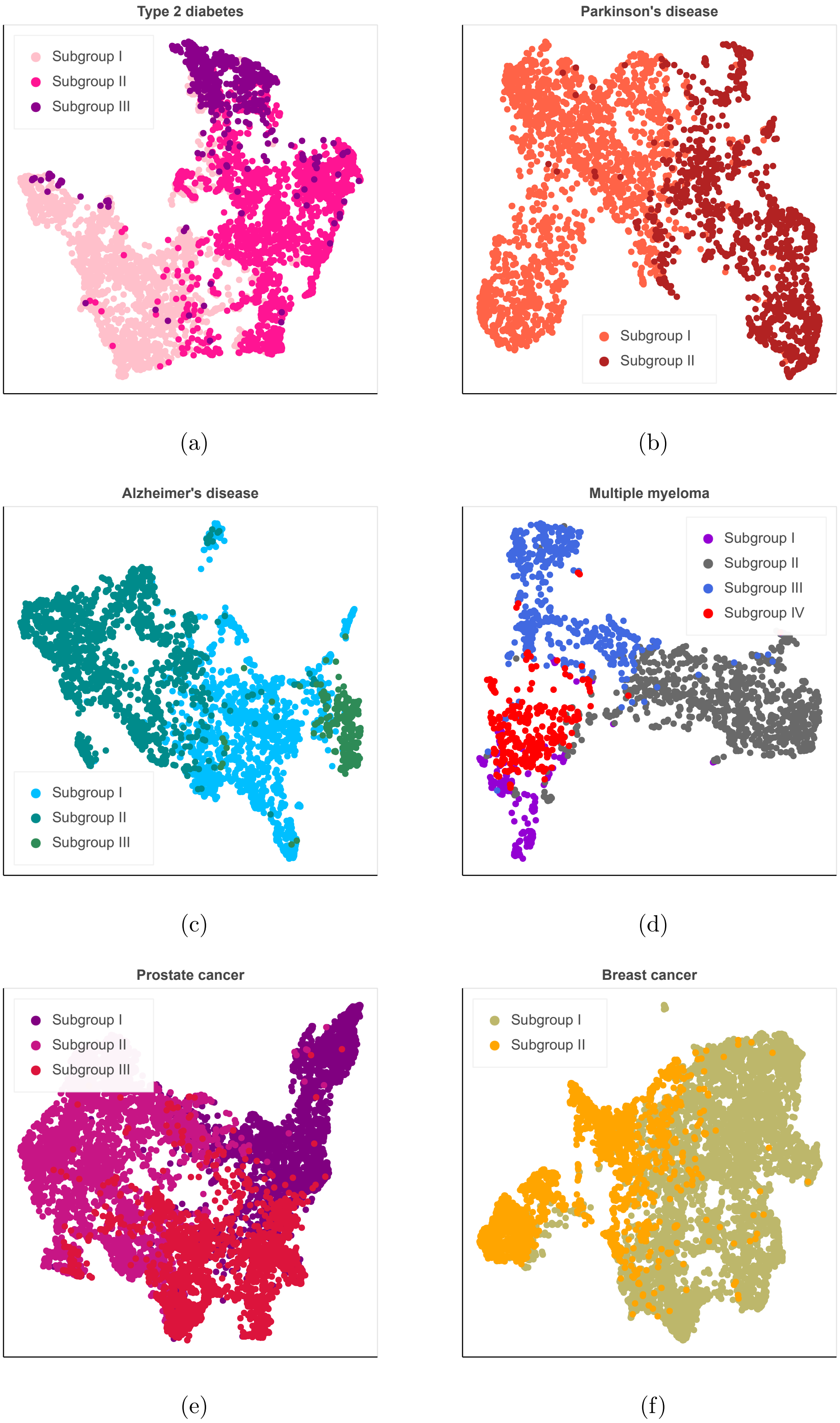
To describe subgroup I, we have investigated less frequent terms besides the first five and find out that 30% of them display the *Unlisted chemotherapy* CPT term ($p < 0.001$), which accounts for 85% of all terms in EHR cohort. Moreover, *Surgery* is performed on 57% of patients in subgroup I ($p < 0.001$), which accounts for 71% of all terms in the cohort. These findings suggest that these patients may have a more advanced disease, as evidenced by lack of screening terms among the five most frequent. As a result, they typically undergo chemotherapy treatment used for more advanced cancer, whereas for early-stage cancer primary surgery (lumpectomy, mastectomy) with or without radiation therapy is preferred. Moreover, this group includes subjects that have already received surgical treatments and thus can show either BC in relapse or be disease-free.

In conclusion, it would be important to better characterize what the general term "unlisted chemotherapy" specifically refers to in terms of more specific treatments to better understand the clinical characteristics of patients. Moreover, hormonal profiles of patients can provide important datapoints in terms of determining treatment options for different molecular subtypes of disease. The two groups considered are possibly two consecutive groups, where patients in subgroup II are newly diagnosed and patients in subgroup I may have had BC in the past and survived, either showing relapse, a disease-free survival endpoint, or a late referral at diagnosis.

Fold-2 disease cohorts

Subgroups from Fold-2 disease cohorts are visualized in Figures 1.4 (a)-(f). It is straightforward to observe that the number of clusters only differs for PC and MM between the two folds. In Table 1.11, the analysis of most frequent terms that characterize patients with T2D in Fold-2 leads to very similar results than Fold-1. However, it is worth noting that subgroup II, the one characterized by microvascular diseases, also shows ICD-9 code *Coronary artery atherosclerosis (414.01)*. This suggests that also patients in subgroup II, to some extent, are affected by cardiovascular diseases, that can hence be a common characteristic for more advanced disease courses shared by different phenotypes (Mean age T2D-II 67.17 (14.73)).

Patients with PD are separated into two subgroups also in Fold-2 (see Figure 1.4 b). Moreover, terms in Table 1.12 distinguish the motor and nonmotor subtypes previously found. Alzheimer's disease is again characterized by three groups in Fold-2 (Figure 1.4 c) with the same clinical profiles of Fold-1 subtypes (see Table 1.13). MM encodings from Fold-2 detect 4 instead of the 5 subgroups from



(a)

(b)

(c)

(d)

(e)

(f)

Figure 1.4: Fold-2 complex disorder subgroups. A subsample of 5, 000 patients with T2D is displayed in Figure (a).

Fold-1 (Figure 1.4 d). In particular, subgroups I, II, and III from Fold-2 have similar characteristics to subgroups I, II, and V from Fold-1. Moreover, Fold-2 subgroup IV includes individuals with characteristics similar to those in both subgroups III/IV from Fold-1 (see Table 1.14).

Patients with PC in Fold-2 are split into three subgroups (Figure 1.4 e), where subgroups II and III appears to be a further division of the previously found subgroup II. Interestingly, Fold-2 subgroup III includes significantly younger subjects compared to same fold subgroup II, but sequence lengths are not significantly different. Unfortunately, it is not clear what makes the two subgroups different and further analysis should be performed. Most frequent terms are reported in Table 1.15. Finally, patients with BC in Fold-2 lead to equal number of clusters and clinical phenotypes than Fold-1 (see Figure 1.4 (f), and Table 1.16).

1.4. Conclusions: indications for neurodevelopmental disorders

In this manuscript, we propose a deep learning ConvAE model to infer informative deep encodings of patients from a large and heterogeneous EHR dataset, and leverage these representations for the stratification of cohorts of patients with complex disorders. We demonstrate that when trained on a diverse cohort of patients, we are able to extract latent semantic representations of individuals and stratify them into informative subclusters within broader complex disorder designations.

We perform external validation by first categorizing patients with PD, AD, MM, T2D, PC, BC, ADHD, and CD according to the total number of disease classes. We determine the extent to which externally supplied labels of patient disorders match the categorizations identified by the model. Model performance on this external validation is compared to four different baselines, with ConvAE representations outperforming all baseline encodings. After performing external validation and verifying that clustering of our model representations is effective in identifying the set of diseases of interest, we are able to determine that the same paradigm allows us to detect clinically meaningful subtypes within six of these high-level disease designations. To this aim, subclusters are inspected through the ranking of most frequent occurrences of medical terms.

What emerged from the internal clinical validation of complex disorder subgroups is that sex may play a central role in the heterogeneous manifestations of these conditions (e.g., Alzheimer's disease) and should be taken into account in stratification studies. The effect of sex on the clinical manifestations may be the reason for apparently more homogeneous BC and PC subgroups, which include only females and males, respectively. Second, disease progression, symptom severity, and comorbidities seem to contribute the most to the phenotypic variability of complex disorders. Patients with T2D divides into three subgroups according to comorbidities (i.e., cardiovascular and microvascular problems) and symptom severity (i.e. newly diagnosed with milder symptoms). Individuals with PD show different disease durations and symptoms (i.e., motor, nonmotor). AD profiles distinguish early- and late-onset groups and separate patients with mild neuropsychiatric symptoms and cerebrovascular disease from patients with mild-to-moderate dementia. Patients with MM are characterized by different comorbidities (e.g., amyloidosis, pulmonary diseases) that manifest alongside precise typical signs of MM. Patients with PC and BC separate according to disease progression.

Stratification studies and phenotyping, in particular of complex neurological and neurodevelopmental disorders, can be beneficial to healthcare at different levels of analysis. At the clinical level, establishing the common features of individuals within subgroup may allow us to better identify personalized care regimens that are more effective than blanket approaches. Such methods also allow us to formulate hypothesis about diagnosis and prognosis as well. In a translational perspective, if a new patient is assigned to a particular subtype, clinical characteristics of similar individuals may help to evaluate the risk of treatment side effects and comorbidity onset. Furthermore, the identification of subtypes may foster specialized prevention and intervention programs. At the molecular level, stratified disease cohorts can help in the discovery of new genetic variants in more homogeneous subpopulations. At both the clinical and molecular level, we should aim at identifying the biological factors that govern prognosis, e.g., patients with T2D in subgroup I can evolve into different clinical profiles. Hence, discovering genetic variants associated to these patients can give insights into their disease progression, symptom severity, and comorbidities. Finally, the description of phenotypic profiles may also uphold clinical studies at the population level.

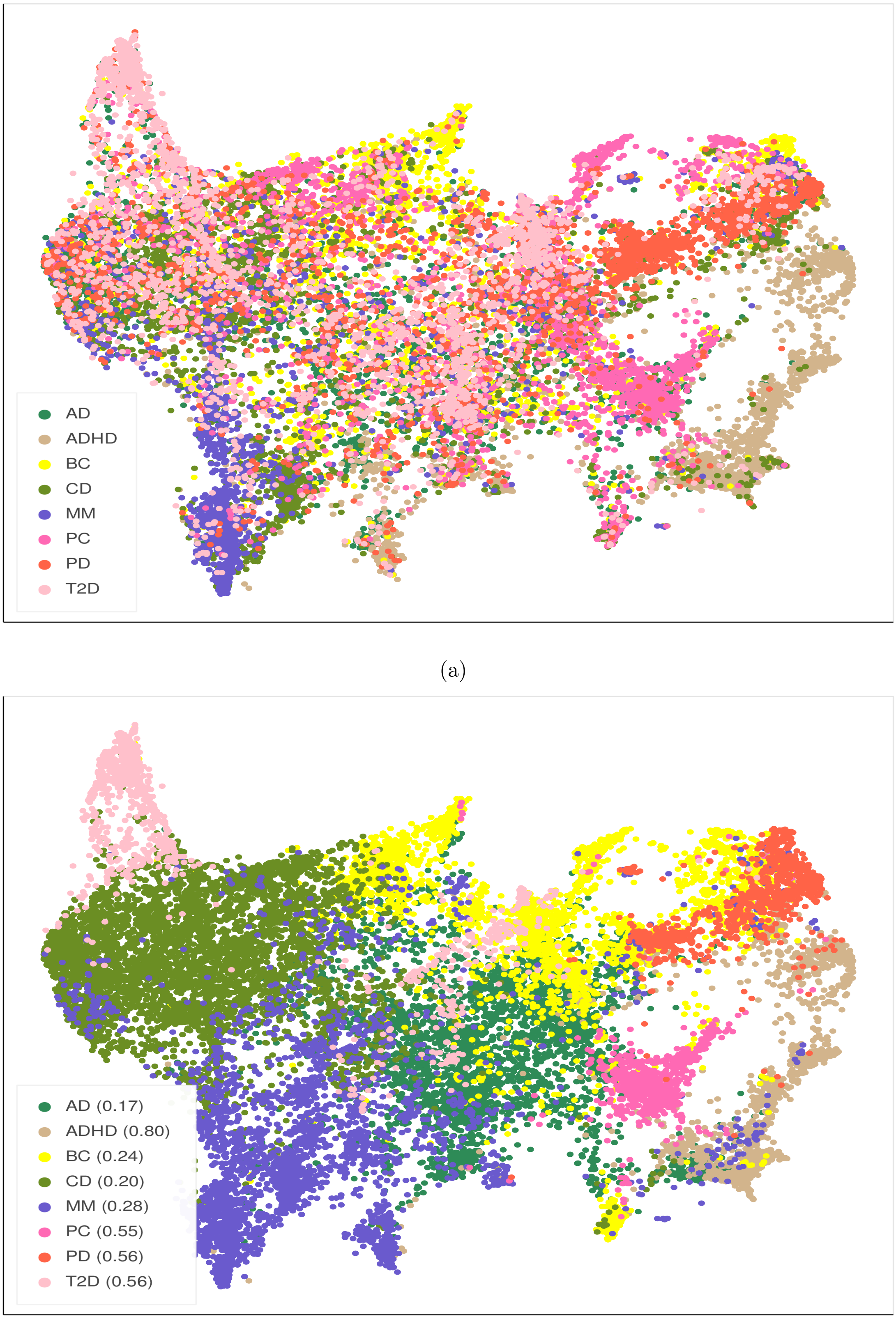
By learning from a large EHR that includes a diverse and heterogeneous population, our method presents opportunity for both generalizability and scalability in the identification and categorization of disease patterns using EHR specific data. Applied to complex disorders, this architecture can be leveraged to detect disease subtypes that reflect disease heterogeneity within patient populations. Moreover, in using the ConvAE model to infer the latent representations of new patients, it is possible to derive the subtype of any particular patient for any particular disease, and track subsequent disease progression and treatment effectiveness. Finally, a fully unsupervised approach is difficult to validate, and translating its discoveries into clinical practice is particularly challenging. Nonetheless, such findings can be leveraged to clarify disease subtypes and improve future disease-specific studies.

Processing EHR with minimum data engineering, on the one hand, preserves all the available information and, to some extent, prevents systematic biases. However, medical records are redundant and sometimes too generic and this complicates phenotyping via enrichment analysis of term occurrences. As an example, we might be interested in the hematopoietic cell transplantation (HCT) technique that is performed on patients with MM to better understand disease severity and progression. Furthermore, it would be interesting to assign a specific type of chemotherapy treatment and hormonal profiles to individuals with BC, in order to uncover the molecular type of the malignancy.

Moving forward, we intend to leverage finer and clearer constructs in the form of pre-processed unstructured terms (i.e., clinical notes) from the EHR into patient sequences. Future work will also focus on including patient genetic profiles in the stratification process.

Among complex disorders, ASCs are a group of neurodevelopmental conditions characterized by heterogeneous manifestations at the genetic, behavioral, neurobiological, and clinical levels, that can vary in the course of an individual developmental trajectory. We have already reported in Section 1.3.1 that encodings of individuals with ADHD are clearly detected during external validation. In the context of neurodevelopmental conditions, this holds promise for translational research purposes.

We intend also to investigate whether the learned representations successfully capture the clinical variability within EHR trajectories of individuals with ASCs. We not only aim to detect subgroups tied to sets of comorbid conditions, but also to inspect pharmacologic treatments, their possible side effects, and screening protocols, to gain insight into tertiary care practices and their efficacy in the management of individuals with ASC.



(a)

(b)

Figure 1.5: UMAP visualization of Fold-1 encodings. (a) complex disorder classes; (b) hierarchical clustering labels.

	Type 2 Diabetes Fold-1			
	Subgroup I (N=18, 325)	**Subgroup II** (N=22, 659)	**Subgroup III** (N=7, 704)	p-value
Female/Male	9, 729/8, 585	11, 053/11, 602	3, 191/4, 510	$< 0.05^a$
Sequence length[1]	96 [3; 4, 849]	192 [9; 4, 849]	64 [5; 3, 767]	$< 0.05^b$
Age[2]	64.77 (14.32)	66.52 (15.13)	69.42 (11.29)	$< 0.05^b$
ICD-9[3]	Hypertension (401.9) - 67% (38%)*	Hypertension (401.9) - 63% (44%)*	Hypertension (401.9) - 77% (18%)*	$^{*} < 0.001^a$
	Hyperlipidemia (272.4) - 53% (43%)*	Postprocedural hypertension (997.91) - 42% (47%)*	Coronary atherosclerosis (vessel) (414.00) - 67% (40%)*	$^{**} < 0.001^a$ (I vs III, II vs III)
	Postprocedural hypertension (997.91) - 49% (44%)*	Edema (782.3) - 38% (54%)*	Coronary artery atherosclerosis (414.01) - 67% (34%)*	
	Edema (782.3) - 33% (37%)*	Hyperlipidemia (272.4) - 36% (36%)*	Hyperlipidemia (272.4) - 61% (21%)*	
	Chest pain (786.50) - 30% (36%)**	Shortness of breath (786.05) - 32% (52%)*	Angina pectoris (413.9) - 47% (47%)*	
Medication	Metformin - 54% (49%)*	Paracetamol - 55% (64%)*	Acetylsalicylic acid - 66% (30%)*	$^{*} < 0.001^a$
	Acetylsalicylic acid 81 mg - 35% (41%)*	Sodium chloride - 53% (66%)*	Sodium chloride - 42% (18%)*	$^{**} < 0.001^a$(I vs II, II vs III)
	Calcium - 28% (47%)*	NA	Clopidrogel - 41% (39%)*	
	Paracetamol - 25% (24%)*	Injection - 45% (62%)*	Acetylsalicylic acid 81 mg - 40% (19%)*	
	Vitamin D - 25% (61%)*	Oxycodone - 42% (66%)**	Metformin - 38% (32%)*	
Lab test	Glucose - 54% (30%)*	Glucose - 80% (55%)*	Hematocrit - 71% (18%)*	$^{*} < 0.001^a$
	Creatinine - 50% (29%)*	Leukocytes - 78% (57%)*	Mean corpuscular hemoglobin concentration - 70% (18%)*	$^{**} < 0.001^a$ (I vs II, II vs III)
	Urea nitrogen - 47% 28%*	Hematocrit - 77% (57%)*	Mean corpuscular volume - 70% (18%)*	
	Chloride - 47% (28%)*	Erythrocytes - 77% (56%)*	Mean corpuscular hemoglobin - 70% (18%)*	
	Sodium - 46% (29%)**	Creatinine - 77% (55%)*	Erythrocytes - 70% (17%)*	
CPT	ECG; interpretation, report - 41% (26%)*	Blood count - 75% (72%)***	ECG; interpretation, report - 80% (22%)†	$^{*} < 0.001^a$
	Calcium - 39% (28%)**	Calcium - 70% (61%)**	Blood count - 65% (25%)*	$^{**} < 0.001^a$ (I vs II, II vs III)
	Hemoglobin A1C - 39% (34%)**	Potassium - 69% (66%)*	Lipid panel - 47% (21%)*	$^{***} < 0.01^a$
	Glucose (reagent strip) - 37% (31%)*	Urea nitrogen - 69% (63%)*	Glucose - 44% (15%)*	
	TSH - 35% (44%)*	Sodium - 69% (66%)*	Hemoglobin A1C - 40% (15%)**	

[1] Median [minimum; maximum]; [2] Mean (standard deviation); [3] from ICD-9 on in-group and (total) percentages; [a] Multiple pairwise chi-squared test; [b] Multiple pairwise t-test

NA = Not Available; ECG = Electrocardiogram; TSH = Thyroid Stimulating Hormone

Table 1.5: Top 5 most frequent terms for the three subgroups of Type 2 Diabetes cohort (Fold-1) in the ICD-9, medication, laboratory test, current procedural terminology ensembles. Each clinical term is followed by relative and total frequencies. Significant comparisons between groups are reported via corrected p-values. When all comparisons are significant, overall significance is reported.

	Parkinson's Disease Fold-1		
	Subgroup I	**Subgroup II**	
	(N=1,368)	(N=1,684)	p-value
Female/Male	596/771	734/950	0.98[a]
Sequence length[1]	26 [3; 1,088]	128 [12; 3,104]	< 0.05[b]
Age[2]	70.76 (13.52)	75.17 (14.11)	< 0.05[b]
ICD-9[3]	Essential tremor (`333.1`) - 21% (56%)*	Hypertension (`401.9`) - 39% (81%)*	* < 0.001[a]
	Anxiety state (`300.00`) - 20% (45%)^{ns}	Postprocedural hypertension (`997.91`) - 32% (83%)*	
	Constipation (`564.00`) - 19% (34%)*	Constipation (`564.00`) - 29% (66%)*	
	Generalized anxiety disorder (`300.02`) - 17% (45%)^{ns}	Edema (`782.3`) - 29% (85%)*	
	Counseling (`V65.40`) - 16% (42%)^{ns}	Other malaise and fatigue (`780.79`) - 25% (72%)*	
Medication	Carbidopa/Levodopa combination - 51% (51%)*	Levodopa - 45% (57%)^{ns}	* < 0.001[a]
	Levodopa - 42% (43%)^{ns}	Carbidopa - 45% (58%)**	** < 0.05[a]
	Carbidopa - 40% (42%)**	Cabidopa/Levodopa combination - 40% (49%)*	
	Carbidopa 25 mg - 37% (57%)*	Paracetamol - 39% (92%)*	
	Levodopa 100 mg - 37% (58%)*	Injection - 31% (88%)*	
Lab test	Mean corpuscular hemoglobin - 3% (4%)*	Glucose - 60% (97%)*	* < 0.001[a]
	Leukocytes - 3% (4%)*	Urea nitrogen - 60% (97%)*	
	Mean platelet volume - 3% (4%)*	Creatinine - 59% (97%)*	
	Width - 3% (4%)*	Leukocytes - 59% (96%)*	
	Erythrocytes - 3% (4%)*	Potassium - 59% (97%)*	
CPT	Unlisted psychiatric service or procedure - 25% (47%)^{ns}	Calcium - 52% (92%)*	* < 0.001[a]
	MRI (brain, brain stem) - 13% (36%)*	Blood count - 52% (94%)*	
	Surgery - 11% (24%)*	ECG; interpretation, report - 51% (95%)*	
	Tonsillectomy and adenoidectomy (< age 12) - 8% (27%)*	Urea nitrogen - 48% (96%)*	
	Tonsillectomy and adenoidectomy (≥ age 12) - 7% (26%)*	Chloride - 47% (96%)*	

[1] Median [minimum; maximum]; [2] Mean (standard deviation); [3] from ICD-9 on in-group and (total) percentages; [a] Multiple pairwise chi-squared test; [b] Multiple pairwise t-test

ns = not significant; ECG = Electrocardiogram; MRI = Magnetic Resonance Imaging

Table 1.6: Top 5 most frequent terms for the two subgroups of Parkinson's disease cohort (Fold-1).

	Alzheimer's Disease Fold-1			
	Subgroup I (N=399)	**Subgroup II** (N=1,170)	**Subgroup III** (N=1,632)	p-value
Female/Male	351/48	792/378	1,249/381	$< 0.001^{a}$
Sequence length[1]	23 [3; 448]	160 [9; 4,465]	128 [9; 4,576]	$< 0.01^{b}$ (I vs II/III)
Age[2]	54.45 (22.48)	84.96 (9.61)	72.51 (19.75)	$< 0.01^{b}$
ICD-9[3]	Routine gynecological examination (V72.31) - 53% (50%)*	Hypertension (401.9) - 51% (48%)*	Postprocedural hypertension (997.91) - 41% (55%)†	$^{*} < 0.001^{a}$
	Postprocedural hypertension (997.91) - 34% (11%)^{ns}	Dementia without behavioral disturbance (294.10) - 39% (58%)*	Hypertension (401.9) - 38% (50%)*	$^{**} < 0.05^{a}$
	Counseling (V65.40) - 34% (20%)**	Postprocedural hypertension (997.91) - 35% (34%)†	Constipation (564.00) - 33% (61%)*	$^{***} < 0.001^{a}$ (I vs II, II vs III)
	Osteoporosis (733.00) - 28% (17%)***	Edema (782.3) - 29% (40%)*	Encounter influenza immunization (V04.81) - 32% (79%)*	$^{\dagger} < 0.01$ (II vs III)
	Family history osteoporosis (V17.81) - 28% (21%)**	Unspecified fall (E888.9) - 28% (44%)**	Cough (786.2) - 29% (60%)**	$^{\star} < 0.001^{a}$ (I vs II/III)
Medication	Calcium - 25% (12%)*	Sodium chloride - 58% (69%)***	Calcium - 32% (60%)*	$^{*} < 0.05^{a}$ (I vs III, II vs III)
	Vitamin D - 24% (13%)**	Paracetamol - 52% (57%)***	Vitamin D - 30% (68%)**	$^{**} < 0.001^{a}$ (I vs II, II vs III)
	Ergocalciferol - 24% (14%)**	Acetylsalicylic acid - 39% (62%)***	Ergocalciferol - 29% (70%)**	$^{***} < 0.001^{a}$
	Estradiol - 15% (50%)***	Docusate sodium - 39% (64%)***	Paracetamol - 27% (42%)***	$^{\dagger} < 0.001^{a}$ (I vs II/III)
	Ethinyl estradiol - 15% (56%)***	Heparin sodium - 36% (77%)***	Donepezil - 25% (54%)†	
Lab test	Chlamydia/Gonorrhoeae amplified DNA - 10% (37%)*	Mean corpuscular volume - 74% (50%)*	Leukocytes - 53% (50%)*	$^{*} < 0.001^{a}$
	Syphilis (rapid plasma reagin) - 6% (6%)**	Creatinine - 74% (51%)*	Glucose - 53% (50%)*	$^{**} < 0.001^{a}$ (I vs II/III)
	HIV 1 - 5% (16%)**	Erythrocytes - 74% (50%)*	Erythrocytes - 53% (50%)*	
	Hepatitis C virus ab - 4% (11%)*	Mean corpuscular hemoglobin concentration - 74% (50%)*	Hematocrit - 52% (50%)*	
	Leukocytes - 4% (1%)*	Glucose - 74% (49%)*	Mean corpuscular hemoglobin concentration - 52% (49%)*	
CPT	Psychiatric service/procedure - 24% (17%)*	Blood count - 75% (56%)***	Calcium - 52% (50%)**	$^{*} < 0.001^{a}$ (I vs II, II vs III)
	Calcium, ionized - 24% (96%)^{ns}	ECG - 73% (57%)***	TSH - 44% (62%)**	$^{**} < 0.01^{a}$
	Calcium - 24% (6%)***	Potassium - 66% (54%)***	Urea nitrogen - 43% (49%)**	$^{***} < 0.001^{a}$
	Vitamin D - 24% (16%)*	Partial Thromboplastin Time Test - 66% (72%)***	Blood count - 42% (44%)**	
	Calcium, total - 23% (8%)**	Calcium - 65% (44%)***	Potassium - 40% (46%)**	

[1] Median [minimum; maximum]; [2] Mean (standard deviation); [3] from ICD-9 on in-group and (total) percentages; [a] Multiple pairwise chi-squared test; [b] Multiple pairwise t-test

ns = not significant; NA = Not Available; ECG = Electrocardiogram; ab = antibodies; TSH = Thyroid-stimulating hormone

Table 1.7: Top 5 most frequent terms for the three subgroups of Alzheimer's disease cohort (Fold-1).

	Multiple Myeloma Fold-1					
	Subgroup I	**Subgroup II**	**Subgroup III**	**Subgroup IV**	**Subgroup V**	
	(N=469)	(N=138)	(N=703)	(N=234)	(N=340)	p-value
Female/Male	236/233	74/63	336/367	129/105	144/196	0.003^a (IV vs V)
Sequence length[1]	372 [28; 4, 849]	17.50 [4; 256]	768 [15; 4, 849]	64 [5; 1, 527]	160 [22; 4, 625]	$< 0.005^b$, 0.4^b (I vs V)
Age[2]	69.67 (14.00)	64.38 (15.20)	63.40 (12.57)	65.39 (17.10)	63.86 (12.32)	$< 0.005^b$ (I vs II/III/IV/V)
ICD-9[3]	Edema (782.3) - 53% (31%)*	Disease of salivary glands (527.9) - 28% (9%)***	Edema (782.3) - 57% (50%)*	Postprocedural hypertension (997.91) - 45% (16%)*	Oth inflammatory/toxic neuropathy (357.89) - 34% (22%)**	
	Anemia (285.9) - 47% (31%)***	Disorders of bone and cartilage (733.99) - 17% (5%)***	Other malaise and fatigue (780.79) - 55% (53%)***	Shortness of breath (786.05) - 37% (13%)*	Disease of salivary glands (527.9) - 27% (21%)***	
	Shortness of breath (786.05) - 45% (31%)*	Hypertension (401.9) - 9% (2%)***	Diarrhea (787.91) - 55% (58%)***	Hypertension (401.9) - 34% (14%)***	Unsp inflammatory/toxic neuropathy (357.9) - 24% (18%)**	
	Pleura effusion (511.9) - 45% (58%)***	Other malaise and fatigue (780.79) - 9% (2%)***	Anemia (285.9) - 50% (49%)***	Other malaise and fatigue (780.79) - 34% (11%)***	Anemia (285.9) - 23% (11%)***	
	Fever (780.60) - 43% (32%)***	Counseling (V65.40) - 8% (2%)*	Shortness of breath (786.05) - 48% (50%)*	Cough (786.2) - 32% (13%)***	Counseling (V65.40) - 22% (15%)*	
Medication	Paracetamol - 69% (42%)**	Vitamin D - 7% (2%)**	Calcium - 66% (54%)***	Calcium - 33% (9%)***	Calcium - 42% (17%)***	
	Sodium chloride - 65% (41%)**	Oxycodone - 7% (2%)***	Injection - 56% (54%)*	Vitamin D - 27% (13%)**	Dexamethasone - 28% (14%)**	
	Oxycodone - 48% (36%)***	Fentanyl - 7% (2%)***	Dexamethasone - 54% (55%)**	Vitamin D3 - 23% (14%)*	Bortezomib - 26% (17%)***	
	Fentanyl - 48% (39%)***	Ergocalciferol - 6% (2%)*	Sodium chloride - 54% (50%)***	Cholecalciferol - 23% (14%)*	Iron - 25% (17%)***	
	Heparin - 48% (46%)***	Acetylsalicylic acid 81 mg - 6% (2%)*	Ondansetron - 52% (56%)***	Ergocalciferol - 23% (12%)*	Acetylsalicylic acid 81 mg - 23% (16%)*	
Lab test	Erythrocytes - 83% (28%)***	Hemoglobin - 9% (1%)*	Chloride - 93% (47%)*	Protein - 27% (5%)***	Hematocrit - 91% (22%)*	
	Glucose - 83% (28%)***	Lymphocytes - 8% (1%)***	Glucose - 93% (46%)***	Glucose - 24% (4%)***	Platelets - 90% (22%)*	
	Urea nitrogen - 83% (28%)***	Leukocytes - 8% (1%)**	Potassium - 93% (47%)***	Erythrocytes - 24% (4%)***	Erythrocytes - 90% (22%)***	
	Mean corpuscular hemoglobin - 83% (28%)**	Mean platelet volume - 8% (1%)**	Mean corpuscular hemoglobin - 93% (46%)**	Leukocytes - 24% (4%)**	Lymphocytes - 90% (23%)***	
	Mean corpuscular hemoglobin concentration - 83% (28%)*	Mean corpuscular hemoglobin concentration - 8% (1%)*	Width - 93% (47%)**	Creatinine - 23% (4%)***	Eosinophils - 89% (23%)***	
CPT	Blood count - 79% (33%)***	Diagnostic/interventional CT - 36% (7%)*	Calcium - 78% (46%)***	Calcium - 42% (8%)***	Gammaglobulin - 63% (24%)**	
	Calcium - 73% (28%)***	PET limited area (Head/neck) - 32% (7%)*	Blood count - 75% (51%)**	Calcium, ionized - 31% (8%)**	Albumin - 60% (19%)***	
	ECG; interpretation, report - 72% (39%)***	PET-CT (skull base to mid-thigh) - 22% (7%)*	Albumin - 73% (49%)***	Surgery - 30% (14%)***	Calcium, ionized - 59% (17%)***	
	Potassium - 71% (34%)*	Tumor imaging PET-CT - 12% (39%)***	Blood count - 72% (51%)***	ECG; interpretation, report - 30% (8%)***	Lactate dehydrogenase - 58% (22%)*	
	PTT - 71% (43%)**	CT thorax (no contrast) - 9% (3%)***	Blood count - 72% (45%)***	Calcium, total - 30% (7%)***	Beta-2 microglobulin - 55% (26%)***	

[1] Median [minimum; maximum]; [2] Mean (standard deviation); [3] from ICD-9 on in-group and (total) percentages; [a] Multiple pairwise chi-squared test; [b] Multiple pairwise t-test

$^{*}p < 0.05^a$; $^{**}p < 0.01^a$; $^{***}p < 0.001^a$

ns = not significant; NA = Not Available; ECG = Electrocardiogram; CT = Computed Tomography; PET = Positron Emission Tomography; PTT = Partial Thromboplastin Time

Table 1.8: Top 5 most frequent terms for the five subgroups of Multiple Myeloma cohort (Fold-1).

	Malignant neoplasm of prostate Fold-1		
	Subgroup I	**Subgroup II**	
	(N=6,916)	(N=1,606)	p-value
Sequence length[1]	96 [9; 4,817]	20 [3; 672]	< 0.05[b]
Age[2]	69.64 (12.98)	69.78 (10.56)	0.68[b]
ICD-9[3]	Hypertension (401.9) - 40% (92%)*	Nocturia (788.43) - 29% (33%)**	* < 0.001[a]
	Postprocedural hypertension (997.91) - 30% (93%)*	Elevated PSA (790.93) - 18% (27%)*	** < 0.01[a]
	Hyperlipidemia (272.4) - 28% (95%)*	Impotence of organic origin (607.84) - 18% (35%)*	
	Edema (782.3) - 24% (94%)*	Urinary frequency (788.41) - 15% (27%)*	
	Personal history of PC (V10.46) - 20% (97%)*	Hypertension (401.9) - 14% (8%)*	
Medication	Paracetamol - 44% (98%)*	Midazolam - 17% (12%)*	* < 0.001[a]
	Oxycodone - 40% (98%)*	Tadalafil - 14% (35%)*	
	Fentanyl - 38% (93%)*	Fentanyl - 12% (7%)*	
	Ondansetron - 33% (97%)*	Sildenafil - 12% (33%)*	
	Injection - 32% (99%)*	Tamsulosin - 10% (12%)*	
Lab test	Glucose - 66% (96%)*	PSA total - 19% (18%)[ns]	* < 0.05[a]
	Leukocytes - 63% (98%)*	PSA post-prostatectomy - 17% (25%)**	** < 0.001[a]
	Creatinine - 63% (99%)*	Testosterone - 16% (26%)**	
	Urea nitrogen - 63% (99%)*	Glucose - 11% (4%)*	
	Potassium - 62% (99%)*	PSA free - 10% (27%)**	
CPT	Calcium - 53% (98%)*	PSA total - 73% (26%)*	* < 0.001[a]
	Urea nitrogen - 50% (98%)*	PSA free - 51% (27%)*	
	Potassium - 49% (98%)*	Testosterone total - 29% (32%)*	
	Chloride - 49% (98%)*	Surgery - 22% (14%)*	
	Sodium - 49% (98%)*	Ultrasound post-voiding residual urine/bladder capacity - 18% (29%)*	

[1] Median [minimum; maximum]; [2] Mean (standard deviation); [3] from ICD-9 on in-group and (total) percentages; [a] Multiple pairwise chi-squared test; [b] Multiple pairwise t-test

ns = not significant; PSA = Prostate-Specific Antigen

Table 1.9: Top 5 most frequent terms for the two subgroups of prostate cancer cohort (Fold-1).

	Malignant neoplasm of breast (female) Fold-1		
	Subgroup I	**Subgroup II**	
	(N=5,971)	(N=1,993)	p-value
Sequence length[1]	128 [6; 4,849]	26 [3; 2,010]	< 0.05[b]
Age[2]	66.67 (14.27)	62.86 (13.73)	< 0.05[b]
ICD-9[3]	Hypertension (401.9) - 34% (89%)*	Lump or mass in breast (611.72) - 27% (29%)*	* < 0.001[a]
	Edema (782.3) - 28% (93%)*	Abnormal mammogram (793.80) - 23% (37%)*	
	Postprocedural hypertension (997.91) - 26% (94%)*	Other screening mammogram (V76.12) - 19% (38%)*	
	Other malaise and fatigue (780.79) - 25% (93%)*	Carcinoma in situ of breast (233.0) - 15% (27%)[ns]	
	Constipation (564.00) - 25% (93%)*	Hypertension (401.9) - 13% (11%)*	
Medication	Paracetamol - 50% (92%)*	Propofol - 27% (19%)*	* < 0.001[a]
	Ondansetron - 46% (87%)*	Fentanyl - 26% (16%)*	
	Fentanyl - 45% (84%)*	Lidocaine - 25% (21%)*	
	Oxycodone - 43% (91%)*	Midazolam - 22% (18%)*	
	Propofol - 40% (81%)*	Ondansetron - 21% (13%)*	
Lab test	Glucose - 67% (97%)*	Leukocytes - 7% (3%)*	* < 0.001[a]
	Leukocytes - 67% (97%)*	Glucose - 6% (3%)*	
	Erythrocytes - 66% (97%)*	Platelets - 6% (3%)*	
	Width - 65% (97%)*	Erythrocytes - 6% (3%)*	
	Mean corpuscular hemoglobin - 65% (97%)*	Mean corpuscular hemoglobin - 6% (3%)*	
CPT	Calcium - 58% (95%)*	Mammography - 35% (32%)*	* < 0.001[a]
	Blood count - 55% (95%)*	Ultrasound - 32% (27%)**	** < 0.05[a]
	Urea nitrogen - 52% (97%)*	Surgery - 30% (19%)*	
	Potassium - 47% (97%)*	Mastectomy, partial - 28% (22%)*	
	Sodium - 47% (97%)*	Mammography, bilateral - 26% (39%)*	

[1] Median [minimum; maximum]; [2] Mean (standard deviation); [3] from ICD-9 on in-group and (total) percentages; [a] Multiple pairwise chi-squared test; [b] Multiple pairwise t-test

ns = not significant

Table 1.10: Top 5 most frequent terms for the two subgroups of breast cancer cohort (Fold-1).

	Type 2 Diabetes Fold-2			
	Subgroup I (N=22,791)	**Subgroup II** (N=19,621)	**Subgroup III** (N=6,347)	p-value
Female/Male	12,121/10,652	9,247/10,374	2,651/3,692	$< 0.001^a$
Sequence length[1]	96 [3; 4,849]	221 [9; 4,849]	64 [5; 4,749]	$< 0.05^b$
Age[2]	64.64 (14.50)	67.17 (14.73)	69.76 (11.35)	$< 0.05^b$
ICD-9[3]	Hypertension (401.9) - 66% (46%)*	Hypertension (401.9) - 64% (38%)*	Hypertension (401.9) - 78% (15%)*	* $< 0.001^a$
	Hyperlipidemia (272.4) - 53% (54%)*	Postprocedural hypertension (997.91) - 42% (41%)*	Coronary artery atherosclerosis (414.01) - 70% (29%)*	
	Postprocedural hypertension (997.91) - 48% (54%)*	Edema (782.3) - 40% (48%)*	Coronary atherosclerosis (vessel) (414.00) - 67% (33%)*	
	Edema (782.3) - 32% (45%)*	Hyperlipidemia (272.4) - 35% (30%)*	Hyperlipidemia (272.4) - 57% (16%)*	
	Chest pain (786.50) - 31% (46%)^{ns}	Coronary artery atherosclerosis (414.01) - 34% (43%)*	Angina pectoris (413.9) - 49% (40%)*	
Medication	Metformin - 54% (60%)*	Paracetamol - 58% (59%)***	Acetylsalicylic acid - 73% (27%)***	* $< 0.001^a$ (I vs II/III)
	Acetylsalicylic acid 81 mg - 35% (50%)**	Sodium chloride - 57% (60%)***	Sodium chloride - 47% (16%)***	** $< 0.001^a$(I vs II, II vs III)
	Calcium - 27% (55%)***	NA	Clopidrogel - 43% (34%)***	*** $< 0.001^a$
	Injection - 26% (35%)***	Injection - 47% (56%)***	Intracoronary nitroglicerin - 41% (55%)***	
	Paracetamol - 25% (30%)***	Glucagon - 45% (66%)***	Metformin - 39% (27%)***	
Lab test	Glucose - 52% (36%)*	Glucose - 84% (50%)*	Hematocrit - 81% (17%)**	* $< 0.001^a$
	Creatinine - 48% (34%)*	Urea nitrogen - 82% (51%)*	Mean corpuscular hemoglobin concentration - 80% (17%)**	** $< 0.001^a$ (I vs II/III)
	Urea nitrogen - 46% 33%*	Hematocrit - 82% (52%)**	Hemoglobin - 80% (17%)**	*** $< 0.05^a$
	Chloride - 46% (33%)*	Leukocytes - 82% (52%)*	Erythrocytes - 80% (16%)**	
	Potassium - 44% (34%)*	Creatinine - 81% (50%)*	Width - 79% (16%)***	
CPT	Calcium - 43% (37%)*	Blood count - 76% (63%)**	ECG; interpretation, report - 80% (18%)**	* $< 0.001^a$ (I vs II, II vs III)
	ECG; interpretation, report - 43% (35%)**	Potassium - 70% (57%)**	Blood count - 74% (23%)**	** $< 0.001^a$
	Hemoglobin A1C - 41% (44%)***	Calcium - 70% (52%)*	Lipid panel - 50% (19%)**	*** $< 0.01^a$
	Urea nitrogen - 38% (35%)***	ECG; interpretation, report - 69% (48%)**	Glucose - 48% (14%)***	
	Glucose - 37% (38%)**	Sodium - 69% (56%)**	Potassium - 44% (12%)**	

[1] Median [minimum; maximum]; [2] Mean (standard deviation); [3] from ICD-9 on in-group and (total) percentages; [a] Multiple pairwise chi-squared test; [b] Multiple pairwise t-test

ns = not significant; *NA* = Not Available; ECG = Electrocardiogram

Table 1.11: Top 5 most frequent terms for the three subgroups of Type 2 Diabetes cohort (Fold-2).

	Parkinson's Disease Fold-2		
	Subgroup I	**Subgroup II**	
	(N=1,851)	(N=1,220)	p-value
Female/Male	799/1,051	535/684	0.73[a]
Sequence length[1]	35 [3; 3,855]	160 [12; 4,369]	< 0.05[b]
Age[2]	71.39 (12.76)	75.65 (15.25)	< 0.05[b]
ICD-9[3]	Anxiety state (`300.00`) - 24% (65%)*	Hypertension (`401.9`) - 45% (64%)*	$^{*} < 0.01$[a]
	Constipation (`564.00`) - 23% (57%)**	Postprocedural hypertension (`997.91`) - 31% (57%)*	$^{**} < 0.05$[a]
	Essential tremor (`333.1`) - 22% (79%)*	Edema (`782.3`) - 30% (59%)*	
	Generalized anxiety disorder (`300.02`) - 20% (65%)**	Constipation (`564.00`) - 26% (43%)**	
	Counseling (`V65.40`) - 19% (62%)^{ns}	Other malaise and fatigue (`780.79`) - 26% (53%)*	
Medication	Carbidopa/Levodopa combination - 49% (68%)*	Carbidopa - 47% (43%)**	$^{*} < 0.001$[a]
	Levodopa - 42% (58%)**	Levodopa - 46% (42%)**	$^{**} < 0.05$[a]
	Carbidopa - 41% (57%)**	Paracetamol - 45% (77%)*	
	Carbidopa 25 mg - 37% (75%)*	Sodium chloride - 42% (90%)*	
	Levodopa 100 mg - 36% (75%)*	Carbidopa/Levodopa combination - 36% (32%)*	
Lab test	Glucose - 9% (15%)*	Erythrocytes - 77% (85%)*	$^{*} < 0.001$[a]
	Leukocytes - 9% (15%)*	Mean corpuscolar hemoglobin - 75% (86%)*	
	Creatinine - 9% (15%)*	Glucose - 75% (85%)*	
	Erythrocytes - 9% (15%)*	Width - 75% (86%)*	
	Urea nitrogen - 8% (15%)*	Leukocytes - 75% (85%)*	
CPT	Unlisted psychiatric service or procedure - 29% (70%)*	Blood count - 64% (84%)*	$^{*} < 0.001$[a]
	Surgery - 17% (48%)*	Calcium - 62% (78%)*	
	MRI (brain, brain stem) - 16% (58%)^{ns}	Urea nitrogen - 60% (85%)*	
	Calcium - 12% (22%)*	ECG - 59% (82%)*	
	Tonsillectomy and adenoidectomy (< age 12) - 9% (43%)*	Potassium - 58% (86%)*	

[1] Median [minimum; maximum]; [2] Mean (standard deviation); [3] from ICD-9 on in-group and (total) percentages; [a] Multiple pairwise chi-squared test; [b] Multiple pairwise t-test

ns = not significant; ECG = Electrocardiogram; MRI = Magnetic Resonance Imaging

Table 1.12: Top 5 most frequent terms for the two subgroups of Parkinson's disease cohort (Fold-2).

	Alzheimer's Disease Fold-2			
	Subgroup I (N=1422)	**Subgroup II** (N=1461)	**Subgroup III** (N=267)	p-value
Female/Male	1004/413	997/463	265/2	< 0.001[a] (I vs III, II vs III)
Sequence length[1]	64 [3; 4785]	160 [16; 4416]	29 [7; 316]	< 0.01[b]
Age[2]	71.64 (19.53)	84.45 (10.61)	42.12 (16.22)	< 0.01[b]
ICD-9[3]	Postprocedural hypertension (`997.91`) - 40% (48%)*	Hypertension (`401.9`) - 49% (59%)*	Routine gynecological examination (`V72.31`) - 81% (53%)*	* < 0.001[a]
	Hypertension (`401.9`) - 35% (41%)*	Dementia without behavioral disturbance (`294.10`) - 39% (70%)*	Counseling (`V65.40`) - 57% (24%)*	** < 0.001[a] (I vs III, II vs III)
	Edema (`782.3`) - 29% (49%)**	Postprocedural hypertension (`997.91`) - 33% (40%)*	Postprocedural hypertension (`997.91`) - 54% (12%)*	
	Encounter influenza immunization (`V04.81`) - 28% (58%)*	Unspecified fall (`E888.9`) - 28% (58%)*	Osteoporosis (`733.00`) - 48% (20%)*	
	Constipation (`564.00`) - 28% (45%)^{ns}	Constipation (`564.00`) - 28% (47%)^{ns}	Family history of osteoporosis (`V17.81`) - 46% (25%)*	
Medication	Calcium - 30% (50%)*	Sodium chloride - 51% (76%)*	Vitamin D - 42% (15%)*	* < 0.001[a]
	Ergocalciferol - 28% (56%)*	Paracetamol - 45% (67%)*	Calcium - 42% (13%)*	** < 0.005[a]
	Vitamin D - 28% (53%)*	Docusate sodium - 34% (80%)*	Ergocalciferol - 40% (15%)*	*** < 0.001[a] (I vs II, II vs III)
	Donepezil - 25% (45%)**	Heparin sodium - 34% (95%)***	Ethinyl estradiol - 27% (65%)*	† < 0.02[a] (I vs III, II vs III)
	Injection - 22% (42%)*	Acetylsalicylic acid - 33% (70%)*	Iron - 21% (12%)†	
Lab test	Glucose - 36% (30%)*	Erythrocytes - 83% (70%)*	Chlamydia/Gonorrhoeae amplified DNA - 16% (47%)*	* < 0.001[a]
	Erythrocytes - 36% (29%)*	Hematocrit - 82% (71%)*	HIV 1 - 4% (11%)^{ns}	** < 0.001[a] (I vs II, II vs III)
	Width - 35% (29%)*	Mean platelet volume - 82% (71%)*	Syphilis (rapid plasma reagin) - 4% (3%)**	
	Platelets - 35% (30%)*	Urea nitrogen - 82% (72%)*	Hepatitis B (sag) - 4% (10%)^{ns}	
	Hemoglobin - 35% (29%)*	Mean corpuscular hemoglobin concentration - 82% (71%)*	Hepatitis C virus ab - 4% (8%)^{ns}	
CPT	Calcium - 42% (35%)*	Blood count - 79% (74%)***	Calcium - 40% (6%)*	* < 0.001[a] (I vs II, II vs III)
	TSH - 38% (48%)**	Potassium - 70% (74%)**	Calcium, ionized - 39% (14%)***	** < 0.001[a] (I vs III, II vs III)
	ECG; interpretation, report - 34% (32%)***	Calcium - 70% (59%)*	Calcium, total - 39% (10%)†	*** < 0.001[a]
	Vaccine/toxoid - 29% (67%)***	Urea nitrogen - 69% (71%)***	Vitamin D - 38% (16%)***	† < 0.001[a] (I vs II, I vs III)
	Urea nitrogen - 29% (29%)***	Glucose - 69% (74%)***	Psychiatric service/procedure - 36% (18%)***	

[1] Median [minimum; maximum]; [2] Mean (standard deviation); [3] from ICD-9 on in-group and (total) percentages; [a] Multiple pairwise chi-squared test; [b] Multiple pairwise t-test

ns = not significant; NA = Not Available; ECG = Electrocardiogram; sag = surface antigen; ab = antibodies; TSH = Thyroid-stimulating hormone

Table 1.13: Top 5 most frequent terms for the three subgroups of Alzheimer's disease cohort (Fold-2).

	Multiple Myeloma Fold-2				
	Subgroup I	**Subgroup II**	**Subgroup III**	**Subgroup IV**	
	(N=191)	(N=891)	(N=502)	(N=299)	p-value
Female/Male	108/83	473/418	230/272	168/130	0.005^a (III vs IV)
Sequence length[1]	29 [3; 779]	318 [6; 4849]	507 [16; 4849]	160 [20; 3835]	$< 0.008^b$, 0.68^b (II vs III)
Age[2]	61.87 (18.76)	64.49 (12.44)	67.24 (14.97)	67.30 (15.40)	$< 0.008^b$ (I vs III/IV, II vs III/IV)
ICD-9[3]	Other malaise and fatigue (780.79) - 19% (5%)***	Oth inflammatory/toxic neuropathy (357.89) - 48% (79%)**	Edema (782.3) - 56% (38%)**	Postprocedural hypertension (997.91) - 55% (25%)*	
	Disease of salivary glands (527.9) - 15% (7%)***	Anemia (285.9) - 40% (50%)**	Fever (780.60) - 51% (40%)**	Hypertension (401.9) - 49% (27%)***	
	Constipation (564.00) - 15% (5%)**	Other malaise and fatigue (780.79) - 38% (48%)*	Anemia (285.9) - 49% (34%)**	Edema (782.3) - 43% (17%)**	
	Fever (780.60) - 15% (4%)**	Unsp inflammatory/toxic neuropathy (357.89) - 38% (76%)***	Shortness of breath (786.05) - 49% (38%)***	Hyperlipidemia (272.4) - 35% (37%)**	
	Counseling (V65.40) - 14% (5%)**	Diarrhea (787.91) - 36% (49%)*	Postprocedural hypertension (997.91) - 44% (34%)*	Other malaise and fatigue (780.79) - 35% (15%)***	
Medication	Oxycodone - 12% (3%)**	Calcium - 59% (61%)***	Paracetamol - 76% (50%)***	Calcium - 39% (14%)***	
	Lidocaine - 1% (4%)***	Dexamethasone - 45% (59%)***	Sodium chloride - 75% (51%)**	Vitamin D - 39% (27%)*	
	Acetylsalicylic acid 81 mg - 9% (3%)**	Bortezomib - 42% (70%)***	Oxycodone - 63% (47%)**	Ergocalciferol - 35% (27%)*	
	Dexamethasone - 9% (3%)***	Acetylsalicylic acid 81 mg - 38% (65%)**	Ondansetron - 56% (46%)**	Paracetamol - 31% (12%)*	
	Paracetamol - 9% (2%)***	Injection - 37% (47%)***	Diphenhydramine - 54% (52%)*	Cholecalciferol - 31% (24%)***	
Lab test	Leukocytes - 15% (2%)***	Width - 91% (57%)***	Sodium - 88% (31%)***	Glucose - 60% (12%)***	
	Erythrocytes - 14% (2%)***	Mean platelet volume - 91% (57%)***	Mean corpuscular volume - 88% (31%)***	Leukocytes - 58% (12%)***	
	Hematocrit - 14% (2%)*	Mean corpuscular hemoglobin - 91% (57%)***	Chloride - 87% (31%)***	Creatinine - 57% (12%)*	
	Mean corpuscular volume - 14% (2%)***	Hemoglobin - 91% (56%)***	Urea nitrogen - 87% (31%)***	Protein - 57% (12%)***	
	Platelets - 13% (2%)***	Hematocrit - 90% (57%)*	Leukocytes - 87% (30%)***	Urea nitrogen- 57% (12%)***	
CPT	Diagnostic/interventional CT - 23% (6%)**	Calcium - 71% (53%)***	Blood count - 85% (39%)***	Calcium - 57% (14%)***	
	PET limited area (Head/neck) - 19% (6%)***	Albumin - 70% (58%)***	ECG; interpretation, report - 78% (46%)**	ECG; interpretation, report - 46% (16%)**	
	Surgery - 17% (7%)***	Gammaglobulin - 69% (68%)**	PTT - 76% (53%)***	Urea nitrogen - 46% (13%)**	
	Psychiatric service/procedure - 15% (6%)*	Blood count - 66% (56%)***	Potassium - 76% (39%)***	Calcium, total - 42% (13%)***	
	PET-CT (skull base to mid-thigh) - 12% (6%)**	Lactate dehydrogenase - 64% (67%)***	Sodium - 75% (39%)***	TSH - 42% (27%)**	

[1] Median [minimum; maximum]; [2] Mean (standard deviation); [3] from ICD-9 on in-group and (total) percentages; [a] Multiple pairwise chi-squared test; [b] Multiple pairwise t-test

$^*p < 0.05^a$; $^{**}p < 0.01^a$; $^{***}p < 0.001^a$

ns = not significant; *NA* = Not Available; ECG = Electrocardiogram; CT = Computed Tomography; MRI = Magnetic Resonance Imaging; PET = Positron Emission Tomography; TSH = Thyroid-stimulating hormone; PTT = Partial Thromboplastin Time

Table 1.14: Top 5 most frequent terms for the four subgroups of Multiple Myeloma cohort (Fold-2).

	Malignant neoplasm of prostate Fold-2			
	Subgroup I (N=2,703)	**Subgroup II** (N=3,846)	**Subgroup III** (N=2,096)	p-value
Sequence length[1]	28 [3; 1,051]	96 [64; 4,849]	96 [56; 3,392]	$< 0.05^{b}$, 0.19^{b} (II vs III)
Age[2]	68.71 (12.46)	70.92 (11.92)	67.83 (14.49)	$< 0.01^{b}$, 0.07 (I vs III)
ICD-9[3]	Nocturia (788.43) - 28% (50%)*	Hypertension (401.9) - 44% (56%)***	Postprocedural hypertension (997.91) - 47% (4%)*	$^{*} < 0.001^{a}$
	Elevated PSA (790.93) - 20% (49%)*	Personal history of PC (V10.46) - 28% (77%)*	Hypertension (401.9) - 44% (31%)***	$^{**} < 0.05^{a}$)
	Urinary frequency (788.41) - 17% (45%)**	Hyperlipidemia (272.4) - 25% (47%)*	Hyperlipidemia (272.4) - 41% (42%)*	$^{***} < 0.001$ (I vs II/III)
	Impotence of organic origin (607.84) - 16% (52%)*	Postprocedural hypertension (997.91) - 23% (40%)*	Edema (782.3) - 34% (38%)*	
	Hypertension (401.9) - 14% (13%)***	Edema (782.3) - 23% (47%)*	Shortness of breath (786.05) - 30% (45%)*	
Medication	Midazolam - 15% (18%)*	Paracetamol - 68% (81%)*	Acetylsalicylic acid 81 mg - 26% (40%)*	$^{*} < 0.001^{a}$
	Fentanyl - 13% (12%)**	Oxycodone - 61% (82%)*	Calcium - 22% (42%)*	$^{**} < 0.001^{a}$ (I vs II, II vs III)
	Tadalis - 12% (47%)***	Fentanyl - 58% (79%)**	Vitamin D - 21% (57%)*	$^{***} < 0.001^{a}$ (I vs II/III)
	Ciprofloxacin - 11% (23%)*	Ondansetron - 50% (82%)*	Ergocalciferol - 20% (59%)*	$^{\dagger} < 0.01^{a}$
	Tamsulosin - 11% (23%)†	Morphine - 50% (92%)*	Acetylsalicylic acid - 19% (26%)†	
Lab test	PSA total - 20% (33%)*	Glucose - 84% (68%)*	Glucose - 47% (21%)*	$^{*} < 0.001^{a}$
	Glucose - 20% (11%)*	Erythrocytes - 84% (74%)*	Urea nitrogen - 46% (21%)*	$^{**} < 0.001^{a}$ (I vs II/III)
	Protein - 18% (14%)*	Leukocytes - 84% (72%)*	Creatinine - 46% (22%)*	
	Bilirubin - 17% (14%)*	Urea nitrogen - 84% (72%)*	Protein - 45% (28%)*	
	Testosterone - 16% (39%)**	Mean corpuscular hemoglobin concentration - 84% (76%)*	Chloride - 44% (21%)*	
CPT	PSA total - 68% (40%)*	Potassium - 71% (79%)*	PSA total - 51% (23%)*	$^{*} < 0.001^{a}$
	PSA free - 52% (44%)*	Sodium - 71% (79%)*	PSA free - 41% (27%)*	$^{**} < 0.001^{a}$ (I vs II/III)
	Surgery - 25% (25%)**	Calcium - 71% (72%)*	ECG; interpretation, report - 41% (32%)**	
	Testosterone - 23% (43%)**	Chloride - 71% (79%)*	Calcium - 40% (22%)*	
	Ultrasound post-voiding residual urine/bladder capacity - 28% (48%)*	Urea nitrogen - 70% (77%)*	Surgery - 34% (26%)**	

[1] Median [minimum; maximum]; [2] Mean (standard deviation); [3] from ICD-9 on in-group and (total) percentages; [a] Multiple pairwise chi-squared test; [b] Multiple pairwise t-test

ECG = Electrocardiogram; PSA = Prostate-Specific Antigen

Table 1.15: Top 5 most frequent terms for the three subgroups of prostate cancer cohort (Fold-2).

	Malignant neoplasm of breast (female) Fold-2		
	Subgroup I	**Subgroup II**	
	(N=5,601)	(N=2,237)	p-value
Sequence length[1]	128 [6; 4,785]	25 [3; 480]	< 0.05[b]
Age[2]	66.98 (14.57)	61.94 (13.25)	< 0.05[b]
ICD-9[3]	Hypertension (401.9) - 37% (89%)*	Lump or mass in breast (611.72) - 26% (33%)*	* < 0.001[a]
	Edema (782.3) - 29% (93%)*	Abnormal mammogram (793.80) - 22% (43%)*	** < 0.05[a]
	Postprocedural hypertension (997.91) - 28% (93%)*	Other screening mammogram (V76.12) - 19% (44%)*	
	Shortness of breath (786.05) - 25% (95%)*	Carcinoma in situ of breast (233.0) - 15% (32%)**	
	Other malaise and fatigue (780.79) - 24% (93%)*	Hypertension (401.9) - 11% (11%)*	
Medication	Paracetamol - 50% (89%)*	Propofol - 28% (23%)*	* < 0.001[a]
	Fentanyl - 45% (80%)*	Fentanyl - 28% (20%)*	
	Ondansetron 44% (83%)*	Midazolam - 24% (22%)*	
	Oxycodone - 42% (88%)*	Lidocaine - 23% (23%)*	
	Propofol - 38% (77%)*	Ondansetron - 23% (17%)*	
Lab test	Leukocytes - 69% (97%)*	Leukocytes - 6% (3%)*	* < 0.001[a]
	Glucose - 69% (97%)*	Glucose - 6% (3%)*	
	Hematocrit - 67% (97%)*	Width - 5% (3%)*	
	Erythrocytes - 67% (97%)*	Mean corpuscular hemoglobin concentration - 5% (3%)*	
	Width - 66% (97%)*	Erythrocytes - 5% (3%)*	
CPT	Calcium - 57% (93%)*	Mammography - 33% (36%)*	* < 0.001[a]
	Blood count - 56% (95%)*	Ultrasound - 32% (32%)*	
	Urea nitrogen - 52% (97%)*	Surgery - 30% (21%)*	
	ECG; interpretation, report - 49% (96%)	Mastectomy, partial - 28% (25%)*	
	Potassium - 48% (97%)*	Ultrasound, breast(s) - 24% (40%)*	

[1] Median [minimum; maximum]; [2] Mean (standard deviation); [3] from ICD-9 on in-group and (total) percentages; [a] Multiple pairwise chi-squared test; [b] Multiple pairwise t-test

ECG = Electrocardiogram

Table 1.16: Top 5 most frequent terms for the two subgroups of breast cancer cohort (Fold-2).

Chapter 2

Behavioral embeddings for the stratification of Autism Spectrum Conditions

Abstract

Autism Spectrum Conditions (ASCs) are a set of neurodevelopmental clinical conditions characterized by impairment in social communication/interaction and restricted and repetitive patterns of behavior/interests. ASC heterogeneity encompasses differences at multiple levels of analysis (e.g., behavioral, genetic) throughout developmental trajectories. Disentangling heterogeneity within ASC population can help develop effective individual treatments and discover high-impact biomarkers that may be masked in case-control studies.

In this chapter, we present a computational approach to represent autistic individuals by means of low-dimensional vectors from their behavioral/cognitive trajectories for stratification studies. Each term of the trajectories is obtained from test scores with specific refinements, that can include general scores (e.g., IQ), as well as "deeper" indices (e.g., cognitive assessment subtests), according to the dimensions of interest.

Word embedding models, such as Word2vect, are applied to encode collections of test scores from the battery of standard tests administered to children and adolescents with ASC referred to the Laboratory of Observation, Diagnosis, and Education (ODFLab). The approach is then used to group together the individual trajectories, offering a patient stratification by means of hierarchical clustering.

Results on the ODFLab dataset show that the new embeddings better leverage the information from sparse, longitudinal datasets, respect to standard quantitative behavioral data processing. Moreover, we were able to identify ASC subgroups characterized by different age at diagnosis, mild-to-severe core symptoms, and varying cognitive impairment.

2.1. Introduction

With the term Autism we refer to a set of behaviorally-defined neurodevelopmental conditions characterized by difficulties in social communication, social interactions and repetitive and restricted behavior and interests. The Diagnostic and Statistical Manual of Mental Disorders, Fifth Edition (DSM-5) [67] describes autism in terms of a *spectrum disorder* with three levels of severity in the two core areas of communication/interactions and behavior/interests. The etiology of autism is still unknown and it is believed to be influenced by multiple genetic, environmental and lifestyle factors, which ensures great heterogeneity that encompasses several levels of analysis (e.g., behavior, genetics, development).

To account for within individual differences at the genetic, neural, behavioral, cognitive and clinical levels, it has been proposed [68] to approach autism in a neurodiversity framework, referring to a set of conditions entailing differences. Indeed, it is argued that the term “disorder” refers to a dysfunctional state which cannot be ascribed to a specific cause or mechanism, whereas the term “difference” can be used in the context of an individual that shows an atypical trait, which not necessarily affects functioning or well-being, when compared to a population norm. Nevertheless, some forms of autism need to be considered as disorders, especially the cases characterized by severe impairments or in co-occurrence with other conditions (e.g., intellectual disability or language disorders), whereas other forms (e.g., above average Intelligence Quotient and no history of language delay) are possibly better addressed as differences, in that, in an autism-friendly environment, individuals may function even at a higher level than typical individuals. The debate on whether to refer to autism as a disorder or a condition is still ongoing, however, to emphasize its multiple-level variability, in the following, we will adopt the term Autism Spectrum Conditions (ASCs).

According to the Centers for Disease and Control Prevention, in the United States, ASC prevalence is estimated at about 1 in 59 children aged 8 years, and it is 4 times more common among males than females [69]. The male/female ratio decreases to $1-2:1$ when moderate to severe intellectual disability is considered [70]. In Europe, a pilot project funded by the European Parliament and managed by the European Commission (i.e., Autism Spectrum Disorder in the European Union - ASDEU) involving 12 European countries reports a prevalence of 1 in 89 children aged $7-9$ years[1]. Within the ASDEU project, the Italian ASC prevalence study indicates a prevalence of 1 in 87 (1.15%) school-aged children [71]. Although ASDEU has led to important findings at the European level, paving the way for the development of future policies, neither the European multi-national study nor the Italian one reports data on male/female ratio. At the population level, studies in Asia, Europe and North America estimate ASC prevalence at $1-2\%$ for individuals with age ranging from 0 to 20 years[2]. Large-scale population-based studies also indicate a male preponderance of the condition with a ratio of $2-3:1$ [4].

Behavioral phenotype refers to a set of behavior and cognitive characteristics that occur in individuals with a known genetic abnormality [72]. As an example, Fragile X syndrome, caused by trinucleotide repeat expansion in the X chromosome gene FMR1, is characterized by behavioral phenotype associated with cognitive impairment, hyperactivity, social anxiety, and autistic-like features. Nevertheless, from the other direction, interest in behavioral phenotypes has grown for idiopathic complex conditions, such as ASCs, in order to understand the underlying biological pathways involved [73].

The advent of computational psychiatry [74, 75], which stems from the increasing availability of healthcare data, has encouraged the application of computational techniques, such as machine learning (ML) and deep learning, for data-driven discoveries in clinical settings, in an attempt to uncover the biological foundations of psychiatric conditions and develop new effective treatments [3]. These techniques have also been applied to ASC data, and what has emerged is that autism spectrum heterogeneity presumably reflects the existence of multiple subgroups tied to diverse genetic

[1] http://asdeu.eu/wp-content/uploads/2016/12/ASDEUExecSummary27September2018.pdf (Accessed on August 31, 2019)

[2] https://www.cdc.gov/ncbddd/autism/data.html (Accessed on August 31, 2019)

and environmental etiologies and that, although convergent to the behavioral autistic profile, show a range of clinical phenotypic characteristics under behavioral, linguistic and cognitive levels.

The heterogeneous display of clinical manifestations and phenotypic characteristics in autism complicates diagnosis and treatments. Hence, understanding autism heterogeneity is key to 1) the development of more accurate diagnostic tools, which take into account both the behavioral and the genetic profiles; 2) fast and early diagnoses; and 3) timely and personalized treatments and interventions.

In this study, we propose a data analysis protocol that processes ASC psychometric data from cognitive and behavioral assessment instruments via ML techniques from Natural Language Processing (NLP) that aim at representing subjects as multidimensional vectors that include aspects of autism heterogeneity (i.e., developmental and behavioral profiles). ML clustering techniques are then applied for stratification purposes, i.e., to create stable clusters of more homogeneous clinical and behavioral characteristics. In autism research, deep embedding representations are still an unexplored method; a recent work [76] combines quantitative screening data and the embedding of categorical values from diagnostic tools as input to a multi-layer neural architecture for autism screening. However, to the best of our knowledge, the study we present is the first attempt to learn encodings of individual ASC profiles, and leverage these encodings, in a completely unsupervised setting, which is designed to learn patterns from data in the absence of an *a priori* knowledge. Deep representations of individuals with ASC for behavioral phenotype stratification is a novel approach that can be groundbreaking to elucidate autism heterogeneity.

2.2. Aspects of autism heterogeneity

We introduce here a review of autism research on the heterogeneous characterization of ASC symptomatology. Furthermore, we describe the latest approaches of ML and deep learning to ASC data in the context of how they can be leveraged for the investigation of autism heterogeneity. Within this framework, we present a novel application of ML towards phenotypic stratification of autism.

Heterogeneity within individuals with ASC has been explored according to different clinical, biological, and developmental aspects. Furthermore, within each area, differences are showing at multiple levels of analysis (e.g., genetic variants detected in both coding and non-coding regions from sequencing studies).

2.2.1. Comorbidities

Coexisting conditions are registered in 70% of individuals [4]. In particular, developmental conditions, once included in autism diagnostic criteria, are now enumerated in the set of medical and psychiatric disorders that co-occur in individuals with ASCs. They include 1) atypical language development and abilities, that show great variability within individuals, varying according to age and comorbid conditions; and 2) motor abnormalities, that are reported in $\sim 79\%$ of ASC cases. Individuals with age < 6 years can show nonconforming language and delay in comprehension, together with difficulties in expressive phonology and grammar, whereas older individuals show nonconforming pragmatics, semantics and morphology. Motor abnormalities comprise motor delay, hypotonia, catatonia, and deficits in coordination, motor planning, gait, and balance. Among developmental conditions, Attention Deficit Hyperactivity Disorder (ADHD) is usually concomitant with ASCs ($28 - 44\%$) [4].

A retrospective study that evaluates the prevalence of medical and psychiatric conditions in $14,381$ individuals with ASC from four hospitals in the Boston area, leverages Electronic Health Records (EHRs) to detect comorbidities significantly associated with ASC compared to the general population [77]. For individuals younger than 35, epilepsy is found in 19.4% of cases, schizophrenia 2.4%, bowel disorders 11.7%, inflammatory bowel disease 0.8%, type I diabetes mellitus 0.8%, autoimmune disorders 0.7%, central nervous system and craniofacial anomalies 12.4%, sleep disorders 1.1%, muscular dystrophy 0.47%, and single gene disorders (e.g., Fragile X Syndrome, Tuberous Sclerosis) $< 1\%$. With the exception of autoimmune disorders, all comorbidities have significant higher counts in ASC cohort compared to the general population ($\sim$ 2M patients).

2.2.2. Sex/gender bias

Male bias in ASC prevalence and more frequent co-occurrence of intellectual disabilities (IDs) and different symptom manifestations reported in females [78] may suggest the presence of factors (e.g., genetic, epigenetic, environmental) that lead to female protection and/or male vulnerability. Numerous studies have focused on the biological basis of male preponderance and characteristics of ASC [70], and different theories, not necessarily mutually exclusive, have been proposed.

The *Greater Variability model* states that males show a greater genetic variability than females resulting in an increased incidence of ASC diagnoses and a decreased severity in symptoms. The *Liability-threshold model* reports that females meeting the diagnostic threshold for ASC carry higher mutational load (i.e., the total genetic burden from deleterious mutations) than males, also carried by relatives. Females with ASC have more genes with *de novo* mutations and genes central to biological pathways more likely to be affected by deletions and duplications [70]. Within the liability-threshold model, the *Sex-chromosome theory* proposes X chromosome in females to be protective. Although ASC-linked genes largely involve autosomes, some of them reside on the X chromosome generating the idea that X-inactivation, which occurs when one copy of X chromosomes in females is silenced, may be protective. Sex hormones have also been accounted for influencing biological structures and processes linked to ASC, such as immune system and neurotransmitter signalling. The *Extreme Male Brain theory* states that males with ASC show more extreme profiles in the systemizing and empathizing dimensions, key in defining male and female brain, possibly due to different levels of fetal testosterone [79].

Together with the investigation of the genetic, epigenetic and environmental factors that may be responsible for the difference in incidence and symptom severity between males and females with ASCs, *diagnostic biases* towards males have also been explored and the theory that ascribes male prevalence in part to underidentified female cases is gaining ground. Most research findings involving biological processes fail to be replicated and are usually based on small sample sizes or remain understudied. Moreover, large-scale population-based studies report a lower male/female ratio and do not find an association between gender ratio and ID suggesting that male prevalence could be less pronounced. It is argued that behavioral differences and manifestations in females with ASC are overlooked by current diagnostic and psychometric instruments, in that, despite unbalanced gender ratio is now widely acknowledged, they could already be influenced by sex and gender because of the longstanding male predominance in case identifications. Modifying the tools to leverage finer level constructs, e.g., social-emotional reciprocity and social anxiety instead of deficits in social communication and social interactions, can help pin down the differences in ASC manifestations and provide better diagnostic instruments in relation to sex and gender [78].

It has also been observed [80], that differences between males and females with ASC can stem from sociocultural influences and expectations. Social and peer pressure can induce females with ASC to camouflage core deficits and non-conforming behavior and gender stereotypes have been considered as well as one of the possible causes of misdiagnosis and overlook. Males with ASC seem to be recognized more easily in the familial and school settings because their behavior is identified as more in contrast with gender role expectations, whereas female behavior stereotypes cause females with ASC to be perceived as showing more accentuated typical behavioral characteristics (e.g., being shy, passive).

2.2.3. Genetic variants and biological pathways

At the genetic level, ASCs display a strong heritable component, in fact, hundreds of genes seem to contribute to ASC risk and heritability is estimated to be between $50-80\%$ [5, 6]. Variants associated with ASCs include all possible mutations, from single-nucleotide variants, that manifest as an alteration of single base pairs, to insertion and deletions, that involve thousands of base pairs (copy number variants). Together with inherited variants, also *de novo* variants (i.e., variants detected in probands and not present in parental genome) contribute to the etiology of autism. It is estimated that rare genetic variants are the most frequent, causal in $10-30\%$ of cases, but it has been observed that also common genetic variation contribute to ASC risk, even though they typically confer a much lower risk in an individual [81]. Recessive mutations have also been investigated in a large ASC cohort study [82] leading to the identification of biallelic disruption mutations of known recessive neurodevelopmental genes as well as not previously implicated genes which are involved in the serotonergic circuit.

Transcriptome analysis (RNA-seq) is an important complement to genomic studies because it allows to inspect both coding and non-coding transcriptional activities and it has been leveraged to understand the common mechanisms underlying ASC. This approach yields information about transcript abundance suggesting potential link between genomic variations and transcriptional dysregulation and can be applied to messenger RNA, that conveys genetic information, small or long non-coding RNA, which regulates gene expression, and it can also investigate alternative splicing, which is a regulatory process that allows a single gene to code for different proteins. Notable findings have uncovered that RNA splicing regulatory programs are disrupted in ASC affecting neuronal activity and that differences in the expression of neuronal genes in a cohort with ASCs respect to controls are thought to be responsible for the disruption of cortex generation during development [5]. Single-cell sequencing, a technique that generates expression profiles at the single cell level, has also been used to investigate transcriptomic changes in specific cell types of the brain (e.g., excitatory neurons, interneurons, astrocytes). A recent study [83] reports that the expression of genes involved in synaptic signaling pathways of cortical neurons is affected in individuals with ASC and that dysregulation of genes in cortical neurons correlates with clinical severity of autism.

Recent advancement in computational tools and sequencing techniques has led to the investigation of the possible effects of non-coding variants in the context of large sample cohorts of individuals with ASC. A study on the contribution of several non-coding genetic regions to ASC susceptibility has been able to distinguish cases ($N = 2,182$ children with ASC) from a separate, unrelated control group. This was done applying a logistic regression classifier that performed best when trained on simple repeat sequences, i.e., sequences of repeating base pairs in intergenic regions of DNA, that have been linked to neuronal differentiation and brain development [84]. Another recent study has applied a deep learning framework, namely convolutional neural networks (CNNs), to predict the functional and pathogenic impact of *de novo* non-coding mutations on $2,002$ transcriptional and 232 post-transcriptional mechanisms. The architecture is trained on biochemical data marking interactions between DNA- and RNA-binding proteins and their targets from $1,790$ simplex families. As a result, a higher impact of *de novo* mutations is observed in probands as compared to unaffected siblings. Moreover, the underlying processes and pathways that are more likely to be affected by these mutations have been identified via a network-based statistical approach that aims at detecting genes and pathways relevant to mutations that disrupt transcriptional and post-transcriptional regulations. According to this study variants from the non-coding genome of ASC probands likely affect neuronal development, synaptic transmission and chromatin regulation [85].

2.2.4. Behavioral and developmental trajectories

At the behavioral level, the broadening of the diagnostic definition in DSM-5 and the elimination of previous subcategories (e.g., pervasive developmental disorder not otherwise specified, and Asperger's disorder) to the advantage of the term *spectrum* extends the concept of behavior heterogeneity [95]. In addition to the behavioral differences reported between males and females with ASC, also age, language development, and cognitive abilities intervene in ASC behavioral presentations. Moreover, all these factors affect the age at which atypical development is recognized, further complicating action, intervention, and support consistent with the severity of manifestations [4]. For this reason, particular attention has turned to the recognition of early signs of atypical development [96] with the aim to lower the diagnostic age from $3-4$ years to $18-24$ months and provide better acknowledgment and care [97].

Research on developmental trajectories provides evidence of a wide range of growth patterns in individuals with ASC. The identification of behavioral differences at multiple levels (e.g., core symptoms,

Reference	Subject (N)	Dataset	Initial age (yrs)	Time points (N)	Instrument scores	Subgroups (N)	Method
[86]	6, 975	DDS[a]	2 – 4	≥ 4	CDER[b] interview core symptom	6	Group-based latent trajectory modeling
[87]	345	Autism center (NC, USA) Clinic (Chicago)	2 – 15	2 – 8	ADOS[c] severity	4	Latent-class growth curve analysis
[88]	129	Local clinics, community (WI, USA)	2.5	3 – 4 (5.5 yrs, final age)	ADOS severity	4	Latent-class growth curve analysis
[89]	421	Canadian multisite study	2 – 4	4 (4 yrs, final age)	ADOS severity VABS[d] Composite	2 – 3	Semi-parametric group-based approach
[90]	203	-	1 – 4	3	ADOS total	5	Latent-class growth curve analysis
[7]	85	4 Autism centers (NC, USA) 1 Autism clinic (Chicago)	2	5 (19 yrs, final age)	ADI-R[e] (social/communication, RSM[f], insistence on sameness) VABS Composite V/NV IQ[g]	3	Growth curve analysis
[91]	105	Community (MD, USA)	3	4 (7.99 yrs, final age)	VABS Composite	2	Growth mixture model
[92]	106	51 public psychiatric units (France)	4.4	4 (20.6 yrs, final age)	VABS (Communication, Socialization, Daily living skills)	2	Semi-parametric group-based approach
[93]	116	4 Autism centers (NC, USA) 1 autism clinic (Chicago)	9	≥ 2 (18 yrs, final age)	ABC[h] total (maladaptive behavior)	3	Growth curve analysis
[94]	403 (high-risk) 163 (low-risk)	4 ASD diagnostic/treatment centers (Canada)	6 – 12 months	4 (36 months, final age)	VABS Composite	3	Semi-parametric group-based approach

Table 2.1: Studies on developmental trajectories in ASCs.

[a] DDS – California Department of Developmental Services

[b] CDER – Client Development Evaluation Report

[c] ADOS – Autism Diagnostic Observation Schedule

[d] VABS – Vineland Adaptive Behavior Scales

[e] ADI-R – Autism Diagnostic Interview - Revised

[f] RSM – Repetitive Sensory Motor

[g] V/NV IQ – Verbal/Nonverbal Intelligence Quotient

[h] ABC – Aberrant Behavior Checklist

symptom severity, adaptive behavior) throughout development and the role of language, co-occurring conditions, and cognitive abilities in shaping ASC progression is crucial for the characterization of ASC phenotype. To this aim, several studies have tried to separate trajectories into informative groups in the attempt to provide fine grained taxonomies (see Table 2.1). The groups identified are often further characterized respect to correlates, i.e., language, cognition, and co-occurring traits. From these studies it emerges a significant heterogeneity between trajectory subgroups that differentiate their growth (i.e., worsening, stable, improving) and level of impairment (i.e., severe, moderate, mild), compared to typical development. Language and non-verbal Intelligence Quotient (IQ) levels are often significant predictors of class membership. A possible limitation to these studies is that, with the only exception of [86], where $6,975$ children with ASC were followed from 4 to 14 years of age (≥ 4 evaluations), all other studies rely on small sample sizes and few temporal data points (see Table 2.1).

2.2.5. Computational approaches and big data

The use of ML and deep learning for autism research stems from the increasing availability of massive quantities of healthcare data from sources such as electronic medical records and genome sequencing. This has recently led researchers to apply, alongside *supervised* learning techniques, *unsupervised* approaches in an attempt to pin down distinct clusters to discriminate autistic manifestations. Examples of unsupervised methods are *clustering* algorithms, in which external labels are assigned to samples that are more similar according to a metric, and more recently developed deep learning models, such as Generative Adversarial Networks [98] that generate new sets of data capturing the distribution of existing ones.

Supervised learning In this setting, predictive models are applied to labeled data (e.g., ASC diagnosis), which can be cross-sectional (i.e., single time point) or longitudinal, on features pertaining to biology, medical imaging, or behavioral psychology [99]. Particular attention has focused on highlighting a set of key diagnostic behavioral features that should be considered to recognize individuals with ASC, in order to reduce assessment time, provide an early diagnosis, and facilitate the diagnostic process. In [100], Random Forest (RF) classifiers are trained on Autism Diagnostic Observation Schedule (ADOS) and Autism Diagnostic Interview - Revised (ADI - R) behavioral scores and the top ranked features that best discriminate between ASC diagnoses and controls in children between 18 and 84 months are extracted. Results suggest that the best features for younger subjects are non verbal (eye contact, gestures, facial expressions), whereas for older subjects they are focused on verbal communication and interaction. Other studies have attempted to select a subset of diagnostic items from ADI-R, Social Responsiveness Scale (SRS), and ADOS applying classification methods and feature selection in discriminating children with ASC from normotypical controls to develop quick and effective ASC diagnostic instruments. Chosen methods are Alternating Decision Trees and Support Vector Machines (SVMs) [101, 102, 103].

To identify early biomarkers for ASC recognition, Li and colleagues [104] have applied multi-channel CNNs to brain imaging data trained on discriminating children with and without ASC at 24 months of age. ML models have also been leveraged in genetics. Krishnan and colleagues at Princeton University and Simons Foundation have trained a weighted linear SVM on a brain-specific gene-interaction functional network [105]. The model learned to discriminate known ASC-linked genes from genes annotated to non-mental-health disorders and highlighted the network patterns in which ASC-linked genes are involved. Genes whose interaction paths resemble those learned by the model are then selected. To validate the results, an exome-sequencing study on *de novo* mutations of $2,517$ simplex families from the Simons Simplex Collection is used and it shows that genes with likely-gene-disrupting mutations are significantly enriched in the top of the new candidate gene ranking. Furthermore, gene expression signatures of spatiotemporal regions of the developing brain are reported to be enriched among the newly-predicted genes. The study lists hundreds of new candidate genes converging on key pathways (e.g., synaptic transmission, neuronal function) and that are likely to be linked to early development phenotypes (e.g., enteric nervous system development, fetal prefrontal, temporal and cerebellar cortex development). These genetic discoveries add to the thousands of genes that have already been linked to ASC, increasing genetic heterogeneity and further complicating therapeutic translations.

Unsupervised learning Unsupervised ASC stratification aims at identifying data-driven patterns in the absence of explicit labeling. Most works focus on clustering analysis of symptom scores. The identified clusters are then further externally validated by cognition, behavioral and emotional functioning measures [99]. Subtype stratifications define informative clinical profiles for further genetic evaluations or to investigate the effectiveness of a specific treatment [106] for more personalized interventions. Clustering and profiling of individuals with ASC usually rely on ADI-R, ADOS core symptom and Vineland Adaptive Behavior Scale evaluations.

Clustering algorithms on ADI-R/ADOS scores has led to the identification of 4 and 2 severity-based subclusters in two studies that subsequently pair the clustering with genetic [107] and Single-Nucleotide Polymorphism (SNP) frequencies [108], respectively. Hu and colleagues have identified 4 subgroups from ADI-R scores on several domains (e.g., language, nonverbal communication, social interactions), characterized by 1) severe language deficits; 2) milder symptoms across the domains; 3) noticeable savant skills; 4) intermediate severity across domains, with lower frequency of savant skills. Veatch et al. have revealed the presence of two subgroups of individuals with ASC, based on ADI-R, ADOS, and Vineland scores. One group is described as “more severe”, i.e., with increasing severity in social and communication scores and motor skills; the other is described as “less severe”.

Interestingly, a work by Lombardo et al. [109] applies hierarchical clustering to the evaluation of mentalizing domain, a cognitive phenomenon linked to difficulties in social-communicative features, and it is one of the few stratification studies, together with [108], that replicates its findings in an independent dataset, reporting the existence of 5 subclusters showing different patterns. Another approach [20] leverages EHRs and hierarchical clustering for the identification of ASC subgroups according to their comorbidities towards the discover of different genetic and environmental contributions to etiology. In particular, 4 different subgroups are identified, with prevalence of 1) epilepsy and recurrent seizures; 2) gastrointestinal and auditory disorders; 3) psychiatric conditions (e.g., bipolar disorder, anxiety, depression); and 4) not better specified comorbidities.

A work by Feczko and colleagues [110] combines supervised and unsupervised techniques to identify subgroups within an ASC cohorts of 47 children between the ages of 9 and 13. Measures from seven different tasks (e.g., Delay Discounting, Spatial Span) are input to a RF classification algorithm to predict individuals with ASC from children with Typical Development (TD; $N = 58$). The RF algorithm produces a proximity matrix, that shows how often a pair of subjects are grouped into the same terminal node by the RF algorithm. This proximity matrix is then considered as a graph and a community detection algorithm is used to identify subgroups in both ASC and TD groups. Feczko et al. report three ASC subgroups and four TD groups that show similar cognitive profiles based on measures of individual impulsivity, visuospatial working memory capacity, and vigilance, among others, suggesting that typical variation in cognitive profiles may impact ASC manifestations. Moreover, no differences in severity of symptoms have been found between ASC subgroups. To externally validate the subgroups, differences in functional brain organization from resting-state functional connectivity MRI between ASC groups are investigated. Results identify one group with altered visual processing mechanisms, a second group with altered attention mechanisms, and a third with both.

2.2.6. Subject embeddings for behavioral phenotyping

To shed light on ASC etiology a possible approach requires: large sample sizes, stratified cohorts, and clinical domain expertise.

- *Large datasets*: small sample studies are usually characterized by low statistical power (i.e., the probability to correctly reject the null hypothesis when it is false), unless the true size of an effect is large and can reliably be observed in reduced samples. Button and colleagues [111] have estimated statistical power in neuroscience (e.g., neuroimaging studies, animal models) deriving the best effect size estimate from a meta-analysis. They have calculated the power of each selected study to detect the effect indicated by the meta-analysis and reported that it does not reach 20%.

  A simulation of sample effect size estimates at different sample sizes across a range of true population effects for a hypothetical case-control study in [95] reports that small sample studies are more likely to show biased effect size estimates compared to the population, due to sampling variability, and that this is attenuated only in large sample size studies ($n > 1000$) irrespective of the true effect. Moreover, these simulations also suggest that small true effect sizes (Cohen's $d = 0.1$) lead to inflated effect estimates ($\sim$ 350% inflation on average) potentially due to skewed distribution in the direction of the effect, particularly in small sample studies ($n < 50$). If the true population effect increases ($d > 0.5$) this phenomenon is attenuated and, notably, very little inflation is detected in large sample sizes ($n > 100$).

- *Stratified cohorts*: for disorders and conditions displaying heterogeneity, case-control studies might suffer from the lack of stratified population samples and fail to identify the likely underlying genetic and phenotypic heterogeneity. It has been shown [14], that increased phenotypic heterogeneity affects the effect size of findings in Genome Wide Association Studies. In particular, genotypic data for $2,938$ control individuals, and $1,936$ individuals with Type 1 Diabetes (T1D), and $1,924$ individuals with Type 2 Diabetes (T2D) are analyzed to investigate the effect of phenotypic heterogeneity on the 20 SNPs most significantly associated with T1D or T2D. Two sets of experiments are run progressively increasing the numerosity of T1D or T2D samples, respectively and associations are investigated via chi-squared tests. Results report that, increasing the numerosity of one stratum over the other, the magnitude of the association of the SNPs declines substantially. At a relatively low degree of stratum "inflation" ($\sim 20/30\%$) most of the significantly associated SNPs become equivocal. This entails that, even in the case of a large sample size, if a highly stratified population is considered as homogeneous, effects can be overlooked.

- *Clinical validation*: validation of results by clinical domain experts is central to translational discoveries. Moreover, it is necessary to avoid methodological and clinical inaccuracies and errors that will eventually lead to fallacies in the results [112].

In ASC studies, population heterogeneity is not necessarily reflected by the sample drawn. In fact, if strata exist, samples can be biased by the enrichment of certain strata of the population over others. As an example, this is presumably what happens in most case-control studies in which the enrichment of male over females leads to male-biased conclusions. The use of a large sample size should guarantee a sufficient sampling that reflects the true prevalence of all the population strata [95]. The lack of knowledge of clear ASC subtypes prevents the construction of stratified cohorts and, although this can be overcome by large datasets, it urges stratification studies that investigate ASC behavioral phenotype subgroups that can be leveraged to uncover their genetic contributions. In this unsupervised framework, the design of the experiment and the validation and interpretation of results should carefully be backed up by clinical experts to effectively provide insights into ASC heterogeneity.

Tackling ASC heterogeneity at the cognitive, and behavioral level not only allows to increase the power of analysis at the genetic level, possibly uncovering new genetic variants linked to ASC, but also favors the identification of stronger associations to subgroups of individuals. Unsupervised approaches applied to observational psychological measures also facilitate the development of fine-grained instruments and the identification of additional specifiers that can help identify reliable subtypes avoiding diagnostic biases (e.g., sex/gender). Moreover, phenotypic refinements may facilitate early diagnoses and the development of targeted interventions towards personalized treatment [113].

Observational data that give insights into ASC manifestations are, as an example, psychometric measures collected via administered tests or recorded video of interventions and home-setting interactions from which it is possible to extract behavioral features, such as eye contact and social smile. Assessment instruments are highly heterogeneous, because they include different data types (e.g., numerical, ordinal, categorical), and they address ASC phenotypes at multiple levels of analysis (e.g., cognitive, behavioral, core symptoms). Moreover, they can be longitudinal, i.e., administered at multiple time points throughout development, to keep track of the progression of the condition, the impact on individual's well being, and to provide feedback on interventions and treatments.

A downside to this plethora of collections is the large amount of missing information usually accompanying ASC data. This, not only regards the absence of labels for subject stratification within ASC, but also input data, that are often incomplete, and small sized. The data collection process is time consuming and requires lots of resources. Furthermore, it is difficult to generate unique large psychological datasets, gathering information from multiple clinical sites, due to the lack of shared assessment protocols that uniform the great variability of observational instruments. ML can be leveraged to overcome the issue of missing information, and to simultaneously grasp the information provided by the longitudinal interdependencies and the multidimensional heterogeneity of clinical scores. In particular, NLP methods and clustering algorithms can be applied to ASC data to, first, provide a subject representation of the developmental and behavioral profile, second, it can assign individuals with ASC to the most similar subgroup according to their profile.

In recent years, NLP has been employed in healthcare, in particular for the secondary use of EHRs [114]. Patients medical records are sequences of both structured (e.g., diagnostic codes, medications, laboratory tests) and unstructured (e.g., clinical notes) data, that provide a snapshot of a person's health status, acting as fingerprints of disease progressions. In this context, clinical terms can be treated as words to provide a representation of patients that reflects their health status and disease course. Text mining methods, based on word counts in documents, and word embeddings, i.e., the deep representation of words in a vector space that aims at maintaining their similarity to other words and their semantic and syntactic attributes in texts, have been applied to EHRs. The goal is to produce patient representations able to capture the temporal development of patient's status for patient similarity [24] and/or to extract patient phenotypes [23]. Moreover, these methods can also be used in stratification studies, see Chapter 1.

In Section 2.3, we combine NLP, and ML unsupervised clustering techniques in a novel data analysis protocol, designed to learn behavioral embeddings of individuals with ASCs with a generalizable and scalable stratification method. This method can then be leveraged to 1) investigate genetic heterogeneity; 2)identify fine-grained, and unbiased ASC profiles to provide early diagnoses; 3) predict ASC characteristics throughout development; and 4) help develop personalized treatments and interventions. Ideally, when neuropsychological and behavioral tests are administered to an individual referred to the healthcare system that displays characteristics compatible with ASC, these information are encoded via trained models and the subject is assigned to the most similar cluster of individuals showing homogeneous characteristics and condition progression.

Data used to test the stratification model are from 204 children and adolescent with ASCs (169 males, 35 females) from the Laboratory of Observation, Diagnosis and Education (ODFLab - University of Trento, Italy). Features of the ASC ODFLab dataset are detailed in Section 2.3.3. Data from cognitive, behavioral, and screening instruments are considered, and each subject is represented by a sequence of terms created from administered instruments and the correspondent scores, ordered according to the date of assessment. The analysis protocol we have developed allows the processing of longitudinal information as well as multiple instrument scores and it includes methods for behavioral term embeddings and clustering algorithms. In the dataset, the minimum number of encounters (i.e., sessions during which a battery of tests is administered to a subject) is 1 and the maximum number is 5, with median equal to 1, whereas the number of administered instruments throughout development ranges from 3 to 24 with median 5. The unsupervised framework we have developed (see Section 2.3.2) is data-specific (i.e., it is not tested on an independent dataset), due to the reduced sample size of our dataset. Most importantly, the clinical interpretation and validation of ASC subgroups is provided in Section 2.4.1.

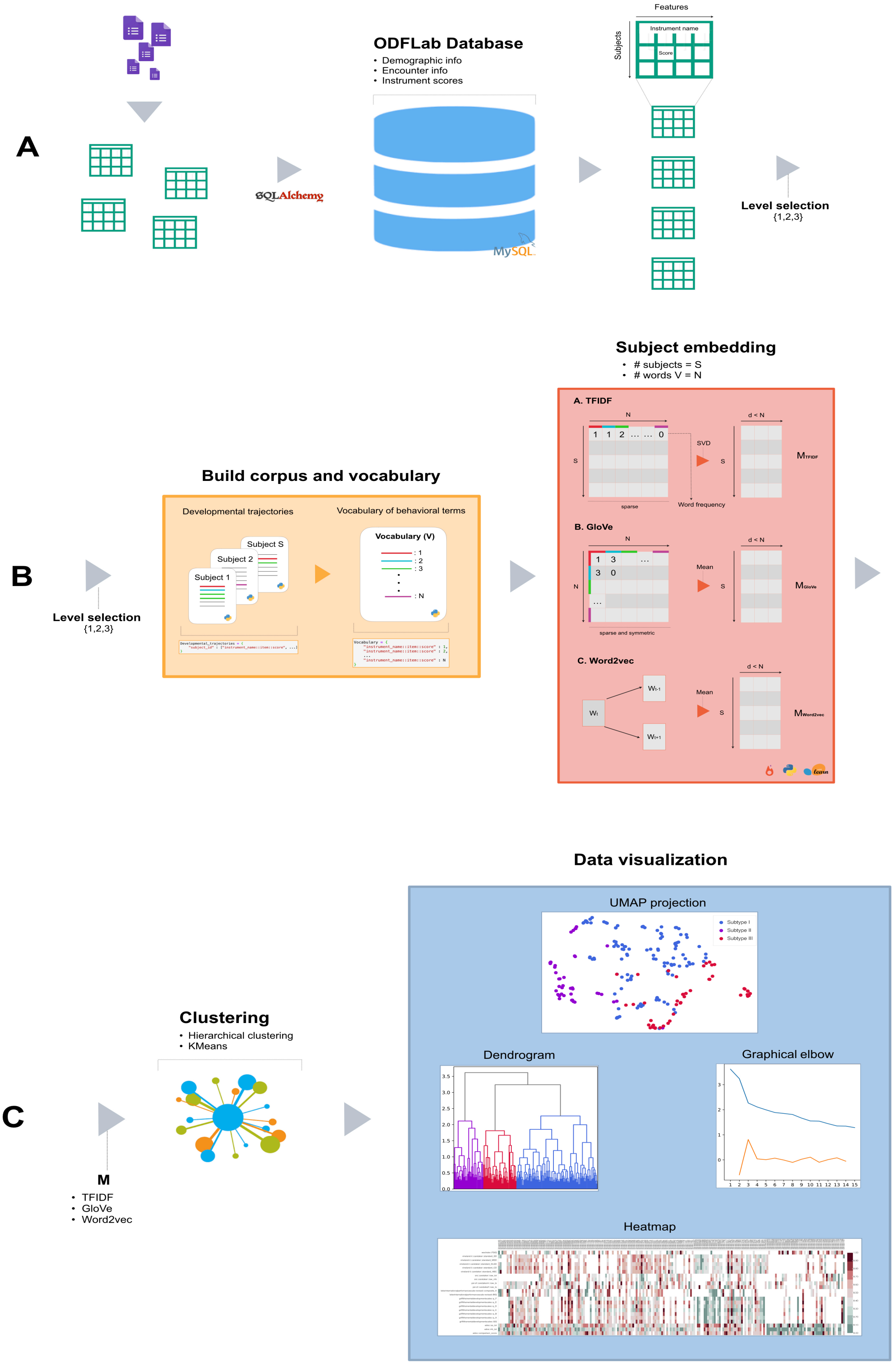
2.3. Material and methods

In the following, we describe our analysis framework for behavioral phenotyping (see Figure 2.1). First, we introduce how ASC behavioral phenotype can be assessed through psychometric instruments, both at a general and a more specific level. Then, we outline the pipeline implemented for the stratification of subjects with ASC. The modules include: 1) pre-processing steps applied to ASC behavioral data; 2) initialization of raw representations of individuals with ASC; 3) embedding methods for latent patient representations that are designed to capture the multidimensional aspects of ASC assessment and/or the progression of cognitive and behavioral characteristics throughout time; 4) clustering algorithms for stratification; 5) visualization tools; and 6) *post hoc* analyses.

2.3.1. Behavioral phenotype

Multiple depth levels

We introduce the battery of standard tests that are administered to assess behavioral phenotypes of children and adolescents with ASC and that provide features of the ODFLab dataset. We adopt a multi-scale approach considering the hierarchical structure of each instrument. At the finer detail level, the descriptions are formed by single subtests for the assessment of specific skills or behavioral aspects. Progressively, we aggregate them to form more general composite measures of latent, theoretical constructs. We enable our pipeline to select the level of detail of interest for the characterization of a behavioral profile. Based on instruments' structure, we identify three depth levels that yield to progressively more general descriptors. As an example, the cognitive profile of an individual between the ages of 6 and 16 can be assessed with the Wechsler Intelligence Scale for Children - Fourth Edition (WISC-IV) via the Full Scale Intelligence Quotient composite index (FSIQ). FSIQ summarizes the general cognitive ability of an individual, hence it can be considered as the highest level of information, i.e., Level 3. However, four composite scales contribute to the FSIQ, namely, Verbal Comprehension, Perceptual Reasoning, Working Memory, and Processing Speed (Level 2), each of which investigates different cognitive abilities. Furthermore, each composite index is obtained from multiple subtests



[Courtesy of Nicole Bussola, PhD candidate]

Figure 2.1: Behavioral phenotyping pipeline. (A) Pre-processing steps; (B) Raw subject representations and embedding methods; (C) Clustering algorithms and visualization.

that refer to the assessment of specific tasks at a finer level of analysis, i.e., Level 1. As an example, Verbal Comprehension is derived from Similarities, Vocabulary, Information, Comprehension, and Word Reasoning subtests.

In the following, we mainly consider *scaled* scores for instrument subtests, i.e., each subtest raw score is converted to a standardized scale, and *standard* scores for composite indices, that indicate the performance of an individual compared to a population norm. To describe behavioral phenotypes, we include a battery of tests that measure: 1) core symptom severity; 2) cognitive functioning; 3) adaptive behavior skills; 4) social impairment severity; and 5) level of dysfunctional interactions with child's caretaker(s) and stressful situations.

Assessment instruments

Autism Diagnostic Observation Schedule - Second Edition (ADOS-2) ADOS-2 is a clinician-administered observational assessment with two diagnostic algorithms that comprise the core behavioral domains Social Affect (SA) and Restricted and Repetitive Behaviors (RRB). SA domain measures social-communication as a lack of typical behaviors that are pervasive across social contexts. In particular, reduced use of gestures or eye contact, and reduced frequency of appropriate social responses. RRB measures atypical behavior, e.g., hand flapping, sensory examination of materials, and excessive reference to a particular topic. ADOS-2 provides a standardized continuous measure of autism spectrum symptom severity, the Calibrated Severity Score (CSS), that it is not influenced by individual characteristics such as age and language skills, and it is used to compare symptom severity across individuals of different developmental levels [115]. ADOS-2 has 5 age-related modules, with separate algorithms according to the language skill level. Toddler Module is designed for the assessment of children from 12 to 30 months of age as an early diagnostic tool; Module 1 is administered to children with no verbal skills or few verbal skills; Module 2 is constructed for children with less than, or more than 5 years of age; Module 3 is for verbally fluent children and adolescents, typically younger than 16 years; and Module 4 is for verbally fluent adolescents and adults. Although CSS can be employed as a score for univocal comparisons of symptom severity, the nature of the symptoms underlying an individual's CSS may vary greatly. In fact, the highest level of severity can be assigned to individuals that present significant social-communication impairments, but exhibit few repetitive behaviors, or vice versa. Therefore, it is critical to include SA and RRB scores in ASC studies to provide a clearer picture of individual characteristics, although an effect of child characteristics should be taken into account in the absence of calibrated scores [116]. This in-depth analysis can go further and consider a finer level of items. As an example, SA domain score for ADOS-2, Module 1, algorithm "few to no words", is the sum of scores in items *verbalizations, gestures, eye contact, facial expressions, gaze and social overture, shared enjoyment, showing, initiation of joint attention, response to joint attention*, and *quality to overtures.* Scores in Level 1 items are integers ranging from 0 to 2. In summary, Level 1 includes the items that constitute directly observed behaviors, and can vary according to modules. Level 2 corresponds to SA and RRB domains for modules Toddler, 1, 2, and 3, and Communication, Reciprocal Social Interaction (RSI), and Stereotyped Behaviors and Restricted Interests domains for Module 4. Level 3 comprises composite CSS for all modules, except for Module Toddler and 4, which report SA-RRB score, that is the sum of the two core domain scores, and composite Communication-RSI score, respectively.

Griffiths Mental Development Scales (GMDS) GMDS is a well-established instrument to measure child level of psychological and motor general development from 0 to 8 years [117]. This procedure provides a Level 2/3 General Developmental Quotient (GQ) and 6 separate Level 1 subscales that assess functioning on areas of development (i.e., Locomotor, Personal-Social, Hearing and Language, Eye and Hand Coordination, Performance, and Practical Reasoning). GMDS quotient and subscales have a theoretical mean of 100 and a standard deviation of 16 and GQ score is the mean of the scores from the six subscales. A score less than 68 is considered to be associated with child's impairment in a specific domain.

Leiter International Performance Scale - Revised (Leiter-R) Leiter-R is a nonverbal measure of cognitive functioning that can be administered to individuals between the ages of 2 and 20 years. Differently from IQ-based cognitive measurements, Leiter-R is based on fluid reasoning, which is an index less influenced by individual characteristics, such as language and education [118]. The short-form for the Visualization and Reasoning batteries comprises, at Level 1, the subtests *figure ground, form completion, sequential order (SO),* and *repeated patterns (RP).* As a whole, they compose Level 2/3 Brief IQ (BIQ), whereas SO and RP contributes to the *Fluid Reasoning* composite index that is included in all levels.

Wechsler tests Wechsler tests are individually-administered tests that measure intelligence and cognitive abilities in children, adolescents, and adults. There are different test versions according to subject's age, but they all provide a Level 3 index of general cognitive ability, the FSIQ, that results from the combination of Level 1 subtests. *Wechsler Preschool and Primary Scale of Intelligence (WPPSI)* measures the cognitive development of children between 2 and 7 years of age. Edition III (WPPSI-III) [119] includes 14 subtests (Level 1) for verbal, performance, and processing speed tasks that can vary according to two age ranges, $2.6 - 3.11$ and $4.0 - 7.3$ years. These subtests are combined in composite Verbal, Performance, General Language, and Processing Speed (only for age range $2.6 - 3.11$) indices, that form Level 2. A previous edition of WPPSI [120] only includes 12 subtests, six for Verbal composite index, and six for Performance.

For children and adolescents from 6 years to 16 years and 11 months of age, we consider the *Wechsler Intelligence Scale for Children (WISC)* editions III and IV [121, 122]. They comprise 13 and 15 subtests, respectively, at Level 1, that constitute Verbal and Performance composite indices (Level 2). Finally, *Wechsler Adult Intelligence Scale*, editions Revised (WAIS-R) and IV, is designed for adolescents and adults from 16 years of age and above. WAIS-R [123] has 11 subtests, of which 6 contribute to the Verbal scale, and 5 to the Performance scale. WAIS-IV [124] has 15 subtests that contribute to four composite indices, i.e., Verbal Comprehension, Perceptual Reasoning, Working Memory, and Processing Speed. All composite scores of Wechsler tests have mean 100 and standard deviation 15, scores between $90 - 109$ are considered average. Subtest scaled scores have mean 10 and standard deviation 3. To uniform and simplify scores between different editions and age-related tests, only subscales and composite indices in common to all Wechsler tests are considered.

Vineland Adaptive Behavior Scales - Second Edition Vineland-II [125] evaluates the adaptive behavior skills of children and adolescents up to 18 years of age. It is based on caregiver's report of observed behavior and it provides an overall composite standard score (Level 3), i.e., the Adaptive Behavior Composite. It contains 4 domains, each with 2/3 subdomains. The main domains are Communication, Daily Living Skills, Socialization, and Motor Skills. Domain standard scores are considered as Level 2, whereas scaled subdomain scores constitute Level 1. Standard composite scores have mean 100 and standard deviation 15. Standard scores < 70 are considered of interest. Scaled scores have mean 15 and standard deviation 3 (possible range $1-24$). Vineland-II scores are considered irrespective of caretaker's sex.

Parental Stress Index - Short Form (PSI-SF) PSI-SF is a 36-item self-report questionnaire of parenting stress for caretakers of children up to 12 years of age [126]. It consists of three Level 1 subscales: Parental Distress (PD), Parent-Child Dysfunctional Interaction (PCDI), and Difficult Child (DC). PD scale measures caretaker's level of distress related to conflicts with a partner, social support, and stress resulting from life restrictions due to child rearing. PCDI subscale reflects caretaker's dissatisfaction about interaction with the child and their perception of the child in comparison to other children. DC scale assesses caretaker's perception of child's regulatory abilities (e.g., temperament, defiance, non-compliance, demandingness). We consider raw scores on these domains, which range from 12 to 60. Moreover, their sum generates a Level 2/3 Total Stress score, which ranges from 36 to 180, and it is considered significant if greater than 85. Higher scores indicate higher levels of stress. Caretaker's sex is considered to account for possible differences in stress levels. This because research on stress level differences between males and females in their relationship with a child with ASC is still ongoing and requires further investigation. In fact, together with sex differences in caretaker stress levels, with females significantly more stressed than males [127], also no significant differences between stress levels [128] have been reported. Stress levels and caretaker-child interaction scores can be informative for structuring interventions and developing programs that help caretakers to cope with stress and focus on the key factors to improve child well being and their relationship with the child. Moreover, DC domain is a measure of child regulatory abilities that should be included in the behavioral characterization of an individual with ASC.

Social Responsiveness Scale (SRS) SRS identifies the presence and severity of social impairment within the autism spectrum for children of age ranging from 4 to 18 years [129]. It is a self-report questionnaire for child's caretaker(s) that comprises 5 Level 1 subscales, i.e., Social Awareness, Social Cognition, Social Communication, Social Motivation, Restricted Interests and Repetitive Behaviors. Raw subscale scores are considered, and their sum contributes to the Total raw score (Level 2/3) that quantifies the severity of social deficits in ASC. The Total score is the sum of the responses to 65 questions, scored on a Likert Scale (i.e., from 1 “not true” to 4 “almost always true”), the higher the score the more severe is the impairment in social interactions. Scores are considered irrespective of caretaker’s sex.

2.3.2. Pipeline

We introduce here the data analysis protocol (or pipeline) implemented for the analysis of the ODFLab data, adapting the modeling framework presented in Chapter 2 for the Mount Sinai health system’s data warehouse. A pipeline consists of data processing elements executed in series that perform different tasks, from data pre-processing to visualization of results. The pipeline is designed for exploratory purposes. Our aim is to derive meaningful embeddings of words and patients where test score data are available. Further, we leverage them within an unsupervised machine learning approach, viable also when labels on endpoints are lacking. In the following, we outline the steps that have been implemented (see Figure 2.1); code is available at https://github.com/landiisotta/behavioral_phenotyping.

Data pre-processing

Our implementation is designed for a relational database of behavioral data stored in structured tables that include patient scores to test items from each instrument presented in Section 2.3.1. Table rows are anonymized subject IDs, and columns store individual information that can be mandatory (i.e., sex, date of birth, and date of assessment), or optional (i.e., age), along with subtest and composite index scores. All tables are first dumped and then filtered according to the study design. In particular, either entire tables or single subjects can be dismissed. After the filtering process, we store demographics and encounter details that we derive from tables for each patient. Patient demographics include sex, date of birth, and number of encounters, i.e., clinical referrals during which all or part of the battery of tests is administered. Encounter data comprise patient sex, date of birth, and list of administered instruments with the corresponding date of assessment.

Raw individual representations

After the desired level is set, the behavioral dataset can be created. Specifically, each subject is represented by the list of tokens generated combining: 1) instrument name; 2) item name; and 3) corresponding score of every administered instrument scores. As an example, let us consider depth-level 3 and a subject p that has received two ADOS-2 Module 3 assessments, one at 8, and the other at 12 years of age. First, our model selects the *comparison score* column from the table corresponding to the diagnostic instrument. Second, it concatenates the information and returns a string of text. Let us suppose that the comparison scores for p are 4 and 6, respectively, i.e., the subject worsen their symptom severity from “low” to “moderate”. Then subject p is represented as the ordered vector:

$$(\texttt{ados::comparison\_score::4}, \texttt{ados::comparison\_score::6})$$

More generally, let L be the chosen level of depth for N administered instruments $\{I_1, ..., I_N\}$. Each instrument I_k has a set of $\{h_k\}$ indices depending on the level L, i.e.,

$$I_k = \{j_{k1}, \ldots, j_{kh_k}\},\ k = 1, \ldots, N$$

If we consider a subject p and we indicate with N_p the number of the instruments administered to them at encounter t ($N_p \leq N$), we obtain that, for each encounter, p is represented as the ordered sequence:

$$p = (w_{tj_{11}}, \ldots, w_{tj_{1h_1}}, \ldots, w_{tj_{N_p1}}, \ldots, w_{tj_{N_ph_{N_p}}})_t,\ t \in E \tag{2.1}$$

where E is the sequence of encounters in chronological order and N_p may vary respect to t. Subjects are hence represented by behavioral term sequences that vary in length. The complete set of tokens $\{w_{\cdot\cdot}\}$, for all $t,\ k$, builds up vocabulary V. Too short sequences, i.e., with $|p| < 3$, are dropped.

Embeddings

The patient embedding module enables the application of three different algorithms, i.e., Time Frequency - Inverse Document Frequency (TFIDF), Global vectors (GloVe) for word representations, and Word2vec, for the representation of subjects with ASC leveraging their behavioral sequential records.

Time Frequency - Inverse Document Frequency This approach makes use of word counts to reflect individual term co-occurrences, as introduced in Section 1.2.6. As a result, it does not take into account the temporal dimension of data, in that each behavioral term count is derived from the entire sequence irrespective of the time at which each term is collected. Applied to our framework, TFIDF aims to weight terms to reflect their importance to a subject in a cohort. Terms that appear more frequently in general are weighted less than less general words. This can be effective in the context of autism heterogeneity research in that we need to weight as more relevant the characteristics that are unique to patients, instead of the ones that are shared by the entire cohort. After weights are generated for a desired level, each subject is represented as a sparse vector of dimension equal to vocabulary size. Then, Singular Value Decomposition is applied to the TFIDF matrix obtained for dimensionality reduction (SVD-TFIDF), as described in Section 1.2.6.

GloVe [130] The Global vectors model is an unsupervised method based on statistics of word occurrences to learn meaningful word representations [130]. Let us consider the set of behavioral term sequences of all subjects as a corpus D and each patient as a sequence of terms t. Let $|V|$ be the vocabulary dimension, i.e., the number of unique terms t in corpus D. For each term $t_i \in V$ we build a count vector $(X_{ij})_j$ for $j = 1, \ldots, |V|$ that stores the number of times a term t_j occurs in the context of word t_i. The *context* is a window of fixed size k that moves over the sequences masking the words outside the context to count co-occurrences only in that subsequence. We define $X = (X_{ij})_{ij}$ as the matrix of word-word co-occurrence counts of dimension $|V| \times |V|$, where $X_i = \sum_h X_{ih}$ is the number of times any word appears in the context of word t_i. Finally, let $\mathbb{P}_{ij} = \frac{X_{ij}}{X_i}$ be the probability that a word t_j appears in the context of t_i. We introduce now the ratio of co-occurrence probabilities for words $t_i,\ t_j,\ t_h$ as $\frac{\mathbb{P}_{ih}}{\mathbb{P}_{jh}}$. This ratio is large if word t_h is related to (i.e., appears in the context of) t_i, but not t_j; it is small if t_h is related to t_j, but not t_i; and it is close to 1 if t_h is related/not related to both t_i and t_j. By construction, the ratio is able to distinguish, relevant words from irrelevant. To learn word representations we need to find the solutions to

$$F(w_i, w_j, \tilde{w}_h) = \frac{\mathbb{P}_{ih}}{\mathbb{P}_{jh}} \tag{2.2}$$

where w and $\tilde{w}$ are word and context vectors of a fixed dimension d, respectively. Each term is both considered as a target word and a context word. We require $F(\cdot)$ to 1) encode information in a linear space; 2) output a scalar; 3) be symmetric; and 4) not ill-defined. Hence, equation (2.2) becomes

$$w_i^T \tilde{w}_h + b_i + \tilde{b}_h = \log\left(1 + X_{ih}\right)$$

where $b_i,\ \tilde{b}_h$ are the bias for $w_j,\ \tilde{w}_h$, respectively. The word representations are learned via a weighted least square regression model with cost function

$$J = \sum_{i,j=1}^{|V|} f(X_{ij})(w_i^T \tilde{w}_j + b_i + \tilde{b}_j - \log\left(1 + X_{ij}\right))^2$$

where $f(\cdot)$ is a weight function that prevents the overweight of frequent co-occurrences. In particular, the weight function is defined as

$$f(x) = \begin{cases} \left(\frac{x}{x_{\max}}\right)^{\alpha} & \text{if } x < x_{\max} \\ 1 & \text{otherwise} \end{cases}$$

where $x_{\max}$ is a cut-off parameter, i.e., the maximum co-occurrence frequency that can be considered, and α a constant.

As TFIDF, also GloVe method leverages term co-occurrences. However, the context structure and the weight process, which attempt to capture word meanings according to their relative position, can be leveraged in this study to grasp the temporal information of behavioral development. Patient raw representations (2.1) are used as input to the GloVe module. First, the terms occurring in the same encounter are shuffled, as done in [23, 28], because the order in which instruments are administered is negligible within encounters and only the temporal order of encounters is of interest. Afterwords, the co-occurrence matrix is built and the model trained to return learned word representations. Finally, patient encodings are derived averaging the vectors of terms that constitute individual behavioral sequences.

Word2vec [22] The Word2vec method architectures are designed to learn distributed representations of words by neural networks. In particular, we use the *continuous Skip-gram model* that has the best performance in learning the semantic relationships between words [22]. Given terms in a vocabulary $t_1, \ldots, t_{|V|}$, the goal of the Skip-gram model is to find word embeddings that are useful to predict the surrounding words in a document D, i.e., to maximize the average log probability

$$\frac{1}{|V|}\sum_{i=1}^{|V|}\sum_{-k\leq j<k,\ j\neq 0}\log\left(p(t_{i+j}|t_i)\right) \tag{2.3}$$

where k is the context window dimension. Given a term t_I and a context term t_O, the probability in Eq. (2.3) is the Softmax function

$$p(t_O|t_I) = \frac{\exp\left(w_{t_O}^T w_{t_I}\right)}{\sum_{t=1}^{|V|}\exp\left(w_t^T w_{t_I}\right)}$$

where w_{t_I} and w_{t_O} are the input (i.e., center word) and output (i.e., context word) vector representations, respectively. The model is trained by minimizing the negative log likelihood.

This approach, in learning the semantic relationship between words in a context, is particularly suited to our purpose of learning latent representations of clinical terms able to reflect the proximity of terms occurring before or after throughout the longitudinal developmental trajectory. To apply this module, first, we shuffle the behavioral terms within encounters as previously done. Second, we fix a context dimension k and we build the center-context pairs for the prediction task. The input layer is the one-hot encoded center word, i.e., a vector of length $|V|$ with 1 at the position corresponding to the term and 0 elsewhere. The hidden layer stores the word embeddings of fixed dimension, and the output layer generates a vector of length $|V|$, which represents the probability mass function that for a perfect prediction is concentrated in the position corresponding to the context vector. After training, the embeddings learned for each term in subject sequences are averaged to output individual representations (i.e., subject embeddings).

Clustering

This module enables clustering via either hierarchical clustering algorithm or k-means. For a detailed explanation of the former with Ward's method see Section 1.2.4. Clustering with k-means tries to separate samples in k groups of equal variance minimizing the within-cluster sum-of-squares. More specifically, given $x_1, \ldots, x_n$ elements, each sample is assigned to the closest cluster, i.e., the one that minimizes

$$J = \sum_{j=1}^{k}\sum_{i=0}^{n}||x_i - \mu_j||^2 \tag{2.4}$$

where μ_j is a cluster centroid, and it is randomly initialized and then replaced by the mean of all points in the cluster. As in hierarchical clustering, k-means requires the number of clusters to be defined *a priori*. The Elbow Method (see Section 1.2.4) can also be applied in this context, and the number of clusters corresponds to the point after which the objective function in Eq. (2.4) starts decreasing in a linear fashion. More specifically, if the number of clusters increases, the quantity J tends to zero, until we reach $k = n$, i.e., number of clusters equal to the number of points. The best number of clusters is the one for which adding a cluster would result in a minimum decrease of the objective function.

Subject embeddings can be input to both clustering algorithms and the optimum number of clusters is returned along with the *Silhouette score*, a clustering performance score, defined as the average over the entire dataset $\{x_i\},\ i = 1, \ldots, n$ of:

$$s(i) = \begin{cases} 1 - \frac{a(i)}{b(i)} & \text{if } a(i) < b(i) \\ 0 & \text{if } a(i) = b(i) \\ \frac{b(i)}{a(i)} - 1 & \text{if } a(i) > b(i) \end{cases}$$

where, for each data point x_i in cluster C_i, we define $a(i)$ as the average Euclidean distance between x_i and all the other elements in the cluster, i.e., a measure of how well x_i is assigned to its cluster, and $b(i)$ as the smallest mean distance of x_i from another cluster $C_k \neq C_i$, i.e., the neighboring cluster.

$$a(i) = \frac{1}{|C_i| - 1} \sum_{j \in C_i,\ i \neq j} ||x_i - x_j||$$
$$b(i) = \min_{k \neq i} \frac{1}{|C_k|} \sum_{j \in C_k} ||x_i - x_j||$$

The average of $s(i)$, for $i = 1, \ldots, n$, becomes a measure of how well the data have been clustered. In fact, $-1 \leq s(i) \leq 1$ is close to 1, when $a(i) << b(i)$, i.e., $x(i)$ is well matched to cluster C_i, and it is close to -1 when $b(1) << a(i)$, i.e., $x(i)$ would have been better clustered if it was assigned to its neighboring cluster.

Visualization

After each subject has been assigned to their cluster, our pipeline uses a visualization module to explore possible candidates for stratification. Specifically, the following plots are produced: 1) Uniform Manifold Approximation and Projection - UMAP [34] for dimensionality reduction, that enables the visualization of the individual representations with their clustering labels; 2) Elbow Method curve, both for hierarchical clustering and k-means; 3) dendrogram for hierarchical clustering, i.e., a hierarchical tree that shows the distances at which single elements and clusters are merged during hierarchical clustering; and 4) heatmap, a table with instrument items as rows, and subject IDs and cluster labels as columns. Each table cell of the heatmap is progressively colored according to the scaled value in the corresponding instrument item. Repeated measure scores, when available, are averaged.

Post hoc comparisons

Statistical analysis module includes two parts: first we test the effect of possible confounders on clustering, in order to investigate if other variables, and not solely the learned representations, play a role in the detection of ASC subgroups. Second, we compare instrument scores, used to construct behavioral terms, between subgroups of individuals to discover possible significant differences and behaviorally characterize ASC subtypes. Confounders are *sex, age, mean age at assessment, number* *of encounters, behavioral sequence length* and *percentage of missing data.* For sex categorical variable and missing information we run chi-squared multiple comparisons in order to investigate whether sex count and available data distributions over subgroups are independent. The remaining variables are compared via pairwise two-tailed t-tests. Mean age at assessment, number of encounters, and behavioral sequence length, if significantly different between subgroups, can suggest that subject representations have failed to learn behavioral term multidimensional and temporal characteristics, but are driven by presence/absence of terms according to assessment age, and the number of terms in a sequence. For behavioral phenotype characterization instrument scores are compared through pairwise two-tailed t-tests. Repeated measure scores are averaged before comparisons and missing data, if present, are dropped.

2.3.3. Implementation details

The behavioral phenotype of children and adolescents with ASC referred to ODFLab (University of Trento, Northern Italy) is used to test the pipeline presented in Section 2.3.2. The laboratory started its activity in 2003 and since then more that $1,000$ subjects with different developmental conditions (e.g. ASCs, ADHD, Developmental Dyslexia - DD) have been referred to the Laboratory for diagnosis and/or treatment.

We first present the preliminary investigations that were done to estimate ASC data availability and the following implementation of an efficient laboratory database for ASC data. Furthermore, we describe the characteristics of the dataset for the present study and the model parameteres we use for the analyses. All modules are implemented in Python, version 3.6.8, except for post hoc analyses that are implemented in R, version 3.4.3 [131]. For embedded representations and clustering algorithms we use `scikit-learn` and `pytorch` libraries [29, 30]. Computations were run on a MacBook Pro featuring a 2.8 GHz Intel Core i7 processor.

Data availability The implementation of a structured laboratory database for the secondary use of ASC data is the result of an initial investigation to provide a reliable summary of all the available information, both as tables and text files, reported by clinicians after every assessment. We focused on subjects who had received an ASC diagnosis ($N = 283$), either from the laboratory or previous to their referral. We aimed to determine the availability of longitudinal assessments. Moreover, we tried to quantify the amount of missing information and whether this could be, to some extent, recovered. We implemented a Python script that, simply counting, for each administered instrument, the corresponding files in anonymized patient folders on the laboratory server, was able to provide a general view of the longitudinality of the data collection process, along with the available information.

The instruments we included were: 1) diagnostic (e.g., ADOS); 2) instruments for cognitive assessment; 3) caretaker-administered; and 4) observational codes. We found that only 55 subjects out of 283 had at least 2 different ADOS evaluations, already in tabular format, and 29 had at least 2 cognitive evaluation scores reported in tables. It emerged that several scores were only reported in

	Instrument	Entries
ADOS-2	MT	14
	M1	97
	M2	40
	M3	96
	M4	50
Observational codes	EAS	18
Developmental cognitive scales	GMDS	141
Nonverbal cognitive functioning	Leiter-R	92
Intelligence scales	WPPSI	1
	WPPSI-III (2.6 – 3.11)	4
	WPPSI-III (4.0 – 7.3)	23
	WISC-III	16
	WISC-IV	66
	WAIS-R	8
	WAIS-IV	21
Caretaker-administered	PSI-SF	285
	SRS	209
	Vineland-II	141

ADOS-2: Autism Diagnostic Observation Schedule - Second Edition, MT-M4: Module Toddler - Module 4; EAS: Emotional Availability Scales [132]; GMDS: Griffiths Mental Development Scales; Leiter-R: Leiter International Performance Scale - Revised; WPPSI: Wechsler Preschool and Primary Scale of Intelligence; WISC: Wechsler Intelligence Scale for Children; WAIS: Wechsler Adult Intelligence Scalel; PSI-SF: Parental Stress Index - Short Form; SRS: Social Responsiveness Scale; Vineland-II: Vineland Adaptive Behavior Scales - Second Edition

Table 2.2: ODFLab database instruments and correspondent number of entries.

the evaluation text documents, which are available for all subjects, but that were not transcribed to a computer-readable format. In conclusion, the number of repeated measures could be improved. For these reasons, we decided to start the process of missing information recovery and database construction to keep laboratory data up-to-date.

ODFLab database To manage ASC data in an user-friendly and straightforward way we created Google Form questionnaires (GForms) to collect anonymized data from children, adolescents, and adults with ASC referred to the laboratory during past years, up to now. In particular, this plan has been devised to recover data incompletely stored or missing altogether, and to store scores from new assessments. The advantages of using a GForm framework to enter patient data are several:

- It is possible to create “Required” items, i.e., the form cannot be closed until the corresponding information is provided. In this way, it is possible to minimize missing information;

- Depending on the data type, a “Response validation” can be added, with warning messages that guide the user throughout the process. The format of mandatory items can be described via regular expressions, which are, in formal language theory, a sequence of characters that define a pattern. In this way, demographic information, which is required for all instruments in clinical setting can be uniformly collected;

- Data collected via GForms can be transformed into a table, that is continuously updated and can be leveraged for secondary analysis, or as a building block of a relational database;
- When the form is divided in sections, if-statements can be implemented. This proves to be useful for reporting ADOS-2 modules that are characterized by different algorithms, depending on specific subject's characteristics (e.g., age, language level). Because each algorithm has different subtests, if algorithm A, whether than B, is selected, the user will only see the correct subtests for that dimension.

Google Spreadsheets generated by the forms populate a relational database stored on the server of the University of Trento and managed by the database management system MySQL. To interact with the database (e.g., run queries, create new tables, populate existing tables) we use the object-relational mapping `sqlalchemy` [133] for Python programming language. GForms are available for instruments listed in Table 2.2, displayed along with their current number of entries[3]. Data insertion started in June 2017 and it is still ongoing; multiple instrument scores from 255 individuals have been so far included in the database.

Dataset Of the 255 subjects in the database we drop adults (i.e., first assessment age $\geq$ 18) and children with solely observational code scores (i.e., EAS, see [132]). After filtering out 49 individuals, the study dataset contains 206 children and adolescents with scores in at least one instrument from the battery presented in Section 2.3.1. ADOS-2 Module 4 has also been dropped after filtering.

Behavioral term sequences Together with the three levels described in Section 2.3.1, we include a fourth instrument depth level (Level 4), which is a custom level that reflects the specific clinical guidelines of ODFLab and that provides an individual clinical profile for ASC evaluation, see Table 2.3. This new level also maximizes the items shared among different editions and modules of the same instruments. As an example, at Level 4 we only select the FSIQ index from all Wechsler intelligence scales (i.e., revisions and age-related versions) to directly compare individuals and maximize the number of available measurements of general intellectual ability.

In Section 2.4, we present results from the pipeline only applied to Level 4 behavioral phenotypes to detect fine-grained subgroups that are not biased by data structural heterogeneity. In fact, because results from multiple instrument revised editions and age-related versions are tied to unique indices and subtests, structural heterogeneity stems from the great number of behavioral terms that are less represented in cohort behavioral sequences (i.e., they involve fewer subjects than more general terms). Level 4 instrument depth can be a good trade-off between fine-grained individual characterization and statistically relevant patterns of terms. Moreover, because it is tailored to clinical requirements, it might provide a more useful characterization towards translational goals.

[3] Accessed on September, 2019

Instrument	Index	
ADOS-2	SA	
	RRB	
	CSS	
GMDS	A	
	B	
	C	
	E	
	F	
	GQ	
Leiter-R	Brief IQ	
	FR	
Wechsler	FSIQ	
Vineland-II	CD	
	DLS	
	SD	
	MS	
	ABC	
PSI-SF	caretaker	TS
SRS	RIRB	
	Total	

ADOS-2: Autism Diagnostic Observation Schedule - Second Edition, SA: Social Affect, RRB: Restricted and Repetitive Behaviors, CSS: Calibrated Severity Score; GMDS: Griffiths Mental Development Scales, A: Locomotor, B: Personal-Social, C: Hearing and Language, D: Eye and Hand Coordination, E: Performance, F: Practical Reasoning; Leiter-R: Leiter International Performance Scale - Revised, Brief IQ: Brief Intelligence Quotient, FR: Fluid Reasoning; FSIQ: Full Scale Intelligence Quotient; CD: Communication Domain, DLS: Daily Living Skills, SD: Socialization Domain, MS: Motor Skills, ABC: Adaptive Behavior Composite; PSI-SF: Parental Stress Index - Short Form, TS: Total Stress; SRS: Social Responsiveness Scale, RIRB: Restricted Interests and Repetitive Behaviors

Table 2.3: Level 4 features for behavioral phenotyping.

After behavioral term sequences are constructed, as described in Section 2.3.2, we drop subjects with sequences shorter than 3 terms ($N = 2$). The final cohort comprises 204 de-identified behavioral records collected from the 27th of January 2010 to the 12th of September 2019 of 35 females and 169 males, with mean age equal to 11.24 ($sd = 5.10$). The median number of encounters is 1 with range from 1 to 5. The median number of assessments, i.e., the total number of administered instruments, is 5, with minimum 1 and maximum 24. Mean age at assessment is 6.84 ($sd = 3.58$). Behavioral sequences at Level 4 have mean length 15.86 ($sd = 11.93$) and median 13.5, with minimum fixed length 3 and maximum 71. Vocabulary size is $1,186$.

Embeddings The embedding dimension is set at 10 for all models.

- *TFIDF* matrix is normalized with Euclidean norm and input to Truncated SVD function to obtain patient encodings from term co-occurrence weighted counts.
- *GloVe* model co-occurrence matrix X is constructed so that terms that are d words apart contribute $\frac{1}{d}$ to the total count. In this way, we weight less distant word pairs, that correspond to temporally distant behavioral terms, in order to account both for the developmental and the multidimensional aspects that characterize ASC trajectories. Distance d depends on the context window size, which is set at 10 words to the left and 10 to the right. The cut-off parameter x_{max} is 10 and $\alpha = \frac{3}{4}$. The model is trained using AdaGrad [134] with learning rate 0.025 for 100 epochs with batch size equal 20. GloVe method generates two sets of word embeddings, W and $\tilde{W}$, from selected words and context words, respectively. In our case, X is symmetric and hence the word embedding sets only differ as a result of random initialization. As suggested by Pennington et al. [130], we consider the sum $W + \tilde{W}$ as our word vectors. For each patient the term encodings corresponding to the terms in their sequence are averaged to create individual representations. All parameters are chosen according to the details reported in [130], except for batch size, x_{max}, and learning rate that have been reduced consistent with a smaller dataset.
- *Word2vec* model is implemented with `pytorch` library [30] and it is trained using Stochastic Gradient Descent and backpropagation with learning rate 0.001 for 150 epochs on all data with batch size 20. Window context size for context prediction is set at 10. Parameters are empirically set to reduce computational cost and account for a small dataset.

Language-inspired embeddings Each language has specific rules of grammar (i.e., syntax), which provide sentences with meaning. Moreover, each word carries its own literal meaning (i.e., denotation) and suggestive meaning (i.e., connotation). These aspects should be taken into account in multi-language embedding. However, the EHR (see Chapter 1) and behavioral sequence embeddings discussed in this Chapter are in principle not tied to a particular language; each word is rather a code than part of a specific natural language vocabulary. The sentences we consider are not tied to any rule of syntax, in that sequence order does not follow predetermined structures. On the contrary, code semantics (i.e., meaning) is more similar to words in natural languages, i.e., they have connotations and annotations.

In this body of work, we have processed normalized codes (e.g., ICD-9) according to international standards, and structured behavioral observations, leveraging instruments that have been translated into several languages with a 1-1 item-level correspondence. For this reason, the denotation of each term is independent of the language used to collect it. The processing of free-text clinical notes or psychological assessment reports would indeed require a different treatment. Code connotations, on the other hand, are inherited from the context and contribute to the general meaning of the sentence.

As such, in both applications our models try to capture the semantic relationship between terms via CNNs, word context, and term co-occurrences. We argue that term context and progression contributes to the sequence general meaning and regulate the embedded representation. As an example, the term ICD-9 code *Diarrhea (787.91)* can have different connotations. If it is next to an anti-cancer medication, such as *Bortezomib*, it can indicate a side effect. If the same ICD-9 code is found in the context of a diagnosis of *Autistic disorder, current or active state (299.00)*, it can refer to a common comorbid condition [20]. Accordingly, a low GMDS scale C score (i.e., hearing and language) and a high GMDS subscale A score (i.e., locomotor) indicates an autistic subject who has compensated the lack of communication and language with motor abilities. If, however, the same GMDS C score is accompanied by a high ADOS-2 SA score and a high RRB score, it indicates an individual affected by a severe form of autism that need very substantial care and support [115].

Clustering and visualization Most unsupervised approaches to ASC stratification leverage hierarchical clustering as clustering algorithm. We hence select the same approach, with Ward's linkage criterion and Euclidean distance. Best number of clusters is selected by the Elbow Method, with cluster number ranging from 2 to 15. In the experiments, for cluster visualizations, UMAP number of neighbors is set at 5 and minimum distance at 0.0.

Representation evaluation

The unsupervised approach and the small sample size, together with the presence of missing information, complicate the implementation of a validation framework. In particular, the lack of a groud-truth (i.e., known ASC subtypes) calls for an unsupervised approach, that does not allow to validate our models. Moreover, the small sample size hinders the possibility to test the trained unsupervised embeddings on a separate, unseen dataset. In fact, splitting our behavioral sequences into training and test, not only means that the training process would have a smaller dataset to learn representations from, but also, that the test dataset may contain words that are absent from the training dataset and that would be lost. Given the exploratory nature of our study, we rely on clinical validation to inspect the subgroups we obtain from different models. Moreover, clusters obtained from the three methods are compared with the Fowlkes-Mallows index (FMI), as described in Section 1.3.2. In addition, in order to prove the promise of our approach, we compare its performance to a baseline clustering of quantitative ASC feature data.

Method	N clusters	Subgroup numerosites	Silhouette score
SVD-TFIDF	2	$(137, 67)$	0.29
GloVe	3	$(65, 10, 67, 11, 24, 20, 7)$	0.19
Word2vec	3	$(115, 42, 47)$	0.12
Features	5	$(155, 35, 5, 1, 8)$	0.10

Table 2.4: Hierarchical clustering results for the implemented methods.

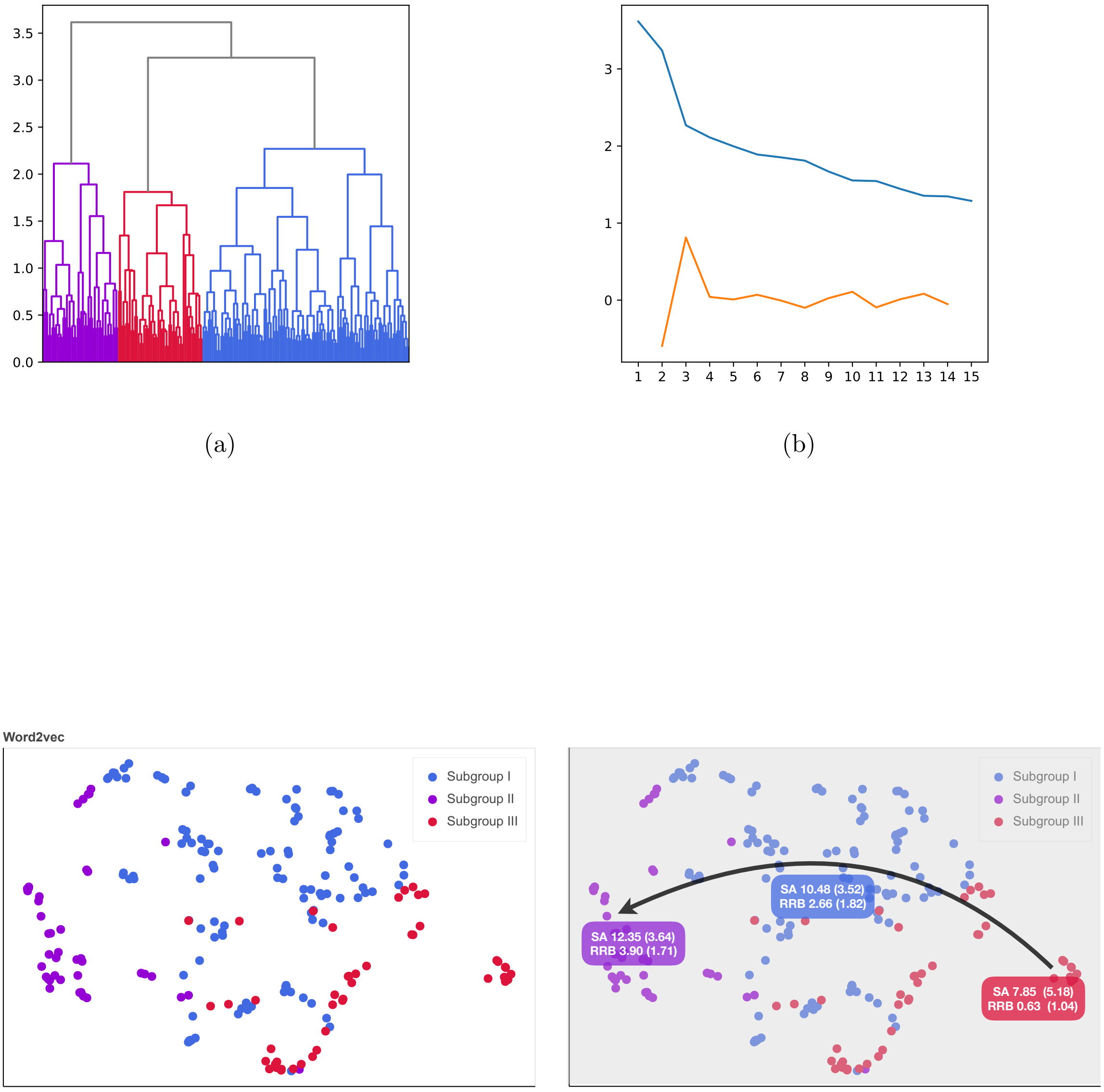
Quantitative features baseline Model baseline relies on clustering of vectors of instrument scores ordered according to the association to a particular age period. Age groups have been defined together with a clinician to reflect the developmental phases of an individual from infancy to late adolescence. Specifically, we define 5 age groups, i.e., group 1 that includes children from 0 to 2.5 years of age; group 2 with children older than 2.5 and up to 6.0 years; group 3 that corresponds to individuals older than 6.0, up to 13.0 years of age; group 4 that includes adolescents, i.e., more than 13.00 and less than 17.0 years of age; and group 5 with individuals with age greater equal than 17.0 years. For each phase, we select the set of instrument that are usually administered and we consider Level 4 index scores. The resulting dataset has subjects as rows and instrument indices as features, that are lexicographically ordered within each phase, and temporally ordered according to age periods. The resulting dataset has 204 observations and 110 features. When the same instrument was administered more than once in the same time period, the scores are averaged. Scaled scores are then clustered via hierarchical clustering.

Clinical validation and post hoc comparisons We want to investigate if behavioral embeddings can be leveraged to represent subjects with ASC and divide them into clinically relevant subgroups, compared to typical quantitative approach. To do this, we validate the groups obtained from each implemented model first ruling out the influence of possible confounders, second comparing quantitative average scores between groups. The significant clinical patterns are highlighted and compared. Multiple comparisons in post hoc analysis are corrected with Holm-Bonferroni method and level of significance is set at 0.05.

2.4. Results and discussion

In the following, we report the number of clusters found by hierarchical clustering performed on each method encodings and the quantitative feature dataset (see Table 2.4). Moreover, we describe the clinical validation of the results (see Tables 2.5-2.9). Finally, all clustering performances are compared via FMI scores (see Table 2.10) to assess consistency of the results.

2.4.1. Clinical validation



(a)

(b)

(c)

Figure 2.2: Word2vec encoding subgroups. Dendrogram (a), blue (I), purple (II), red (III); Elbow curve and second derivatives (b); and UMAP encoding projections with subgroup labels and trajectory of ADOS-2 core symptoms [Mean (sd); SA = Social Affect; RRB = Restricted and Repetitive Behaviors] (c) are displayed. In plot (b) the blue curve is the Elbow curve, i.e., the variance explained at the next iteration, and the orange curve is the second derivative curve, that is maximum in presence of inflection points.

		Word2vec			
		Subgroup I	**Subgroup II**	**Subgroup III**	
		(N=115)	(N=42)	(N=47)	p-value
	Current age	10.70 (4.82)	11.28 (6.27)	12.32 (4.66)	ns
	N encounters	1.57 (0.94)	1.43 (0.77)	1.15 (0.36)	I vs III**
	Female/Male	16/99	13/29	6/41	ns^a
	Assessment age	5.90 (3.42)	6.34 (3.14)	7.62 (3.47)	< 0.01
	Sequence length	17.03 (12.84)	17.55 (10.10)	8.66 (4.68)	III vs I/II***
	Missing values	42%	34%	65%	$< 0.001^a$
ADOS-2	CSS	6.16 (1.64)	6.54 (1.31)	4.30 (2.76)	III vs I/II***
	RRB	2.66 (1.82)	3.90 (1.71)	0.63 (1.04)	< 0.001
	SA	10.48 (3.52)	12.35 (3.64)	7.85 (5.18)	< 0.01
GMDS	A	79.82 (15.39)	63.28 (20.66)	81.73 (24.76)	II vs I/III**
	B	73.13 (17.05)	52.36 (23.02)	70.55 (30.52)	II vs I/III*
	C	69.26 (28.92)	46.97 (33.20)	68.91 (40.00)	II vs I*
	D	77.58 (18.42)	54.85 (20.10)	67.55 (29.30)	II vs I***
	E	92.12 (21.10)	67.74 (24.95)	86.27 (35.19)	II vs I/III*
	F	75.42 (23.41)	53.10 (30.11)	72.55 (34.27)	II vs I**
	GQ	77.3 (17.11)	56.41 (23.11)	73.73 (31.17)	II vs I/III*
Leiter-R	BIQ	87.25 (22.40)	68.90 (25.01)	85.86 (12.27)	I vs II*
	Fluid reasoning	89.83 (20.36)	72.00 (22.60)	83.71 (10.31)	I vs II*
PSI-SF	Total stress (F)	89.21 (24.34)	92.32 (25.82)	85.31 (20.18)	ns
	Total stress (M)	83.35 (22.07)	84.92 (19.55)	82.17 (19.46)	ns
SRS	RIRB	17.20 (7.79)	16.38 (6.84)	18.36 (8.30)	ns
	Total	92.51 (30.62)	90.43 (27.87)	105.10 (45.68)	ns
Vineland-II	ABC	70.12 (12.29)	55.02 (14.65)	62.73 (20.26)	I vs II***
	CD	70.16 (12.09)	51.93 (16.17)	64.64 (20.41)	I vs II***
	DLSD	75.57 (13.62)	59.81 (16.48)	67.27 (20.53)	I vs II***
	MSD	77.56 (10.73)	62.69 (15.91)	71.14 (11.16)	I vs II***
	SD	71.16 (12.82)	57.30 (14.11)	61.55 (16.14)	I vs II***
Wechsler	FSIQ	85.84 (19.62)	96.00 (27.45)	98.43 (21.64)	I vs III*

$^{*}p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$; ^{a} χ^2 test; ns = non significant comparisons

ADOS-2: Autism Diagnostic Observation Schedule - Second Edition, SA: Social Affect, RRB: Restricted and Repetitive Behaviors, CSS: Calibrated Severity Score; GMDS: Griffiths Mental Development Scales, A: Locomotor, B: Personal-Social, C: Hearing and Language, D: Eye and Hand Coordination, E: Performance, F: Practical Reasoning; Leiter-R: Leiter International Performance Scale - Revised, Brief IQ: Brief Intelligence Quotient; PSI-SF: Parental Stress Index - Short Form, Total Stress (F/M): female/male; SRS: Social Responsiveness Scale, RIRB: Restricted Interests and Repetitive Behaviors; ABC: Adaptive Behavior Composite; CD: Communication Domain, DLSD: Daily Living Skills Domain, MS: Motor Skills Domain, SD: Socialization Domain; FSIQ: Full Scale Intelligence Quotient

Table 2.5: Confounder and feature statistics for Word2vec encodings [Mean (sd)].

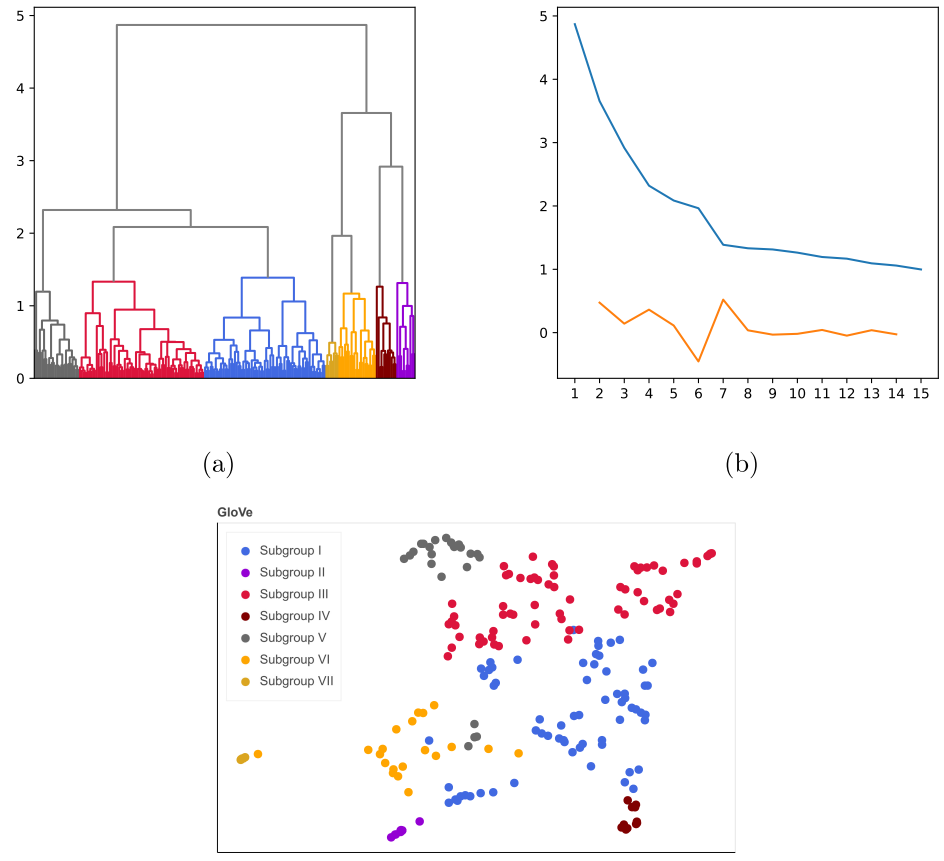
Word2vec Individual Word2vec representations are divided into three subgroups (see Figure 2.2 c). Investigating the external factors that may have influenced subgroup structure we report that subgroup III shows a significantly higher number of missing values (65%), lower sequence lengths ($M = 8.66,\ sd = 4.68,\ p < 0.001$), and lower average number of encounters ($M = 1.15,\ sd = 0.36,\ p < 0.001$) respect to the other subgroups. The identification of subgroup III can have been influenced by these factors and, more generally, by lack of information. However, this group is also characterized by individuals that have received a late diagnosis, i.e., they report a significantly higher mean age of assessment ($M = 7.62,\ sd = 3.47$) respect to the other subgroups. Hence, the investigation of the associated behavioral phenotype can lead to relevant discoveries of the cognitive and symptom severity profiles associated to delayed diagnosis, that can inform screening practices, lowering the age at which specific manifestations are detected. Investigating the male/female ratio, we observe that only 13% of individuals in subgroup III are females, whereas subgroup II reports the highest percentage (31%).

The three subgroups can be described by their symptom severity profiles, i.e., mild (subgroup III), moderate (subgroup I), and severe (subgroup II) core symptom severity (see Figure 2.2 (c) and Table 2.5). Furthermore, we observe that subgroup II includes low-functioning individuals (GMDS GQ < 70) with impairment in social communication and interaction and strongly characterized by restricted and repetitive behaviors. SA and RRB domains of ADOS-2 have significantly higher scores compared to the other two groups. Moreover, group II presents a concordance between caretaker-reported impairment in communication, daily living skills, socialization, motor skills and the developmental level reported by GMDS scales. Individuals in subgroup II are also characterized by greater nonverbal cognitive impairment (Leiter BIQ and GMDS GQ indices, respectively) compared to the other groups and significant deficits in motor skills ($M = 63.28,\ sd = 20.66$).

Subgroup I includes individuals with moderate functioning, and difficulties in social interactions and repetitive behaviors, although not as severe as those characterizing subgroup II (SA score significantly lower than subgroup II and higher than subgroup III). Moreover, they report language difficulties, i.e., GMDS Hearing and Language subscale has mean score of 69.29 ($sd = 28.92$) and it is significantly lower than the other subscale scores within the same group (two-tailed t-test comparisons, $p < 0.05$), except for practical reasoning. This suggests that these subjects activate compensatory mechanisms involving other developmental functioning dimensions to overcome verbal difficulties.

Subgroup III reports mild ASC symptom severity and higher cognitive functioning (FSIQ $M = 98.43,\ sd = 21.64$). These subjects have low RRB scores and moderate social impairment (see Table 2.5). Interestingly, 57% of subjects report a verbal Wechsler assessment, which suggests that they are characterized by greater language and communication abilities respect to the other subgroups. Moreover, they report a lower number of Vineland-II assessments (17% of subjects), which is usually not administered to caretakers of high-functioning individuals (see interactive heatmap in Supplementary Material folder).

In conclusion, we identify two subgroups (I/II) with similar ASC behavioral characteristics and different cognitive profiles, i.e., subgroup I less impaired than subgroup II (all GMDS scales comparisons are significant). The third subgroup includes high-functioning individuals who received a



(a)

(b)

(c)

Figure 2.3: GloVe encoding subgroups. Dendrogram (a), blue (I) - gold (VII); Elbow curve and second derivatives (b); and UMAP encoding projections with subgroup labels (c) are displayed.

late diagnosis and that show mild ASC symptoms. Less severe core symptoms are usually observed in high-functioning and older individuals, whom have probably developed compensatory mechanisms and camouflaging of ASC behaviors throughout development. As reported in Section 2.2, females with ASC are more likely to show intellectual disabilities respect to males, either because of protective factors, camouflaging, or male diagnostic bias. This is particularly evident for subgroups II and III, with less females included in the high-functioning group (subgroup III) than in the low-functioning subgroup II. This suggests the need for a deeper investigation of subgroup III profiles in females with ASC, possibly considering larger numerosities of females with ASC showing similar subgroup characteristics.

GloVe GloVe individual representations are distributed in seven groups (see Figure 2.3 c) that display less clear patterns than Word2vec representation groups. Mean assessment age is not significantly different between subgroups VI and VII, and sequence lengths mean differences are not significant for the group profiles we discuss in the following. This entails reliability of the clinical profiles attributable to the subgroups found. However, it is worth saying that none of the pairwise mean score comparisons between groups was significant, except for ADOS-2 scores. This is possibly due to the reduced sample

		GloVe							
		Subgroup I	**Subgroup II**	**Subgroup III**	**Subgroup IV**	**Subgroup V**	**Subgroup VI**	**Subgroup VII**	
		(N=65)	(N=10)	(N=67)	(N=11)	(N=24)	(N=20)	(N=7)	p-value
	Current age	9.81 (4.12)	14.36 (4.56)	10.35 (5.98)	12.93 (5.70)	12.79 (4.47)	13.65 (3.73)	13.40 (3.42)	*ns*
	N encounters	1.54 (0.92)	1.30 (0.48)	1.55 (0.97)	1.18 (0.40)	1.33 (0.56)	1.15 (0.37)	1.29 (0.49)	*ns*
	Female/Male	11/54	0/10	16/51	1/10	4/20	3/17	0/7	-
	Assessment age	5.41 (2.84)	10.12 (3.11)	5.89 (3.24)	8.18 (3.52)	7.36 (3.72)	9.78 (3.12)	10.65 (2.10)	all$<$ 0.001 except VI vs VII
	Sequence length	19.42 (12.96)	4.40 (1.26)	18.03 (12.08)	8.18 (3.46)	12.04 (5.58)	7.45 (2.68)	10.00 (4.04)	I vs II/IV/VI, III vs II/VI*
	Missing values	33%	81%	37%	67%	51%	69%	66%	all$<$ 0.001^a except I vs III, IV vs VI/VII, VI vs VII
ADOS-2	CSS	5.46 (1.29)	3.36 (1.96)	6.22 (2.35)	5.85 (0.55)	6.93 (1.49)	5.20 (2.67)	8.60 (0.70)	I vs II/V/VII, II vs III/V/VII, VII vs III/IV/VI*
	RRB	2.68 (1.77)	0.00 (0.00)	3.33 (2.31)	1.38 (1.26)	2.89 (1.69)	1.24 (1.00)	1.40 (0.52)	I vs II/VI, II vs III/V, III vs IV/VI, V vs VI*
	SA	10.35 (3.52)	5.73 (2.94)	11.02 (4.49)	8.92 (2.78)	11.45 (4.06)	8.39 (4.91)	13.80 (1.62)	I vs II, II vs III/V/VII*
GMDS	A	74.13 (17.84)	*NA*	76.92 (21.10)	*NA*	69.45 (17.08)	*NA*	*NA*	*ns*
	B	68.38 (20.91)	*NA*	67.00 (24.12)	*NA*	58.27 (16.86)	*NA*	*NA*	*ns*
	C	59.67 (30.19)	*NA*	68.57 (34.90)	*NA*	49.09 (27.44)	*NA*	*NA*	*ns*
	D	69.77 (19.95)	*NA*	71.35 (25.29)	*NA*	64.45 (19.98)	*NA*	*NA*	*ns*
	E	88.56 (24.07)	*NA*	83.34 (27.92)	*NA*	67.36 (20.90)	*NA*	*NA*	*ns*
	F	66.87 (25.75)	*NA*	72.32 (30.89)	*NA*	52.44 (22.01)	*NA*	*NA*	*ns*
	GQ	71.27 (20.11)	*NA*	73.05 (24.19)	*NA*	54.50 (20.05)	*NA*	*NA*	*ns*
Leiter-R	BIQ	85.28 (25.45)	*NA*	87.19 (21.33)	87.67 (22.03)	75.57 (16.07)	89.14 (33.84)	66.00 (16.43)	*ns*
	Fluid reasoning	87.16 (22.25)	*NA*	90.07 (21.18)	88.33 (18.61)	79.64 (16.95)	89.43 (26.48)	66.50 (14.48)	*ns*
PSI-SF	Total stress (F)	86.48 (22.03)	*NA*	89.94 (28.70)	94.86 (19.72)	82.53 (13.13)	99.00 (18.03)	104.00 (29.73)	*ns*
	Total stress (M)	78.02 (19.42)	*NA*	88.20 (22.13)	81.57 (19.04)	80.79 (15.62)	93.33 (24.75)	90.60 (27.90)	*ns*
SRS	RIRB	16.77 (8.23)	*NA*	16.13 (7.86)	23.06 (7.17)	19.25 (5.81)	16.28 (6.25)	19.17 (6.65)	*ns*
	Total	94.17 (36.21)	*NA*	94.30 (36.10)	115.44 (27.71)	94.40 (30.30)	82.22 (28.85)	97.17 (29.08)	*ns*
Vineland-II	ABC	64.78 (14.09)	*NA*	63.55 (15.54)	*NA*	66.40 (18.56)	*NA*	*NA*	*ns*
	CD	64.72 (14.76)	*NA*	62.57 (16.95)	*NA*	69.90 (21.08)	*NA*	*NA*	*ns*
	DLSD	73.41 (13.90)	*NA*	67.10 (18.15)	*NA*	68.90 (15.88)	*NA*	*NA*	*ns*
	MSD	75.40 (12.64)	*NA*	69.50 (15.32)	*NA*	76.33 (14.50)	*NA*	*NA*	*ns*
	SD	68.03 (14.60)	*NA*	64.49 (15.41)	*NA*	65.90 (12.79)	*NA*	*NA*	*ns*
Wechsler	FSIQ	90.83 (21.06)	99.33 (20.96)	95.57 (18.00)	98.13 (21.60)	75.83 (24.54)	92.14 (19.31)	72.83 (18.50)	*ns*

$^*p < 0.05$; ^{a} χ^2 test; ns = non significant comparisons; NA = not available

ADOS-2: Autism Diagnostic Observation Schedule - Second Edition, SA: Social Affect, RRB: Restricted and Repetitive Behaviors, CSS: Calibrated Severity Score; GMDS: Griffiths Mental Development Scales, A: Locomotor, B: Personal-Social, C: Hearing and Language, D: Eye and Hand Coordination, E: Performance, F: Practical Reasoning; Leiter-R: Leiter International Performance Scale - Revised, Brief IQ: Brief Intelligence Quotient; PSI-SF: Parental Stress Index - Short Form, Total Stress (F/M): female/male; SRS: Social Responsiveness Scale, RIRB: Restricted Interests and Repetitive Behaviors; ABC: Adaptive Behavior Composite; CD: Communication Domain, DLSD: Daily Living Skills Domain, MS: Motor Skills Domain, SD: Socialization Domain; FSIQ: Full Scale Intelligence Quotient

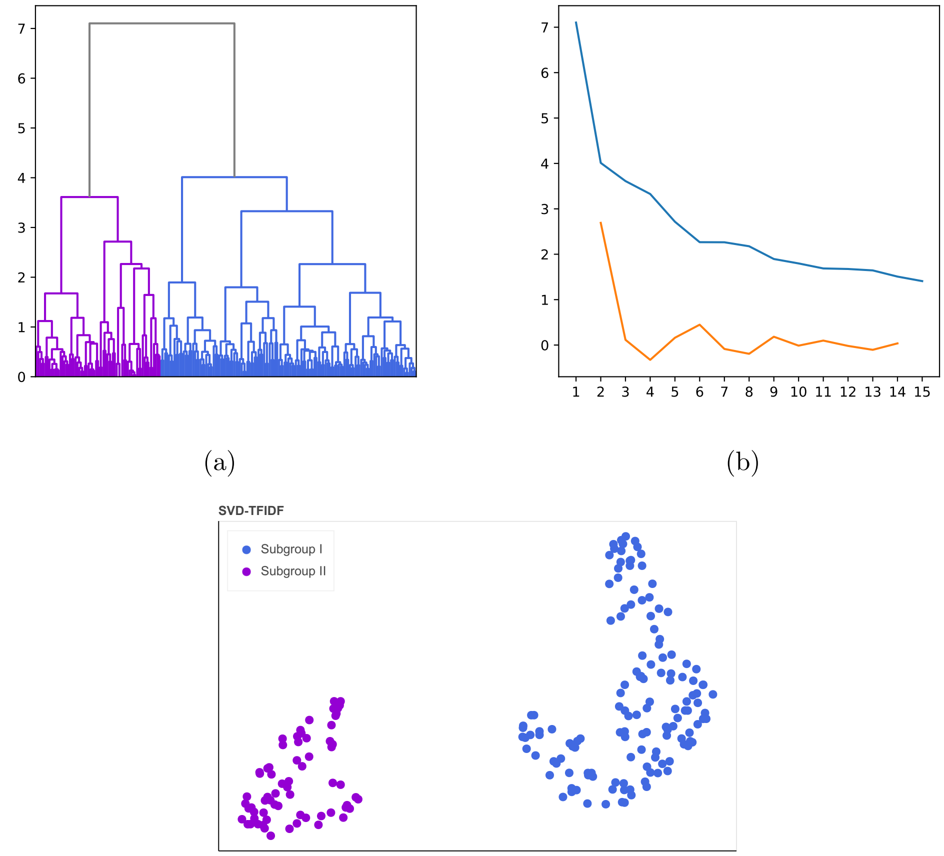
Table 2.6: Confounder and feature statistics for GloVe encodings [Mean (*sd*)].

size, however, for this reason, the following clinical validation is purely qualitative and based on the observation of mean scores reported in Table 2.6.

We find that subgroups VI and VII are characterized by late diagnoses that identify social communication/interaction impairment and reduced repetitive behaviors. On the other hand, subgroups II and IV include late diagnosed high-functioning individuals that display mild core symptoms. Finally, subgroups I, III, and V include younger subjects (see Table 2.6), with moderate ASC symptoms and heterogeneous developmental profiles (i.e., cognitive functioning and adaptive skills). An interesting aspect that emerges from subgroup investigation concerns caretaker stress levels. High level of stress are reported, by both female and male caretakers, for older subjects with mildly-impaired or within the norm cognitive levels and reduced social-communication skills. This is evident for subgroups VI and VII, where reported total stress scores are higher than all other subgroups. Specifically, subgroup VII shows a distinct social impairment (SA score $M = 13.80,\ sd = 1.62$), mildly-impaired nonverbal cognitive functioning (BIQ score $M = 66.00,\ sd = 16.43$), and total caretaker stress greater than 90 (PSI-SF cut off is set at 85). Moreover, subgroup VI includes subjects that show less social impairment (SA score $M = 8.39,\ sd = 4.91$), higher cognitive functioning (BIQ score $M = 89.14,\ sd = 33.84$), and total stress scores similar to subgroup VII (i.e., > 90.00). At the clinical level, social-communication impairment, especially in older subjects, influences caretaker stress more than the presence of intellectual disabilities or the combination of stereotypical movements and good cognitive functioning (i.e., scores > 70) as in subgroups III and V. This can be ascribed to the lack of successful interactions between caregiver and child and to the prolonged unsuccessful effort of caretakers in engaging with the child before they receive an ASC diagnosis. A correct and timely identification of such situations is key to the development of clinical practices that aim at alleviating caretaker stress and improve caretaker-child interaction. Subgroup III shows adaptive skill scores lower than 70 in all domains, whereas subgroup V reports low (i.e., < 70) GMDS scores in all subscales. Between subgroups I and III, subgroup I display less motor skills than subgroup III and similar profiles according to the other domains.

In conclusion, this result suggests the presence of further subgroups within individuals with similar core symptom and general cognitive functioning profiles (e.g., subgroups II, IV, VI, VII). Deeper characterizations of individuals with ASC by means of other clinical markers (e.g., genetic profiles, biological pathways, neural systems) can further clarify the phenotypic differences detected by the psychometric instruments. Moreover, we observe that the inclusion of caretaker aspects of child rearing in stratification studies to early identify stressful situations arising from caretaker-child interaction can be crucial to lower the diagnostic age for individuals showing ASC symptoms with high cognitive functioning.

SVD-TFIDF The SVD of TFIDF matrix produces individual representations that are separated by hierarchical clustering into two clearly distinct subgroups (see Figure 2.4 c). Observing the heatmap of instrument scores for each subject (see Supplementary Material folder) it is evident that subgroup II ($N = 67$) includes high-functioning, late diagnosed individuals. On the contrary, subgroup I ($N = 137$)



(a)

(b)

(c)

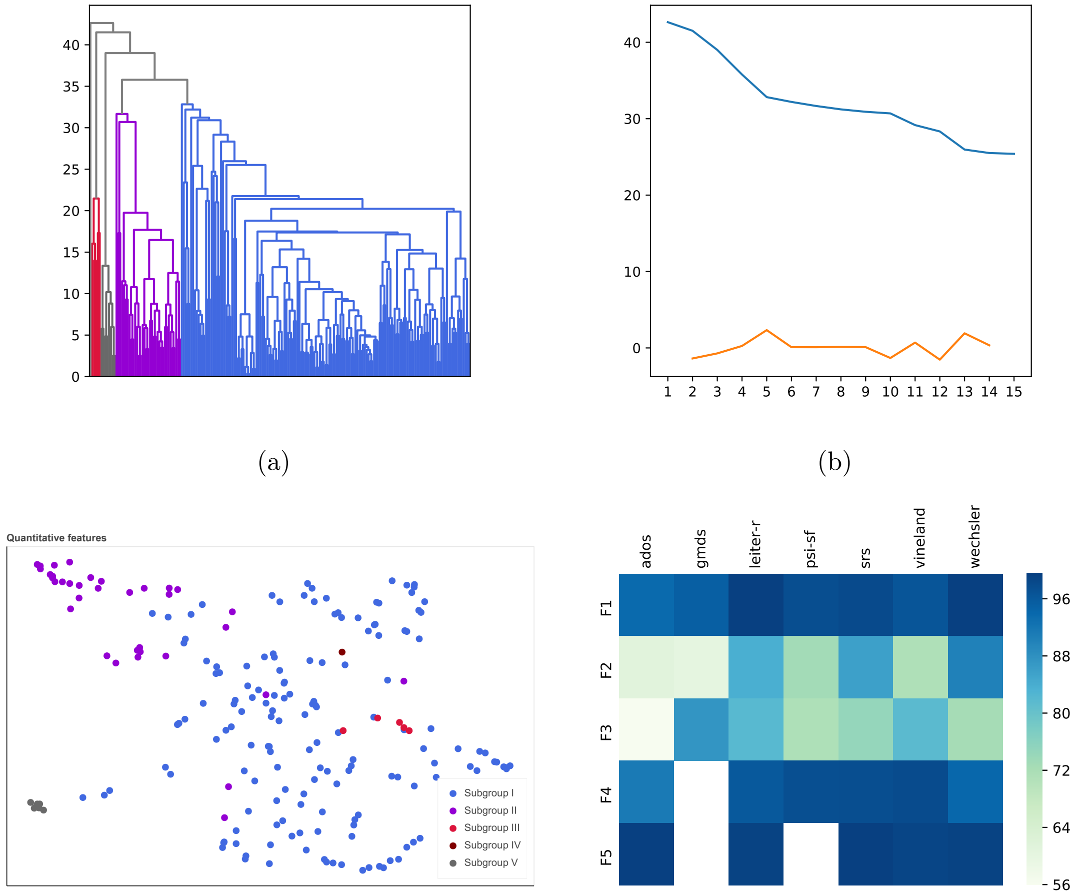
Figure 2.4: SVD-TFIDF encoding subgroups. Dendrogram (a), blue (I) and purple (II); Elbow curve and second derivatives (b); and UMAP encoding projections with subgroup labels (c) are displayed.

		SVD-TFIDF		
		Subgroup I	**Subgroup II**	
		(N=137)	(N=67)	p-value
	Current age	10.09 (5.11)	13.60 (4.22)	< 0.001
	N encounters	1.58 (0.94)	1.15 (0.36)	< 0.001
	Female/Male	29/108	6/61	*ns*
	Assessment age	5.57 (3.01)	9.91 (2.90)	< 0.001
	Sequence length	19.15 (12.00)	7.30 (2.97)	< 0.001
	Missing values	33%	70%	$< 0.001^a$
ADOS-2	CSS	6.00 (1.69)	5.55 (2.58)	0.09
	RRB	3.12 (1.92)	1.07 (1.16)	< 0.001
	SA	11.10 (3.76)	8.52 (4.43)	< 0.001
GMDS	A	75.19 (19.30)	*NA*	-
	B	66.89 (22.15)	*NA*	-
	C	62.76 (32.51)	*NA*	-
	D	70.20 (22.38)	*NA*	-
	E	84.54 (25.99)	*NA*	-
	F	68.44 (28.21)	*NA*	-
	GQ	70.90 (22.32)	*NA*	-
Leiter-R	BIQ	83.67 (23.44)	84.28 (22.80)	0.48
	Fluid reasoning	86.91 (21.83)	82.94 (18.56)	0.88
PSI-SF	Total stress (F)	90.91 (24.66)	85.57 (22.02)	0.16
	Total stress (M)	84.55 20.99	81.17 (21.16)	0.29
SRS	RIRB	16.06 (7.64)	19.11 (7.52)	0.12
	Total	90.41 (29.79)	101.47 (39.18)	0.25
Vineland-II	ABC	64.72 (15.36)	*NA*	-
	CD	63.91 (16.42)	*NA*	-
	DLSD	69.88 (16.73)	*NA*	-
	MSD	72.17 (14.40)	*NA*	-
	SD	65.97 (14.90)	*NA*	-
Wechsler	FSIQ	86.16 (20.02)	93.43 (22.10)	< 0.05

[a] χ^2 test; *NA* = not available

ADOS-2: Autism Diagnostic Observation Schedule - Second Edition, SA: Social Affect, RRB: Restricted and Repetitive Behaviors, CSS: Calibrated Severity Score; GMDS: Griffiths Mental Development Scales, A: Locomotor, B: Personal-Social, C: Hearing and Language, D: Eye and Hand Coordination, E: Performance, F: Practical Reasoning; Leiter-R: Leiter International Performance Scale - Revised, Brief IQ: Brief Intelligence Quotient; PSI-SF: Parental Stress Index - Short Form, Total Stress (F/M): female/male; SRS: Social Responsiveness Scale, RIRB: Restricted Interests and Repetitive Behaviors; ABC: Adaptive Behavior Composite; CD: Communication Domain, DLSD: Daily Living Skills Domain, MS: Motor Skills Domain, SD: Socialization Domain; FSIQ: Full Scale Intelligence Quotient

Table 2.7: Confounder and feature statistics for SVD-TFIDF matrix encodings [Mean (*sd*)].



(a)

(b)

(c)

(d)

Figure 2.5: Quantitative feature subgroups. Dendrogram (a), blue (I) - grey (V); Elbow curve and second derivatives (b); UMAP encodings projections with subgroup labels (c); and percentages of missing data in F1 - F5 time periods (d) are displayed.

includes all the remaining individuals and their corresponding behavioral profiles, with no distinctions. Subgroup I shows all levels of core symptom severity and cognitive functioning. Individuals in subgroup II report a significantly higher FSIQ than subgroup I ($M = 93.43,\ sd = 22.10,\ p < 0.05$). Moreover, subjects in subgroup II are older than subjects in subgroup I ($M = 9.91,\ sd = 2.90$) and their core symptoms are less severe (SA, RRB $p < 0.001$).

As observed for Word2vec representations, the percentage of females in subgroup II (10%) is lower than subgroup I (27%), again suggesting a possible diagnostic male bias, and/or camouflaging effect in high-functioning females with less severe ASC symptoms. Although subgroup II is consistent with a known clinical profile, all tests performed to check for possible confounders are significant (see Table 2.7). This, together with the noisy subgroup I, seems to invalidate the effectiveness of TFIDF representations to isolate meaningful subgroups.

Quantitative features Quantitative feature data are divided into 5 time periods (F1 - F5), according to subject development phases. They are processed via hierarchical clustering and return 5 highly unbalanced subgroups (see Figure 2.5 c). As displayed by the heatmap in Supplementary Material

	ADOS-2	GMDS	Leiter-R	PSI-SF	SRS	Vineland-II	Wechsler
F1	93.63 (2.08)	95.10 (1.01)	99.51 (0.00)	97.55 (0.49)	98.04 (0.00)	96.57 (0.00)	99.51
F2	61.27(0.00)	60.29 (0.00)	83.82 (0.00)	72.79 (0.74)	85.78 (0.00)	71.08 (0.00)	90.20
F3	55.88 (0.00)	87.32 (0.17)	81.86 (0.00)	71.32 (1.72)	74.51 (0.00)	81.57 (4.31)	72.55
F4	91.18 (0.00)	*NA*	96.08 (0.00)	97.30 (0.25)	97.55 (0.00)	97.84 (0.59)	94.12
F5	99.51 (0.00)	*NA*	99.02 (0.00)	*NA*	99.51 (0.00)	98.73 (0.39)	99.02

NA = not available

ADOS-2: Autism Diagnostic Observation Schedule - Second Edition; GMDS: Griffiths Mental Development Scales; Leiter-R: Leiter International Performance Scale - Revised; PSI-SF: Parental Stress Index - Short Form; SRS: Social Responsiveness Scale

Table 2.8: Percentages of missing values for quantitative feature data.

	Quantitative features				
	Subgroup I (N=155)	Subgroup II (N=35)	Subgroup III (N=5)	Subgroup V (N=8)	p-value
Current age	11.73 (5.16)	9.18 (4.56)	7.20 (0.89)	14.01 (3.84)	I vs II*
N encounters	1.36 (0.72)	1.43 (0.74)	3.80 (1.10)	1.38 (0.52)	III vs I/II/V***
Female/Male	26/129	7/28	0/5	2/6	-

$^{***}p < 0.001$

Table 2.9: Confounder statistics for quantitative feature data.

folder, we have imputed missing information with mean feature scores. Between-group comparisons of current age, number of encounters and female/male counts are reported in Table 2.9. The percentages of missing data in each time period, averaged over every instrument scales, are reported in Table 2.8. The amount of missing data, see Figure 2.5 (d) is greater than 90% for all features except the ones spanning periods F2, F3 (< 90%). Clinical profile inspection does not identify uniform behavioral profiles. This result suggests that considering quantitative features for stratification of longitudinal trajectories, either requires a great amount of data to overcome missing information, or a complete longitudinal dataset, which can be very uncommon because of the amount of time and clinical activity required to build it.

FMI scores for clustering performance comparisons are reported in Table 2.10. We expect higher FMI scores for hierarchical clustering of subject encodings derived from NLP methods, i.e., Word2vec, GloVe, and SVD-TFIDF. We expect low scores for comparisons to quantitative feature clustering. On

	SVD-TFIDF	GloVe	Word2vec
GloVe	0.51	-	-
Word2vec	0.49	0.30	-
Features	0.63	0.42	0.55

Table 2.10: Fowlkes-Mallows index scores for clustering comparisons.

the contrary, the lowest score ($FMI = 0.30$) is reported when comparing Word2vec and GloVe representation clusterings. This result suggests that pair of individuals encoded by one method are more likely to be clustered separately when represented by the other method. This can be a consequence of the different number of clusters identified, i.e., 3 and 7, respectively. Moreover, higher scores in the other comparisons may be caused by the presence of high-dimensional subgroups ($N > 100$) found both with SVD-TFIDF and quantitative feature representations. This increases the probability of finding the same pairs of individuals in the same cluster when we compare SVD-TFIDF and quantitative features to other clusterings.

In conclusion, from these results it arises that, in this context, longitudinal behavioral phenotypes are better detected by Word2vec and GloVe methods. In particular, clinically relevant and known subtypes are identified by the Word2vec method, whereas GloVe method appears to identify less clear subtypes that should be further investigated for novel behavioral and developmental phenotypes.

Despite being differently characterized, ASC subgroups found via clustering of Word2vec and GloVe representations show characteristics in common to previous studies on stratification of behavioral aspects of ASC profiles. From both Word2vec and GloVe representations we obtain age-specific subgroups of late diagnosed individuals that are characterized by mild ASC symptoms (i.e., reduced repetitive and restricted behaviors and moderate social-communication impairment) and high cognitive functioning. Moreover, we identify subgroups including younger subjects with moderate or severe symptoms and moderate/low cognitive functioning. These findings are in line with ASC subgroups identified in [106, 107, 108].

In particular, all three studies detect groups that vary according to symptom severity and cognitive functioning. Moreover, Veatch and colleagues [108] report a more severe subgroup with increasing severity in motor skills, which is comparable to subgroup I of GloVe encodings that includes subjects with impaired motor skills and moderate symptom severity, and to subgroup II of Word2vec encodings that include individuals with severe core symptoms and deficits in motor skills. Word2vec representations lead to the identification of a subgroup of moderate symptom severity (subgroup I) with language difficulties, and a subgroup of moderate social impairment and good language/communication skills. These findings are comparable to results from [107], where a subgroup of severe language deficits is identified, together with a group characterized by milder symptoms across all considered domains (e.g., nonverbal communication, social interactions, play skills), including language domain.

Results from Stevens et al. [106] differentiate 5 ASC subtypes according to: high-functioning individuals; increased language/skill domains; low-functioning individuals with impairment in all domains except for motor skills; and high skill development in play, language, academic, and motor domains, and variable skill development in social, adaptive, cognitive, and executive domains. This result is comparable to what we have obtained from Word2vec representations. In fact, subgroup I of Word2vec model is characterized by significantly higher adaptive behavior skills, although with moderate symptom severity, while subgroup II shows impaired adaptive behavior, and severe symptoms.

Finally, the higher number of females in low-functioning subgroups, respect to other groups, for both Word2vec and GloVe representations, is in line with the reports of more frequent co-occurrence of intellectual disabilities and different symptom manifestations detected in females with ASC [78]. Moreover, in this study, we observe a reduced proportion of females in late-diagnosed mild core symptom subgroups of subjects with ASC and high cognitive functioning. This supports the hypothesis of a diagnostic bias towards males or female camouflaging of less severe core deficits [80], which can both lead to the overlook of female cases when not accompanied by severe symptoms and intellectual disabilities.

2.5. Conclusions

This work has introduced a data-driven behavioral profiling tool leveraging NLP methods for stratifying individuals with ASCs. To the best of our knowledge, this approach is novel and shows promise as a way to provide personalized care and treatment, test the efficacy of an intervention, and inform diagnosis. Moreover, every behavioral term embedding learned from heterogeneous datasets can be used in transfer learning settings. For example, to initialize embedding matrices or patient encodings for other secondary uses (e.g., predictive modeling). Finally, identifying informative and more homogeneous subtypes within ASCs, leveraging both the longitudinality and the multidimensionality (e.g., cognitive, adaptive behavior, symptomatology) of clinical measures, can help framing genetic studies that, moving away from the case-control paradigm, may lead to the discovery of biomarkers whose effect can be overlooked in more heterogeneous and not stratified cohorts.

In this experiment, results from NLP methods applied to behavioral data, compared to results from the quantitative behavioral dataset, suggest that behavioral embeddings can be considered as a promising tool able to leverage all the available information from sparse longitudinal datasets. Specifically, within NLP algorithms, Word2vec and GloVe representations lead to more encouraging results than SVD-TFIDF, possibly due to the use of *context of words* for term representation learning. While TFIDF matrix stores term co-occurrences from behavioral term sequences irrespective to time, GloVe returns a word-word co-occurrence matrix built from a fixed context surrounding a term. On the other hand, Word2vec algorithm learns term embeddings from a classification task that aims at predicting the context of every word. Fixing the context dimension, as to represent an individual clinical encounter, should allow to consider the multidimensionality of terms derived from the battery of tests and provide latent representations of individuals with ASC that take into account all the aspects of a behavioral profile. Sliding the context throughout the developmental trajectory, when available, should help grasping the developmental interdependencies of terms, which shape an individual clinical profile.

Such an approach can be leveraged in the context of developmental psychology to extract latent representation of developmental behavioral phenotypes in individuals with ASCs and other neurodevelopmental conditions. An important aspect of the method presented is that the stratification process can be addressed using all the available information that is generally very sparse. Our approach is novel in the context of psychological data and behavioral phenotyping, however, a shortcoming is the small sample size of its real-world application. This may hinder the ability to detect all possible ASC subtypes, due to their underrepresentation in the cohort, furthermore, it complicates the application of other NLP techniques that leverage deep learning methods, such as LSTM recurrent networks, also applied to text data for word embedding (e.g., [135]). In the future, we plan to apply our pipeline on larger datasets and to integrate other NLP methods to derive patient behavioral encodings and compare them to the existing ones. We plan to validate models and compare performances (i.e., on the number of detected clusters, clustering performance scores, and clinical descriptions of subgroups) on new datasets *unseen* by the models (i.e., test sets) and to tune model parameters. Moreover, we plan to conduct a more rigorous investigation of the possible changes in condition severity throughout development withing each identified subgroup. We also intend to include treatment information to predict its efficacy and indicate the behavioral profile of individuals that report best intervention effects. Finally, subgroups validation should also include external information about genetic profiles, through gene expression data, and brain activity, determined by fMRI, and EEG, in order to investigate how behaviorally-detected subgroups may reflect heterogeneity at different levels, e.g., genetic, neural systems.

Chapter 3

Clinical heterogeneity in Autism Spectrum Conditions: a big data approach

Abstract

In this chapter, we investigate the generalizability of the patient representations learned in Chapter 1. We selected the clinical sequences of subjects with ASC within the Mount Sinai EHR collection (for a total of 1,439 subjects) and we stratified their latent representations to discover more homogeneous subgroups. We found three separate subgroups of individuals that differ according to age at diagnosis, clinical practices (e.g., hearing assessment), treatments (e.g., anti-seizure drugs) and comorbid conditions (e.g., constipation, convulsions). These subgroups can inform individual risk to adverse reactions, successful screening procedures, and co-occurrent condition prognosis.

3.1. Introduction

This chapter focuses on the use of de-identified Electronic Health Records (EHRs), presented in Section 1.2.1, for the stratification of Autism Spectrum Conditions (ASCs). We have extensively discussed ASCs in Sections 2.1 and 2.2, where supervised and unsupervised learning methods for classification and stratification studies were introduced. Here, we further develop the concept of ASC heterogeneity at the genetic, behavioral, and neurobiological levels. Making sense of ASC heterogeneity is key to 1) provide early diagnosis, through early sign detection; 2) identify the characteristics that may present differently in females and males to improve the diagnostic process; 3) develop more personalized and effective treatments; and 4) discover high-impact biomarkers from more homogeneous cohorts of individuals.

Currently, ASC diagnosis relies on behavioral observation combined with clinical judgment and, when not concomitant with severe comorbidities, e.g., Intellectual Disabilities (IDs), it might be overlooked (e.g., males bias and female camouflaging effect), or delayed (e.g., late-diagnosed subgroups as identified in Chapter 2). Heterogeneity throughout development results in distinct developmental profiles, which complicate the implementation of effective treatements [95]. The investigation of clinical practices in tertiary care can be useful to identify the clinical characteristics shown and the treatments received by subjects with ASC within health facilities. The clinical features extracted and their developmental patterns can be used to identify relevant subgroups of individuals showing similar profiles and help practitioners in their challenging management process [136].

EHRs, with their *structured* and *unstructured* records provide an invaluable source of information about individual health histories and patient interaction with the healthcare system. In the context of ASCs, diagnostic codes (i.e., ICD-9 codes) and clinical notes (i.e., unstructured records) within an EHR have been leveraged to detect subgroups of individuals with ASC according to their comorbid conditions [20, 137], and automate the extraction of diagnostic criteria and individual classification [138]. In the work of Doshi-Velez and colleagues [20], ICD-9 codes are aggregated into 45 categories with at least 5% prevalence in a sample of $4,927$ individuals with ASC from Boston Children's Hospital. Individual's age ranges from birth to 15 years and their clinical record sequence is represented as a co-occurrence vector divided into 30 6-month windows. These vectors are clustered via hierarchical clustering (Euclidean distance, Ward's method) and three high-morbidity subgroups and one with no significantly elevated comorbidities are obtained. Comorbidities from the three high-morbidity groups are then investigated. One subgroup is characterized by seizures, gastrointestinal (GI) conditions, cerebral palsy, and disorders of the visual pathways. Subgroup 2 includes GI disorders, disorders and infection of the ear, cardiac disorders, and congenital anomalies. Subgroup 3 is characterized by psychiatric disorders (e.g., episodic mood disorders, bipolar disorder, depression, anxiety dissociative and somatic symptom disorders, conduct disorders, and hyperkinetic syndrome of childhood). Similarly, Lingren et al. [137] have developed an automated cohort selection algorithm that leverages ICD-9 codes and clinical notes to more clearly select a sample of individuals who not only received an ASC diagnosis, but also met the clinical diagnostic standards. Patient count vectors of aggregated ICD-9 codes are clustered via k-means clustering and result in $3-4$ subgroups from multiple site datasets (e.g., Boston Children's Hospital, Cincinnati Children's Hospital Medical Center). As in the previous study, the groups identified show similar dominant categories: 1) psychiatric problems (i.e., anxiety disorder, hyperkinetic syndrome, obsessive compulsive disorder, and depression); 2) developmental disorders (i.e., dyslexia, lack of coordination, disorders of ear, skin); and 3) epilepsy and recurrent seizures. In another study [138], Word2vec Skip-gram model (see Section 2.3.2) is used to create an automated lexicon for diagnostic criteria extraction and classify individuals with ASC from EHR clinical notes.

In these studies, engineered (i.e., aggregation into categories) ICD-9 codes and diagnostic clinical notes are used, and vectors of term counts are used to represent patients. Monitoring of condition progression is only addressed in [20] by considering the co-occurrence of terms restricted to subsequences respect to time (i.e., 6-month windows). The examination of comorbid conditions that are significantly more likely in individuals with ASC than in the general population is key to the investigation of whether there are actual subtypes tied to distinct etiologies, with different genetic and environmental contributions. Reducing EHR histories to engineered diagnostic codes and limit the investigation to comorbidity detection tasks can help overcome structural problems. EHR sequences are noisy and redundant, in particular, patients medical histories include generic terms such as *Virus* among medications, or *Counseling not otherwise specified (`V65.40`)* among ICD-9 codes. Moreover, redundancy not only stems from the repetition of codes at every encounter (e.g., diagnostic codes), but also from terms referring to the same practice that are coded differently; e.g., *Lisdexamfetamine* is a medication that is used to treat Attention Deficit and Hyperactivity Disorder (ADHD), which can also be reported in EHRs with its correspondent brand name *Vyvanse*. Nonetheless, structured codes can 1) inform a patient status at a finer level, e.g., among gastrointestinal problems we can distinguish *Nausea (787.02), Vomiting alone (787.03)*, i.e., without nausea, and *Persistent vomiting* (536.2); 2) be combined to assess the risk for adverse drug reactions, or better explain a clinical screening process, e.g., presence of *Audiometry* procedure and *Expressive language disorder* in subjects with ASC indicate a common screening procedure in children with ASC [139].

This analysis is presented as a case study, where we leverage the latent patient representations of individuals with ASC to stratify patient medical records. We select individuals from Fold-2 test set EHRs (see Chapter 1) of the Mount Sinai Health System's data warehouse (MSDW). These records include ICD-9 codes, medications, laboratory tests, and Current Procedural Terminologies (CPTs). The encodings selected are clustered via hierarchical clustering and clinical validation of the results is performed. We aim to show that the model, trained on a domain-free dataset, successfully learns informative encodings from EHR trajectories of individuals with ASC. We process $1,439$ subjects who have received an ASC diagnosis and we identify three separate subgroups. We obtain results comparable to [20, 137] in terms of comorbid conditions with less engineered features (no aggregation or time frame definition). Moreover, adding medication, laboratory test and procedure codes, we are able to better contextualize individual health status (e.g., medication side effects), and also gain a general overview of the tertiary care diagnostic process and intervention (e.g., ADHD treatements and screening procedures). Finally, we have considered the complete EHR sequences in the analyses to leverage all the available information.

The distinction of subjects with ASC from an urban tertiary care service according to comorbidities, treatments, and drug adverse reaction risk can inform the way autism is recognized in a specific context and support clinician in the diagnostic and management process towards the development of personalized treatements and more effective clinical practices. Clinical findings from stratification studies in autism research should be integrated with findings from different research fields (e.g., genetics, neurobiology) to contribute to the uncovering of ASC etiology [4].

3.2. Autism Spectrum Condition records

We select individuals with ASC from the $\sim 1.6M$ MSDW records presented in Section 1.2.5. We identify patients that include in their EHR histories ICD-9 diagnostic codes related to ASC and also report corresponding SNOMED Clinical Term identifiers. In particular, among ICD-9 terms, we check for: *Autistic disorder, current or active/residual state (299.00, 299.01)*, *Other specified pervasive developmental disorders (299.80, 299.81)*, and *Unspecified pervasive developmental disorders (299.90, 299.91)*. Among SNOMED codes we check for *Active infantile autism, Residual infantile autism,* and *Pervasive developmental disorder* codes. We obtain $2,908$ individuals with both diagnostic concepts in the entire dataset, of which 1,439 are included in Fold-2 test set (see Section 1.2.5). Number of females in the cohort is 443 and number of males is 996. Mean age at 2015 is 21.47 ($min/max = [7.00; 40.00]$) years, and mean age at diagnosis is 17.39 ($sd = 16.40$) years. Of the individuals included, 97% report *Autistic disorder, current or active state (299.00)* among ICD-9 terms. The remaining subjects report diagnostic terms that are no longer included in the Diagnostic and Statistical Manual of Mental Disorders, Fifth Edition (DSM-5) [67] description of autism spectrum. In particular, Pervasive Developmental Disorders - Not Otherwise Specified (PDD-NOS) is a term used in both ICD-9 and more recent ICD-10 editions[1] and refers to children that show impairments in socio-pragmatic communication and absent repetitive and restricted behaviors. In line with DSM-5 manual, which introduces the term Autism Spectrum Disorder, the new ICD-11 edition will eliminate PDD-NOS diagnosis[2]. Moreover, *Other specified pervasive developmental disorders* correspond to Asperger's disorder and SNOMED concept *Residual infantile autism*, which is introduced in and limited to DSM-III, encompasses individuals who once met criteria for infantile autism, but no longer do so [140]. This entails that only a fraction of subjects with PDD-NOS will meet the diagnostic criteria of autism and receive an autism diagnosis. Because we include adults in the cohort, who may have received a diagnosis before the release of DSM-5 guidelines, we have decided to include all terms in the selection process to avoid the elimination of subjects who could be otherwise diagnosed with ASD according to DSM-5. We shorten the encoded sequences at the time of the first diagnosis, as done in Section 1.3.2, to reduce the noise and we average the subsequences into a unique latent vector for each individual. These vectors are then input to hierarchical clustering (with Euclidean distance and Ward's method) and the best number of clusters is selected via the Elbow method. We check differences between groups for confounders, i.e., age at diagnosis, current age, and sequence lengths, via pairwise t-tests with Bonferroni correction. The distributions between subgroups of male/female counts and diagnostic code frequencies as clinical descriptions are investigated via pairwise chi-squared tests with Bonferroni correction.

3.3. Results

We obtain three distinct clusters of individuals (see Figure 3.1) in which a high percentage of patients $\geq 95\%$ shows a diagnosis of *Autistic disorder, current or active state (299.00)*, see Table 3.1. Interestingly, *Autistic disorder, residual state (299.01)* is more common for subgroups that include older individuals, i.e., subgroups II/III, whereas PDD-NOS diagnosis is more frequent in children, i.e., individuals in subgroup I. Residual state autistic disorder converts to *Autistic disorder (F84.0)* in ICD-10 version. This suggests that clinicians might be more prone to overlook ASC symptoms in older individuals, possibly because of the comensatory mechanisms developed during adolescence and adulthood. Moreover, PDD-NOS diagnosis for children can suggest a difficulty in the early assessment of ASC symptoms. We do not find individuals with the diagnostic code corresponding to Asperger's disorder.

[1] https://www.who.int/classifications/icd/en/bluebook.pdf (Accessed on October 20, 2019)

[2] https://icd.who.int/browse11/l-m/en (Accessed on October 20, 2019)

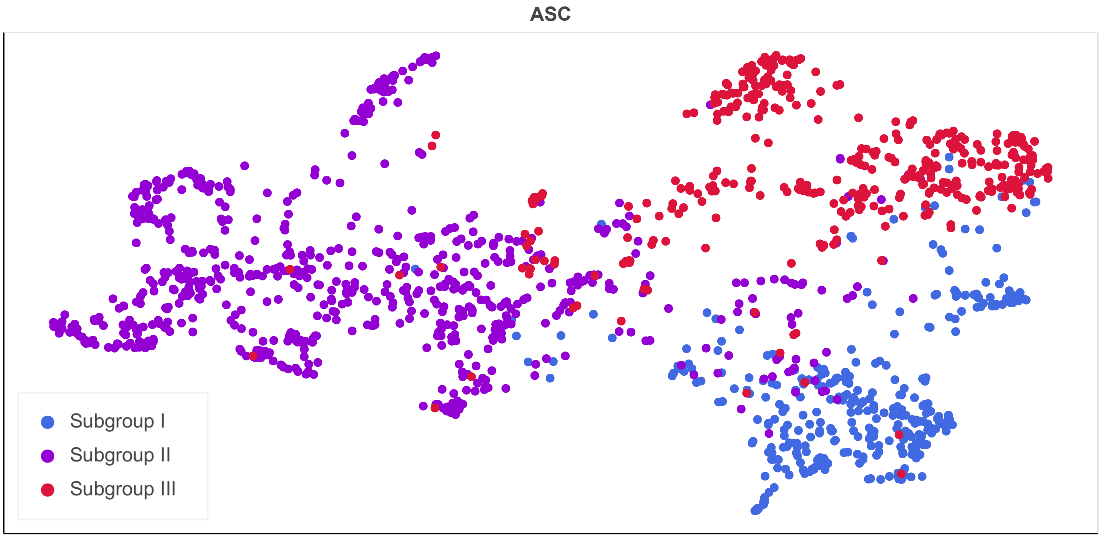


Figure 3.1: ASC subgroups of 1,439 individual EHRs from the Mount Sinai Health System's data warehouse.

	MSDW ASC cohort			
	Subgroup I	Subgroup II	Subgroup III	
	(N=346)	(N=717)	(N=376)	p-value
Age at diagnosis^2	4.12 [2.30; 7.69]	16.68 [8.61; 35.10]	10.85 [5.04; 17.94]	< 0.001
Current age^1	11.67 (7.65)	28.18 (17.89)	17.70 (13.33)	< 0.001
Sequence length^2	64.00 [5; 2, 112]	128.00 [9; 4, 625]	18.00 [3; 992]	< 0.001
Female/Male	68/278	276/441	99/277	II vs I/III^{***a}
Autistic disorder, current or active state (299.00)	0.98	0.96	0.98	ns^a
Autistic disorder, residual state (299.01)	0.46	0.62	0.57	I vs II/III^{*a}
PDD-NOS (299.90)	0.22	0.11	0.11	I vs II/III^{***a}

^{1} Mean (standard deviation); ^{2} Median [minimum; maximum]; $^*p < 0.05$; $^{***}p < 0.001$; ^{a} χ^2 test; ns = non significant

PDD-NOS = Pervasive developmental disorder, not otherwise specified

Table 3.1: Characteristics of ASC subgroups. Two-sided multiple pairwise t-tests and chi-squared tests with Bonferroni correction are performed for subgroup comparisons.

	Most frequent disorders			
	Subgroup I	Subgroup II	Subgroup III	
Constipation (564.00)	0.18	0.26	0.13	II vs I/III*
Anxiety state (300.00)	0.08	0.25	0.25	I vs II/III***
Diarrhea (787.91)	0.32	0.23	0.07	< 0.05
Generalized anxiety disorder (300.02)	0.07	0.21	0.24	I vs II/III***
ADHD (314.01)	0.11	0.10	0.20	III vs I/II**
Developmental speech or language disorder (315.39)	0.31	0.11	0.18	< 0.01
Malaise and fatigue (780.79)	0.11	0.20	0.13	II vs I/III*
Hallucinations (780.1)	0.02	0.07	0.11	I vs II/III**
Asthma, unspecified type, unspecified (493.90)	0.41	0.19	0.07	< 0.001

$^*p < 0.05$; $^{**}p < 0.01$; $^{***}p < 0.001$; ADHD = Attention deficit hyperactivity disorder

Table 3.2: Percentage of comorbidities in ASC subgroups. Multiple pairwise chi-squared with Bonferroni correction are performed for subgroup comparisons.

Subgroup I includes youngest patients at the time of diagnosis ($M = 6.34$, $min/max = [2.30; 7.69]$) and medium sequence lengths. Individuals in subgroup II are adults (mean age at diagnosis $M = 24.29$, $min/max = [8.61; 35.10]$) with significantly longer sequence lengths. Subgroup III clusters adolescents (mean age at diagnosis $M = 14.40$, $min/max = [5.04; 17.94]$) with shortest sequences ($M = 54.74$, $min/max = [3; 992]$). Subgroup II includes 64% of females from the entire cohort, that corresponds to the 38% of subjects in the group. Results from the comparison of age at diagnosis, current age, and sequence length are significant (Table 3.1). This can be explained by the presence in the EHRs of medical terms that are specifically age related, e.g., *Visual reinforcement audiometry* procedure, which is specifically designed to assess hearing in infants from 6 months to 3 years. Moreover, also the list of medications differs between children and adults, e.g., *Oxycodone*, an opioid medication used for treatement of moderate to severe pain, is tipically administered to adults. As expected, 95% of oxycodone terms in the dataset is clustered in subgroup II. The identification of age-related subgroups is similar to what we report in Section 2.4 when describing behaviorally defined subgroups. In that case, we identify subgroups of high-functioning adults. Mean sequence lengths between groups, although significantly different, do not vary according to age, with shorter sequences for younger subjects, as we may have expected. On the contrary, sequence lengths for subgroup I have mean equal 124.05, $min/max = [5; 2,112]$ terms, whereas they are shorter for subgroup III. These considerations support the possibility that sequences of terms are properly represented and that this result should be read within a neurodevelopmental framework, which is reflected in the presence/absence of specific clinical terms.

Subgroup validation consists in reporting the most frequent terms among ICD-9 codes, medications, laboratory tests and CPTs and test if they distribute independently among groups through chi-squared tests with Bonferroni correction. First, we report most frequent terms for the entire dataset (see Table 3.2). We observe that subgroup I is characterized by GI problems from the presence of *Diarrhea (787.91)* diagnosis, language impairment, and respiratory problems, from the presence of *Asthma (493.90)* diagnosis. Subgroup II reports a high percentage of individuals with *Constipation (564.00)*, and anxiety disorders (*Anxiety (300.00, 300.02)* terms). Subgroup III shares with subgroup II significantly high percentages of individuals with anxiety. Moreover, 20% of patients have ADHD as a comorbid condition, and 11% suffers from *Hallucinations (780.1)*. Anxiety state, detected in both subgroup II and III, is common in individuals with ASC and may contibute to aggressive, explosive, or self-injurious behavior. Anxiety treatment includes pharmacotherapy, cognitive behavioral therapy, or accommodation to address sensory sensitivities. Maladaptive behaviors in children with ASC may occur in response to anxiety or frustration, e.g., from core symptoms or GI problems [141], that might be true for subgroup II, where anxiety state is concomitant with constipation.

We now validate the single subgroups at the clinical level and then compare their comorbidities to the those from the literature. In Tables 3.3-3.6, we list most frequent clinical terms (with a maximum of 10) that are statistically significant at the $\alpha = 0.05$ level for each subgroup after Bonferroni correction. From the term lists we dropped uninformative concepts, such as diagnostic and procedural general terms, e.g., *General medical examination for administrative purposes (V70.3)*, *Unlisted special service,* *procedure or report.* Moreover, we drop terms related to patient anamnesis (i.e., medical history), e.g., vaccinations received in the context of standard immunization protocols of children and adolescents, because not informative for the clinical ASC profiles. Furthermore, also too rare terms ($< 1\%$) are eliminated, e.g., *Arylsulfatase* laboratory test.

Subgroup I Individuals grouped in cluster I show common diseases of childhood, along with ASC-related conditions and medical procedures that point to communication disorders and sensory differences. In particular, we identify *Otitis media (382.9)*, *Throat pain (784.1)*, *Dermatitis (691.8)*, *Asthma (493.92)*, and *Mumps (072.9)* as common diseases of childhood. We also find associated clinical terms, such as *Inhalation spacing device* among asthma medications, or *Cetirizine* to treat dermatitis. Asthma is one of the most common chronic diseases of childhood. In the United States, over six million individuals less than 18 years of age have asthma[3]. Moreover, allergic disease is associated with the development, severity, and persistence of asthma. Up to 80% of children with atopic dermatitis develop asthma and/or allergic rhinitis later in childhood [142]. This characterization is supported by the presence of allergen-specific IgE tests (see Table 3.5).

Among ASC-related terms, we find *Chronic gingivitis (523.10)* and *Expressive language disorder (315.31).* Atypical sensory behaviors/aberrant sensory perception is common in children with ASC [143]. They include preoccupation with sensory features of objects, over- or underresponsiveness to environmental stimuli, or paradoxical responses to sensory stimuli. This can hinder the development of self-care skills, including dental hygiene. The presence of periodontal disease and CPT terms *Fluoride, Teeth radiologic examination*, and *Dentoalveolar structure*, suggest the presence of sensory differences, which require the implementation of strategies to address daily-living routines. The presence of communication disorders and hearing tests (e.g., *Speech audiometry*) is consistent with the practice of performing an audiologic evaluation on children with language delays [139]. Language development is tied to social/relational aspects. In typical infants, gestural communication is followed by verbal communication. Between the age of six and twelve months, children consistently respond to their names. Individuals with impaired language development may fail in the task of responding appropriately to external verbal stimuli. Within early signs of autism surveillance, hearing is assessed to rule out a possible cause of impaired communication skills.

Interestingly, *Lead, capillary blood* test (see Table 3.5) also refers to tests of potential treatable etiologies for developmental problems. In general, neurologic effects of lead poisoning in children include neurobehavioral deficits. Also low-level lead poisoning may affect the cognitive and behavioral development of children [144]. Furthermore, within atypical sensory differences, children with ASC may develop pica (i.e., repeated eating of nonfood substances), or a preoccupation with licking nonfood items, and therefore may be at increased risk for lead poisoning [145]. Eating disorders may also be the cause of GI problems, such as diarrhea, which affects at a higher percentage individuals in subgroup I.

[3] http://www.cdc.gov/asthma/most_recent_data.htm (Accessed on October 21, 2019)

Subgroup II Subgroup II includes patients with GI problems, intellectual disabilities (IDs), and cardiovascular symptoms. IDs affect $\sim 45\%$ of individuals with ASC [4], whereas among GI problems, common symptoms include *Constipation (564.00)*, that we find for 26% of subjects in this subgroup (see Table 3.3). In Table 3.4, we find drugs corresponding to constipation treatments (e.g., *Docusate sodium*). Among frequent terms in subgroup II, we also find *Convulsions (780.39)* term, which refers to single seizure episodes and not necessarily to epilepsy. However, some comorbid conditions, such as epilepsy, may develop later during adolescence or adulthood [4] and this can be consistent with seizure epileptic manifestations in subgroup II. Among medication terms, anti-seizure medication *Lorazepam* is administered, which can also be used to treat anxiety. Hallucinations, detected in 7% of subjects in subgroup II, can be an adverse reaction to the anti-seizure drug. Among cardiovascular problems we find *Hypertension (401.9)* and *Chest pain (786.50)* and related treatments, i.e., *Magnesium, Lidocaine*, that are administered to treat ventricular arrhythmias and abnormal heart rhythm. Constipation, but, most importantly, cardiovascular symptoms, may be related to antipsychotic drugs that can be administered to individuals with ASC to treat irritability, including aggression, self-injurious behavior, and quickly changing moods [146]. Among most frequent administered drugs, not displayed in Table 3.4, we find that antipsychotic *Risperidone* is administered to 11% of patients in subgroup II, accounting for the 73% of all terms in the cohort.

Subgroup III Individuals in subgroup III are characterized by the co-occurrence of ADHD (20%) and medication-associated side effects. Together with ADHD, *Attention deficit with no hyperactivity (314.00)* is diagnosed to 9% of individuals in subgroup III and it accounts for 52% of all diagnostic terms. ADHD comorbidity is found in $30-50\%$ individuals with ASC and, in such cases, it has been observed an increased risk for greater severity of psychosocial problems. Children with ASC aged $4-8$ years, whose caregivers report significant symptoms of ADHD, show lower cognitive functioning, more severe social impairment, and greater delays in adaptive functioning than children with ASD-only [147]. Among medications, we find an alpha-2-adrenergic agonist, i.e., *Guanfacine*, and *Lisdexamfetamine* stimulant. According to the US clinical guidelines for ADHD treatment [148], effects of nonstimulants on core ADHD symptoms are weaker than stimulant medications. Guanfancine is usually reserved for children and adolescents who respond poorly to a trial of stimulants. Given their recent use, their efficacy have been less assessed and they have not been approved for use in preschool-aged children. On the other hand, stimulants are most commonly prescribed for ADHD treatments. Among amphetamines, lisdexamfetamine can be administered to children from 6 years of age and prove to be effective in lowering core symptom severity [149]. However, registered adverse effects include, among others, abdominal pain, and, more rarely, hallucinations. Both conditions are found among ICD-9 codes in subgroup III. Another characteristic that is observed with more frequency in individuals with ASC and may be caused by medication-associated side effects is obesity, that we find coded as *Overweight (278.02)* in subgroup III, see Table 3.3.

ICD-9[1]		
Subgroup I	**Subgroup II**	**Subgroup III**
Chronic gingivitis, plaque induced (523.10) - 16% (76%)	Constipation (564.00) - 26% (62%)	ADHD (314.01) - 20% (40%)
Otitis media (382.9) - 16% (64%)	Malaise and fatigue (780.79) - 20% (62%)	Unspecified fall (E888.9) - 9% (17%)
Throat pain (784.1) - 16% (46%)	Nausea (787.02) - 19% (68%)	Attention deficit with no hyperactivity (314.00) - 9% (52%)
Dermatitis (691.8) - 14% (71%)	Postprocedural hypertension (997.91) - 18% (82%)	Vomiting (787.03) - 6% (9%)
Asthma (493.92) - 11% (67%)	Hypertension (401.9) - 17% (96%)	Persistent vomiting (536.2) - 6% (10%)
Mumps (072.9) - 11% (70%)	Chest pain (786.50) - 14% (71%)	Shortness of breath (786.05) - 5% (10%)
Hearing loss (389.9) - 10% (39%)	Intellectual disabilities (319) - 13% (80%)	Abdominal pain (789.00) - 5% (9%)
Nocturnal enuresis (788.36) - 10% (44%)	Esophageal reflux (530.81) - 13% (70%)	Overweight (278.02) - 3% (13%)
Impacted cerumen (380.4) - 8%(49%)	Convulsions (780.39) - 11% (70%)	Edema (782.3) - 3% (5%)
Expressive language disorder (315.31) - 8% (56%)	Dizziness and giddiness (780.4) - 10% (69%)	Disorder of parathyroid gland (252.9) - 3% (91%)

[1] in-group and (total) percentages; ADHD = Attention deficit hyperactivity disorder

Table 3.3: Most frequent ICD-9 terms in ASC subgroups.

Medication[1]		
Subgroup I	**Subgroup II**	**Subgroup III**
Inhalation spacing device - 14% (73%)	Oxycodone - 18% (95%)	Guanfacine - 6% (50%)
Cetirizine - 11% (59%)	Acetaminophen 325 mg - 17% (92%)	Miralax - 4% (13%)
Amoxicillin 400 mg - 11% (67%)	Vitamin D - 16% (76%)	Lisdexamfetamine - 3% (68%)
Ear drops - 8% (47%)	Ergocalciferol - 14% (76%)	Lisdexamfetamine - 3% (75%)
Montelukast - 8% (53%)	Docusate sodium 100 mg - 14% (95%)	Electrolyte IV line - 3% (12%)
Budesonide - 8% (70%)	Magnesium - 14% (84%)	Polyethylene glycol - 3% (11%)
Carbamide peroxide - 7% (58%)	Lidocaine - 12% (88%)	Cholesterol - 3% (10%)
Loratadine 5mg - 7% (89%)	Lorazepam - 11% (86%)	Tuberculin - 3% (10%)
Budesonide - 7% (72%)	Morphine - 11% (94%)	Benadryl - 3% (7%)
Prednisolone sodium - 7% (68%)	Lorazepam - 11% (90%)	Paracetamol - 3% (5%)

[1] in-group and (total) percentages.

Table 3.4: Most frequent medication terms in ASC subgroups.

Laboratory test[1]	
Subgroup I	**Subgroup II**
Lead, capillary blood - 7% (92%)	Amylase - 29% (92%)
Peanut ige - 5% (90%)	Gamma-glutamyltransferase - 27% (92%)
Timothy ige - 5% (71%)	pH - 25% (86%)
Der. pteronyssinus ige - 5% (71%)	Lactate dehydrogenase - 22% (94%)
Cat dander ige - 5% (71%)	CO2 - 21% (90%)
Cashew nut ige - 5% (94%)	Lactate - 17% (94%)
Brazil nut ige - 5% (94%)	Lipoprotein lipase - 17% (93%)
Walnut ige - 5% (94%)	aPTT panel - 17% (95%)
Hazel nut ige - 5% (94%)	PT/INR - 17% (94%)
Birch ige - 5% (76%)	Blood gas testing - 15% (97%)

[1] in-group and (total) percentages;

aPTT = Activated Partial Thromboplastin Time;

PT/INR = prothrombin time/international normalized ratio

Table 3.5: Most frequent laboratory test terms in ASC subgroups. Subgroup III does not include any laboratory test terms.

	CPT[1]	
Subgroup I	**Subgroup II**	**Subgroup III**
Fluoride - 18% (77%)	Urea nitrogen - 48% (88%)	Skin test; tubercolosis - 3% (10%)
Tympanometry (impedance testing) - 12% (67%)	Chloride - 45% (89%)	Blood count; HGB - 3% (4%)
Speech audiometry threshold - 9% (74%)	Sodium - 45% (90%)	Enzyme - 3% (6%)
Speech audiometry threshold - 9% (75%)	Potassium - 44% (90%)	Blood count - 1% (5%)
VRA - 8% (66%)	Glucose - 44% (90%)	Radiologic examination, chest - 1% (3%)
Percutaneous tests allergenic extracts - 8% (78%)	Creatinine - 44% (89%)	Blood count - 1% (2%)
Radiologic examination, teeth - 8% (74%)	Phosphatase, alkaline - 40% (89%)	Radiologic examination, chest, 2 views (frontal/lateral) - 1% (1%)
Dentoalveolar structure - 8% (68%)	Phosphatase, alkaline - 39% (89%)	-
Hearing disorder assessment - 7% (64%)	Albumin - 38% (87%)	-
Audiometry - 5% (76%)	TSH - 37% (82%)	-

[1] in-group and (total) percentages; CPT = Current Procedural Terminology; VRA = Visual Reiforcement Audiometry; TSH = Thyroid Stimulating Hormone; HGB = hemoglobin

Table 3.6: Most frequent CPT terms in ASC subgroups.

3.4. Discussion

Comparing comorbidities in the entire dataset with findings from [20] we observe that two of the subgroups found by Doshi-Velez and colleagues report higher prevalence of autism diagnosis, whereas the third subgroup includes more individuals with Asperger's syndrome. Moreover, PDD-NOS proportions are low in all subgroups. Our result, despite the wider age range in the cohort considered, has a comparable distribution of diagnosis, if we take into account that we do not have Asperger's syndrome code within our EHRs. In [20], two subgroups are characterized by GI disorders, as we also report for subgroups I/II, whereas the third subgroup is characterized by psychiatric disorders, including anxiety. In our case, we find anxiety diagnostic terms both in subroup II and III, although a higher proportion is found in subgroup III, which also shows psychotic symptoms that can be associated with a wide variety of primary psychiatric disorders (i.e., hallucinations). We observe that the general characterization of subgroups appears to mirror, to some extent, the one reported in the literature.

Investigating the most significantly frequent terms in each subgroup, Doshi-Velez and collaborators describe subgroup 1 as characterized by seizures (77.5%) and IDs (60%). Subgroup 2 is characterized by cardiac and auditory disorders, asthma, and congenital anomalies involving the ear, eye, and cranial nerve. Both groups are described as showing more severe autistic symptoms. Finally, subgroup 3 has the highest rate of individuals with Asperger's syndrome and the lowest rate of IDs (27.8%). Comorbidities for this group are mainly psychiatric disorders [20]. In this study, we have identified three subgroups characterized by: 1) communication disorders and atypical sensory differences; 2) GI problems, IDs, and cardiovascular symptoms; 3) ADHD and pharmacologic treatment side effects. We have also argued that cardiovascular symptoms in subgroup II can be linked to risperidone side effects.

Comparing the results to subgroup comorbidities from [20], what emerges more evidently is that not taking into account all the available codes included in EHRs can mask what drives the presence of co-occurrent conditions. More specifically, cardiac problems and psychiatric conditions can be treatement side effects (e.g., anti-seizure, antipsychotic drugs), whereas hearing problems can actually refer to screening clinical practices that arise from caretaker concerns due to communication difficulties.

Our findings show that our ConvAE architecture successfully learns latent representations of individuals with different conditions. In this application, we confirm ASC heterogeneity at the clinical level and provide a useful report on how important it is to include all available clinical terms for a comprehensive description of clinical trajectories of individuals with ASC. In fact, drug prescriptions and assessment/screening protocols may influence the characterization of individual trajectories. Moreover, we observe that the presence of a higher percentage of females (38%) in subgroups II, which includes older subjects at diagnosis, cannot be completely explained by the presence of IDs (13%). Results in Section 2.4, report older individuals at diagnosis as those that are more hardly identified because of the compensatory mechanisms that they may develop throughout development, and for this reason fewer females are included in such subgroups. By means of EHRs, it appears that more females detected at a older age are identified, hence, combining clinical stratification results with findings from developmental psychology may advance the uncovering of the neurodevelopmental differences that differentiate males and females with ASC.

3.5. Conclusions

We have presented here a novel approach to EHRs of individuals with ASC, leveraging latent representations of clinical trajectories to identify informative subgroups within ASC. To this aim, we do not limit such an approach to diagnostic codes for comorbidity detections, but we have also included medications, laboratory tests and procedures. We obtain three distinct subgroups that characterize individuals with ASC according to comorbidities, but also to screening procedures and medication side effects. Ideally, identifying new patients with latent representations similar to those used for clustering can inform individual risk to adverse reactions, successful screening procedures, and prognosis of co-occurrent conditions.

Limitations to our approach mainly reside in the temporal resolution and investigation of clinical trajectories. Although our model proposes to process multi-dimensional longitudinal clinical histories, we should take into account that high-age diagnosis might include individuals that have been diagnosed during infancy and treated elsewhere, before arriving at the hospital facility. In the future, we aim to combine whole-sequence analysis and time structure by dividing sequences into age ranges of developmental milestones. More specifically, we propose to divide individual EHRs into fixed age-range subsequences and perform a clustering analysis on this stratified cohort. Then, we aim to investigate whether clusters are differently characterized and observe how individuals might move from a cluster to another throughout development. In conclusion, future work will also include filtering of less informative and redundant terms to investigate how this might translate in different results that should then be compared to the ones obtained here, for a further validation of our method efficiently deals with noise.

References

[1] Jensen, P. B., Jensen, L. J., & Brunak, S. (2012). Mining electronic health records: towards better research applications and clinical care. *Nature Reviews Genetics*, *13*, 395.

[2] Xiao, C., Choi, E., & Sun, J. (2018). Opportunities and challenges in developing deep learning models using electronic health records data: a systematic review. *Journal of the American Medical Informatics Association*, *25*(10), 1419–1428.

[3] Miotto, R., Wang, F., Wang, S., Jiang, X., & Dudley, J. T. (2017). Deep learning for healthcare: review, opportunities and challenges. *Briefings in Bioinformatics*, *19*(6), 1236–1246.

[4] Lai, M. C., Lombardo, M. V., & Baron-Cohen, S. (2014). Autism. *Lancet*, *383*, 896–910. doi:10.1016/S0140-6736(13)61539-1

[5] Quesnel-Vallières, M., Weatheritt, R. J., Cordes, S. P., & Blencowe, B. J. (2018). Autism spectrum disorder: insights into convergent mechanisms from transcriptomics. *Nature Reviews Genetics*, *20*(1), 51–63. doi:10.1038/s41576-018-0066-2

[6] Bai, D., Yip, B. H. K., Windham, G. C., Sourander, A., Francis, R., Yoffe, R., … Sandin, S. (2019). Association of genetic and environmental factors with autism in a 5-country cohort. *JAMA Psychiatry*, *76*(10), 1035–1043. doi:10.1001/jamapsychiatry.2019.1411

[7] Lord, C., Bishop, S., & Anderson, D. (2015). Developmental trajectories as autism phenotypes. *American Journal of Medical Genetics Part C: Seminars in Medical Genetics*, *169*(2), 198–208. doi:10.1002/ajmg.c.31440

[8] Langston, J. W. (2006). The parkinson's complex: Parkinsonism is just the tip of the iceberg. *Annals of Neurology*, *59*(4), 591–596. doi:10.1002/ana.20834

[9] Greenland, J. C., Williams-Gray, C. H., & Barker, R. A. (2018). The clinical heterogeneity of parkinson's disease and its therapeutic implications. *European Journal of Neuroscience*, *49*(3), 328–338. doi:10.1111/ejn.14094

[10] de Mel, S., Lim, S. H., Tung, M. L., & Chng, W. J. (2014). Implications of heterogeneity in multiple myeloma. *BioMed Research International*, *2014*, 1–12. doi:10.1155/2014/232546

[11] Pearson, E. R. (2019). Type 2 diabetes: A multifaceted disease. *Diabetologia*, *62*(7), 1107–1112. doi:10.1007/s00125-019-4909-y

[12] Cutting, G. R. (2014). Cystic fibrosis genetics: from molecular understanding to clinical application. *Nature Reviews Genetics*, *16*(1), 45–56.

[13] Alexandrov, V., Brunner, D., Menalled, L. B., Kudwa, A., Watson-Johnson, J., Mazzella, M., … Kwak, S. (2016). Large-scale phenome analysis defines a behavioral signature for Huntington's disease genotype in mice. *Nature Biotechnology*, *34*(8), 838–44.

[14] Manchia, M., Cullis, J., Turecki, G., Rouleau, G. A., Uher, R., & Alda, M. (2013). The Impact of Phenotypic and Genetic Heterogeneity on Results of Genome Wide Association Studies of Complex Diseases. *PLoS ONE*, *8*(10), e76295. doi:10.1371/journal.pone.0076295

[15] Baytas, I. M., Xiao, C., Zhang, X., Wang, F., Jain, A. K., & Zhou, J. (2017). Patient Subtyping via Time-Aware LSTM Networks. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 65–74). doi:10.1145/3097983.3097997

[16] Li, L., Cheng, W. Y., Glicksberg, B. S., Gottesman, O., Tamler, R., Chen, R., ... Dudley, J. T. (2015). Identification of type 2 diabetes subgroups through topological analysis of patient similarity. *Science translational medicine*, *7*(311), 311ra174.

[17] Chen, R., Sun, J., Dittus, R. S., Fabbri, D., Kirby, J., Laffer, C. L., ... Malin, B. (2018). Patient Stratification Using Electronic Health Records from a Chronic Disease Management Program. *IEEE Journal of Biomedical and Health Informatics*. doi:10.1109/JBHI.2016.2514264

[18] Taslimitehrani, V., Dong, G., Pereira, N. L., Panahiazar, M., & Pathak, J. (2019). Developing EHR-driven heart failure risk prediction models using CPXR(Log) with the probabilistic loss function. *Journal of Biomedical Informatics*, *60*(1), 797.

[19] Perotte, A., Elhadad, N., Hirsch, J. S., Ranganath, R., & Blei, D. (2015). Risk prediction for chronic kidney disease progression using heterogeneous electronic health record data and time series analysis. *Journal of the American Medical Informatics Association*, *22*(4), 872–880.

[20] Doshi-Velez, F., Ge, Y., & Kohane, I. (2013). Comorbidity Clusters in Autism Spectrum Disorders: An Electronic Health Record Time-Series Analysis. *Pediatrics*, *133*(1), e54–e63. doi:10.1542/peds.2013-0819

[21] Lawton, M., Ben-Shlomo, Y., May, M. T., Baig, F., Barber, T. R., Klein, J. C., ... Hu, M. T. M. (2018). Developing and validating Parkinson's disease subtypes and their motor and cognitive progression. *Journal of Neurology, Neurosurgery & Psychiatry*, *89*(12), 1279–1287. doi:10.1136/jnnp-2018-318337

[22] Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient Estimation of Word Representations in Vector Space. arXiv: 1301.3781 [cs.CL]

[23] Glicksberg, B. S., Miotto, R., Johnson, K. W., Shameer, K., Li, L., Chen, R., & Dudley, J. T. (2017). Automated disease cohort selection using word embeddings from Electronic Health Records. In *Biocomputing 2018* (pp. 145–156). doi:10.1142/9789813235533_0014

[24] Zhu, Z., Yin, C., Qian, B., Cheng, Y., Wei, J., & Wang, F. (2016). Measuring Patient Similarities via a Deep Architecture with Medical Concept Embedding. In *2016 IEEE 16th International Conference on Data Mining* (pp. 749–758). doi:10.1109/icdm.2016.0086

[25] Suo, Q., Ma, F., Yuan, Y., Huai, M., Zhong, W., Zhang, A., & Gao, J. (2017). Personalized disease prediction using a CNN-based similarity learning method. In *2017 IEEE International Conference on Bioinformatics and Biomedicine* (pp. 811–816). doi:10.1109/BIBM.2017.8217759

[26] Zhang, X., Chou, J., Liang, J., Xiao, C., Zhao, Y., Sarva, H., ... Wang, F. (2019). Data-Driven Subtyping of Parkinson's Disease Using Longitudinal Clinical Records: A Cohort Study. *Scientific Reports*, *9*(1), 797.

[27] Miotto, R., Li, L., Kidd, B. A., & Dudley, J. T. (2016). Deep patient: an unsupervised representation to predict the future of patients from the electronic health records. *Scientific Reports*, *6*, 26094. doi:10.1038/srep26094

[28] Choi, Y., Chiu, C. Y. I., & Sontag, D. (2016). Learning low-dimensional representations of medical concepts. In *AMIA Summits on Translational Science Proceedings* (Vol. 2016, pp. 41–50).

[29] Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, *12*, 2825–2830.

[30] Paszke, A., Gross, S., Chintala, S., Chanan, G., Yang, E., DeVito, Z., ... Lerer, A. (2017). Automatic differentiation in pytorch. In *NeurIPS Autodiff Workshop.*

[31] Kingma, D., & Adam, J. B. (2014). Adam: A Method for Stochastic Optimization. In *Proc. 3rd International Conference on Learning Representations* (pp. 1–15). arXiv:1412.6980.

[32] Dugger, S. A., Platt, A., & Goldstein, D. B. (2017). Drug development in the era of precision medicine. *Nature Reviews Drug Discovery*, *17*(3), 183–196. doi:10.1038/nrd.2017.226

[33] Hochreiter, S., & Schmidhuber, J. (1997). Long Short-Term Memory. *Neural Computation*, *9*(8), 1735–1780. doi:10.1162/neco.1997.9.8.1735

[34] McInnes, L., Healy, J., & Melville, J. (2018). UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction. arXiv: 1802.03426 `[stat.ML]`

[35] McConville, R., Santos-Rodriguez, R., Piechocki, R. J., & Craddock, I. (2019). N2D: (Not Too) Deep Clustering via Clustering the Local Manifold of an Autoencoded Embedding. arXiv: 1908.05968 `[cs.LG]`

[36] DeFronzo, R. A., & Ferrannini, E. (1991). Insulin Resistance: A Multifaceted Syndrome Responsible for NIDDM, Obesity, Hypertension, Dyslipidemia, and Atherosclerotic Cardiovascular Disease. *Diabetes Care*, *14*(3), 173–194. doi:10.2337/diacare.14.3.173

[37] Vallon, V., & Komers, R. (2011). Pathophysiology of the diabetic kidney. *Comprehensive Physiology*, *1*(3), 1175–1232.

[38] Thiruvoipati, T. (2015). Peripheral artery disease in patients with diabetes: Epidemiology, mechanisms, and outcomes. *World Journal of Diabetes*, *6*(7), 961–969. doi:10.4239/wjd.v6.i7.961

[39] Malaguarnera, L., Cristaldi, E., & Malaguarnera, M. (2010). The role of immunity in elderly cancer. *Critical Reviews in Oncology/Hematology*, *74*(1), 40–60. doi:10.1016/j.critrevonc.2009.06.002

[40] Delamaire, M., Maugendre, D., Moreno, M., Goff, M. C. L., Allannic, H., & Genetet, B. (1997). Impaired Leucocyte Functions in Diabetic Patients. *Diabetic Medicine*, *14*(1), 29–34. doi:10.1002/(sici)1096-9136(199701)14:1<29::aid-dia300>3.0.co;2-v

[41] Fowlkes, E. B., & Mallows, C. L. (1983). A Method for Comparing Two Hierarchical Clusterings. *Journal of the American Statistical Association*, *78*(383), 553–569. doi:10.1080/01621459.1983.10478008

[42] Meilă, M. (2007). Comparing clusterings – an information based distance. *Journal of Multivariate Analysis*, *98*(5), 873–895. doi:10.1016/j.jmva.2006.11.013

[43] Li, C. I., Uribe, D. J., & Daling, J. R. (2005). Clinical characteristics of different histologic types of breast cancer. *British Journal of Cancer*, *93*(9), 1046–1052. doi:10.1038/sj.bjc.6602787

[44] de Lau, L. M. L., & Breteler, M. M. B. (2006). Epidemiology of Parkinson's disease. *The Lancet Neurology*, *5*(6), 525–535. doi:10.1016/s1474-4422(06)70471-9

[45] Jain, S., Lo, S. E., & Louis, E. D. (2006). Common Misdiagnosis of a Common Neurological Disorder. *Archives of Neurology*, *63*(8), 1100–1104. doi:10.1001/archneur.63.8.1100

[46] Alves, G., Wentzel-Larsen, T., & Larsen, J. P. (2004). Is fatigue an independent and persistent symptom in patients with Parkinson disease? *Neurology*, *63*(10), 1908–1911. doi:10.1212/01.wnl.0000144277.06917.cc

[47] Siciliano, M., Trojano, L., Santangelo, G., Micco, R. D., Tedeschi, G., & Tessitore, A. (2018). Fatigue in Parkinson's disease: A systematic review and meta-analysis. *Movement Disorders*, *33*(11), 1712–1723. doi:10.1002/mds.27461

[48] Scheltens, P., Blennow, K., Breteler, M. M. B., de Strooper, B., Frisoni, G. B., Salloway, S., & der Flier, W. M. V. (2016). Alzheimer's disease. *The Lancet*, *388*(10043), 505–517. doi:10.1016/s0140-6736(15)01124-1

[49] Nichols, E., Szoeke, C. E. I., Vollset, S. E., Abbasi, N., Abd-Allah, F., Abdela, J., … Murray, C. J. L. (2019). Global, regional, and national burden of Alzheimer's disease and other dementias, 1990–2016: a systematic analysis for the Global Burden of Disease Study 2016. *The Lancet Neurology*, *18*(1), 88–106. doi:10.1016/s1474-4422(18)30403-4

[50] Manji, H., Jäger, H. R., & Winston, A. (2013). HIV, dementia and antiretroviral drugs: 30 years of an epidemic. *Journal of Neurology, Neurosurgery & Psychiatry*, *84*(10), 1126–1137. doi:10.1136/jnnp-2012-304022

[51] Lyketsos, C. G., Lopez, O., Jones, B., Fitzpatrick, A. L., Breitner, J., & DeKosky, S. (2002). Prevalence of Neuropsychiatric Symptoms in Dementia and Mild Cognitive Impairment. *JAMA*, *288*(12), 1475–1483. doi:10.1001/jama.288.12.1475

[52] Snyder, H. M., Corriveau, R. A., Craft, S., Faber, J. E., Greenberg, S. M., Knopman, D., ... Carrillo, M. C. (2015). Vascular contributions to cognitive impairment and dementia including Alzheimer's disease. *Alzheimer's & Dementia*, *11*(6), 710–717. doi:10.1016/j.jalz.2014.10.008

[53] Birks, J. S., & Harvey, R. J. (2018). Donepezil for dementia due to Alzheimer's disease. *Cochrane Database of Systematic Reviews*, *6*(6), CD001190. doi:10.1002/14651858.cd001190.pub3

[54] Cowan, A. J., Allen, C., Barac, A., Basaleem, H., Bensenor, I., Curado, M. P., ... Fitzmaurice, C. (2018). Global Burden of Multiple Myeloma. *JAMA Oncology*, *4*(9), 1221–1227. doi:10.1001/jamaoncol.2018.2128

[55] Kyle, R. A., Gertz, M. A., Witzig, T. E., Lust, J. A., Lacy, M. Q., Dispenzieri, A., ... Greipp, P. R. (2003). Review of 1027 Patients With Newly Diagnosed Multiple Myeloma. *Mayo Clinic Proceedings*, *78*(1), 21–33. doi:10.4065/78.1.21

[56] Berk, J. L. (2005). Pleural effusions in systemic amyloidosis. *Current Opinion in Pulmonary Medicine*, *11*(4), 324–328. doi:10.1097/01.mcp.0000162378.35928.37

[57] Rajkumar, S. V., Gertz, M. A., & Kyle, R. A. (1998). Primary systemic amyloidosis with delayed progression to multiple myeloma. *Cancer*, *82*(8), 1501–1505. doi:10.1002/(sici)1097-0142(19980415)82:8<1501::aid-cncr11>3.0.co;2-8

[58] Perera, E., Revington, P., & Sheffield, E. (2010). Low grade marginal zone B-cell lymphoma presenting as local amyloidosis in a submandibular salivary gland. *International Journal of Oral and Maxillofacial Surgery*, *39*(11), 1136–1138. doi:10.1016/j.ijom.2010.05.001

[59] Attal, M., Lauwers-Cances, V., Hulin, C., Leleu, X., Caillot, D., Escoffre, M., ... Moreau, P. (2017). Lenalidomide, Bortezomib, and Dexamethasone with Transplantation for Myeloma. *New England Journal of Medicine*, *376*(14), 1311–1320. doi:10.1056/nejmoa1611750

[60] Argyriou, A. A., Iconomou, G., & Kalofonos, H. P. (2008). Bortezomib-induced peripheral neuropathy in multiple myeloma: a comprehensive review of the literature. *Blood*, *112*(5), 1593–1599. doi:10.1182/blood-2008-04-149385

[61] Miguel, J. F. S., Schlag, R., Khuageva, N. K., Dimopoulos, M. A., Shpilberg, O., Kropff, M., ... Richardson, P. G. (2008). Bortezomib plus Melphalan and Prednisone for Initial Treatment of Multiple Myeloma. *New England Journal of Medicine*, *359*(9), 906–917. doi:10.1056/nejmoa0801479

[62] Ravaglia, S., Corso, A., Piccolo, G., Lozza, A., Alfonsi, E., Mangiacavalli, S., ... Costa, A. (2008). Immune-mediated neuropathies in myeloma patients treated with bortezomib. *Clinical Neurophysiology*, *119*(11), 2507–2512. doi:10.1016/j.clinph.2008.08.007

[63] Dispenzieri, A., & Kyle, R. A. (2005). Neurological aspects of multiple myeloma and related disorders. *Best Practice & Research Clinical Haematology*, *18*(4), 673–688. doi:10.1016/j.beha.2005.01.024

[64] Fitzmaurice, C., Akinyemiju, T. F., Lami, F. H. A., Alam, T., Alizadeh-Navaei, R., Allen, C., ... Naghavi, M. (2018). Global, Regional, and National Cancer Incidence, Mortality, Years of Life Lost, Years Lived With Disability, and Disability-Adjusted Life-Years for 29 Cancer Groups, 1990 to 2016. *JAMA Oncology*, *4*(11), 1553–1568. doi:10.1001/jamaoncol.2018.2706

[65] Gore, J. L., Kwan, L., Lee, S. P., Reiter, R. E., & Litwin, M. S. (2009). Survivorship Beyond Convalescence: 48-Month Quality-of-Life Outcomes After Treatment for Localized Prostate

Cancer. *JNCI: Journal of the National Cancer Institute*, *101*(12), 888–892. doi:10.1093/jnci/djp114

[66] Sørlie, T., Tibshirani, R., Parker, J., Hastie, T., Marron, J. S., Nobel, A., . . . Botstein, D. (2003). Repeated observation of breast tumor subtypes in independent gene expression data sets. *Proceedings of the National Academy of Sciences*, *100*(14), 8418–8423. doi:10.1073/pnas.0932692100

[67] American Psychiatric Association. (2013). *Diagnostic and Statistical Manual of Mental Disorders* (5th ed.). doi:10.1176/appi.books.9780890425596

[68] Baron-Cohen, S. (2017). Editorial Perspective: Neurodiversity – a revolutionary concept for autism and psychiatry. *Journal of Child Psychology and Psychiatry*, *58*(6), 744–747. doi:10.1111/jcpp.12703

[69] Baio, J., Wiggins, L., Christensen, D. L., Maenner, M. J., Daniels, J., Warren, Z., . . . Dowling, N. F. (2018). Prevalence of autism spectrum disorder among children aged 8 years – autism and developmental disabilities monitoring network, 11 sites, United States, 2014. *MMWR Surveillance Summaries*, *67*(6), 1–23. doi:http://dx.doi.org/10.15585/mmwr.ss6706a1

[70] Ferri, S. L., Abel, T., & Brodkin, E. S. (2018). Sex Differences in Autism Spectrum Disorder: a Review. *Current Psychiatry Reports*, *20*(2), 9. doi:10.1007/s11920-018-0874-2

[71] Narzisi, A., Posada, M., Barbieri, F., Chericoni, N., Ciuffolini, D., Pinzino, M., . . . Muratori, F. (2018). Prevalence of Autism Spectrum Disorder in a large Italian catchment area: a school-based population study within the ASDEU project. *Epidemiology and Psychiatric Sciences*, 1–10. doi:10.1017/s2045796018000483

[72] Simonoff, E. (2015). Intellectual disability. In A. Thapar, D. S. Pine, J. F. Leckman, S. Scott, M. J. Snowling, & E. Taylor (Eds.), *Rutter's Child and Adolescent Psychiatry* (6th ed., Chap. 54, pp. 719–737). doi:10.1002/9781118381953.ch54

[73] Le Couteur, A., & Szatmari, P. (2015). Autism spectrum disorder. In A. Thapar, D. S. Pine, J. F. Leckman, S. Scott, M. J. Snowling, & E. Taylor (Eds.), *Rutter's Child and Adolescent Psychiatry* (6th ed., Chap. 51, pp. 661–682). doi:10.1002/9781118381953.ch51

[74] Wang, X. J., & Krystal, J. H. (2014). Computational psychiatry. *Neuron*, *84*(3), 638–654. doi:10.1016/j.neuron.2014.10.018

[75] Huys, Q. J. M., Maia, T. V., & Frank, M. J. (2016). Computational psychiatry as a bridge from neuroscience to clinical applications. *Nature Neuroscience*, *19*(3), 404–413. doi:10.1038/nn.4238

[76] Wang, H., Li, L., Chi, L., & Zhao, Z. (2019). Autism Screening Using Deep Embedding Representation. In J. M. F. Rodrigues, P. J. S. Cardoso, J. Monteiro, R. Lam, V. V. Krzhizhanovskaya, M. H. Lees, . . . P. M. A. Sloot (Eds.), *Computational science – iccs 2019. lecture notes in computer science* (Vol. 11537, pp. 160–173). doi:10.1007/978-3-030-22741-8_12

[77] Kohane, I. S., McMurry, A., Weber, G., MacFadden, D., Rappaport, L., Kunkel, L., . . . Churchill, S. (2012). The Co-Morbidity Burden of Children and Young Adults with Autism Spectrum Disorders. *PLoS ONE*, *7*(4), e33224. doi:10.1371/journal.pone.0033224

[78] Lai, M. C., Lombardo, M. V., Auyeung, B., Chakrabarti, B., & Baron-Cohen, S. (2015). Sex/Gender Differences and Autism: Setting the Scene for Future Research. *Journal of the American Academy of Child & Adolescent Psychiatry*, *54*(1), 11–24. doi:10.1016/j.jaac.2014.10.003

[79] Baron-Cohen, S. (2002). The extreme male brain theory of autism. *Trends in Cognitive Sciences*, *6*(6), 248–254. doi:10.1016/s1364-6613(02)01904-6

[80] Kreiser, N. L., & White, S. W. (2013). ASD in Females: Are We Overstating the Gender Difference in Diagnosis? *Clinical Child and Family Psychology Review*, *17*(1), 67–84. doi:10.1007/s10567-013-0148-9

[81] Vorstman, J. A. S., Parr, J. R., De-Luca, D. M., Anney, R. J. L., Jr, J. I. N., & Hallmayer, J. F. (2017). Autism genetics: opportunities and challenges for clinical translation. *Nature Reviews Genetics*, *18*(6), 362–376. doi:10.1038/nrg.2017.4

[82] Doan, R. N., Lim, E. T., Rubeis, S. D., Betancur, C., Cutler, D. J., Chiocchetti, A. G., ... Yu, T. W. (2019). Recessive gene disruptions in autism spectrum disorder. *Nature Genetics*, *51*(7), 1092–1098. doi:10.1038/s41588-019-0433-8

[83] Velmeshev, D., Schirmer, L., Jung, D., Haeussler, M., Perez, Y., Mayer, S., ... Kriegstein, A. R. (2019). Single-cell genomics identifies cell type-specific molecular changes in autism. *Science*, *364*(6441), 685–689. doi:10.1126/science.aav8130

[84] Varma, M., Paskov, K. M., Jung, J. Y., Chrisman, B. S., Stockham, N. T., Washington, P. Y., & Wall, D. P. (2018). Outgroup Machine Learning Approach Identifies Single Nucleotide Variants in Noncoding DNA Associated with Autism Spectrum Disorder. In *Biocomputing 2019* (pp. 260–271). doi:10.1142/9789813279827_0024

[85] Zhou, J., Park, C. Y., Theesfeld, C. L., Wong, A. K., Yuan, Y., Scheckel, C., ... Troyanskaya, O. G. (2019). Whole-genome deep-learning analysis identifies contribution of noncoding mutations to autism risk. *Nature Genetics*, *51*(6), 973–980. doi:10.1038/s41588-019-0420-0

[86] Fountain, C., Winter, A. S., & Bearman, P. S. (2012). Six Developmental Trajectories Characterize Children With Autism. *Pediatrics*, *129*(5), e1112–e1120. doi:10.1542/peds.2011-1601

[87] Gotham, K., Pickles, A., & Lord, C. (2012). Trajectories of Autism Severity in Children Using Standardized ADOS Scores. *Pediatrics*, *130*(5), e1278–e1284. doi:10.1542/peds.2011-3668

[88] Venker, C. E., Ray-Subramanian, C. E., Bolt, D. M., & Weismer, S. E. (2013). Trajectories of Autism Severity in Early Childhood. *Journal of Autism and Developmental Disorders*, *44*(3), 546–563. doi:10.1007/s10803-013-1903-y

[89] Szatmari, P., Georgiades, S., Duku, E., Bennett, T. A., Bryson, S., Fombonne, E., ... Thompson, A. (2015). Developmental Trajectories of Symptom Severity and Adaptive Functioning in an Inception Cohort of Preschool Children With Autism Spectrum Disorder. *JAMA Psychiatry*, *72*(3), 276–283. doi:10.1001/jamapsychiatry.2014.2463

[90] Visser, J. C., Rommelse, N. N. J., Lappenschaar, M., Servatius-Oosterling, I. J., Greven, C. U., & Buitelaar, J. K. (2017). Variation in the Early Trajectories of Autism Symptoms is Related to the Development of Language, Cognition, and Behavior Problems. *Journal of the American Academy of Child & Adolescent Psychiatry*, *56*(8), 659–668. doi:10.1016/j.jaac.2017.05.022

[91] Farmer, C., Swineford, L., Swedo, S. E., & Thurm, A. (2018). Classifying and characterizing the development of adaptive behavior in a naturalistic longitudinal study of young children with autism. *Journal of Neurodevelopmental Disorders*, *10*(1), 1. doi:10.1186/s11689-017-9222-9

[92] Baghdadli, A., Michelon, C., Pernon, E., Picot, M. C., Miot, S., Sonié, S., ... Mottron, L. (2018). Adaptive trajectories and early risk factors in the autism spectrum: a 15-year prospective study. *Autism Research*, *11*(11), 1455–1467. doi:10.1002/aur.2022

[93] Anderson, D. K., Maye, M. P., & Lord, C. (2011). Changes in Maladaptive Behaviors From Midchildhood to Young Adulthood in Autism Spectrum Disorder. *American Journal on Intellectual and Developmental Disabilities*, *116*(5), 381–397. doi:10.1352/1944-7558-116.5.381

[94] Sacrey, L. A. R., Zwaigenbaum, L., Bryson, S., Brian, J., Smith, I. M., Raza, S., ... Garon, N. (2018). Developmental trajectories of adaptive behavior in autism spectrum disorder: a high-risk sibling cohort. *Journal of Child Psychology and Psychiatry*, *60*(6), 697–706. doi:10.1111/jcpp.12985

[95] Lombardo, M. V., Lai, M. C., & Baron-Cohen, S. (2019). Big data approaches to decomposing heterogeneity across the autism spectrum. *Molecular Psychiatry*, *24*(10), 1435–1450. doi:10.1038/s41380-018-0321-0

[96] Lord, C., Luyster, R., Guthrie, W., & Pickles, A. (2012). Patterns of developmental trajectories in toddlers with autism spectrum disorder. *Journal of Consulting and Clinical Psychology*, *80*(3), 477–489. doi:10.1037/a0027214

[97] Elder, J., Kreider, C., Brasher, S., & Ansell, M. (2017). Clinical impact of early diagnosis of autism on the prognosis and parent-child relationships. *Psychology Research and Behavior Management*, *10*, 283–292. doi:10.2147/prbm.s117499

[98] Goodfellow, I. J., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., ... Bengio, Y. (2014). Generative adversarial nets. In *Proceedings of the 27th International Conference on Neural Information Processing Systems* (Vol. 2, pp. 2672–2680). Montreal, Canada: MIT Press.

[99] Wolfers, T., Floris, D. L., Dinga, R., van Rooij, D., Isakoglou, C., Kia, S. M., ... Beckmann, C. F. (2019). From pattern classification to stratification: towards conceptualizing the heterogeneity of Autism Spectrum Disorder. *Neuroscience & Biobehavioral Reviews*, *104*, 240–254. doi:10.1016/j.neubiorev.2019.07.010

[100] Abbas, H., Garberson, F., Glover, E., & Wall, D. P. (2018). Machine learning approach for early detection of autism by combining questionnaire and home video screening. *Journal of the American Medical Informatics Association*, *25*(8), 1000–1007. doi:10.1093/jamia/ocy039

[101] Wall, D. P., Dally, R., Luyster, R., Jung, J.-Y., & DeLuca, T. F. (2012). Use of Artificial Intelligence to Shorten the Behavioral Diagnosis of Autism. *PLoS ONE*, *7*(8), e43855. doi:10.1371/journal.pone.0043855

[102] Kosmicki, J. A., Sochat, V., Duda, M., & Wall, D. P. (2015). Searching for a minimal set of behaviors for autism detection through feature selection-based machine learning. *Translational Psychiatry*, *5*(2), e514. doi:10.1038/tp.2015.7

[103] Bone, D., Bishop, S. L., Black, M. P., Goodwin, M. S., Lord, C., & Narayanan, S. S. (2016). Use of machine learning to improve autism screening and diagnostic instruments: effectiveness, efficiency, and multi-instrument fusion. *Journal of Child Psychology and Psychiatry*, *57*(8), 927–937. doi:10.1111/jcpp.12559

[104] Li, G., Liu, M., Sun, Q., Shen, D., & Wang, L. (2018). Early Diagnosis of Autism Disease by Multi-channel CNNs. In Y. Shi, H. I. Suk, & M. Liu (Eds.), *Machine Learning in Medical Imaging. MLMI 2018. Lecture Notes in Computer Science* (Vol. 11046, pp. 303–309). doi:10.1007/978-3-030-00919-9_35

[105] Krishnan, A., Zhang, R., Yao, V., Theesfeld, C. L., Wong, A. K., Tadych, A., ... Troyanskaya, O. G. (2016). Genome-wide prediction and functional characterization of the genetic basis of autism spectrum disorder. *Nature Neuroscience*, *19*(11), 1454–1462. doi:10.1038/nn.4353

[106] Stevens, E., Dixon, D. R., Novack, M. N., Granpeesheh, D., Smith, T., & Linstead, E. (2019). Identification and analysis of behavioral phenotypes in autism spectrum disorder via unsupervised machine learning. *International Journal of Medical Informatics*, *129*, 29–36. doi:10.1016/j.ijmedinf.2019.05.006

[107] Hu, V. W., & Steinberg, M. E. (2009). Novel clustering of items from the Autism Diagnostic Interview-Revised to define phenotypes within autism spectrum disorders. *Autism Research*, *2*(2), 67–77. doi:10.1002/aur.72

[108] Veatch, O. J., Veenstra-VanderWeele, J., Potter, M., Pericak-Vance, M. A., & Haines, J. L. (2014). Genetically meaningful phenotypic subgroups in autism spectrum disorders. *Genes, Brain and Behavior*, *13*(3), 276–285. doi:10.1111/gbb.12117

[109] Lombardo, M. V., Lai, M. C., Auyeung, B., Holt, R. J., Allison, C., Smith, P., ... Baron-Cohen, S. (2016). Unsupervised data-driven stratification of mentalizing heterogeneity in autism. *Scientific Reports*, *6*, 35333. doi:10.1038/srep35333

[110] Feczko, E., Balba, N. M., Miranda-Dominguez, O., Cordova, M., Karalunas, S. L., Irwin, L., ... Fair, D. A. (2018). Subtyping cognitive profiles in Autism Spectrum Disorder using a Functional Random Forest algorithm. *NeuroImage*, *172*, 674–688. doi:10.1016/j.neuroimage.2017.12.044

[111] Button, K. S., Ioannidis, J. P. A., Mokrysz, C., Nosek, B. A., Flint, J., Robinson, E. S. J., & Munafò, M. R. (2013). Power failure: why small sample size undermines the reliability of neuroscience. *Nature Reviews Neuroscience*, *14*(5), 365–376. doi:10.1038/nrn3475

[112] Bone, D., Goodwin, M. S., Black, M. P., Lee, C. C., Audhkhasi, K., & Narayanan, S. (2014). Applying Machine Learning to Facilitate Autism Diagnostics: Pitfalls and Promises. *Journal of Autism and Developmental Disorders*, *45*(5), 1121–1136. doi:10.1007/s10803-014-2268-6

[113] Harris, J. C. (2019). The Necessity to Identify Subtypes of Autism Spectrum Disorder. *JAMA Psychiatry*. doi:10.1001/jamapsychiatry.2019.1928

[114] Esteva, A., Robicquet, A., Ramsundar, B., Kuleshov, V., DePristo, M., Chou, K., . . . Dean, J. (2019). A guide to deep learning in healthcare. *Nature Medicine*, *25*(1), 24–29. doi:10.1038/s41591-018-0316-z

[115] Lord, C., Rutter, M., DiLavore, P. C., Risi, S., Gotham, K., & Bishop, S. (2012). *Autism Diagnostic Observation Schedule, Second Edition (ADOS-2)*. 2nd ed. Torrance, CA: Western Psychological Services.

[116] Hus, V., Gotham, K., & Lord, C. (2012). Standardizing ADOS Domain Scores: Separating Severity of Social Affect and Restricted and Repetitive Behaviors. *Journal of Autism and Developmental Disorders*, *44*(10), 2400–2412. doi:10.1007/s10803-012-1719-1

[117] Griffiths, R. (1996). *The Griffiths mental development scales from birth to two years, manual, the 1996 revision*. Association for Research in Infant and Child Development. Henley: Test Agency.

[118] Roid, G. H., & Miller, L. J. (1997). *Leiter international performance scale – revised (Leiter-R)*. Wood Dale, IL: Stoelting.

[119] Fancello, G., & Gianchetti, C. (2008). *Wechsler preschool and primary scale of intelligence – III*. Florence: Giunti O.S.

[120] Wechsler, D. (1989). *WPPSI-R: Wechsler preschool and primary scale of intelligence – revised*. San Antonio, TX: Psychological Corporation.

[121] Wechsler, D. (1991). *WISC-III: Wechsler intelligence scale for children: Manual*. San Antonio, TX: Psychological Corporation.

[122] Wechsler, D. (2003). *Wechsler intelligence scale for children – Fourth Edition (WISC-IV)*. San Antonio, TX: Psychological Corporation.

[123] Wechsler, D. (1981). *WAIS-R manual: Wechsler adult intelligence scale – revised*. San Antonio, TX: Psychological Corporation.

[124] Wechsler, D. (2008). *Wechsler adult intelligence scale – Fourth Edition (WAIS-IV)*. San Antonio, TX: NCS Pearson.

[125] Sparrow, S. S., Balla, D. A., & Cicchetti, D. V. (2005). *Vineland II: Vineland adaptive behavior scales*. Circle Pines, MN: AGS Publishing.

[126] Abidin, R. R. (1990). *Parenting stress index – short form*. Charlottesville, VA: Pediatric Psychology Press.

[127] Tehee, E., Honan, R., & Hevey, D. (2009). Factors Contributing to Stress in Parents of Individuals with Autistic Spectrum Disorders. *Journal of Applied Research in Intellectual Disabilities*, *22*(1), 34–42. doi:10.1111/j.1468-3148.2008.00437.x

[128] Ozturk, Y., Riccadonna, S., & Venuti, P. (2014). Parenting dimensions in mothers and fathers of children with Autism Spectrum Disorders. *Research in Autism Spectrum Disorders*, *8*(10), 1295–1306. doi:10.1016/j.rasd.2014.07.001

[129] Constantino, J. N. (2013). Social Responsiveness Scale. In F. R. Volkmar (Ed.), *Encyclopedia of Autism Spectrum Disorders*. New York, NY: Springer.

[130] Pennington, J., Socher, R., & Manning, C. (2014). Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing* (pp. 1532–1543).

[131] R Core Team. (2017). *R: A Language and Environment for Statistical Computing.* R Foundation for Statistical Computing. Vienna, Austria. Retrieved from https://www.R-project.org/

[132] Biringen, Z., Robinson, J. L., & Emde, R. N. (2000). Appendix B: The emotional availability scales (3rd ed.; an abridged infancy/early childhood version). *Attachment & human development*, *2*(2), 256–270.

[133] Bayer, M. (2015). *SQLAlchemy: The Python SQL Toolkit and Object Relational Mapper.* Retrieved from http://www.sqlalchemy.org

[134] Duchi, J., Hazan, E., & Singer, Y. (2011). Adaptive subgradient methods for online learning and stochastic optimization. *Journal of Machine Learning Research*, *12*, 2121–2159.

[135] Melamud, O., Goldberger, J., & Dagan, I. (2016). context2vec: Learning Generic Context Embedding with Bidirectional LSTM. In *Proceedings of The 20th SIGNLL Conference on Computational Natural Language Learning* (pp. 51–61). doi:10.18653/v1/K16-1006

[136] Volkmar, F., Siegel, M., Woodbury-Smith, M., King, B., McCracken, J., & State, M. (2014). Practice Parameter for the Assessment and Treatment of Children and Adolescents With Autism Spectrum Disorder. *Journal of the American Academy of Child & Adolescent Psychiatry*, *53*(2), 237–257. doi:10.1016/j.jaac.2013.10.013

[137] Lingren, T., Chen, P., Bochenek, J., Doshi-Velez, F., Manning-Courtney, P., Bickel, J., ... Savova, G. (2016). Electronic Health Record Based Algorithm to Identify Patients with Autism Spectrum Disorder. *PLoS ONE*, *11*(7), e0159621. doi:10.1371/journal.pone.0159621

[138] Leroy, G., Gu, Y., Pettygrove, S., & Kurzius-Spencer, M. (2017). Automated Lexicon and Feature Construction Using Word Embedding and Clustering for Classification of ASD Diagnoses Using EHR. In *Natural Language Processing and Information Systems - 22nd International Conference on Applications of Natural Language to Information Systems, NLDB 2017, Proceedings* (Vol. 10260 LNCS, pp. 34–37). Lecture Notes in Computer Science. doi:10.1007/978-3-319-59569-6_4

[139] Johnson, C. P., & Myers, S. M. (2007). Identification and Evaluation of Children With Autism Spectrum Disorders. *Pediatrics*, *120*(5), 1183–1215. doi:10.1542/peds.2007-2361

[140] Volkmar, F. R. (2013). Residual Autism. In F. R. Volkmar (Ed.), *Encyclopedia of Autism Spectrum Disorders.* doi:10.1007/978-1-4419-1698-3_1584

[141] Vasa, R. A., Mazurek, M. O., Mahajan, R., Bennett, A. E., Bernal, M. P., Nozzolillo, A. A., ... Coury, D. L. (2016). Assessment and Treatment of Anxiety in Youth With Autism Spectrum Disorders. *Pediatrics*, *137*(Supplement 2), S115–S123. doi:10.1542/peds.2015-2851j

[142] Eichenfield, L. F., Hanifin, J. M., Beck, L. A., Lemanske, R. F., Sampson, H. A., Weiss, S. T., & Leung, D. Y. M. (2003). Atopic Dermatitis and Asthma: Parallels in the Evolution of Treatment. *Pediatrics*, *111*(3), 608–616. doi:10.1542/peds.111.3.608

[143] Filipek, P. A., Accardo, P. J., Ashwal, S., Baranek, G. T., Cook, E. H., Dawson, G., ... Volkmar, F. R. (2000). Practice parameter: Screening and diagnosis of autism: Report of the Quality Standards Subcommittee of the American Academy of Neurology and the Child Neurology Society. *Neurology*, *55*(4), 468–479. doi:10.1212/wnl.55.4.468

[144] Liu, J., Liu, X., Wang, W., McCauley, L., Pinto-Martin, J., Wang, Y., ... Rogan, W. J. (2014). Blood Lead Concentrations and Children's Behavioral and Emotional Problems. *JAMA Pediatrics*, *168*(8), 737–745. doi:10.1001/jamapediatrics.2014.332

[145] Shannon, M., & Graef, J. W. (1996). Lead Intoxication in Children with Pervasive Developmental Disorders. *Journal of Toxicology: Clinical Toxicology*, *34*(2), 177–181. doi:10.3109/15563659609013767

[146] Barnard, L., Young, A. H., Pearson, J., Geddes, J., & O'Brien, G. (2002). A systematic review of the use of atypical antipsychotics in autism. *Journal of Psychopharmacology*, *16*(1), 93–101. doi:10.1177/026988110201600113

[147] Leitner, Y. (2014). The Co-Occurrence of Autism and Attention Deficit Hyperactivity Disorder in Children–What Do We Know? *Frontiers in Human Neuroscience*, *8*, 268. doi:10.3389/fnhum.2014.00268

[148] ADHD: Clinical Practice Guideline for the Diagnosis, Evaluation, and Treatment of Attention-Deficit/Hyperactivity Disorder in Children and Adolescents. (2011). *Pediatrics*, *128*(5), 1007–1022. doi:10.1542/peds.2011-2654

[149] Cortese, S., Adamo, N., Giovane, C. D., Mohr-Jensen, C., Hayes, A. J., Carucci, S., . . . Cipriani, A. (2018). Comparative efficacy and tolerability of medications for attention-deficit hyperactivity disorder in children, adolescents, and adults: A systematic review and network meta-analysis. *The Lancet Psychiatry*, *5*(9), 727–738. doi:10.1016/s2215-0366(18)30269-4

Conclusions

The advent of precision medicine is based on the increasing availability of big data in healthcare; in this context, artificial intelligence methods are offering novel predictive methods to reduce the impact of heterogeneity, personalizing at the individual level prognosis and estimates of efficacy of treatments. Autism Spectrum Conditions (ASCs), and more in general neurodevelopmental conditions, often converge phenotypically to a similar diagnostic profile, but they show wide heterogeneity in terms of clinical manifestations and response to treatment along time.

This fact supports the importance of unsupervised learning techniques for patient stratification, in order to identify informative subgroups of more homogeneous characteristics and a similar disease evolution within a larger patient cohort. With a translational perspective, making sense of heterogeneity would help develop new personalized treatments and interventions and detect new biomarkers, whose effect might be masked in the absence of data enriched in population strata.

To this aim, this thesis introduces a set of novel unsupervised learning methods; we apply such methods to represent individuals with ASCs and complex disorders in a latent space that includes encodings of multi-source clinical and behavioral-cognitive information and their differentiation throughout time. To extract these latent representations of patients we combined Convolutional Neural Networks and Autoencoders in an end-to-end architecture that learns to reconstruct the original sequence of clinical terms. These representations are then clustered towards discovering of informative subgroups. In particular, we find that individuals with ASCs can be clustered into representations that leverage Electronic Health Records (EHRs) and behavioral-cognitive profiles. The approach is demonstrated first on the structured EHR sequences of a cohort of $\sim 1.6M$ subjects from which patient clinical representations are learned. The encodings successfully contextualize concepts into clear subgroups of disorders. From these representations we select individuals with complex conditions, in particular ASCs, and investigate how the subgroups identified differed in terms of medication prescriptions, procedures, laboratory tests, and comorbid conditions. We have found that individual subgroups within ASC can inform clinical practices, e.g., screening policies, co-occurrent diseases, and adverse reaction to treatments (e.g., stimulant side effects). On the other hand, when complex disorders, such as Multiple Myeloma, and Alzheimer's disease, were selected, we have found that clinical trajectories differed according to disease progression, symptom severity, and comorbidities (e.g., cardiovascular and microvascular problems coexisting with Type 2 Diabetes Mellitus).

Second, word embedding methods, i.e., Word2vec and Global vectors for word representations, have been applied to behavioral profiles of children and adolescents with ASC referred to an academic reference centre (ODFLab at the University of Trento, Italy). This is a novel approach, where each subject is represented by their ordered sequence of instrument scores collected from a battery of tests designed for diagnostic purposes and intervention monitoring. Such an approach not only makes use of the deep embedding of clinical terms related to behavioral and cognitive functioning throughout development at multiple levels of refinement, but also provides quantitative scores that can be used to assess symptom severity and level of functioning. Results show promise in the detection of subgroups with different symptom severity and diverse developmental profiles.

The lack of known etiologies and the behavioral definition of ASC phenotype complicates clinical activity in cognitive psychology. The diagnostic process is time-consuming and sometimes challenging, in that not all individuals present with a clear ASC symptomatology. Interventions are not always effective, because they are not tailored to every subject with ASC. Stratifying ASC populations based on behavioral manifestation of symptoms not only carries promises in research studies, but also anticipates translational perspectives. Ideally, behavioral characteristics of every subject with suspected ASC can be tied to specific subgroups supporting clinicians in providing a diagnosis, developing effective treatments, and informing personalized prognosis.

The EHR data considered in Chapters 1 and 3 come from the Mount Sinai Data Warehouse and they are analyzed as EHR sequences where each term carries a literal meaning in English well defined by standard ICD-9 diagnosis codes, medications normalized to RxNorm, standard CPT-4 codes, and lab tests normalized to LOINC (e.g., `icd9::Alzheimer's disease::331.0` indicates a patient with Alzheimer's diagnosis). As the clinical terms follow international standards, the pipeline can be transferred "as is" to clinical data collected from different healthcare facilities, provided that term correspondences are available. In case other ontologies or medical terms descriptive systems are used, these should be pre-processed using tools based on annotation systems to preserve their generalizability. Free-text clinical notes, whose meaning is tied to the language used, could be preprocessed with tools such as the Open Biomedical Annotator.

Differently, data from Chapter 2 are assembled from clinical instruments administered in Italian: e.g. `gmds::gq::70` indicates an autistic subject with GMDS General quotient of 70. Still, these scales are highly standardized: the same codes are actually used across international versions of the tests in a 1-1 item-level correspondence, and thus they could can be back translated to English to connect embeddings at the international scale.

Notably, the two embedding spaces that have been built in the two applications have different scopes of analysis. The EHR-based embedding space learns term connotations from context to provide meaning to and characterize patient sequences. We already mentioned (see Chapter 2) that the ICD-9 term *Diarrhea (787.91)* next to an anti-cancer medication, such as *Bortezomib*, can indicate treatment side effects. At the same time, next to a diagnosis of *Autistic disorder, current or active state (299.00)* can refer to common comorbid conditions. The EHR embeddings should be trained to encompass at best the possible contextual meanings of terms.

While the same core machine learning pipelines are used to develop the embeddings for patient EHRs and for behavioral profiles of autistic children, the embeddings are kept separate and are supported by diverse ontology terms, thus, there is no direct need to correct for different languages.

However, there is a logical connection between EHR embeddings and those developed from the ODFLab collection. Indeed, the behavioral embedding space from ODFLab data can be combined as a vertical specification of fine-grained phenotypes specific to ASCs. Finally, the development of a more complete system could be based on the standard acquisition of terms from different ontologies for which correspondences can be developed, preserving their generalizability and enabling multi-center data collection.

A drawback of unsupervised machine learning studies for complex conditions is the lack of a priori information that can be used to validate models. In fact, model validation mainly relies on clinical inspection of enriched terms or quantitative scores. More effort should be spent in the development of reliable unsupervised approaches for translational purposes: we have presented here one of the first attempts in that direction.

Acknowledgments

First, I sincerely would like to thank my supervisors, Professor Paola Venuti and Professor Cesare Furlanello for the opportunities and mentorship they have provided. They have insightfully guided my research and helped me develop fundamental skills in both computational modeling and developmental psychology. They have also patiently put up with my stubbornness throughout the whole three-year PhD. Furthermore, I would like to thank Margherita Francescatto, PhD who has monitored my work and made suggestions on how to improve it.

I would like to acknowledge the contribution of my supervisor at the Institute for Next Generation Healthcare (INGH), Riccardo Miotto, PhD, with whom I have developed and worked hard on what has become the first and third chapter of this thesis. He has been a great mentor and encouraged me to develop the computational skills and research methods necessary for the study. Sincere thanks also to Joel T. Dudley, PhD, that welcomed me at his Institute allowing me to work on Mount Sinai health system's data warehouse. Moreover, a crucial role has been played by my co-worker at INGH Sarah Cherng, PhD, to whom I am grateful for her precious help and insightful suggestions. Furthermore, the ConvAE architecture has been developed together with Hao-Chih Lee, PhD, who has provided me with valuable recommendations during the model implementation process.

The data that have been used in the second chapter are the result of the outstanding clinical work of Arianna Bentenuto, PhD at the Laboratory of Observation, Diagnosis, and Education (ODFLab, University of Trento). Furthermore, she helped me with the clinical validation of both the behavioral and clinical subgroups within ASC cohorts. Her insights are always valuable and her help has been fundamental. At FBK, I would like to thank Valerio Maggio, PhD who has reviewed the code of the behavioral phenotyping pipeline and helped me improve my computational skills. Also, thanks to Ernesto Arbitrio who helped me built the ODFLab database.

Furthermore, I want to acknowledge the contribution of Doctor Giulia Landi and Doctor Giacomo Mori to the clinical validation of complex disorder subgroups. Their precious observation demonstrate how much can be accomplished when working together with physicians in an interdisciplinary environment.

Last, but not least, I want to thank my talented PhD colleague Nicole Bussola. Despite being constantly hard working on digital pathology models she found the time to draw the beautiful analysis protocol pictures that can be found in chapters 1 and 2.

Related work

Papers

1. Landi, I., Giannotti, M., Venuti, P., & de Falco, S. (2019). Maternal and family predictors of infant psychological development in at-risk families: A multilevel longitudinal study. *Research in nursing & health*, 2020;43, 17–27. DOI:10.1002/nur.21989

2. Venuti, P., Bentenuto, A., Cainelli, S., Landi, I., Suvini, F., Tancredi, R., Igliozzi, R., & Muratori, F. (2017). A joint behavioral and emotive analysis of synchrony in music therapy of children with autism spectrum disorders. *Health Psychology Report*, 5(2), 162-172.

Under review

3. Mazzoni, N., Landi, I., Ricciardelli, P., Actis-Grosso, R., & Venuti, P. Motion or emotion? Recognition of neutral and emotional point-light and full-light body stimuli in children with high-functioning and low-functioning Autism Spectrum Disorder. *Frontiers in Psychology.*

Conference papers (with Review Committee)

4. Landi, I., Miotto, R., Lee, H., Danieletto, M., Laganà, A., Furlanello, C., & Dudley, J. T. "Medical Sequence Encoding for the Stratification of Complex Disorders". AMIA 2019 Informatics Summit, March 2019, San Francisco. (Selected for podium presentation)

5. Isotta Landi, Riccardo Miotto, Paola Venuti, Joel T. Dudley and Cesare Furlanello. "Electronic Health Record encodings for neurodevelopmental disorder stratification". 3rd Annual MAQC Society Conference, 2019. 3rd place, best poster presentation.

6. Jurman, G., Maggio, V., Landi, I., Francescatto, M., Chierici, M., De Domenico, M., & Furlanello, C. "omicsCNN: a general deep learning framework for omics data modeling and classification". Human Genome Meeting Trainee Symposium, March 2018, Yokohama, Japan. (Award and selected for podium presentation)

7. Arianna Bentenuto, Anna Peripoli, Isotta Landi, Noemi Mazzoni, Teresa Del Bianco, Ilaria Basadonne, Liliana Carrieri, Stefano Cainelli, and Paola Venuti. "Males and Females with ASD: a gender comparison of cognitive and behavioral profiles". International Society for Autism Research Annual Meeting, 2018.

8. Teresa Del Bianco, Isotta Landi, Noemi Mazzoni, Ilaria Basadonne, Arianna Bentenuto, and Paola Venuti, "Towards a Clarification of Attention to Faces in Atypical Development: Sustained Attention to the Face is Task-Dependent in Autism Spectrum Disorder". International Society for Autism Research Annual Meeting, 2018.

9. Noemi Mazzoni, Jessica Rizzotto, Elena Tonelli, Arianna Bentenuto, Teresa Del Bianco, Ilaria Basadonne, Isotta Landi, and Paola Venuti. "From Emotion Production to Emotion Comprehension: An Emotional Training Proposal for Children with ASD". International Society for Autism Research Annual Meeting, 2018.

10. Ferrari, S., Reggianini, C., Gualtieri, F., Landi, I., Florio, D., & Petrone, A. M. C. "Impact of migration-related trauma on self-harm and suicide in migrants confined to prison: A prospective cohort study". 26th European Congress of Psychiatry/European Psychiatry 48S (2018) S453-S764; S603.

11. Giuseppe Jurman, Valerio Maggio, Diego Fioravanti, Ylenia Giarratano, Isotta Landi, Margherita Francescatto, Claudio Agostinelli, Marco Chierici, Manlio De Domenico, and Cesare Furlanello. “Convolutional neural networks for structured omics: OmicsCNN and the OmicsConv layer”. NIPS Workshop on Machine Learning in Computational Biology, 2017.

12. Arianna Bentenuto, Stefano Cainelli, Isotta Landi, Ferdinando Suvini, Roberta Igliozzi, Filippo Muratori, and Paola Venuti. “A joint behavioral and emotive analysis of synchrony in music therapy in children with ASD”. XI Autism-Europe International Congress, 2016.

13. Noemi Mazzoni, Teresa Del Bianco, Isotta Landi, Paola Ricciardelli, Rossana Actis-Grosso, and Paola Venuti. “The Level of Intelligence Modulates the Recognition of Point-Light Displays in Children with Autism Spectrum Disorder (ASD): A Comparison Between High Functioning and Low Functioning ASD”. International Meeting for Autism Research, 2016.

Preprints

14. Jurman, G., Maggio, V., Fioravanti, D., Giarratano, Y., Landi, I., Francescatto, M., Agostinelli, C., Chierici, M., De Domenico, M., & Furlanello, C. (2017). Convolutional neural networks for structured omics: OmicsCNN and the OmicsConv layer. arXiv preprint arXiv:1710.05918.

In preparation

- Landi, I., Miotto, R., Glicksberg, B. S., Lee, H., Cherng, S., Landi, G., Danieletto, M., Dudley, J. T., & Furlanello, C. Unsupervised deep representations of electronic health records for data-driven phenotyping and complex disorder stratification. *To be submitted to Nature Machine Intelligence.*

- Landi, I., Bentenuto, A., Maggio, V., Furlanello, C., & Venuti, P. Behavioral-cognitive embeddings for the stratification of Autism Spectrum Conditions. *To be submitted to PLOS One*

- Pasqualotto, A., Landi, I., & Venuti, P., Role of the WISC subtests in predicting Developmental Dyslexia severity. *To be submitted to Journal of Learning Disabilities*