Universiteit Leiden

Master Computer Science

Unsupervised machine learning methods to understand the social and psychological effects of prescription opioids

Name:	Ramya Tumkur Rameshchandra
Student ID:	s3306593
Date:	25/06/2024
Specialisation:	Artificial Intelligence
1st supervisor:	Prof. Dr. Marco Spruit
2nd supervisor:	Dr. Mitra Baratchi

Master's Thesis in Computer Science

Leiden Institute of Advanced Computer Science (LIACS)
Leiden University
Niels Bohrweg 1
2333 CA Leiden
The Netherlands

Acknowledgements

Working on this research project has been a profound learning experience. Firstly, I would like to express my sincere gratitude to my supervisor, Prof. Dr. Marco Spruit, for the invaluable guidance, support, and encouragement throughout the research journey. His expertise and patience have been instrumental in shaping the direction and quality of this work. I also thank Dr. Mitra Baratchi, my second supervisor, for offering her profound knowledge and critical insights that greatly enriched my research. I also thank Dr. Dennis Mook-Kanamori for suggesting the research topic and providing critical feedback that was vital in refining this work. A special thanks to Henk de Jong for helping me set up the working environment and answering my endless questions.

A heartfelt thanks goes out to my family and friends without whose unconditional love and support this would not have been possible. Lastly, I would like to express my appreciation to anyone who took an interest in my academic journey.

Abstract

Since the 1990s, there has been a rapid increase in overuse, abuse, and overdose deaths, along with the significant medical, social, psychological, demographic, and economic consequences associated with prescription opioids. Social and psychological effects are of particular interest because they extend beyond individual addiction to impact families, communities, and social systems, leading to issues such as mental health disorders, social isolation, and economic hardship. In this work, association rule mining is used on the ATC, ICPC codes, and patient demographics to draw interesting relationships. Specifically, Apriori and FP-growth algorithms were used to find frequent itemsets from which association rules were derived. A positive correlation was found between women and opioid prescriptions. Fentanyl and morphine were found to be positively correlated with dementia. Fentanyl, tramadol, and morphine were associated with memory, concentration, and orientation disorders, while any opioid prescription was also associated with sleep disorders. Tramadol and oxycodone prescriptions were associated with the loss or death of a partner. A negative association between prescription opioids and tobacco abuse was found. Furthermore, MCA, autoencoders, and temporal LSTMs were used for dimensionality reduction with the end goal of patient subtyping. However, a large number of clusters were found that could not be translated to clinical relevance. This is the first research to focus on using unsupervised machine learning to understand the social and psychological effects of prescription opioids in the Netherlands. The results underpin existing research, as well as uncover novel associations that can potentially fuel further research.

Contents

1 Introduction

The Leiden University Medical Center in Leiden is a university medical center for research, education, and patient care. The LUMC ELAN network is a collection of GP data and has been used consistently in risk prediction and population health analysis. An Electronic Health Record (EHR) is a digital version of a patient's medical chart. EHRs are real-time, patient-centric records that make information available instantly and securely to authorized users.
Opioids are a class of powerful pain-relieving medications. Opioids work by binding to specific receptors in the brain, spinal cord, and other parts of the body, reducing the perception of pain and producing feelings of relaxation and euphoria. In a clinical setting, opioids are commonly used for pain management. They are prescribed to alleviate moderate to severe pain resulting from various medical conditions, surgical procedures, injuries, or chronic pain conditions such as cancer-related pain or neuropathic pain. Additionally, opioids may be administered during anesthesia to induce sedation and manage pain during surgical procedures. Opioid agonists such as methadone and buprenorphine are used to treat opioid use disorder by reducing cravings and withdrawal symptoms. These medications activate opioid receptors to control cravings without producing the same euphoric effects as other opioids. Opioids are used more liberally to manage severe pain and suffering in end-of-life palliative care [32].
Opioid Use Disorder (OUD) is a pattern of opioid use that leads to clinically significant impairment or distress, as defined by the Diagnostic and Statistical Manual of Mental Disorders (DSM-5) [33, 69]. It is characterized by loss of control over opioid use, physical dependence, tolerance, and withdrawal symptoms when stopping [9]. OUD exists on a spectrum from mild to severe, based on the number of diagnostic criteria met. It is a multifaceted condition influenced by various factors. Genetic predisposition plays a role and certain individuals are more susceptible to addiction due to genetic variations that affect the opioid response. Environmental influences such as stress, trauma, and social instability can contribute significantly, as can co-occurring mental health disorders such as depression or anxiety, which can lead individuals to self-medicate with opioids. Overprescribing practices of healthcare providers, particularly for the treatment of chronic pain, have also fueled the opioid crisis and the rise of OUD. In addition, social attitudes towards substance use and cultural norms surrounding pain management play a role.
Since the 2000s, more than a million people have died in the United States from drug overdoses, the majority of which were due to opioids [1]. The rise in opioid use in the United States can be mainly attributed to three reasons: Overprescription of opioid medications in the 1990s, an increase in heroin use in the 2010s, and a dramatic spike in synthetic opioid (fentanyl) overdose deaths since around 2013. Although not as severe as in the United States, the trend in the Netherlands indicates an increase in prescription opioids [49]. The overall number of prescription opioid users in the Netherlands nearly doubled from 4,109 per 100,000 inhabitants in 2008 to 7,489 per 100,000 inhabitants in 2017. This increase was primarily driven by a quadrupling of the number of oxycodone users. During the same time period, the number of opioid-related hospital admissions tripled, while the number of patients in addiction care for opioid use disorders (other than heroin) nearly doubled. Since 2014, the number of opioid-related deaths has tripled. This change in opioid use is concerning. Opioid use disorder has been shown to disrupt all aspects of users' lives. OUD is associated with reduced physical health and reduced emotional well-being [33]. There have been ties between OUD and issues with personal relationships, crime, financial instability, and career (education) [7].
The Nederlands Huisartsen Genootschap [2] is the scientific association of and for general prac-

titioners (huisartsen) in the Netherlands. The NHG translates the International Classification of Primary Care (ICPC) code list to adapt it for use in Dutch general practice settings. ICPC codes are used to register and code patient episodes, symptoms, diagnoses, and procedures in the EHRs. The ICPC stands as the predominant international standard for systematically organizing and documenting clinical data in primary care settings [6].
It is evident that opioid use has far-reaching effects and the impacts need to be analyzed to understand the full consequences. However, little research has been done on the social and psychological effects of prescription opioids in the Netherlands. Issues that affect the social aspect of the patient's concerns are coded using 'Z' ICPC while those corresponding to psychological issues are coded using 'P' ICPC. The social effects include financial, food, housing, and work problems, to name a few. The psychological issues include depression, insomnia, sexual dissatisfaction, chronic alcohol abuse, tobacco use, among others. The complete list of P and Z ICPC codes is listed in Appendix A[1]. This thesis focuses on the social and psychological effects of prescription opioids. The purpose of this project is to examine the effects by using unsupervised machine learning algorithms to find patterns in patient data. The first part of the project focuses on using association rule mining to find the rules that occur most frequently. It can discover interpretable rule patterns that associate patient characteristics with ICPC outcomes. The second part of the project investigates patient subtypes. Subtyping can find distinct subgroups of patients accounting for the inherent heterogeneity in patient populations that is often missed when analyzing the aggregated data. For example, the subtypes of patients prescribed opioids could guide targeted therapies and improve overall treatment effectiveness (precision medicine).

1.1 Research Questions

This study will examine deidentified EHR data from the ELAN network in The Netherlands that span the years 2010-2019. The primary goal of the thesis is: *How can unsupervised machine learning be used to uncover insights from prescription trends and consequent social and psychological effects of opioids?*
Several unsupervised machine learning algorithms were considered in the context of uncovering patterns in opioid use. Association Rule Mining and Clustering were chosen for further investigation. Keeping in line with this, the following subquestions are coined:

1. Is opioid prescription affected by patient attributes such as gender, age, level of education, profession, and postal code?

2. Is there a relationship between the reporting of the ICPC code, the gender of the patient, and the prescription of medications?

3. Can patient subtypes be identified based on clinical characteristics, demographic factors, and diagnoses?

To answer the first and second subquestion, association rule mining is used to discover patterns and most frequently co-occuring items.
Clustering algorithms such as K-means and DBSCAN are used for patient subtyping. Dimensionality reduction is first achieved using LSTM autoencoders before performing clustering.

[1]The ICPC codes on NHG are translated from Dutch to English using Google Translate

1.2 Research Approach

This research follows the CRISP-DM (Cross-Industry Standard Process for Data Mining) framework, a comprehensive and widely adopted methodology for data mining projects [80]. It has 6 key phases: Business understanding, Data understanding, Data preparation, Modeling, Evaluation, and Deployment. It is an iterative process, and each step is performed several times. The general outline is given in figure 1. Several iterations of CRISP-DM were performed for

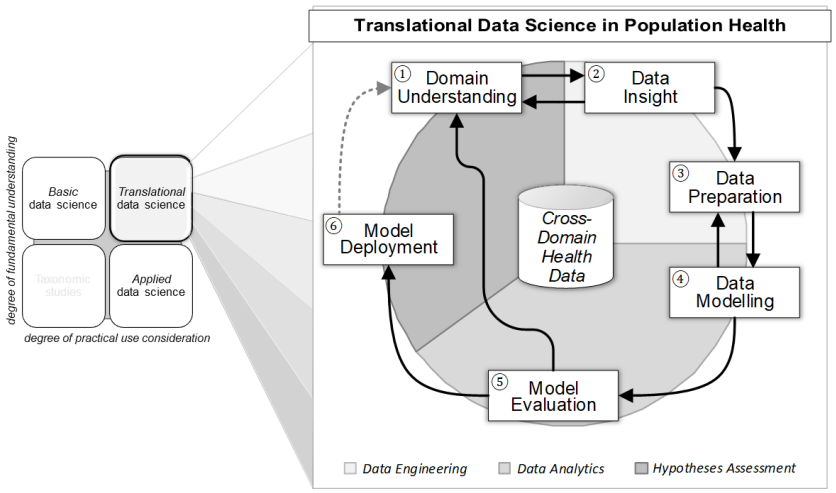


Figure 1: CRISP-DM[79]

this project but not all 6 steps were performed in all of the iterations. The application of CRISP-DM to this work for the first iteration can be elucidated as:

1. Business understanding: Several meetings were held with a physician to understand the motivation and objective of the research. A further literature review was conducted to understand the current state of research and possible knowledge gaps in the study of the effects of prescription opioids.

2. Data Understanding: The required data was discussed and collected with the help of the data manager. The ELAN codebook was studied to understand the information registered in each field of the EHR data. Preliminary data analysis was performed and the findings were verified with the data manager.

3. Data preparation: Data were processed to remove redundancies, incorrect entries, and null values. Important features were extracted and transformed for modeling.

4. Modeling: The experiments were carried out using association rule mining and clustering and are described in Experiments and Results, 6.

5. Model evaluation: The performance of the models and the insights obtained are discussed in Discussion, 7. The research questions are answered and analyzed.

6. Deployment: The models were not deployed in this project. Instead, the application of the results is described.

1.3 Outline

The rest of the report is structured as follows. In the second section 2, the theoretical background is presented. Section 3, elaborates on the related work and contains the literature review on the use of association rule mining in a medical setting and the use of machine learning in patient subtyping. Section 4 explores the data and describes the dataset, privacy, and statistical analysis. Section 5 details the methodology used for the experimental setup. Section 6 describes the results of the experiments, and Section 7 contains a detailed discussion of the results. Section 8 is the conclusion in which we answer the research questions. Lastly, Section 9 contains the limitations and future work.

2 Theoretical Background

This section discusses the concepts and frameworks that relate to this thesis.
Unsupervised learning is a type of machine learning algorithm that is used to discover patterns, relationships, or structures in unlabeled data. Unlike supervised learning, where the algorithm is provided with labeled training data (input-output pairs), unsupervised learning algorithms work with input data that do not have predefined labels or categories. Some of the unsupervised learning techniques include clustering, dimensionality reduction, anomaly detection, association rule mining, and independent component analysis. **Clustering** algorithms partition the input data into groups based on similarity or distance metrics. Popular examples are K-means clustering, hierarchical clustering, and DBSCAN (Density-Based Spatial Clustering of Applications with Noise). For low dimensionality, unsupervised learning can be tackled effectively. But the curse of dimensionality [23] sets in quickly with an increase in the number of features. To help with this, **Dimensionality Reduction** techniques aim at reducing the number of dimensions while preserving the information. PCA (Principle Component Analysis), t-SNE (t-distributed Stochastic Neighbor Embedding), and autoencoders are popular dimensionality reduction techniques. **Anomaly Detection** algorithms identify data points that deviate significantly from the rest of the data. Several techniques such as isolation forests, clustering, statistical methodologies, one-class SVMs, and neural networks can be used for anomaly detection. **Association Rule Mining** identifies frequent itemsets and generates rules using algorithms such as Apriori and FP growth to describe relationships between different attributes. **Independent Component Analysis** is a computational technique used to separate multivariate signals into additive components, assuming that the sources are non-Gaussian and mutually independent.
The goal of this work is to understand the patterns that occur between opioid use and medication in relation to the demographics of the patients. Keeping in line with the aim, unsupervised learning provides an advantage over other methodologies (namely, supervised learning, semi-supervised learning, and reinforcement learning) to provide novel insights. Unsupervised machine learning is particularly useful for descriptive tasks such as identifying meaningful trends and structures, finding groups in results, as opposed to predictive learning[74]. For example, supervised learning can be used to predict mortality in patients with cardiovascular disease [8]. However, it is not particularly useful for identifying the mechanisms of the disease. The complex multifactor interaction between different features helps in a better understanding of the disease and its effects. The experiments are separated into two parts: Association Rule Mining and Patient Subtyping. The related theory described in this section follows the same subdivision.

2.1 Association Rule Mining

Association rule mining is a data mining technique that is used to discover interesting relationships, patterns, or associations in transactional or relational datasets. A pattern is interesting if it is valid with some degree of certainty, novel, potentially useful, and easily understood by humans [44]. Measures of pattern interestingness can be used to guide the pattern discovery process.
ARM extracts rules from transactional datasets composed of a set of $I = \{I_1, I_2, ..., I_n\}$ attributes called *items* and a set $D = \{T_1, T_2, ..., T_m\}$ where $T_{1,2,...,m} \subset I$.
An association rule is generally an implication on a pair of itemsets X, Y where $X, Y \subset I$. They take the form $X \Rightarrow Y$ where X is called the *antecedent* and Y is called the *consequent*

and where $X \neq \emptyset, Y \neq \emptyset, X \cap Y = \emptyset$. It can be interpreted as: If items in X are present in a transaction, items in Y are likely present in the same transaction.
There are broadly two types of interestingness measures in pattern discovery. **Objective measures** are based on the characteristics of the data itself and the statistics that underlie them [67]. Some important objective measures for a rule $X \Rightarrow Y$ are as follows:

- **Support**: The support of a rule can be defined as the percentage of transactions in D that contain both the itemsets X, Y.

$$Support(X \Rightarrow Y) = \frac{|(X \cup Y)|}{|D|}, \qquad range : [0, 1] \tag{1}$$

- **Confidence**: Confidence measures the conditional property and is the percentage of transactions in D containing X that also contain Y.

$$Confidence(X \Rightarrow Y) = \frac{|(X \cup Y)|}{|X|}, \qquad range : [0, 1] \tag{2}$$

- **Lift** is a correlation measure between the antecedent and the consequent. It compares the observed frequency of the rule's occurrence to what would be expected if the antecedent and consequent were independent.

$$Lift(X \Rightarrow Y) = \frac{Support(X \cap Y)}{Support(X) * Support(Y)}, \qquad range : [0, \infty] \tag{3}$$

  A lift value greater than 1 implies a positive correlation; occurrence of X implies the occurrence of Y and vice versa. A lift value less than 1 indicates a negative correlation. A lift value of 1 indicates independence and there exists no correlation between X and Y.

- **Conviction** quantifies the degree of dependence between X and Y. It reflects how much the prediction of the consequent is improved by considering the antecedent, relative to what would be expected if the antecedent and consequent were independent.

$$Conviction(X \Rightarrow Y) = \frac{1 - Support(Y)}{1 - Confidence(X \Rightarrow Y)}, \qquad range : [0, \infty] \tag{4}$$

  Conviction is a positive real number. A value of 1 implies the independence between X and Y. The higher the value of conviction (> 1), greater the correlation between the antecedent and the consequent.

- **Leverage** measures the difference between the observed frequency of co-occurrence of X and Y and the frequency that would be expected if they were independent.

$$Leverage(X \Rightarrow Y) = Support(X \cup Y) - (Support(X) * Support(Y)), \qquad range : [-1, 1] \tag{5}$$

  A positive leverage value suggests a positive association, and a negative leverage value suggests a negative (or avoidance) association between the antecedent and the consequent. A value close to 0 implies independence.

Support and Confidence are considered the fundamental metrics and are used to select and rank the association rules. However, support and confidence cannot fully capture the interestingness. For example, an association rule with high support and confidence values could potentially be misleading if the antecedent and consequent have a high frequency of occurrence in the dataset. Therefore, correlation measures like lift, conviction, and leverage are essential to augment the process of finding interesting rules. A **frequent pattern** is one that satisfies the minimum support condition. **Rare patterns** are those that occur rarely but are important for analysis [54]. **Negative patterns** uncover those rules with a negative correlation between items [20].
Subjective interestingness measures depend on human judgement and domain knowledge. They are considered important if the observations confirm a previous hypothesis or if the observations contradict a current understanding of the domain. The latter is important in fueling new research and formulating novel hypotheses [17].

Association Rule Mining can be broadly described as a two-step process:

1. Find the frequent itemsets: These are itemsets that fulfill the minimum support criteria.

2. Find strong association rules from the itemsets found in the previous step: These rules are those that satisfy the minimum support and minimum confidence criteria.

The first step is extremely time-consuming and resource-intensive compared to the second step. A challenge frequently encountered when extracting frequent itemsets is the generation of an overwhelming number of itemsets, particularly when the support is low. If an itemset is frequent, all its combinatorial subsets are frequent as well. This results in an exponential increase in frequent itemsets. An itemset A is considered to be a **closed frequent itemset** in dataset D if there exists no proper superset B such that B has the same support in D as A. Several frequent itemset mining methodologies exist, with Apriori and FP-growth being two of the more popular algorithms. The next two sections discuss these two algorithms in detail.

2.1.1 Apriori Algorithm

Apriori is a frequent itemset mining algorithm first proposed by R. Agrawal and R. Srikant in 1994 to mine Boolean association rules [13]. It is an iterative approach in which itemsets of size n are analyzed to find itemsets of size n+1. In the first step, the frequent itemset of size 1 is found. The resulting set is denoted by L_1. This set in turn is used to find sets of consequently larger size based on the minimum support: $L_2, L_3, ...L_k$ until no more frequent k-itemsets can be found.
Apriori Property: If an itemset is frequent, all its nonempty subsets must also be frequent. This is the downward closure property of frequent itemsets. It belongs to a class of properties called *antimonotonicity* which implies that if a set fails a test, all its supersets will fail the same test as well.
Apriori algorithm works by iterating the following two steps until no more frequent itemsets can be found:

1. **Join**: A frequent itemset of size k, L_k is generated by joining L_{k-1} with itself. The resulting set is called the candidate set C_k, which is further analyzed. Apriori assumes that items within a transaction or an itemset are ordered lexicographically. Let l_1 and l_2 be two itemsets in L_{k-1} and $l_1[j]$ refers to the jth item in itemset l_1. l_1 and l_2 are joinable

only if their first k-2 items are the same. In addition, condition $l_1[k-1] < l_2[k-1]$ is applied (lexicographic order) to ensure that no duplicates are generated. The resulting itemset formed by joining l_1 and l_2 is $\{l_1[1], l_1[2], ..., l_1[k-2], l_1[k-1], l_2[k-1]\}$.

2. **Prune**: To find the frequent itemset of size k, L_k, from the candidate set of size k, C_k, the database needs to be scanned to find those itemsets that satisfy the minimum support. Since the candidate set could potentially be large, the Apriori property is employed to reduce the number of itemsets for which the support is calculated. Any itemset of size k-1 that is not a frequent itemset cannot be a subset of frequent k-itemsets. Therefore, if any k-1 subset of the candidate set C_k is not in $L_{(k-1)}$, it cannot be frequent and can be removed from the candidate set. Subset testing can be performed quickly using a hash tree of all frequent itemsets.

The pseudocode of the Apriori algorithm is given in algorithm 1.

Though Apriori algorithm is a powerful algorithm designed to find frequent itemsets, it is plagued by two drawbacks:

- It could possibly have to generate a very large number of candidate sets. For example, if $|L_1| = 10^4$, it would need to generate more than 10^7 number of C_2 candidate sets.
- For each candidate set, the database might have to be repeatedly scanned to calculate the support.

FP-growth algorithm provides a solution which is discussed in the next sub-section.

2.1.2 FP-growth Algorithm

Frequent Pattern [43] growth of simply FP-growth algorithm is a highly efficient algorithm for mining frequent itemsets and association rules from large databases. If efficiently discovers frequent itemsets without generating candidate itemsets explicitly, making it particularly effective for large datasets. The FP-Growth algorithm works by first constructing an FP-tree data structure from the input dataset, where each path from the root to a leaf represents a transaction. It then recursively mines frequent itemsets by traversing the FP-Tree and generating conditional pattern bases. This divide-and-conquer approach avoids generating candidate itemsets explicitly, making FP-Growth more efficient than Apriori. The FP-growth algorithm is given in algorithm 2.

2.1.3 Generating Association Rules from Frequent Itemsets

Extracting strong association rules from the frequent itemset is straightforward:

- For each frequent itemset *l*, generate all non-empty subsets.
- For each of these subsets *s*, generate the association rule $s \Rightarrow (l - s)$ if the confidence(equation 2) of the rule is above the minimum confidence threshold.

2.2 Patient Subtyping

Patient Subtyping, also known as patient segmentation or patient profiling, involves categorizing patients into distinct subgroups based on specific criteria. The goal of patient subtyping is to identify homogeneous groups of patients with similar clinical characteristics, disease profiles, risk factors, or treatment responses. By stratifying patients into meaningful subgroups, healthcare providers and researchers can tailor interventions, treatments, and care plans to the specific needs and characteristics of each subgroup, leading to more personalized and effective healthcare delivery [84].
In this project, patient subtyping is performed using clustering. Before clustering, dimensionality reduction techniques are applied to obtain a latent representation of the data.

2.2.1 One-hot Encoding

One hot encoding is a popular technique used in machine learning and data preprocessing to represent categorical variables as binary vectors. In this method, each unique category in the categorical variable is assigned a unique binary value, where only one bit is set to 1 and the rest are set to 0. This binary representation effectively creates a "one-hot" vector for each category, with the presence of a 1 indicating the category's presence and the 0s indicating its absence. This encoding preserves the categorical nature of the variable while allowing machine learning models to effectively leverage the information. One-hot encoding is particularly useful for algorithms that cannot directly handle categorical variables. However, it is important to note that one hot encoding may lead to a significant increase in feature dimensionality, especially for variables with a large number of unique categories, which can impact computational efficiency and model complexity.

2.2.2 Dimensionality Reduction

Dimensionality reduction is a fundamental technique in machine learning and data analysis aimed at reducing the number of features or variables in a dataset while preserving its essential characteristics [59]. By reducing the dimensionality of the data, dimensionality reduction methods can alleviate computational complexity and mitigate the curse of dimensionality by extracting latent features to facilitate better interpretation of the data. By reducing the number of features, dimensionality reduction can also help prevent overfitting and enhance the generalization performance of the models. Dimensionality reduction includes two main methodologies:

- Feature selection: Feature selection methods aim to select a subset of the original features that are most relevant to the prediction task while discarding irrelevant or redundant features. Common feature selection techniques include filter methods, wrapper methods, and embedded methods.

- Feature extraction: Feature extraction methods transform the original features into a lower-dimensional space using techniques such as Principal Component Analysis (PCA), Singular Value Decomposition (SVD), manifold learning, or Autoencoders.

2.2.2.1 MCA

Multiple Correspondence Analysis (MCA) [41] is a powerful multivariate statistical technique used for analyzing categorical data involving more than two variables. It is an extension of Correspondence Analysis (CA), which is designed to analyze the relationships between the

rows and columns of a two-way contingency table. MCA, on the other hand, can handle datasets with more than two categorical variables. The primary goal of MCA is to identify and visualize the underlying structure and associations among the categorical variables in a dataset. It achieves this by transforming the categorical data into a low-dimensional space by computing the principal components of the cross-tabulation matrix of the categorical variables and projecting the data onto these components.

2.2.2.2 Autoencoders

Autoencoders are artificial neural networks capable of learning a *latent representation* or codings of typically a lower dimension from the the input data. It is a type of unsupervised learning method. An autoencoder consists of two main components:

- Encoder: The encoder is a neural network that takes the input data and maps it to a lower-dimensional latent space. The encoder learns to extract the most salient features and patterns from the input, effectively compressing the data.
- Decoder: The decoder is another neural network that takes the latent representation and attempts to reconstruct the original input. The decoder learns to map the latent space back to the original high-dimensional input space.

The training of an autoencoder involves minimizing the reconstruction error between the original input and the reconstructed output. This encourages the autoencoder to learn a compact, meaningful representation of the data. Autoencoders can be extended and specialized for different tasks [40]:

1. Stacked autoencoders: Stacked autoencoders are a deep learning architecture that combines multiple layers of autoencoders, where the output of one autoencoder is used as the input to the next. This hierarchical structure allows the network to learn increasingly abstract and complex representations of the input data, enabling more powerful feature extraction and dimensionality reduction.
2. Convolutional Autoencoders: Convolutional autoencoders are a variant of autoencoders specifically designed for processing and reconstructing images. The encoder and decoder are regular CNN networks composed of convolutional and pooling layers. The encoder typically reduces the dimension of the data while increasing the number of feature maps while the decoder does the reverse. Convolutional autoencoders are highly effective for tasks like image denoising, compression, and anomaly detection.
3. Recurrent Autoencoders: Recurrent autoencoders are a variant of autoencoders that incorporate recurrent neural network (RNN) architectures, such as Long Short-Term Memory (LSTM) or Gated Recurrent Unit (GRU), in the encoder and decoder components, enabling them to effectively handle and reconstruct sequential, text, or time-series data.
4. Denoising Autoencoders: Denoising autoencoders learn useful features from noisy data. This is done by adding noise to the input and training the autoencoder to generate the original input. The noise can be pure Gaussian noise added to the inputs, or it can be randomly switched-off inputs.

5. Sparse Autoencoders: Sparse autoencoders are a variant of the standard autoencoder architecture that introduce sparsity constraints on the latent representation or encoding. It does so by learning a compact, interpretable, and meaningful representation of the input data by encouraging the network to activate only a small subset of the latent units for each input. The constraint is usually achieved by using a sigmoid activation function in the coding layer, use a large coding layer, or by adding a sparsity penalty term to the autoencoder's loss function.

6. Variational autoencoders: Variational Autoencoders (VAEs) are a type of generative autoencoder that introduce a probabilistic approach to the latent representation. VAEs have an encoder network that learns to map the input data to the parameters of a probability distribution in the latent space, typically a Gaussian distribution. The decoder network then attempts to reconstruct the original input from the sampled latent representation. During training, the latent representation is sampled from the distribution learned by the encoder, introducing stochasticity into the model.

Hyperparameter tuning in autoencoders is crucial, as it ensures that the model achieves its best performance in terms of reconstruction accuracy and feature learning [65].
The choice of architecture directly affects the ability of the autoencoder to reconstruct the input data [58]. More complex architectures may capture more intricate patterns but risk overfitting, while simpler architectures may be more efficient, but potentially less expressive.
Batch normalization normalizes the input to each layer in a neural network to have a mean of zero and a standard deviation of one. Accelerates training by allowing higher learning rates and reduces the internal covariate shift [38].
Activation functions introduce non-linearity into neural networks, enabling them to learn complex patterns and relationships in data. Choosing the right activation function depends on the specific problem, the architecture of the network, and the desired output characteristics [78]. Some of the popular activation functions are:

- **Tanh**: The tanh activation function, also called the hyperbolic tangent activation function, is used in artificial neural networks and transforms input values to produce output values between -1 and 1. Mathematically, it can be represented as:

$$tanh(x) = \frac{(e^x - e^{-x})}{(e^x + e^{-x})}$$

- **ReLU**: ReLU (Rectified Linear Unit) is a piecewise linear function that outputs the input directly if it is positive, otherwise it outputs zero. Mathematically, it is expressed as:

$$f(x) = max(0, x)$$

- **SELU**: SELU (Scaled Exponential Linear Unit) is an advanced activation function that tackles shortcomings of some activation functions. Its key innovation is built-in normalization: SELU automatically adjusts layer inputs to maintain a consistent statistical distribution (mean 0, standard deviation 1) throughout the network. This self-normalizing property can simplify neural architectures and potentially enhance performance. Mathematically it can be described as:

$$f(x) = \lambda x, \quad \text{if x} \geq 0$$

$$f(x) = \lambda\alpha(e^x - 1), \quad \text{if x<0}$$

where $\alpha \approx 1.6733$ and $\lambda \approx 1.0507$

2.2.2.3 LSTM

Long Short-Term Memory (LSTM) networks [45] is a type of recurrent neural network (RNN) architecture that is specifically designed to model and capture long-range dependencies and temporal dynamics in sequential data. Unlike traditional Recurrent Neural Networks, which often struggle to learn and retain information over long sequences due to the vanishing gradient problem, LSTMs are equipped with specialized memory cells and gating mechanisms that enable them to effectively handle long sequences. The primary advantage of using LSTMs is

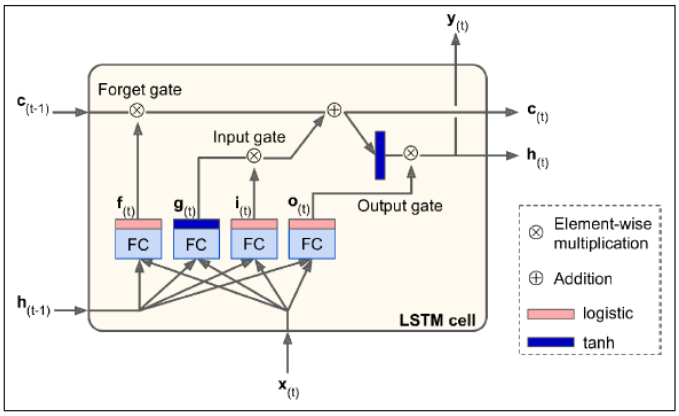


Figure 2: LSTM cell[40]

their ability to store long-term and short-term memory separately, which in turn diminishes the occurrence of exploding and vanishing gradients. The 3 main gates of the LSTM are:

- Forget gate $f_{(t)}$: determines what percentage of the long-term memory needs to be reduced.
- Input gate $i_{(t)}$: determines what percentage of the potential memory to add to the long-term memory.
- Output gate $o_{(t)}$: determines the short-term memory in the current time step.

The equations governing the LSTM computations are given below:

$$
\begin{aligned}
i_{(t)} &= \sigma(W_{xi^T} x_{(t)} + W_{hi^T} h_{(t-1)} + b_i) \\
f_{(t)} &= \sigma(W_{xf^T} x_{(t)} + W_{hf^T} h_{(t-1)} + b_f) \\
o_{(t)} &= \sigma(W_{xo^T} x_{(t)} + W_{ho^T} h_{(t-1)} + b_o) \\
g_{(t)} &= tanh(W_{xg^T} x_{(t)} + W_{hg^T} h_{(t-1)} + b_g) \\
c_{(t)} &= f_{(t)} \bigotimes c_{(t-1)} + i_{(t)} \bigotimes g_{(t)} \\
y_{(t)} &= h_{(t)} = o_{(t)} \bigotimes tanh(c_{(t)})
\end{aligned}
$$

where,

- $W_{xi}, W_{xf}, W_{xo}, W_{xg}$ are weight matrices of each of the four layers for their connection to the input vector $x_{(t)}$.
- $W_{hi}, W_{hf}, W_{ho}, W_{hg}$ are weight matrices of each of the four layers for their connection to the previous short-term state $h_{(t-1)}$.
- b_i, b_f, b_o, b_g are the bias terms for each of the four layers.

2.2.2.4 Temporal LSTMs

Traditional Long Short-Term Memory (LSTM) networks are designed to handle sequential data with uniform time intervals between consecutive elements. However, many real-world sequential datasets, such as patient medical records or business process event logs, exhibit irregular time intervals between the data points. To address this issue, researchers have proposed Time-Aware LSTMs (T-LSTMs) introduced by Baytas et al. [22], which extend the standard LSTM architecture to explicitly incorporate the elapsed time between consecutive elements in the sequence. T-LSTM incorporates the elapsed time between two ATC codes for a patient. T-

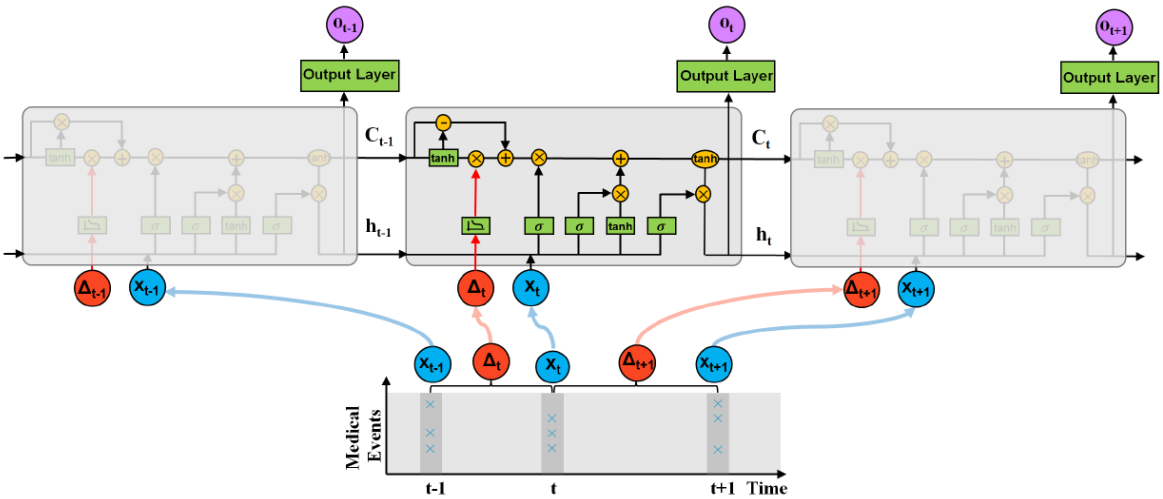


Figure 3: Temporal LSTM[22]

LSTM applies the memory discount by employing the elapsed time between successive elements to weigh the short-term memory content. A nonincreasing function is used:

$$g(\Delta_t) = \frac{1}{log(e + \Delta_t)} \tag{6}$$

The elapsed time is used to adjust the long-term memory component (and consequently the short-term memory at the output gate). This is done by calculating C^S_{t-1}, $\hat{C}^S_{t-1}$, C^T_{t-1} and C^*_{t-1} (given in 7-10) in order, which is then used as the adjusted long term memory feeding into a regular LSTM cell. In addition to forget, input, and output gates, the TLSTM uses the following equations to adjust the cell state.

$$C^S_{t-1} = tanh(W_d C_{t-1} + b_d) \tag{7}$$

$$\hat{C}^S_{t-1} = C^S_{t-1} * g(\Delta_t) \tag{8}$$

$$C^T_{t-1} = C_{t-1} - C^S_{t-1} \tag{9}$$

$$C^*_{t-1} = C^T_{t-1} + \hat{C}^S_{t-1} \tag{10}$$

$$\tilde{C} = tanh(W_c x_t + U_c h_{t-1} + b_c) \tag{11}$$

$$C_t = f_t * C^*_{t-1} + i_t * \tilde{C} \tag{12}$$

$$h_t = o * tanh(C_t) \qquad \text{Current hidden state} \tag{13}$$

2.2.3 Clustering

Clustering is a fundamental technique in data mining and machine learning that involves grouping a set of objects or data points into clusters based on their similarities. It is an unsupervised learning method, meaning that the data are not labeled or classified beforehand. Instead, the algorithm identifies patterns and structures within the data, forming clusters of similar objects. This similarity is typically measured by a distance metric or similarity function, which quantifies the degree of closeness or dissimilarity between data points [86]. Clustering algorithms aim to optimize one or more objective functions, such as minimizing the variance within the cluster or maximizing the variance between clusters.

2.2.3.1 K-means

K-means is one of the most widely used clustering algorithms. It partitions the data points into K clusters with each data point in a cluster being closest to the mean data point of the cluster. It works by minimizing the distance between data points within the cluster. The most common distance metric used is the sum of squared errors:

$$SSE = \sum_{k=1}^{K} \sum_{x_i \in C_k} ||x_i - \mu_k||^2 \tag{14}$$

where K is the number of clusters, C_k is cluster k and, μ_k is the cluster center of C_k.
It is an iterative process that updates the cluster assignment at each step. The algorithm follows as below:

1. **Initialization**: Select K random points as the cluster centers, $\mu_1, \mu_2, ..., \mu_K$
2. **Assignment**: For each data point x_i, compute its distance from each cluster center and assign it to the closest cluster
3. **Update**: Recalculate the center of each cluster as the mean of all data points belonging to the respective clusters
4. **Iterate**: Repeat steps 2 and 3 until there is no change in cluster membership

K-means is one of the simplest and widely used clustering algorithms. It has a time complexity of $O(n * K * t * d)$ where n is the number of data point, K is the number of clusters, t is the number of iterations and, d is the dimension of the data. It is computationally efficient for large datasets but can find local optima in a limited number of permutations [15, 70].
K-means requires K to be pre-defined which can pose a challenge when the structure of the data is unknown. K-means is sensitive to initialization of initial cluster centers and the hyperparameters can significantly impact the clusters found [26]. It finds spherical clusters with similar variance that are linearly separable [72]. Another drawback of K-means is that it is sensitive to outliers and the presence of which can distort the clustering result.
To find the optimal K, there are several methods. The elbow method is one such algorithm. It is a heuristic technique which works by calculating the sum of squares within the cluster (SSE) for each cluster. The SSE is plotted on the y-axis, and the corresponding K value is plotted on the x-axis. The plot is typically a decreasing curve in which SSE decreases rapidly with K initially and gradually tapers. The point at which this occurs, the elbow, is considered optimal K. The rationale behind it is that increasing the number of clusters beyond K does

not significantly improve the modeling of the data.
The silhouette coefficient is a measure of how well each data point fits into its assigned cluster compared to other clusters. It ranges from -1 to 1, where a high value indicates that the data point is well-matched to its own cluster and poorly matched to neighboring clusters. For each value of k, the average silhouette coefficient is calculated using the formula:

$$silhouette\_average = \frac{1}{n} * \sum_{i=1}^{n} \frac{b(i) - a(i)}{max(a(i), b(i))} \tag{15}$$

where n is the number of data points in the cluster, a(i) is the average distance between data point i and all other points in its cluster and, b(i) is the minimum average distance between data point i and points in all other clusters.
The average silhouette score for each cluster is plotted again K. The optimal value of K is the one that maximizes the average silhouette coefficient.
Another method for obtaining the optimal value of K is Bayesian Information Criterion (BIC). BIC adds a coefficient which penalizes large values of K, effectively preventing overfitting or underfitting the model. BIC for a model with K clusters is given by:

$$BIC(k) = n * log(RSS/n) + k * long(n) \tag{16}$$

where n is the total number of data points, RSS is the residual sum of squares.
The optimal value of k is the one that minimizes the BIC score, representing the best trade-off between model fit and complexity.

2.2.3.2 DBSCAN

DBSCAN is a density-based clustering algorithm commonly used for identifying clusters of arbitrary shapes [36]. DBSCAN can find clusters of different variances and does not require the number of clusters to be predetermined. The key parameters of DBSCAN are ϵ (the neighborhood radius) and minPts (the minimum number of points required to form a dense region). The choice of these parameters can significantly impact the clustering results. DBSCAN uses 3 types of points:

- **Core points**: A data point is considered a core point if it has at least a specified number of neighboring points (MinPts) within a specified radius (ϵ).
- **Border points**: A data point is classified as a border point if it lies within the ϵ radius of a core point but does not have enough neighboring points to be considered a core point itself.
- **Noise point**: Data points that are neither core points nor border points are classified as noise points or outliers.

DBSCAN performs the following steps to find clusters:

1. **Neighborhood Computation**: For each point p in the dataset, the algorithm computes the ϵ-neighborhood of p, which is the set of points within a distance ϵ from p
2. **Core Point Identification**: A point p is classified as a core point if its ϵ-neighborhood contains at least minPts points. Core points are potential cluster centers

3. **Cluster Formation**: The algorithm starts with an arbitrary unvisited core point p and retrieves all points that are density-reachable from p. These points form a new cluster, and the process is repeated for all unvisited core points

4. **Noise Handling**: Points that are not density-reachable from any core point are classified as noise or outliers

A point p is **density-reachable** from a point q if there exists a chain of points $p_1, p_2, ..., p_n$ such that $p_1 = q, p_n = p$, and for every $p_i (1 \leq i \leq n)$, the neighborhood of p_i within a given radius ϵ contains at least a minimum number of points (minPts).
DBSCAN can effectively handle noise and outliers. However, it has a computational complexity of $O(n^2)$ in the worst case, where n is the number of data points.

3 Related Work

Opioids are a class of drugs that act on opioid receptors in the body to produce analgesic (pain relief), sedative, and euphoric effects. Opioids can be natural, synthetic, or semi-synthetic compounds and are often prescribed for acute and chronic pain conditions. Opioids exert their effects primarily by binding to specific receptors in the brain and spinal cord, known as mu, delta, and kappa receptors [16]. **Natural Opioids** are derived from the opium poppy plant. Examples include morphine and codeine [64]. **Semi-synthetic opioids** are derived from natural opioids. Some examples are oxycodone, hydrocodone, and heroin. **Synthetic opioids** are synthesized in the laboratory and include fentanyl and methadone. Fentanyl is potent and is often used in anasthesia and to manage severe pain. However, opioids also carry a significant risk of tolerance, dependence, and addiction, as well as potentially life-threatening side effects such as respiratory depression. Therefore, their use must be carefully monitored and managed by healthcare professionals.

Opioid use disorder (OUD) is defined as a problematic pattern of opioid use that leads to clinically significant impairment or distress [33]. OUD encompasses a spectrum of behaviors, from problematic opioid misuse to severe addiction, and is associated with morbidity, mortality, and social burden [3, 27].

Machine learning has been consistently used on medical data for predictive analytics, personalized medicine, clinical decision support, drug discovery, reduced healthcare costs, and clinical research [19]. Several studies have been conducted to understand the effects of prescription opioids on the population. In particular, issues of opioid use disorder and the risk of addiction and overdose in patients have been investigated. One study examined the performance of different machine learning algorithms in predicting the risk of opioid overdose among medicare beneficiaries with opioid prescriptions [57]. It evaluated data from cancer-free Medicaid beneficiaries using multivariate logistic regression (MLR), least absolute shrinkage and selection operator (LASSO), random forest, gradient boost machine (GBM) and deep neural networks (DNN). A total of 268 potential predictors were measured, including sociodemographics, health status, patterns of opioid use, and practitioner-level and regional-level factors. The algorithms, DNN and GBM in particular, were able to accurately predict the risk of opioid overdose, providing valuable information for healthcare providers and policymakers to implement targeted interventions and preventive strategies. A study in the United States [34] used machine learning to understand the biomedical profile of opioid-dependent patients and identify those at risk of developing dependence or overdose from EHR data from millions of patients. A random forest classifier was trained to classify patients according to their likelihood of having a substance dependency diagnosis, using a control group without substance-related conditions matched by age, sex, and certain diseases. It found that the diagnosis of substance dependence was co-occurring with psychiatric comorbidities. Furthermore, a 2020 study analyzed 10 million insurance claims from 550000 patients to model an algorithm for the prediction of the early diagnosis of OUD [76]. Word2Vev was utilized to obtain an embedding representation of medical codes, which was analyzed using the XGBoost algorithm. The model found that the number of days of opioid use, the number of overlaps in opioid prescription, the number of opioid, benzodiazepine, and muscle relaxant prescriptions strongly predicted a positive diagnosis of OUD. The model provided a mean of 14.4 months reduction in diagnosis time. A similar study analyzes the risk of OUD in US adolescents [42]. 3 ML models were used: artificial neural networks, distributed random forest, and gradient-boosting machine. The performance of the models was similar and showed that the ML models were useful in the prediction of OUD. Of

the 41,579 adolescents aged 12 to 17 years in the study sample, 3.7% had OUD.
Although numerous studies have tried to understand the cause of OUD, not many studies have focused on the psychological and social effects of prescription opioids. There are a few studies that analyze the correlation between psychological comorbidity and opioid prescription. However, the social effects have not been studied. Machine learning has not been used to address the problem in most cases.

3.1 Association Rule Mining

The term 'Association Rules' was coined in a 1966 paper which presented a method (GUHA - General Unary Hypotheses Automaton) for automatically generating hypotheses from empirical data [47]. Association rule mining was popularized in [14] as an efficient algorithm to discover significant association rules from large transaction databases.
Association rule mining (ARM) has expanded from its original use of market basket analysis to many real-world applications. It has numerous applications in medicine and healthcare: clinical decision support, disease prediction and diagnosis, treatment optimization, public health surveillance, clinical research, and management of healthcare resources [18]. A study used ARM to discover patterns between blood patterns and disease history in a young adult population with myocardial infarction [55]. It was able to determine that glucose, smoking, triglyceride total cholesterol, and creatine were associated with diabetes and hypertension. Medical data is often distributed across different sites, making it challenging to perform data mining tasks on combined data. To address this, a study [51] introduced a two-phase approach for ARM: 1. Local mining at each site to generate frequent itemsets (co-occurring medical conditions, symptoms, etc.) using the Apriori algorithm, 2. Global mining by combining locally generated frequent itemsets and applying a modified Apriori algorithm to discover global association rules across all sites. The proposed methodology is applied to predict heart disease using association rules from distributed medical databases containing patient records. The discovered association rules reveal relationships between various factors (e.g. age, sex, symptoms, lifestyle) and the risk of developing heart disease. In a similar vein, Apriori was used to generate association rules that revealed relationships between primary diagnoses (the reason for the patient's visit) and secondary diagnoses (co-occurring conditions) [61]. The patterns and associations discovered could contribute to improved patient care, early diagnosis, and better management of complex medical conditions.
Rare rules often represent unique or unexpected associations between items or attributes in the dataset. Discovering such rules can lead to novel insights and uncover hidden patterns that may not be apparent from frequent itemsets alone. They may highlight an anomaly or provide a more nuanced understanding of the dataset. Association rules are considered rare when they have low support but high confidence, and the search space could potentially grow exponentially. A novel tree structure called Rare Pattern Tree (RP-Tree) was introduced to compactly store rare patterns and their support information [46]. The RP-Tree is built by inserting transactions from the database into the tree structure, with each node in the tree representing an item. The algorithm performs a depth-first traversal of the RP-Tree to generate candidate rare patterns. Effective pruning techniques based on the Apriori principle are employed to reduce the search space and eliminate redundant computations. In a similar study, proposed assigning different minimum support thresholds to different items or itemsets based on their rarities or importance [53].
ARM has been used to identify and validate new comorbidities associated with OUD [66].

OUD comorbidities from the FDA Adverse Event Reporting System (FAERS) database were extracted using a FP-growth algorithm in 12 million case reports. The EHR-based case-control study confirmed significantly increased risks for hypothyroidism (AOR=1.45), hyperthyroidism (AOR=1.46), and type 2 diabetes (AOR=1.28) in patients with OUD compared to those without OUD. Similarly, a study used the National Inpatient Sample Database (NIS), which contains inpatient discharge records from approximately 20% community hospitals in the United States to discover frequent patterns and associations between opioid use and various medical conditions [52]. The authors used the Apriori association rule mining algorithm to discover frequent patterns and association rules. The study identified several medical comorbidities strongly associated with opioid use, including chronic pain conditions, mental health disorders, and substance use disorders.

This work focuses on applying Association rule mining on prescription opioid data, which is a novel analysis. Some studies have used ARM to understand the relationship between opioid use, OUD, and other morbid conditions. The work in this thesis attempts to find the social and psychological effects of prescription opioids.

3.2 Patient Subtyping

Subtyping is the task of classifying a disease into distinct subgroups of patients based on specific characteristics of the patient, which can then guide treatment decisions based on the subgroup to which a patient belongs [84]. Patient subtyping enables the development of personalized and more effective treatment strategies tailored to each subgroup's unique characteristics and needs. For example, subtyping alcohol-dependent subjects is crucial in clinical trials to avoid overlooking the usefulness of effective drugs when tested in a heterogeneous population [56]. A study used longitudinal clinical records to subtype patients with Parkinson's disease (PD) according to their disease progression patterns [88]. A Long Short-Term Memory (LSTM) deep learning algorithm was used to represent each patient as a multidimensional time series. Three distinct PD subtypes were identified in 466 patients with idiopathic PD. A study analyzed data from the 2001-2002 National Epidemiologic Survey on Alcohol and Related Conditions (NESARC), which included 43,093 adults in the United States [85]. Latent class analysis (LCA) was used to identify subtypes of nonmedical opioid users based on their patterns of substance use and psychiatric disorders. Multinomial logistic regression procedures were employed to examine the characteristics of the identified subtypes.

Adjusted Clinical Groups (ACG) is a risk adjustment methodology used in healthcare to assess and manage patient populations [4]. Developed by researchers at Johns Hopkins University, ACG is based on the premise that people with similar patterns of healthcare utilization and clinical diagnoses have similar healthcare needs and costs. ACG assigns each patient a set of morbidity and utilization measures called Resource Utilization Bands (RUBs) and Adjusted Clinical Groups (ACGs), respectively. RUBs indicate the level of morbidity or disease burden for each patient, while the ACGs classify patients into categories based on their expected healthcare utilization patterns. ACG has been widely used in healthcare care management, population health management, risk adjustment for payment models, and research. It helps healthcare organizations identify high-risk patients, tailor interventions to individual patient needs, and allocate resources more efficiently.

Extensive work has been done to subtype patients in their opioid usage patterns. One study compared different clustering approaches based on k-means to group longitudinal opioid use trajectories extracted from EHR data [63]. The R package *kml* was used for clustering and

the Calinski and Harabasz criterion to choose the optimal k. LSTM networks were also used to obtain a latent representation on which clustering was performed. The resulting subtypes are interpreted using decision trees to understand how each subtype is influenced by factors such as characteristics of opioid medications, patient diagnoses, procedures, and demographics. Another study examined different subtypes of opioid craving trajectories during medication-assisted treatment for opioid use disorder (MOUD) and studied their association with treatment outcomes [25]. Support Vector Machines (SVM) were trained to predict subtype membership using pretreatment data and early Ecological Momentary Assessment (EMA) reports. The subgroups found differed in critical outcomes, including drug use lapse, stress, and exposure to drug cues.

Studies have been conducted to group patients according to their opioid medication. However, these studies did not consider the temporal aspect of prescriptions. This work uses a temporal LSTM to find latent representation and mine clusters based on prescription trends and the social and psychological issues.

4 Exploratory Data Analysis

The data set used for this project comes from an anonymized primary care data set named the Extramural Leiden Academic Network (ELAN) from LUMC. It contains data from 100+ GP practitioners in the Leiden and The Hague area. Different aspects of patient care are included such as patient demographics, episode, journals, medication, laboratory, and correspondence.

4.1 Dataset

Three data files were requested for the investigation:

- Patient: This file contains details of the demographics of the patient such as gender, birth year, education, occupation, etc.
- Episode: The episode file lists patient visits to the GP, with recorded details including the ICPC code, episode type, and start and end dates of the episode.
- Medication: The medication file includes all the medications prescribed in an episode. It contains information on the date of prescription, ATC codes, label names, dose, and duration, to name a few.

The inclusion criteria for the data set used in this research are patients with diagnosis of ICPC codes P and Z between 01/01/2010 and 31/12/2019. Therefore, the primary filter used are the ICPC codes in patient episodes, in the chosen period of time. Patient details and medication are extracted using the patient ID from the resulting episodes list. The prescription of opioids was ascertained by the occurrence of **N01AH*** and **N02A*** ATC codes.

4.2 Data Privacy

The ELAN dataset is hosted by STIZON dataset [10] which is a Trusted Third Party (TTP) for national healthcare data from various healthcare providers. In accordance with the data transfer agreement, the data are stored on a PHEG departmental drive and are accessed only through the LUMC network.

4.3 Statistical Analysis

The data contain the details of a total of 313776 patients. An overview is provided in Table 1. A preliminary scan of the data revealed a large number of null values in many data fields.

Data	Count
Patient	313776
Medication	1177433
Episode	1003117

Table 1: Overview of data

Tables 2, 3, 4 show the field name and number of non-null count for patient, medication and episode files, respectively.

Field	Description	Non-null Count
PATNR	Patient Number	313776
PRAKNR	Practice Number	313776
Woonverband	Residential number	292757
dWoonverbandsoort	Type of residence	17324
dWoonverbandpositie	Position in patient's living environment	19650
dPostcodecijfers	Postcode	312105
iGeboortejaar	Year of birth	313776
iOverlijdensjaar	Year of death	757
dGeslacht	Gender	313776
Thuisland	Country	8173
dBurgerlijkeStaat	Marital status	15430
Beroep	Profession	1557
dOpleiding	Education	59
dInschrijfdatum	Date of registration with healthcare provider	309696
dUitschrijfdatum	Date of deregistration with healthcare provider	59273
dRedenVertrek	Reason for deregistration	55196

Table 2: Patient - Non-null count

Field	Description	Non-null Count
PATNR	Patient Number	1177433
PRAKNR	Practice Number	1177433
EpisodeID	Episode Number	1119492
dVoorschrijfdatum	Prescription date	1177433
dEinddatum	Date until which medicine was prescribed	1076319
dStopdatum	Date on which medication was stopped	8718
Etiketnaam	Name of medicine	1135055
dPRK	KNMP-Prescription code	1067096
dGPK	KNMP-Generic product code	1067325
dATC	Code ('Anatomisch Therapeutisch Chemisch Classificatie')	1177433
dChronisch	Chronic medication indicator	51096
dDuur	Prescription length in days	51466
dIteraties	Number of permitted repeat prescriptions	2251
dHoeveelheid	Quantity	1177075
Dosiscode	Dosage code	1158590
dSterkte	Strength of drug	1007153
dToedieningomschrijving	Route of medication administration	802138
dVoorschriftICPC	ICPC code of diagnosis for which medication was prescribed	295500
dEpisodeICPC	ICPC code of episode	437826
dSpecialisme	Medical speicality of the prescriber	1044430

Table 3: Medication - Non-null count

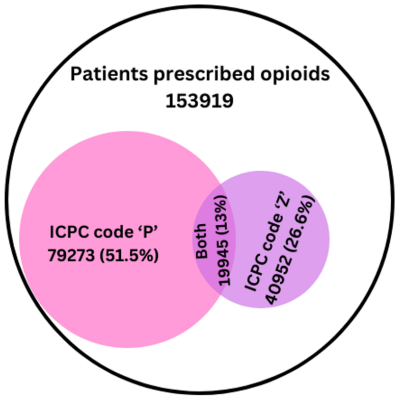
It is evident that the percentage of non-null inputs for several of the fields is very low. Imputation, the process of replacing missing values in a dataset with estimated values, can be

Field	Description	Non-null Count
PATNR	Patient Number	1003117
PRAKNR	Practice Number	1003117
EpisodeID	Episode Number	1003117
dBegindatum	Start date of episode	1003117
dEinddatum	End date of episode	166780
dMutatiedatum	Date of last change in episode registration	900466
dICPC	ICPC code registered during episode	1003117
dEpisodetype	Indicator for the type of episode	992696
dActief	Episode activity indicator	900466
dAttentie	Indicator for the attention value of the episode	55257

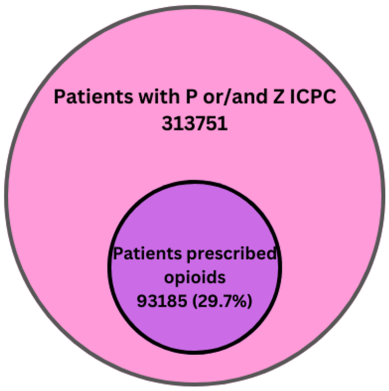
Table 4: Episode - Non-null count

problematic when the non-null count is low. Imputing missing values based on a small sample is unlikely to accurately represent the true underlying distribution of that variable. In this case, common imputation method assumptions [35](multivariate normality, missing at random) are more likely to be violated and any errors in the imputation are propagated, which in turn could create bias in the results. In EHRs, missing data usually occurs due to data entry errors, irregularity in fields recorded, and patient health status, making it difficult to model. Therefore, since the percentage of observed values in many fields falls below 5-10%, imputation is not recommended for this research. Instead, the fields with large null values are dropped and are not considered further.
The next step is to check for duplicate entries. There were 2 duplicate entries in the medication file, which were removed.
To perform further analysis of the data, the patient, medication, and episode files were merged. It is interesting to note that no correlation could be found between the 'Episode ID' fields in the medication and episode files. Therefore, the merging was performed on Patient ID (PATNR field). Furthermore, only the parent ICPC code was considered. For example, an ICPC code of P06.01 would be truncated to P06 and an ICPC code of Z16.03 would be truncated to Z16.
Between 2010 and 2019, a total of 153919 patients received opioid prescriptions, of which 60.5% presented ICPC codes P or Z. Specifically, 26.6% of those patients prescribed opioids were recorded to have at least one count of the Z ICPC code; 51.5% of those patients prescribed opioids were recorded to have at least one count of P ICPC code. Of the 313751 patients with either P/Z ICPC recorded, 93185 were prescribed opioids (29.7% of the patients). The details are shown in the Euler diagrams in Figure 4 .

The patient gender receiving the majority of opioid prescriptions is female as shown in figure 5.
The distribution of birth years of patients prescribed opioids follows a bell curve with majority of the patients being born around the 1960s as shown in figure 6.
In the cohort with P/Z ICPC codes and prescribed opioids, 24 different ATC codes are present as shown in figure 7. Similarly, 70 different ICPC codes are recorded as shown in figure 8.



(a) Percentage distribution of patients prescribed opioids by ICPC codes



(b) Percentage distribution of patients with ICPC codes

Figure 4: Patient Distribution

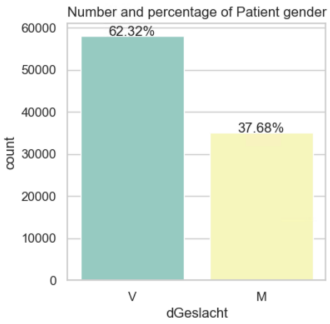


Figure 5: Gender distribution in patients prescribed opioids. 'V' corresponds to female, 'M' corresponds to male

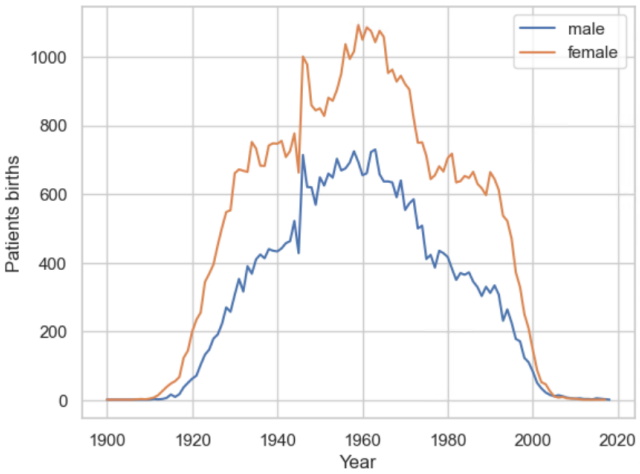


Figure 6: Birth year distribution for male and female patients

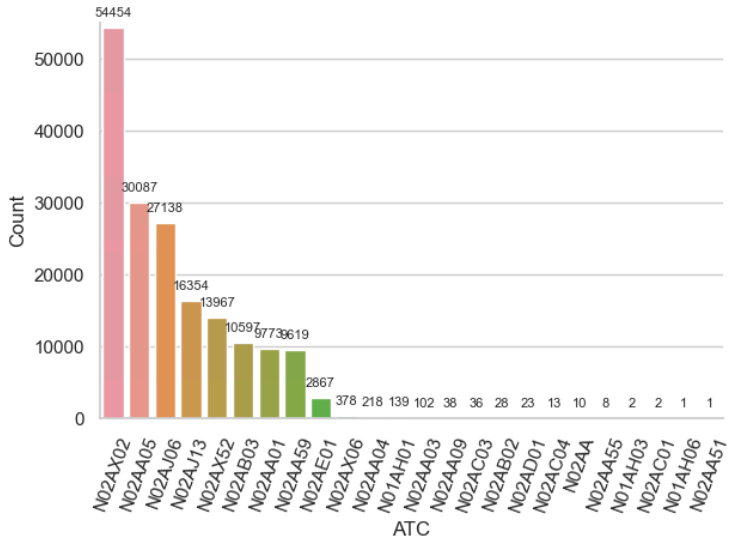


Figure 7: ATC codes

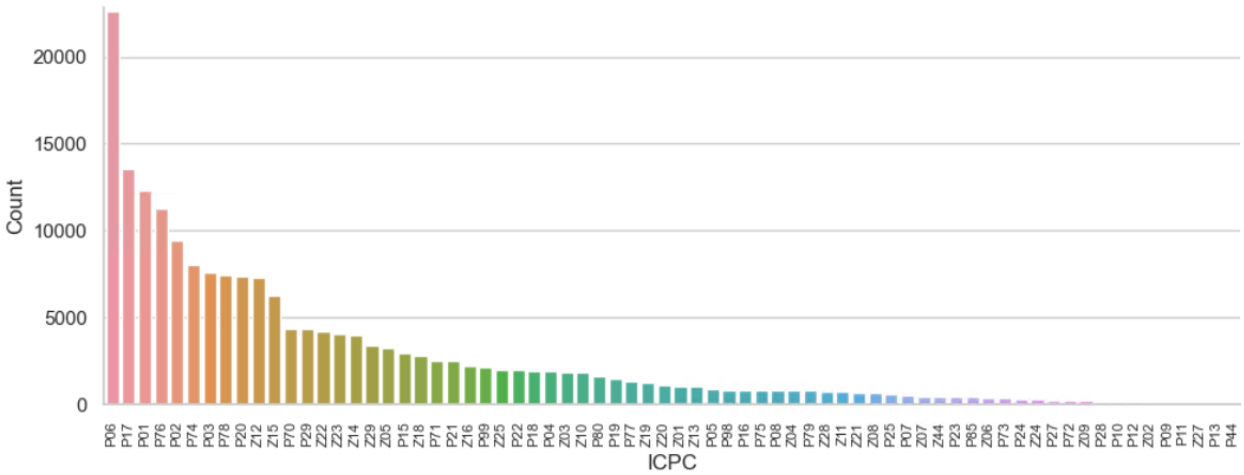


Figure 8: ICPC codes

5 Methodology

This section describes the experimental and assessment setup. It describes the data cleaning and manipulation used to prepare it for analysis. The first section provides an overview of the methods used in association rule mining. The second subsection describes how patient subtyping is achieved.
The analysis is performed using Python programming language. Data analysis used the *Pandas* library extensively. The raw data was provided as CSV files, which are cleaned and summarized in Panda's *DataFrames*. Each row of the dataframe corresponds to a patient and contains details about their demographics, medication, and diagnosis codes.
Association rule mining as well as artificial neural networks require data to be in a certain format. The raw data and the preprocessing steps are described in this section.

5.1 Association Rule Mining

The Python library *MLXTEND* is used to find the frequent itemsets and the corresponding association rules. Data in the form of tuples containing patient gender, ICPC codes and ATC codes is transformed into the required form using the library's *Transaction Encoder*. Frequent itemsets are discovered before finding association rules. A high-level overview of the methodology is given in diagram 9.

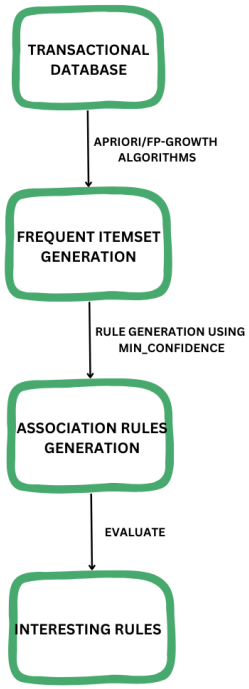


Figure 9: Association Rule Mining for frequent itemsets

5.1.1 Reducing the number of Association Rules

Association rule mining often finds a large number of rules due to its inherent nature of exhaustively searching for all possible associations within the dataset. To find interpretable

rules which can be used for hypotheses discovery, the number of association rules has to be reduced. Domain knowledge and statistical rulings are used to achieve this:

- As seen in table 5, there exist pairs of rules where the Antecedent and the Consequent are interchanged with the same Support and Lift values. In such cases, only the rule with the higher Confindence is retained.
- To reduce redundancy, two rules were applied such that the Antecedent and the Consequent followed certain patterns. Specifically, only those rules were retained wherein both the Antecedent and the Consequent followed either of the two regexes:
  `r'^\(((,\s)?(\'N[a-zA-Z0-9]+)\')+\,?\)\$'` and,
  `r'^\(((,\s)?\'(M|V|P[0-9]+|Z[0-9]+)\')+\,?\)\$'`. This corresponds to entries that contain only ATC codes and a combination of gender and ICPC codes, respectively.
- The opinion of the domain expert was obtained; opioid medications tend to be prescribed with each other. The occurrence of solely ATC codes in both Antecedent and Consequent is expected and can be excluded from the analysis. Therefore, such rules were eliminated.

5.1.2 Association Verification

The association rules discovered could uncover a potential novel relationship. However, it is necessary to ensure that the associations found occur only in the cohort that has been prescribed opioids. To achieve this, association rule mining is also performed in the general population that had P/Z ICPC codes but were not prescribed opioids.
Associations that occur in the prescribed opioid cohort as well as in the general population are not considered interesting. Rules that occur in the former group but not in the latter and vice versa are considered interesting and included in a further detailed analysis. Rules with high lift, leverage, conviction and those with low values are also considered to understand the positive and negative correlations. Rules that include ATCs NA02AX02, N02AX52, N02AA01, N02AB03, NA02AA05, and NA02AJ13 are deemed interesting, as these opiates are potent, addictive or have a high risk of misuse [5, 11].

5.2 Patient Subtyping

The first step of the analysis is to prepare the data. This is achieved using the *sklearn* library. The data is multi-hot encoded so that for each patient the presence or absence of an ATC, gender, or ICPC is denoted by 1 and 0, respectively. The overview of the patient subtyping process is given in diagram 10. Dimensionality reduction techniques like MCA, Autoencoders, and T-LSTM are used to obtain latent representations.

MCA

MCA from the *Prince* library was used. The percentage of variance explained by each component and the scree plot were used to gauge the effectiveness of the method.

Autoencoder

Tensorflow is used to build the autoencoders. The performance of an autoencoder is highly dependent on its hyperparameters. Therefore, hyperparameter tuning is done to find the optimal architecture and parameters. It is performed step by step, and the best performing model

in a step is used for subsequent hyperparameter tuning.
The first step is to find the optimal architecture. A basic autoencoder with one layer in the encoder and one layer in the decoder blocks is compared against stacked autoencoders. The various architectures compared are given below:

- A basic autoencoder with one layer in the encoder and one layer in the decoder.
- A stacked autoencoder with 1 hidden layer with 256 nodes in the encoder and 1 hidden layer in the decoder with 256 nodes.
- A stacked autoencoder with 2 hidden layers. The layers contain 512 and 256 nodes in the encoder and the reverse in the decoder.

An experiment was then performed to check whether batch normalization would improve the performance. The best performing architecture from the previous step was used for the comparison against the same architecture with batch normalization layers between its hidden layers.
The next set of experiments was to find the best activation functions. Tanh, ReLU, and SELU were used for each of the layers in the best performing model found so far.
The last set of hyperparameter tuning experiments was performed to find the best coding length. This is the number of nodes in the layer connecting the encoder and decoder. This constitutes the latent representation of patient data. Coding lengths 10, 30, 50 and, 100 were considered.

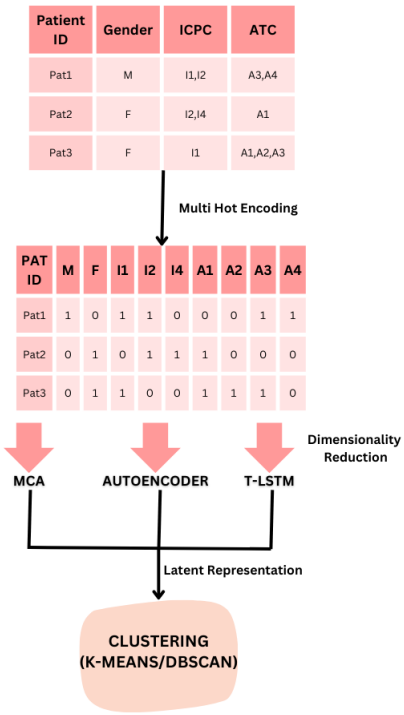


Figure 10: Patient Subtyping

T-LSTM

T-LSTM considers the time interval between episodes in generating patient latent representation. The input data consists of two sequences: ATC/ICPC and time interval. Data are fed into the T-LSTM in batches of size 32. The sequences are divided into batches so that sequences of the same length are in a batch. Each batch is of size 32; each element in a batch is a sequence; each element of the sequence is a one-hot encoded ATC/ICPC or time interval.
Performance of the model for different learning rates and number of nodes in the 3 hidden layers is analyzed.
The latent patient representation is used for clustering.

Clustering

1. **K-means**
   K-means clustering is dependent on the initialization conditions, and therefore the experiments are repeated 20 times and the averages are considered in the following experiments. K-means clustering is performed on patient latent representations with K=2, 3, ..., 30. A plot is drawn with K against the inertia value for that cluster. The value of K that forms the elbow of the plot is noted. To ensure validity, the silhouette score is calculated for each of the K values mentioned above. The value of K with the lowest silhouette score is noted. The BIC score is calculated for each cluster and the cluster with the lowest score is noted. The optimal K is decided on the basis of the elbow, silhouette, and BIC plots.

2. **DBSCAN**
   DBSCAN has two hyperparameters *eps* and *min_samples*. Different values were experimented with to obtain a reasonable number of clusters.

6 Experiments and Results

This section details the experiments performed to extract insights from the EHR data. Two methodologies are employed: Association Rule Mining (ARM) and Clustering which are discussed in detail in different sub-sections. They elaborate on the experimental setup, data manipulation, and analysis performed.

6.1 Association Rule Mining

The library used for the analysis is *Mlxtend*. It requires the data to be in a certain format to apply frequent itemset mining algorithms. Data from each patient is converted into a tuple with each item corresponding to the patient's gender, individual ATC codes, and ICPC codes. *Transactional Encoder* from the *Mlxtend* is then used to pre-process the data. Apriori and FP-growth algorithms were used to obtain the most frequent itemsets with **Support=0.01** and **Confidence=0.01** as the thresholds. The top 8 rules sorted by lift are given in table 5.

Antecedent	Consequent	Support	Confidence	Lift	Leverage	Conviction
(N02AJ06, N02AX52)	(N02AJ13, N02AA59)	0.012	0.316	18.534	0.011	1.438
(N02AJ13, N02AA59)	(N02AJ06, N02AX52)	0.012	0.731	18.534	0.011	3.583
(N02AJ13, N02AJ06)	(N02AA59, N02AX52)	0.012	0.248	14.307	0.011	1.308
(N02AA59, N02AX52)	(N02AJ13, N02AJ06)	0.012	0.719	14.307	0.011	3.383
(N02AJ13, N02AB03)	(N02AA05, N02AX52)	0.011	0.517	11.038	0.010	1.975
(N02AA05, N02AX52)	(N02AJ13, N02AB03)	0.011	0.255	11.038	0.010	1.311
(N02AX52, N02AB03)	(N02AJ13, N02AA05)	0.011	0.549	9.999	0.010	2.099
(N02AJ13, N02AA05)	(N02AX52, N02AB03)	0.011	0.217	9.999	0.010	1.250

Table 5: Top 8 association rules sorted by Lift

Some ATC codes appear more frequently than others. The list is provided below for easier understanding of the rest of the section. The full list of ATC codes is given in the appendix.

- N02AX02: Tramadol
- N02AA05: Oxycodone
- N02AJ06: Codeine and Paracetamol
- N02AJ13: Tramadol and Paracetamol
- N02AX52: Tramadol combinations
- N02AB03: Fentanyl
- N02AA01: Morphine
- N02AA59: Codeine, combinations excl. psycholeptics
- N02AE01: Buprenorphine

Given below is the list of most frequently occurring P/Z ICPC codes. The full list is given in the appendix.

- P06: Insomnia/other sleep disorder
- P17: Tobacco abuse
- P01: Feeling anxious/nervous/tense
- P76: Depression
- P02: Crisis/transient stress response
- P74: Anxiety disorder/anxiety state

- P03: Feeling down/depressed
- P78: Overvoltage
- P20: Memory/concentration/orientation disorders
- Z12: Relationship problem with partner
- Z15: Loss/death of partner
- P70: Dementia
- P29: Other psychological symptoms/complaints
- Z05: Problem with work situation
- P15: Chronic alcohol abuse
- Z18: Problem with child's illness
- P71: Other organic psychosis(s)
- P21: Attention-deficit/hyperactivity disorder
- Z16: Relationship problem with child
- P99: Other mental disorders
- Z25: Problem due to violence
- P22: Other child behavior concerns
- P18: Drug abuse
- P04: Irritable/angry feeling/behavior
- Z03: Housing/neighborhood problem
- P19: Substance abuse
- P12: Enuresis [ex. U04]
- P24: Specific learning difficulty
- Z22: Problem with illness of parents/family
- Z20: Relationship problem with parents/family
- Z23: Loss/death of parents/family
- Z14: Problem with partner's illness
- Z29: Other social problem

Apriori and FP-growth use different methods to find the frequent itemset. Therefore, it is important to ensure that the itemsets are mostly identical. Comparing the two sets found that both the algorithms find the same itemsets.
The number of frequent itemsets and association rules found by apriori and FP-growth algorithms is given in table 6.

Frequent Itemset Count	Association Rules Count
1035	5800

Table 6: Number of frequent itemsets and association rules

The number of association rules was reduced from 5800 to 553 using the rules in section 5.1.1.

6.1.1 Analysis

To ensure the validity of the association rules found, it was compared against the general population. The set of patients that presented with P or Z ICPC codes with no prescription of opioids between 2010-2019 was used. Association rules were found by using Apriori and FP-growth algorithms. It resulted in 51 rules which are shown in table 7.

Antecedent	Consequent	Support	Confidence	Lift	Leverage	Conviction
(P19)	(M)	0.014	0.782	1.664	0.006	2.439
(P15)	(M)	0.014	0.727	1.548	0.005	1.945
(P12)	(M)	0.011	0.660	1.404	0.003	1.559
(P24)	(M)	0.033	0.633	1.348	0.008	1.447
(P21)	(M)	0.037	0.626	1.331	0.009	1.417
(Z18)	(V)	0.014	0.689	1.301	0.003	1.515
(P99)	(M)	0.022	0.596	1.267	0.004	1.311
(P78)	(V)	0.043	0.648	1.223	0.007	1.337
(Z15)	(V)	0.019	0.647	1.222	0.003	1.333
(P01)	(V)	0.071	0.646	1.221	0.013	1.331
(P02)	(V)	0.047	0.640	1.208	0.008	1.308
(P74)	(V)	0.052	0.635	1.199	0.008	1.289
(P03)	(V)	0.043	0.630	1.190	0.006	1.272
(Z16)	(V)	0.012	0.622	1.174	0.001	1.244
(P17)	(M)	0.045	0.551	1.173	0.006	1.182
(Z22)	(V)	0.021	0.620	1.170	0.003	1.237
(Z20)	(V)	0.013	0.612	1.156	0.001	1.214
(P76)	(V)	0.049	0.612	1.156	0.006	1.213
(Z23)	(V)	0.021	0.606	1.145	0.002	1.195
(Z12)	(V)	0.036	0.599	1.132	0.004	1.174
(Z14)	(V)	0.012	0.588	1.110	0.001	1.142
(Z25)	(V)	0.011	0.581	1.097	0.001	1.123
(P22)	(M)	0.025	0.515	1.095	0.002	1.092
(Z29)	(V)	0.019	0.579	1.094	0.001	1.119
(P29)	(V)	0.026	0.579	1.094	0.002	1.118
(P70)	(V)	0.011	0.571	1.079	0.000	1.097
(P20)	(M)	0.022	0.506	1.077	0.001	1.073
(P76)	(P06)	0.010	0.136	1.041	0.000	1.006
(Z05)	(V)	0.015	0.545	1.030	0.000	1.035
(P06)	(V)	0.071	0.543	1.025	0.001	1.029
(P06)	(M)	0.059	0.456	0.971	-0.001	0.975
(Z05)	(M)	0.012	0.454	0.966	-0.000	0.970
(P20)	(V)	0.021	0.493	0.931	-0.001	0.928
(P22)	(V)	0.023	0.484	0.915	-0.002	0.912
(P29)	(M)	0.019	0.420	0.893	-0.002	0.913
(Z29)	(M)	0.014	0.420	0.893	-0.001	0.913
(Z12)	(M)	0.024	0.400	0.851	-0.004	0.883
(P17)	(V)	0.036	0.448	0.845	-0.006	0.852
(Z23)	(M)	0.014	0.393	0.836	-0.002	0.873
(P01)	(P06)	0.012	0.108	0.828	-0.002	0.974
(P76)	(M)	0.031	0.387	0.824	-0.006	0.865
(Z22)	(M)	0.013	0.379	0.808	-0.003	0.854
(P03)	(M)	0.025	0.369	0.785	-0.006	0.840
(P74)	(M)	0.030	0.364	0.775	-0.008	0.833
(P02)	(M)	0.026	0.359	0.764	-0.008	0.827

(P99)	(V)	0.015	0.403	0.762	-0.004	0.788
(P01)	(M)	0.039	0.353	0.750	-0.013	0.818
(Z15)	(M)	0.010	0.352	0.749	-0.003	0.818
(P78)	(M)	0.023	0.351	0.747	-0.007	0.817
(P21)	(V)	0.022	0.373	0.705	-0.009	0.750
(P24)	(V)	0.019	0.366	0.691	-0.008	0.741

Table 7: Association rules for population with P/Z ICPC codes but no opioid prescription

In addition, the patient population who was prescribed opioids but had no ICPC P / Z codes was also analyzed as shown in table 8.

Patient Population	Count
Have P/Z ICPC and prescribed opioids	93185
Have P/Z ICPC and not prescribed opioids	220566
Do not have P/Z ICPC and prescribed opioids	60734

Table 8: Patient population distribution

The association rules of interest found in the patient cohort prescribed opioids and presented with at least one incidence of P or Z ICPC codes are given in table 9.

Rule ID	Antecedent	Consequent	Support	Confidence	Lift	Leverage	Conviction
I1	(P71)	(N02AB03)	0.012	0.436	3.835	0.009	1.572
I2	(V, P70)	(N02AB03)	0.010	0.339	2.980	0.007	1.340
I3	(P70)	(N02AB03)	0.015	0.318	2.795	0.009	1.299
I4	(P70)	(N02AA01)	0.011	0.230	2.192	0.006	1.162
I5	(V, P20)	(N02AB03)	0.011	0.228	2.003	0.005	1.148
I6	(P19)	(M)	0.011	0.725	1.924	0.005	2.267
I7	(P20)	(N02AB03)	0.016	0.204	1.797	0.007	1.114
I8	(P15)	(M)	0.020	0.643	1.707	0.008	1.746
I9	(Z15)	(N02AB03)	0.012	0.183	1.607	0.005	1.085
I10	(P20)	(N02AA01)	0.013	0.160	1.523	0.004	1.065
I11	(V, P20)	(N02AX52)	0.011	0.224	1.492	0.004	1.095
I12	(N02AA05, N02AX02, N02AB03)	(P06)	0.014	0.346	1.461	0.004	1.167
I13	(N02AX52, N02AJ13, N02AX02)	(V, P06)	0.012	0.214	1.455	0.003	1.085
I14	(V, Z15)	(N02AX52)	0.010	0.204	1.363	0.002	1.068
I15	(P71)	(N02AA05)	0.011	0.430	1.331	0.002	1.187
I16	(Z15)	(N02AX52, N02AJ13)	0.010	0.148	1.312	0.002	1.041
I17	(P20)	(N02AX52, N02AJ13)	0.011	0.145	1.285	0.002	1.037
I18	(Z18)	(V)	0.024	0.797	1.280	0.005	1.864
I19	(N02AX52, N02AX02)	(P76)	0.011	0.153	1.269	0.002	1.038
I20	(P22)	(V)	0.016	0.789	1.266	0.003	1.787
I21	(Z15)	(N02AX52)	0.012	0.189	1.262	0.002	1.048
I22	(N02AJ06, N02AX02)	(P06)	0.035	0.294	1.245	0.006	1.082
I23	(Z12)	(P76)	0.011	0.150	1.241	0.002	1.034
I24	(Z15)	(N02AA05)	0.014	0.215	1.226	0.002	1.050
I25	(V, Z15)	(N02AA05)	0.019	0.396	1.226	0.003	1.121
I26	(P06, P17)	(N02AA05)	0.011	0.395	1.226	0.002	1.121
I27	(Z16)	(V)	0.018	0.763	1.224	0.003	1.592
I28	(P21)	(M)	0.012	0.459	1.220	0.002	1.154
I29	(V, Z14)	(N02AA05)	0.010	0.390	1.209	0.001	1.110
I30	(P76, P06)	(N02AA05)	0.010	0.390	1.209	0.001	1.110
I31	(P20)	(N02AJ13)	0.016	0.211	1.207	0.002	1.046
I32	(P17)	(N02AX02, N02AA05)	0.030	0.210	1.204	0.005	1.045
I33	(P06, P17)	(N02AX02)	0.020	0.700	1.198	0.003	1.387
I34	(P76, P01)	(V)	0.011	0.746	1.197	0.001	1.485
I35	(Z15)	(N02AA05)	0.026	0.384	1.191	0.004	1.100
I36	(P17)	(M)	0.065	0.447	1.188	0.010	1.128
I37	(P02)	(P76)	0.013	0.142	1.180	0.002	1.025
I38	(Z14)	(N02AA05)	0.016	0.380	1.179	0.002	1.093
I39	(Z15)	(V)	0.049	0.729	1.170	0.007	1.394
I40	(P20)	(N02AX02, N02AA05)	0.016	0.204	1.170	0.002	1.037
I41	(P17, P01)	(N02AX02)	0.011	0.683	1.169	0.001	1.313
I42	(P06, Z12)	(N02AX02)	0.010	0.683	1.169	0.001	1.313
I43	(Z23)	(V)	0.031	0.728	1.168	0.004	1.387
I44	(Z15)	(N02AX02, N02AA05)	0.013	0.203	1.165	0.001	1.036
I45	(P76, P02)	(V)	0.010	0.726	1.165	0.001	1.376

I46	(P06, M)	(N02AA01)	0.010	0.122	1.163	0.001	1.019
I47	(P20)	(N02AA05)	0.029	0.374	1.160	0.004	1.083
I48	(Z22)	(V)	0.032	0.722	1.158	0.004	1.356
I49	(P06, P78)	(N02AX02)	0.010	0.676	1.157	0.001	1.285
I50	(P01)	(V)	0.095	0.719	1.154	0.012	1.343
I51	(P71)	(M)	0.011	0.432	1.146	0.001	1.097
I52	(P06, P02)	(V)	0.015	0.713	1.145	0.001	1.137
I53	(P74)	(V)	0.059	0.713	1.144	0.007	1.315
I54	(P02)	(V)	0.069	0.709	1.138	0.008	1.297
I55	(P78)	(V)	0.056	0.706	1.134	0.006	1.284
I56	(P80)	(N02AX02)	0.011	0.660	1.130	0.001	1.225
I57	(Z23)	(N02AJ06)	0.014	0.327	1.125	0.001	1.054
I58	(Z22)	(N02AJ06)	0.014	0.327	1.123	0.001	1.053
I59	(N02AJ06, N02AX02)	(P17)	0.019	0.163	1.122	0.002	1.021
I60	(P29)	(N02AA05)	0.016	0.361	1.121	0.001	1.061
I61	(N02AX52)	(V)	0.104	0.698	1.120	0.011	1.247
I62	(Z16, V)	(N02AX02)	0.011	0.654	1.119	0.001	1.202
I63	(P18)	(N02AX02)	0.013	0.652	1.116	0.001	1.196
I64	(P22)	(N02AX02)	0.013	0.650	1.113	0.001	1.190
I65	(P06, P17)	(M)	0.012	0.419	1.113	0.001	1.073
I66	(Z25)	(N02AX02)	0.014	0.649	1.111	0.001	1.186
I67	(Z16)	(N02AX02)	0.015	0.648	1.109	0.001	1.182
I68	(P03)	(N02AJ06, N02AX02)	0.010	0.132	1.107	0.001	1.014
I69	(P17)	(N02AX02)	0.094	0.646	1.106	0.009	1.176
I70	(P03)	(V)	0.056	0.688	1.104	0.005	1.208
I71	(N02AA01)	(M)	0.043	0.414	1.099	0.003	1.064
I72	(Z25)	(V)	0.014	0.677	1.087	0.001	1.168
I73	(N02AX52, N02AX02)	(P17)	0.011	1.158	1.086	0.000	1.014
I74	(P70)	(V)	0.030	0.676	1.084	0.002	1.163
I75	(Z12)	(N02AX02)	0.047	0.631	1.080	0.003	1.127
I76	(Z05)	(M)	0.013	0.404	1.073	0.000	1.046
I77	(P76)	(V)	0.080	0.666	1.069	0.005	1.130
I78	(P74)	(N02AJ06, N02AX02)	0.010	0.128	1.068	0.000	1.009
I79	(V)	(P17)	0.016	0.155	1.066	0.000	1.011
I80	(P29)	(V)	0.030	0.663	1.064	0.001	1.120
I81	(P02)	(V)	0.011	0.120	1.064	0.000	1.008
I82	(Z12)	(V)	0.049	0.658	1.057	0.002	1.104
I83	(N02AJ06)	(V)	0.191	0.658	1.057	0.010	1.104
I84	(Z14)	(V)	0.028	0.656	1.053	0.001	1.096
I85	(Z03)	(V)	0.012	0.652	1.047	0.000	1.084
I86	(N02AA05, N02AB03)	(V)	0.046	0.650	1.044	0.001	1.079
I87	(P20)	(M)	0.031	0.392	1.042	0.001	1.026
I88	(Z29)	(V)	0.023	0.648	1.040	0.000	1.072
I89	(N02AA05, N02AA01, N02AB03)	(M)	0.010	0.387	1.029	0.000	1.017
I90	(P06)	(N02AJ06)	0.070	0.299	1.028	0.001	1.011
I91	(Z18)	(N02AA05)	0.010	0.331	1.028	0.000	1.013
I92	(P78)	(N02AJ06)	0.023	0.299	1.026	0.000	1.011

I93	(N02AA59)	(P01)	0.013	1.135	1.024	0.000	1.003
I94	(P03)	(N02AA05)	0.026	0.329	1.021	0.000	1.010
I95	(P06)	(M)	0.089	0.377	1.001	0.000	1.001
I96	(Z29)	(N02AX02)	0.021	0.584	1.000	0.000	1.000
I97	(Z22)	(N02AA05)	0.014	0.322	0.999	0.000	0.999
I98	(P06)	(V)	0.147	0.622	0.998	-0.000	0.998
I99	(P20)	(V)	0.048	0.607	0.974	-0.001	0.959
I100	(P76)	(P06)	0.027	0.228	0.966	-0.000	0.989
I101	(P01)	(N02AA05)	0.041	0.311	0.964	-0.001	0.983
I102	(Z05)	(V)	0.020	0.595	0.955	-0.000	0.931
I103	(P20)	(N02AX02)	0.043	0.547	0.936	-0.002	0.917
I104	(P99)	(V)	0.013	0.581	0.933	-0.000	0.901
I105	(P76)	(P01)	0.014	0.123	0.932	-0.001	0.989
I106	(Z29)	(M)	0.012	0.351	0.932	-0.000	0.960
I107	(P78)	(N02AX02, N02AA05)	0.012	0.158	0.909	-0.001	0.981
I108	(N02AB03)	(P76)	0.012	0.109	0.908	-0.001	0.987
I109	(P01)	(P06)	0.028	0.214	0.906	-0.002	0.971
I110	(Z12)	(M)	0.025	0.341	0.905	-0.002	0.945
I111	(P29)	(M)	0.015	0.336	0.892	-0.001	0.939
I112	(P17)	(V)	0.080	0.552	0.883	-0.010	0.841
I113	(P76)	(M)	0.040	0.333	0.884	-0.005	0.934
I114	(P78)	(N02AX52)	0.010	0.131	0.879	-0.001	0.978
I115	(N02AA01)	(P17)	0.013	0.126	0.868	-0.002	0.978
I116	(P21)	(V)	0.014	0.540	0.866	-0.002	0.819
I117	(N02AA59)	(M)	0.033	0.326	0.865	-0.005	0.924
I118	(P04)	(V)	0.010	0.515	0.827	-0.002	0.778
I119	(P03)	(M)	0.025	0.311	0.827	-0.005	0.905
I120	(N02AX52)	(M)	0.045	0.301	0.801	-0.011	0.892
I121	(P70)	(N02AX02)	0.020	0.455	0.779	-0.005	0.763
I122	(P78)	(M)	0.023	0.293	0.778	-0.006	0.881
I123	(N02AB03)	(P17)	0.012	0.113	0.776	-0.003	0.963
I124	(P02)	(M)	0.028	0.290	0.770	-0.008	0.878
I125	(N02AX52, N02AA05)	(M)	0.013	0.288	0.766	-0.004	0.876
I126	(P74)	(M)	0.024	0.286	0.760	-0.007	0.873
I127	(P01)	(M)	0.037	0.280	0.744	-0.012	0.866
I128	(Z22)	(M)	0.012	0.277	0.737	-0.004	0.862
I129	(Z23)	(M)	0.011	0.271	0.720	-0.004	0.855
I130	(Z15)	(M)	0.018	0.270	0.717	-0.007	0.854
I131	(N02AX52, N02AJ06)	(M)	0.010	0.257	0.682	-0.004	0.838
I132	(N02AJ06, N02AJ13)	(M)	0.012	0.256	0.680	-0.006	0.838
I133	(P15)	(V)	0.010	0.356	0.572	-0.008	0.585

Table 9: Association rules for population with P/Z ICPC codes and opioid prescription. Rules that also occur in the general population and have lift > 1 are in gray. Rules that also occur in the general population but have a lift < 1 are in green.

Several hypotheses can be extrapolated from the association rules. The associated rules for each hypothesis is given in brackets and refers to the rule ID in table 9:

- Women are prescribed opioids more often than men with 62% of opioid prescriptions going to women [5]
- There exists a positive correlation between opioid usage and sleep disorders [I12, I13, I22, I26, I30, I33, I42, I46, I49, I52, I65, I90, I95, I98, I100]
- Fentanyl and morphine usage corresponds to an increase in dementia, concentration, and memory problems [I2, I3, I4, I5, I7, I10]
- A positive correlation exists between loss of partner/family and opioid prescription [I9, I14, I16, I21, I24, I25, I35, I39, I43, I44, I57]
- Chronic alcohol usage is correlated to opioid use [I8]

6.2 Patient Subtyping

Pre-processing of the data is done using the *Pandas* library. The first step was to obtain the multihot encoded patient data. The gender, ATC codes, and ICPC codes were one-hot encoded. This resulted in a matrix with 96 columns; Each patient is represented by a vector of length 96. Several methods were employed to reduce the dimensionality, which is described below.

6.2.1 MCA

The *Prince* library is used to perform MCA. The experiment was performed with the following hyperparameters:

- **Number of components**: 10
- **Number of iterations**: 10
- **Engine**: sklearn
- **random state**: 42

The percentage of variance captured by each component is given in table 10: A scree plot is

Component	Eigen value	% of variance	% of variance (cumulative)
0	0.310	1.84	1.84
1	0.285	1.69	3.53
2	0.240	1.43	4.96
3	0.231	1.37	6.33
4	0.224	1.33	7.66
5	0.219	1.30	8.96
6	0.213	1.26	10.22
7	0.211	1.26	11.48
8	0.207	1.23	12.71
9	0.205	1.22	13.93

Table 10: Eigen value and % variance summary of MCA of 10 components

used to help determine the optimal number of components or factors to retain in the analysis. It displays the eigenvalues (or variance explained) associated with each principal component or factor, plotted in descending order. It tends to produce a graph with a steep slope followed by a leveling off or 'elbow' in the curve. The components to the left of the 'elbow' represent the substantial portion of the variance. The scree plot for dimensionality reduction using MCA is given in figure 11.

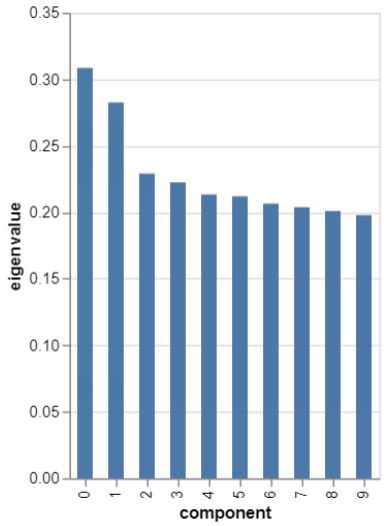


Figure 11: Scree Plot

6.2.2 Autoencoder

Autoencoders are used as a powerful dimensionality tool. The *Keras* library from the *Tensorflow* framework is used to build the autoencoders. *matplotlib*, *seaborn* and *plotly* are the libraries of choice for visualization of the results. Several experiments were performed to find the best autoencoder model for dimensionality reduction. The experiments used *Adam* optimization algorithm is used for gradient descent. *Binary crossentropy* loss is used for learning. Each model was trained for 50 epochs with a batch size of 128.
The experiments to find the optimal model are described below.

6.2.2.1 Architecture

A comparison was made between a basic autoencoder with no hidden layers, an autoencoder with 1 hidden layer, and an autoencoder with 2 hidden layers. In all autoencoders, Relu activation was used except in the output layer in which softmax activation was used. The coding length of 10 was used. In the autoencoder with 1 hidden layer, a hidden layer with 256 neurons was chosen. In the autoencoder with 2 hidden layers, the first hidden layer contained 512 neurons and the second hidden layer contained 256 neurons. The autoencoders with 1 and 2 hidden layers quickly achieve low validation loss. The basic autoencoder is capable of learning, but is outperformed by the other two as shown in figure 12. The autoencoder with one hidden layer is chosen for the following experiments.

6.2.2.2 Batch Normalization

Batch normalization is a technique used in the training of deep neural networks that helps sta-

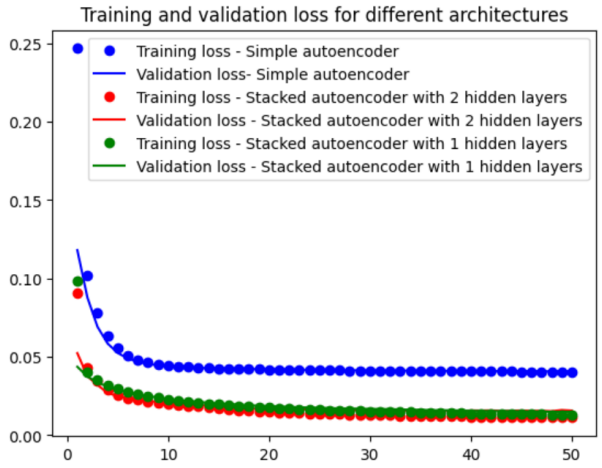


Figure 12: Validation loss for different architectures

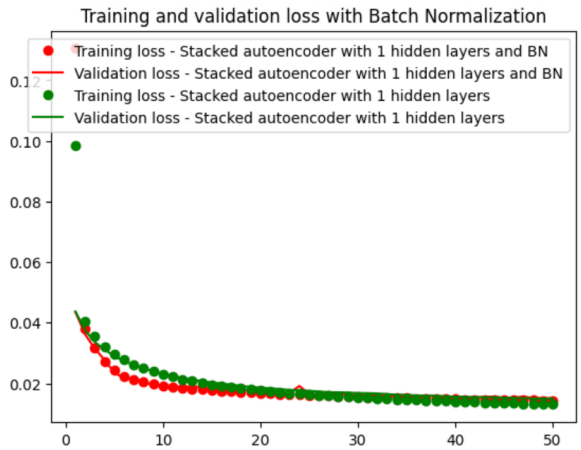


Figure 13: Validation loss for batch normalization

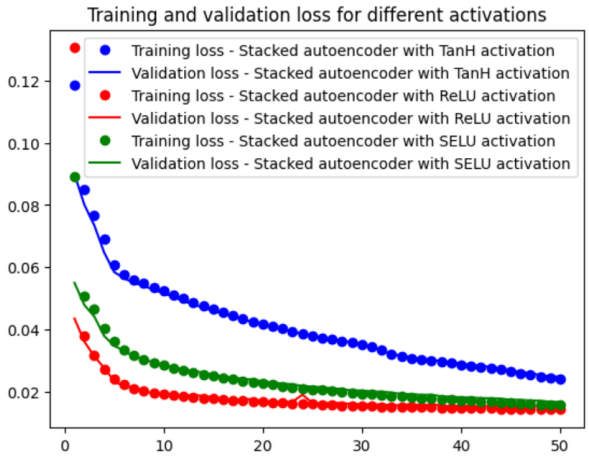


Figure 14: Validation loss for different activation functions

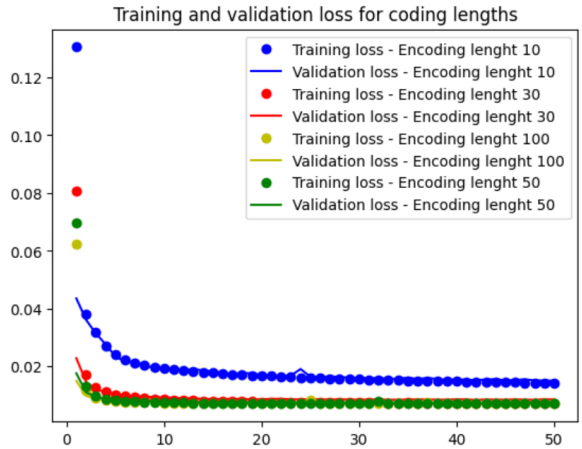


Figure 15: Validation loss of different coding lengths

bilize and accelerate the training process. To determine whether batch normalization improves the performance of an autoencoder with 1 hidden layer, a comparison was made.
Batch normalization did not improve the performance of the model, as shown in Fig. 13, so an autoencoder without batch normalization was used in the following experiments.

6.2.2.3 Activation Function

Activation functions can affect the performance of a model in several ways: affect how gradients are propagated through the network during backpropagation, how quickly a model converges during training, and the overall performance and accuracy of the model. TanH, ReLU, and SELU activations were tested to find the optimal activation function. In 50 epochs, TanH performed the worst, reaching a validation loss of 0.0249. ReLU performed the best, as it learned quickly and reached a validation loss of 0.0150 after 50 epochs, as shown in figure 14. ReLU was used as the activation function of choice for further experiments.

6.2.2.4 Coding Length

The coding length represents how compactly the data can be encoded in this reduced space. A shorter coding length indicates that the data has been compressed effectively, capturing the essential information with fewer dimensions. A shorter coding length implies that the model has fewer parameters to learn, which can result in faster training times, lower memory usage, and potentially better generalization to new data. The performance of the model with coding lengths 10, 30, 50, and 100 was tested. A coding length of 50 was the best performing and a coding length of 10 was the worst performing as shown in the figure 15.

6.2.2.5 Clustering

Two clustering algorithms were tested: K-means and DBSCAN.
To find the optimal K value for k-means clustering, elbow plot(figure 16), silhouette coefficient plot(figure 17), and BIC plots(figure 18) are analyzed.

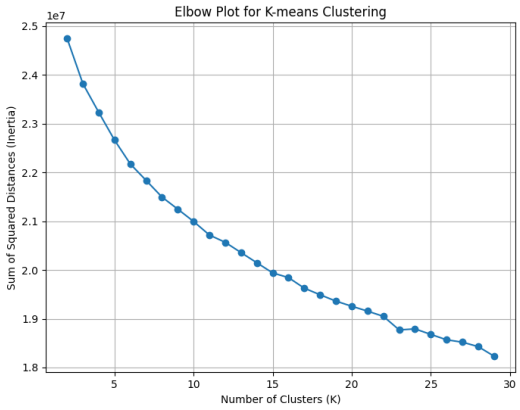


Figure 16: Elbow Plot

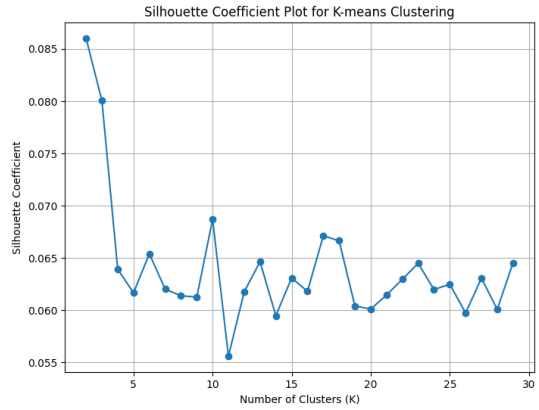


Figure 17: Silhouette coefficient plot

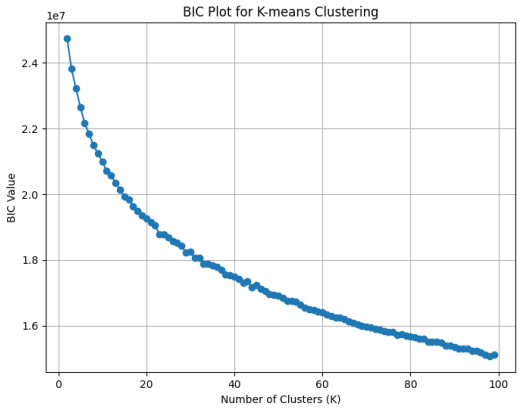


Figure 18: BIC plot

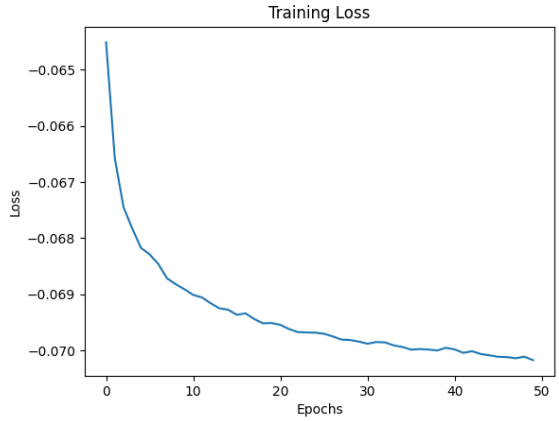


Figure 19: Training loss of T-LSTM

The elbow plot shows that the number of clusters that minimize sum of squared distances is greater than 30. From the Silhouette coefficient plot, the optimal number of clusters cannot be determined. BIC was plotted for clusters of up to size 100. However, the BIC value

decreased for each increase in number of clusters. Larger clusters (e.g., larger than 100) are more likely to include a diverse range of patients, leading to high within-cluster variability. This heterogeneity can mask meaningful patterns and make it difficult to draw precise conclusions or tailor treatments effectively.
DBSCAN can find clusters of arbitrary size and does not require the number of clusters to be predetermined. The silhouette score is a measure of how similar an object is to its own cluster (cohesion) compared to other clusters (separation). It ranges from -1 to 1, where a higher score indicates better-defined clusters. A score of 1 would indicate that the points are very well clustered, with points far away from the nearest neighboring cluster. A value of 0 indicates that the points are in or very close to the decision boundary between two neighboring clusters. A value of -1 indicates that the points are potentially assigned to the wrong clusters, with the points being closer to neighboring clusters than to their own clusters. The hyperparameters epsilon and minimum number of samples and their corresponding number of clusters and average silhouette score are given in table 11. It is evident from the results that

ϵ	min_samples	Number of clusters	Average silhouette score
0.4	5	**1266**	**0.2778**
0.4	50	115	0.0673
0.5	10	578	0.1996
0.7	40	149	0.0860
0.6	40	149	0.0860
0.7	100	57	0.0161
0.7	125	38	-0.0052

Table 11: Number of clusters and average silhouette score for different values of ϵ and min_samples

the latent representations from the autoencoders generate a large number of clusters. The highest average silhouette score was achieved with a cluster size of 1266.

6.2.3 T-LSTM

The Temporal LSTM uses the medication of a patient and the time interval between the prescriptions as input. It was trained for 50 epochs and achieved a training loss of -0.070 as shown in figure 19.

7 Discussion

This section provides in-depth analysis and insight into the experiments performed for this research.

7.1 Association Rule Mining

Fentanyl (NA02B03) is associated with Dementia (Alzheimer's Disease, vascular dementia) and other organic psychosis (delerium). These cases occur in approximately 1% of patients with a high lift value. Memory/concentration/orientation disorders, loss/death of partner, and insomnia/other sleep disorder are also positively correlated with prescription fentanyl. It is also evident that fentanyl prescription co-occurs with being a woman. Fentanyl use is slightly negatively correlated with tobacco abuse, as evidenced by a lift value of 0.776, leverage value -0.003, and conviction value of 0.963.

Tramadol (N02AX02) and the combination of Tramadol with other drugs (N02AX52) are also positively correlated with problems with sleep and memory.These drugs also co-occur with depression (support=0.011, confidence=0.153, lift=1.269). Crisis/transient stress response also occurs with patients prescribed tramadol. Tobacco abuse along with feeling anxious are positively correlated to tramadol (support=0.011, confidence=0.683, lift=1.169, conviction=1.313). Loss/death of a partner has a correlation with the prescription of tramadol. Relationship problem with partner along with sleep disorders co-occur with tramadol prescription. There is a small positive correlation between being a woman, having a relationship problem with a child, and tramadol. There is a small negative correlation between dementia and tramadol prescription (support=0.020, confidence=0.455, lift=0.779, conviction=0.763). Interestingly, there is a negative correlation between being a man and a tramadol prescription.

Morphine (N02AA01) prescription is positively associated with dementia (support=0.010, confidence=0.229, lift=2.191, conviction=1.162). It also co-occurs with memory, concentration and, orientation issues. In contrast to the two drugs mentioned above, morphine is positively associated with sleep disorders in **both** men and women. Morphine use has a small negative correlation with tobacco abuse.

Oxycodone (N02AA05) is associated with sleep disorders, particularly in women. It is also positively correlated with other organic psychosis (lift=1.331). A positive correlation exists between being a woman, loss/death of a partner and prescription of oxycodone (support = 0.019, confidence = 0.396, lift = 1.222, conviction = 1.211). Buprenorphin (N02AE01) and oxycodone co-occur frequently with being women. The prescription of oxycodone is also associated with being a woman and having problems with the health of her partner. Depression, tobacco abuse, and sleep disorders also co-occur in patients prescribed oxycodone. It also has a small positive correlation with patients experiencing the loss or death of a partner or problems with partners illness. A small correlation exists between memory and concentration problems and oxycodone.

Codeine, combinations that exclude psycholeptics (N02AA59) had a positive correlation with sleep disorders, particularly in women. The rest of this paragraph uses codeine to refer to ATC N02AA59. There is also a small positive correlation between codeine and women who abuse tobacco. A small negative correlation exists between men and codeine.

Women who have been prescribed opioids have a positive correlation with experiencing problems with children's behavior, while there is a slight negative correlation in women not prescribed opioids. Among men prescribed opioids, there is a small positive correlation with having

problems at work, while those not prescribed opioids have a small negative correlation. However, in the latter case, the difference in the lift values is not significant enough to draw this conclusion confidently.

In most of the rules, the leverage value is a value slightly greater than 0 while the lift and conviction values are greater than 1. This implies that the rule represents a strong association. However, the items involved have relatively low support in the dataset. Leverage is sensitive to the overall frequency of items, so rare items tend to have lower leverage even if they are strongly associated when they do occur. These imply the occurrence of **rare association rules**.

Chronic alcohol and substance abuse occurs mainly in men. This pattern is also seen in the general population that does not take prescribed opioids. Women are more prone to have problems with child's illness; this pattern is also seen in the general population. Attention deficit/hyperactivity disorder is positively correlated with men in both cohorts. Similarly, there is a small positive correlation between tobacco abuse and men in both cohorts. Women in both cohorts are more likely to report having relationship problems with their children. Similarly, women in both cohorts reported having problems after loss/death or illness of a partner or family member. More women in both cohorts also reported feeling anxious, tense, or nervous in contrast to men. Crisis/transient stress response or feeling down/depressed was also reported by women more often. Women also reported more problems with violence. There is a slight negative correlation between tobacco abuse and women in both cohorts. There is a small negative correlation between men and depression. There is also a negative correlation between women and attention deficit disorders; and men and men and crisis/transient stress response. Men also have a negative correlation to anxiety disorder/anxiety state, feeling anxious/nervous/tense, problems with illness of parents/family, loss/death of parents/family, and loss/death of a partner. This result could also possibly occur due to the underreporting of mental health issues among men [28].

There are certain rules that occur in the general population but do not show in the population prescribed opioids (rules found with thresholds support=0.01 and confidence=0.01). Enuresis was reported frequently for men in the general population. Women in the general population had a negative correlation with specific learning disabilities such as dyslexia, speech development disorder, and motor development disorder. However, men in the general population had a positive correlation with the specific learning disabilities mentioned above. There is a positive correlation between men and the autism spectrum and adjustment disorders in the general population. These do not occur/ do not occur frequently (support< 0.01) in the population prescribed opioids.

One study found that the frequency of opioid use increased steadily with age and was particularly evident in patients with dementia [48]. The use of fentanyl was positively correlated with the loss of a partner. No study was found that evaluated the prescription of fentanyl to grieving patients. It is an interesting area of research that requires further exploration. Opioid is generally prescribed for chronic pain; several studies show a bidirectional relationship between pain and sleep. Chronic pain can lead to poor sleep quality and poor sleep quality can increase pain intensity [24]. This could be a contributing factor to sleep disorders that co-occur with opioid prescriptions. Several studies point to the occurrence of sleep disturbances due to opioid use disorder [37, 29, 31].

Several studies were conducted to assess the effects of tramadol on sleep. Even a single dose of tramadol was found to cause sleep disturbances [83]. Similar results were found in a study that examined sleep quality in tramadol-addicted patients. Tramadol was found to reduce sleep effi-

ciency and increase arousal from sleep [12]. A study also found that tramadol misuse could lead to learning and memory impairment [87]. Tramadol provides an antidepressant effect and was used to treat lower back pain in patients with depression [82]. It was also shown to be effective in the treatment of major depressive disorder [77]. This could possibly explain the positive association between tramadol and depression/anxiety. A study proposed that tramadol may be effective in treating anxiety and depression involving elements of social loss, separation, or betrayal of interpersonal bonds, based on the role of the mammalian "PANIC/separation-distress system" [73]. Tramadol is a weak mu-opioid receptor agonist that also increases serotonin and norepinephrine levels in the limbic system of the brain. Two clinical cases where tramadol was effective in treating depressive mood and reducing excessive alcohol consumption were presented. The authors propose further investigation into the use of tramadol and other "safe" opioid agonists such as buprenorphine as antidepressant treatments, particularly in cases of social stress and interpersonal loss. A study compared tramadol use between men and women in Norway, Denmark, and Sweden [75]. In all three countries, there were more female than male tramadol users during the study period. A similar situation could possibly explain the small negative correlation between being a man and the use of tramadol.

A study investigated the use of low-dose opioids to treat agitation in patients with severe dementia [60]. The study provides evidence for the potential use of long-acting, low-dose morphine as a safe and effective treatment for agitation in advanced dementia, particularly in very old patients (≥ 85 years) where agitation is difficult to control. A study consistently found that morphine suppresses REM sleep, increases wakefulness, and alters NREM sleep patterns, particularly reducing slow-wave sleep [50]. These effects were dose-dependent and observed in male post-addicts under controlled experimental conditions. The findings highlight the significant impact of morphine on human sleep architecture and the sleep-wake cycle.

Oxycodone can potentially cause delirium, especially in elderly patients or those who are opioid-naive [30]. However, evidence suggests that the risk of delirium with oxycodone is comparable to other opioid analgesics, such as morphine [81].

A study analyzed the opioid use pattern based on gender. The results are in line with the results found in this research. Women have been prescribed opioid medications in significantly greater numbers than men [39]. Women were more likely than men to have been prescribed the opioid they used non-medically [21]. A study explored methods to mitigate bias in a machine learning model trained on real clinical psychiatry data [62]. The researchers focused on a model designed to predict future administrations of benzodiazepines based on past data and found a bias of favorable results for men over women. Polysubstance use was common among both genders but more prevalent among men. For women, older age and unmarried status predicted abuse/dependence. For men, lower income and polysubstance use predicted abuse/dependence. The association rules found point to men having a positive correlation with tobacco abuse. A study aimed to quantify the association between tobacco smoking and opioid use/opioid use disorders through a systematic review and meta-analysis of observational studies [71]. The meta-analysis confirmed a significant positive association between tobacco smoking and both opioid use and opioid use disorders. In our research, a positive association was found between tobacco abuse and tramadol, oxycodone, and codeine prescriptions. A small negative correlation was found between tobacco abuse and the use of fentanyl and morphine. Association rule mining was successful in finding correlations between prescription opioids and some side effects which corroborate with existing research. Some of the results, such as the association of opioid prescription with the loss of a partner/family member, are not widely researched, but there are some studies.

7.2 Patient Subtyping

Dimensionality reduction is a critical technique in machine learning as it helps mitigate the curse of dimensionality, improves computational efficiency, and reduces noise and overfitting. MCA works by first converting the categorical variables into an indicator matrix of binary dummy variables. Then it applies correspondence analysis on this matrix to find the principal components or dimensions that best summarize the data. MCA was performed with 10 components. The total variance explained by them is approximately 14%. The original data set has a very high dimensionality and requires many principal components to capture a significant portion of the variance. It is also possible that there is little correlation between the original features, and the variance is spread more evenly across many principal components. A simple test revealed that 60% of the variance was explained by the first 50 components. The original dimensionality of the data is 96, therefore it is evident that sufficient compression is not provided by MCA. It is possible that the data is has a complex structure that cannot be captured by the linear principal components.

The natural next step was to check whether the data were non-linearly separable. Autoencoders use non-linear activation functions in their layers and multiple hidden layers, which allow them to learn complex, non-linear relationships within the data. The depth of the network (number of layers) determines the complexity of features the autoencoder can learn. Deeper networks can capture more intricate patterns and hierarchies in the data. The width of each layer (number of neurons) affects the richness of the representation. More neurons can capture more detailed information, but too many can lead to overfitting. In the experiments, a basic autoencoder with no hidden layers did not achieve a good validation loss. An autoencoder with hidden layers proved to perform better. There was no significant difference in the performance of the autoencoders with 1 or 2 layers. Batch normalization normalizes the input of each layer so that the inputs have a consistent distribution. However, there was no significant difference in performance between an autoencoder with batch normalization and without. The placement and type of activation functions significantly impact the learning process and the quality of the representations. In this work, an autoencoder with TanH activation performed the worst and was the slowest to reduce validation loss. SELU activation function was not able to achieve a low validation loss as quickly as ReLU but at the end of 50 epochs there was no significant difference in performance. The bottleneck (the narrowest part of the network) forces the autoencoder to compress the input data into a lower-dimensional space. The size of this layer is critical because it determines the extent of compression and the balance between retaining essential features and discarding noise. A coding length of 10 performed the worst, while there was no significant difference between coding lengths of 50 and 100. Therefore, a coding length of 50 was used in further experiments.

The elbow plot to find the optimal K was not successful. The sum of squared distances reduced consistently even with 30 clusters. The highest average silhouette score was associated with the data points clustered into 2 clusters. Similarly to the elbow plot, the BIC plot saw a consistent reduction in BIC value with increase in cluster numbers and could not be used to determining the optimal value of K. The number of clusters with the optimal K is greater than 100. Clinicians and healthcare providers need subtypes that are interpretable and actionable. With more than a hundred clusters, it becomes challenging to draw meaningful conclusions and make clinical decisions based on such a fragmented classification. When creating a very large number of clusters, there is a risk of overfitting the model to the specific dataset. This means that the clusters may not generalize well to new, unseen data.

With DBSCAN, ϵ specifies how close points need to be to each other to be considered part of the same cluster, and min_samples specifies the minimum number of points required to form a cluster. It is evident that as the number of clusters increases, the average silhouette score also increases. The clustering favors large number of clusters which points towards a case of overfitting. Clinicians are less likely to use a model that provides an overwhelming number of patient subtypes [68]. Translating research findings into clinical practice is more straightforward when there are fewer, well-defined clusters that can be studied in-depth and across different populations.

Cosine similarity loss is used to measure the similarity between the input and the reconstructed output. Instead of focusing on the absolute difference between these vectors (as in Mean Squared Error), cosine similarity loss focuses on the orientation or angle between the vectors, making it particularly useful in situations like similarity search where the direction of the data vectors is more important than their magnitude. The cosine similarity score 0 indicates orthogonality, while values closer to -1 show greater similarity. The lowest cosine similarity loss achieved by the T-LSTM was approximately -0.07, indicating that the model was unable to find precise latent representations.

8 Conclusion

How can unsupervised machine learning be used to find intelligent insights into prescription trends and consequent social and psychological effects of opioids?
Two methodologies were used to find patterns in the opioid prescription pattern in the Netherlands between 2010 and 2019. Association rule mining proved to be a simple yet powerful tool in understanding the relationship between opioids and ICPC reported by patients. Clustering, however, was not an easy task. Obtaining patient representation was not possible using MCA and autencoder dimensionality reduction techniques.
Research questions 1: Is opioid prescription affected by patient attributes such as gender, age, level of education, profession, and postal code?
The age of the patient could not be accurately determined due to incomplete patient death records and could not be used in the analysis. The level of education and the profession had a large number of nulls, making it difficult to use imputation techniques. The postal code was added to the patient details in association rule mining, but did not appear in the rules derived from frequent itemsets. It also increased the dimensionality of the data without adding useful information to patient representations and was thus excluded. The gender of the patient was crucial to the analysis. Women are prescribed more opioids than men.
Research question 2: Is there a relationship between ICPC code presentation, patient gender and medication prescription?
Several interesting patterns could be found between medication, gender, and diagnosis. Fentanyl and morphine were positively associated with dementia. Fentanyl, tramadol, and morphine were associated with memory, concentration, and orientation disorders. Most opioids, namely, fentanyl, tramadol, morphine, oxycodone, and codeine, co-occurred with sleep disorders. Most of the opioid use is also correlated with women. Depression was reported by patients who were also prescribed tramadol and oxycodone. Tramadol and oxycodone were also prescribed to patients who reported the loss or death of a partner. The use of tramadol was associated with patients having problems with their child(ren) while oxycodone was associated with women having problems with the health of their partner. The use of fentanyl, morphine, and codeine was negatively associated with tobacco abuse. Chronic alcohol and substance abuse occurs mainly in men. This pattern is also seen in the general population that does not take prescribed opioids. Women are more prone to have problems with child's illness, while attention deficit/hyperactivity disorder is more prevalent in men.
Research question 3: Can patient subtypes be identified based on clinical characteristics, demographic factors, and diagnoses?
MCA, autoencoders, and Temporal LSTM were used to reduce the dimensionality of the data. Clinical analysis requires the number of clusters to be small to draw conclusive characteristics of the clusters. K-means and DBSCAN pointed towards a number of clusters > 100. This was not useful and was not further analyzed.

9 Future Work

Exploratory data analysis revealed that most of the fields were empty. These fields correspond to patient demographic details. Including these in the analysis could have a profound impact. Manually filling the fields with the correct information or different imputation techniques could be used to perform a better analysis. Further analysis could also include patient journals and laboratory results.

Patient subtyping could not be performed due to the large number of clusters found. Hierarchical clustering could be used to handle complex, high-dimensional data. Hierarchical clustering with appropriate dissimilarity measures, such as the Gower coefficient, can handle mixed data types. The dendrogram produced by hierarchical clustering can reveal subgroups within larger clusters, potentially uncovering novel disease subtypes.

This work focuses on finding patterns in the trend of prescription opioids and diagnoses. Several hypotheses were discovered. For many of the hypotheses found, a corresponding study could also be found. Further experiments or translational studies could be conducted to determine the veracity of the hypotheses.

References

[1] URL https://www.cfr.org/backgrounder/fentanyl-and-us-opioid-epidemic.

[2] URL https://www.nhg.org/.

[3] URL https://nida.nih.gov/.

[4] URL http://mchp-appserv.cpe.umanitoba.ca/viewConcept.php?printer=Y&conceptID=1304.

[5] URL https://www.westmercia.police.uk/news/west-mercia/news/2023/september/drug-users-warned-about-dangerous-synthetic-opioids.

[6] URL https://www.globalfamilydoctor.com/site/DefaultSite/filesystem/documents/Groups/WICC/InternationalClassificationofPrimaryCareDec16.pdf.

[7] Feb 2016. URL https://www.fbi.gov/news/stories/raising-awareness-of-opioid-addiction.

[8] Mortality prediction of patients with cardiovascular disease using medical claims data under artificial intelligence architectures: Validation study. *JMIR Medical Informatics*, 9, 2021. ISSN 22919694. doi: 10.2196/25000.

[9] Nov 2023. URL https://www.hopkinsmedicine.org/health/conditions-and-diseases/opioid-use-disorder.

[10] Nov 2023. URL https://stizon.nl/.

[11] Prescription opioids drugfacts, May 2023. URL https://nida.nih.gov/publications/drugfacts/prescription-opioids.

[12] E. A. Abdullah, F. A. E. Moussa, M. E. Amin, M. A. Basheer, and A. A. Saleh. Sleep characteristics in patients with tramadol dependence. *International journal of health sciences*, 2022. ISSN 2550-6978. doi: 10.53730/ijhs.v6ns1.7227.

[13] R. Agrawal and R. Srikant. Fast algorithms for mining association rules. 1994.

[14] R. Agrawal, T. Imieliński, and A. Swami. Mining association rules between sets of items in large databases. *ACM SIGMOD Record*, 22, 1993. ISSN 01635808. doi: 10.1145/170036.170072.

[15] M. Ahmed, R. Seraj, and S. M. S. Islam. The k-means algorithm: A comprehensive survey and performance evaluation, 2020. ISSN 20799292.

[16] R. Al-Hasani and M. R. Bruchas. Molecular mechanisms of opioid receptor-dependent signaling and behavior, 2011. ISSN 00033022.

[17] A. S. Al-hegami. Subjective measures and their role in data mining process. *In Proceedings of the 6th International Conference on Cognitive Sytems*, 2004.

[18] W. Altaf, M. Shahbaz, and A. Guergachi. Applications of association rule mining in health informatics: a survey. *Artificial Intelligence Review*, 47, 2017. ISSN 15737462. doi: 10.1007/s10462-016-9483-9.

[19] Q. An, S. Rahman, J. Zhou, and J. J. Kang. A comprehensive review on machine learning in healthcare industry: Classification, restrictions, opportunities and challenges, 2023. ISSN 14248220.

[20] M. L. Antonie and O. R. Zaïane. Mining positive and negative association rules: An approach for confined rules. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 3202, 2004. ISSN 16113349. doi: 10.1007/978-3-540-30116-5_6.

[21] S. E. Back, R. L. Payne, A. N. Simpson, and K. T. Brady. Gender and prescription opioids: Findings from the national survey on drug use and health. *Addictive Behaviors*, 35, 2010. ISSN 03064603. doi: 10.1016/j.addbeh.2010.06.018.

[22] I. M. Baytas, C. Xiao, X. Zhang, F. Wang, A. K. Jain, and J. Zhou. Patient subtyping via time-aware lstm networks. volume Part F129685, 2017. doi: 10.1145/3097983.3097997.

[23] R. E. Bellman. *Adaptive Control Processes*. Princeton University Press, Princeton, 1961. ISBN 9781400874668. doi: doi:10.1515/9781400874668. URL `https://doi.org/10.1515/9781400874668`.

[24] M. J. Brennan and J. A. Lieberman. Sleep disturbances in patients with chronic pain: Effectively managing opioid analgesia to improve outcomes, 2009. ISSN 03007995.

[25] A. J. Burgess-Hull, L. V. Panlilio, K. L. Preston, and D. H. Epstein. Trajectories of craving during medication-assisted treatment for opioid-use disorder: Subtyping for early identification of higher risk. *Drug and Alcohol Dependence*, 233, 2022. ISSN 18790046. doi: 10.1016/j.drugalcdep.2022.109362.

[26] M. E. Celebi, H. A. Kingravi, and P. A. Vela. A comparative study of efficient initialization methods for the k-means clustering algorithm. *Expert Systems with Applications*, 40, 2013. ISSN 09574174. doi: 10.1016/j.eswa.2012.07.021.

[27] H. Y. Chang, H. Kharrazi, D. Bodycombe, J. P. Weiner, and G. C. Alexander. Healthcare costs and utilization associated with high-risk prescription opioid use: A retrospective cohort study. *BMC Medicine*, 16, 2018. ISSN 17417015. doi: 10.1186/s12916-018-1058-y.

[28] B. N. Chatmon. Males and mental health stigma, 2020. ISSN 15579891.

[29] M. D. Cheatle and L. R. Webster. Opioid therapy and sleep disorders: Risks and mitigation strategies. *Pain Medicine (United States)*, 16, 2015. ISSN 15264637. doi: 10.1111/pme.12910.

[30] J. H. Crane and K. J. Suda. Oxycodone induced delirium and agitation in an elderly patient following total right knee arthroplasty. *International Journal of Clinical Pharmacy*, 33, 2011. ISSN 22107703. doi: 10.1007/s11096-011-9553-7.

[31] N. J. Cutrufello, V. D. Ianus, and J. A. Rowley. Opioids and sleep, 2020. ISSN 15316971.

[32] D. Dowell, K. R. Ragan, C. M. Jones, G. T. Baldwin, and R. Chou. Cdc clinical practice guideline for prescribing opioids for pain - united states, 2022. *MMWR. Recommendations and reports : Morbidity and mortality weekly report. Recommendations and reports*, 71, 2022. ISSN 15458601. doi: 10.15585/mmwr.rr7103a1.

[33] A. M. Dydyk. Opioid use disorder, Jan 2024. URL `https://www.ncbi.nlm.nih.gov/books/NBK553166/`.

[34] R. J. Ellis, Z. Wang, N. Genes, and A. Ma'Ayan. Predicting opioid dependence from electronic health records with machine learning. *BioData Mining*, 12, 2019. ISSN 17560381. doi: 10.1186/s13040-019-0193-0.

[35] N. Erler. Missing values in clinical research (ep16). URL `https://rmisstastic.netlify.app/tutorials/Erler_course_MultipleImputation_2018/Erler_slides_MICourse_2018.pdf`.

[36] M. Ester, H. P. Kriegel, J. Sander, and X. Xu. A density-based algorithm for discovering clusters a density-based algorithm for discovering clusters in large spatial databases with noise. 1996.

[37] H. R. Fathi, A. Yoonessi, A. Khatibi, F. Rezaeitalab, and A. Rezaei-Ardani. Crosstalk between sleep disturbance and opioid use disorder: A narrative review. *Addiction health*, 12, 2020. ISSN 2008-4633. doi: 10.22122/ahj.v12i2.249.

[38] C. Garbin, X. Zhu, and O. Marques. Dropout vs. batch normalization: an empirical study of their impact to deep learning. *Multimedia Tools and Applications*, 79, 2020. ISSN 15737721. doi: 10.1007/s11042-019-08453-9.

[39] T. G. Goetz, J. B. Becker, and C. M. Mazure. Women, opioid use and addiction. *FASEB Journal*, 35, 2021. ISSN 15306860. doi: 10.1096/fj.202002125R.

[40] A. Géron. *Hands-on machine learning with scikit-learn and tensorflow: Concepts, tools, and techniques to build Intelligent Systems*. O'Reilly, 2021.

[41] A. H and V. D. *Multiple Correspondence Analysis. Encyclopedia of Measurement and Statistics*. 2007.

[42] D. H. Han, S. Lee, and D. C. Seo. Using machine learning to predict opioid misuse among u.s. adolescents. *Preventive Medicine*, 130, 2020. ISSN 10960260. doi: 10.1016/j.ypmed.2019.105886.

[43] J. Han, J. Pei, and Y. Yin. Mining frequent patterns without candidate generation. *SIGMOD Record (ACM Special Interest Group on Management of Data)*, 29, 2000. ISSN 01635808. doi: 10.1145/335191.335372.

[44] J. Han, M. Kamber, and J. Pei. *Data Mining: Concepts and Techniques*. Morgan Kaufmann, 2023.

[45] S. Hochreiter and J. Schmidhuber. Long short-term memory. *Neural Computation*, 9, 1997. ISSN 08997667. doi: 10.1162/neco.1997.9.8.1735.

[46] N. Hoque, B. Nath, and D. K. Bhattacharyya. An efficient approach on rare association rule mining. volume 201 AISC, 2013. doi: 10.1007/978-81-322-1038-2_17.

[47] P. Hájek, I. Havel, and M. Chytil. The guha method of automatic hypotheses determination. *Computing*, 1, 1966. ISSN 0010485X. doi: 10.1007/BF02345483.

[48] C. Jensen-Dahm, C. Gasse, A. Astrup, P. B. Mortensen, and G. Waldemar. Frequent use of opioids in patients with dementia and nursing home residents: A study of the entire elderly population of denmark. *Alzheimer's and Dementia*, 11, 2015. ISSN 15525279. doi: 10.1016/j.jalz.2014.06.013.

[49] G. A. Kalkman, C. Kramers, R. T. van Dongen, W. van den Brink, and A. Schellekens. Trends in use and misuse of opioids in the netherlands: a retrospective, multi-source database study. *The Lancet Public Health*, 4, 2019. ISSN 24682667. doi: 10.1016/S2468-2667(19)30128-8.

[50] D. C. Kay, R. B. Eisenstein, and D. R. Jasinski. Morphine effects on human rem state, waking state and nrem sleep. *Psychopharmacologia*, 14, 1969. ISSN 00333158. doi: 10.1007/BF00403581.

[51] A. M. Khedr, Z. A. Aghbari, A. A. Ali, and M. Eljamil. An efficient association rule mining from distributed medical databases for predicting heart diseases. *IEEE Access*, 9, 2021. ISSN 21693536. doi: 10.1109/ACCESS.2021.3052799.

[52] Y. M. Kim, P. Kathuria, and D. Delen. Discovering opioid users' medical comorbidities: a data mining approach. *Journal of Substance Use*, 25, 2020. ISSN 14759942. doi: 10.1080/14659891.2019.1659869.

[53] R. U. Kiran and P. K. Reddy. An improved multiple minimum support based approach to mine rare association rules. 2009. doi: 10.1109/CIDM.2009.4938669.

[54] Y. S. Koh and S. D. Ravana. Unsupervised rare pattern mining: A survey. *ACM Transactions on Knowledge Discovery from Data*, 10, 2016. ISSN 1556472X. doi: 10.1145/2898359.

[55] D. G. Lee, K. S. Ryu, M. Bashir, J. W. Bae, and K. H. Ryu. Discovering medical knowledge using association rule mining in young adults with acute myocardial infarction. *Journal of Medical Systems*, 37, 2013. ISSN 01485598. doi: 10.1007/s10916-012-9896-1.

[56] O. M. Lesch and H. Walter. Subtypes of alcoholism and their role in therapy. volume 31, 1996. doi: 10.1093/oxfordjournals.alcalc.a008221.

[57] W. H. Lo-Ciganic, J. L. Huang, H. H. Zhang, J. C. Weiss, Y. Wu, C. K. Kwoh, J. M. Donohue, G. Cochran, A. J. Gordon, D. C. Malone, C. C. Kuza, and W. F. Gellad. Evaluation of machine-learning algorithms for predicting opioid overdose risk among medicare beneficiaries with opioid prescriptions. *JAMA network open*, 2, 2019. ISSN 25743805. doi: 10.1001/jamanetworkopen.2019.0968.

[58] H. Ma, Y. Lu, and H. Zhang. Determining the near optimal architecture of autoencoder using correlation analysis of the network weights. volume 3, 2016. doi: 10.5220/0006039000530061.

[59] L. J. P. V. D. Maaten, E. O. Postma, and H. J. V. D. Herik. Dimensionality reduction: A comparative review. *Journal of Machine Learning Research*, 10, 2009. ISSN 0169328X. doi: 10.1080/13506280444000102.

[60] P. L. Manfredi, B. Breuer, S. Wallenstein, M. Stegmann, G. Bottomley, and L. Libow. Opioid treatment for agitation in patients with advanced dementia. *International Journal of Geriatric Psychiatry*, 18, 2003. ISSN 08856230. doi: 10.1002/gps.906.

[61] S. M.Kang'ethe and P. W. Wagacha. Extracting diagnosis patterns in electronic medical records using association rule mining. *International Journal of Computer Applications*, 108, 2014. doi: 10.5120/18987-0425.

[62] P. Mosteiro, J. Kuiper, J. Masthoff, F. Scheepers, and M. Spruit. Bias discovery in machine learning models for mental health. *Information (Switzerland)*, 13, 2022. ISSN 20782489. doi: 10.3390/info13050237.

[63] S. Mullin, J. Zola, R. Lee, J. Hu, B. MacKenzie, A. Brickman, G. Anaya, S. Sinha, A. Li, and P. L. Elkin. Longitudinal k-means approaches to clustering and analyzing ehr opioid use trajectories for clinical subtypes, 2021. ISSN 15320464.

[64] N. I. of Health (NIH). Livertox: Clinical and research information on drug induced liver injury [internet]. bethesda (md): National institute of diabetes and digestive and kidney diseases; 2012-. pyrazinamide. *National Institute of Diabetes and Digestive and Kidney Diseases*, 2012.

[65] E. Ordway-West, P. Parveen, and A. Henslee. Autoencoder evaluation and hyper-parameter tuning in an unsupervised setting. 2018. doi: 10.1109/BigDataCongress.2018.00034.

[66] Y. Pan and R. Xu. Mining comorbidities of opioid use disorder from fda adverse event reporting system and patient electronic health records. *BMC Medical Informatics and Decision Making*, 22, 2022. ISSN 14726947. doi: 10.1186/s12911-022-01869-8.

[67] L. P. Phan, N. Q. Phan, V. C. Phan, H. H. Huynh, H. X. Huynh, and F. Guillet. Classification of objective interestingness measures. *EAI Endorsed Transactions on Context-aware Systems and Applications*, 3, 2016. doi: 10.4108/eai.12-9-2016.151678.

[68] C. R. Planey and O. Gevaert. Coincide: A framework for discovery of patient subtypes across multiple datasets. *Genome Medicine*, 8, 2016. ISSN 1756994X. doi: 10.1186/s13073-016-0281-4.

[69] C. C. m. professional. Opioid use disorder: What it is, symptoms treatment. URL `https://my.clevelandclinic.org/health/diseases/24257-opioid-use-disorder-oud`.

[70] W. Qian, Y. Zhang, and Y. Chen. Structures of spurious local minima in k-means. *IEEE Transactions on Information Theory*, 68, 2022. ISSN 15579654. doi: 10.1109/TIT.2021.3122465.

[71] A. Rajabi, M. Dehghani, A. Shojaei, M. Farjam, and S. A. Motevalian. Association between tobacco smoking and opioid use: A meta-analysis, 2019. ISSN 18736327.

[72] Y. P. Raykov, A. Boukouvalas, F. Baig, and M. A. Little. What to do when k-means clustering fails: A simple yet principled alternative algorithm. *PLoS ONE*, 11, 2016. ISSN 19326203. doi: 10.1371/journal.pone.0162259.

[73] A. Rougemont-Bücking, F. Gammab, and J. Pankseppc. Use of tramadol in psychiatric care: A comprehensive review and report of two cases. *Swiss Medical Weekly*, 147, 2017. ISSN 14243997. doi: 10.4414/smw.2017.14428.

[74] I. H. Sarker. Machine learning: Algorithms, real-world applications and research directions, 2021. ISSN 26618907.

[75] A. B. Schelde, A. M. S. Sørensen, M. Hindsø, M. B. Christensen, E. Jimenez-Solem, and R. Eriksson. Sex and age differences among tramadol users in three nordic countries. *Danish Medical Journal*, 67, 2020. ISSN 22451919.

[76] Z. Segal, K. Radinsky, G. Elad, G. Marom, M. Beladev, M. Lewis, B. Ehrenberg, P. Gillis, L. Korn, and G. Koren. Development of a machine learning algorithm for early detection of opioid use disorder. *Pharmacology Research and Perspectives*, 8, 2020. ISSN 20521707. doi: 10.1002/prp2.669.

[77] N. A. Shapira, M. L. Verduin, and J. D. DeGraw. Treatment of refractory major depression with tramadol monotherapy [2] (multiple letters), 2001. ISSN 01606689.

[78] S. Sharma, S. Sharma, and A. Athaiya. Activation functions in neural networks. *International Journal of Engineering Applied Sciences and Technology*, 04, 2020. doi: 10.33564/ijeast.2020.v04i12.054.

[79] M. Spruit. *Translational Data Science in Population Health*. Leiden University, Feb. 2023. doi: 10.5281/zenodo.7665858. URL `https://doi.org/10.5281/zenodo.7665858`.

[80] M. Spruit, T. Dedding, and D. Vijlbrief. Self-service data science for healthcare professionals: A data preparation approach. 2020. doi: 10.5220/0009169507240734.

[81] Y. Sugiyama, R. Tanaka, T. Sato, T. Sato, A. Saitoh, D. Yamada, and M. Shino. Incidence of delirium with different oral opioids in previously opioid-naive patients. *American Journal of Hospice and Palliative Medicine*, 39, 2022. ISSN 19382715. doi: 10.1177/10499091211065171.

[82] T. Tetsunaga, T. Tetsunaga, M. Tanaka, and T. Ozaki. Efficacy of tramadol–acetaminophen tablets in low back pain patients with depression. *Journal of Orthopaedic Science*, 20, 2015. ISSN 14362023. doi: 10.1007/s00776-014-0674-4.

[83] B. Walder, M. R. Tramèr, and R. Blois. The effects of two single doses of tramadol on sleep: A randomized, cross-over trial in healthy volunteers. *European Journal of Anaesthesiology*, 18, 2001. ISSN 02650215. doi: 10.1046/j.1365-2346.2001.00772.x.

[84] R. C. Wang and Z. Wang. Precision medicine: Disease subtyping and tailored treatment, 2023. ISSN 20726694.

[85] L. T. Wu, G. E. Woody, C. Yang, and D. G. Blazer. Subtypes of nonmedical opioid users: Results from the national epidemiologic survey on alcohol and related conditions. *Drug and Alcohol Dependence*, 112, 2010. ISSN 03768716. doi: 10.1016/j.drugalcdep.2010.05.013.

[86] D. Xu and Y. Tian. A comprehensive survey of clustering algorithms. *Annals of Data Science*, 2, 2015. ISSN 2198-5804. doi: 10.1007/s40745-015-0040-1.

[87] L. U. Zebedee, M. W. Bariweni, Y. I. Oboma, and I. G. Ilegbedion. Tramadol abuse and addiction: effects on learning, memory, and organ damage. *Egyptian Pharmaceutical Journal*, 21, 2022. ISSN 20909853. doi: 10.4103/epj.epj_58_21.

[88] X. Zhang, J. Chou, J. Liang, C. Xiao, Y. Zhao, H. Sarva, C. Henchcliffe, and F. Wang. Data-driven subtyping of parkinson's disease using longitudinal clinical records: A cohort study. *Scientific Reports*, 9, 2019. ISSN 20452322. doi: 10.1038/s41598-018-37545-z.

Appendices

A ICPC and ATC codes

A.1 ICPC codes

- **P ICPC codes**

Table 12: 'P' ICPC codes

ICPC	Concern
P01	Feeling anxious/nervous/tense
P02	Crisis/transient stress response
P02.01	Post-traumatic stress disorder
P03	Feeling down/depressed
P04	Irritable/angry feeling/behavior
P05	Feeling/acting old
P06	Insomnia/other sleep disorder
P06.01	Sleep apnea syndrome
P07	Libido loss/reduction
P08	Sexual satisfaction loss/reduction
P08.01	Erectile dysfunction
P08.02	Vaginismus
P08.03	Ejaculation praecox
P09	Concerns about sexual preference
P09.01	Gender Incongruence
P10	Stammering/stuttering/tics
P10.01	Stammering/stuttering
P10.02	Tics/stereotypy
P11	Eating problem(s) in child
P12	Enuresis [ex. U04]
P13	Encopresis
P15	Chronic alcohol abuse
P15.01	Alcoholism
P15.02	Delirium tremens
P15.03	Wernicke-Korsakoff syndrome
P15.05	Problematic alcohol use
P15.06	Binge drinking
P16	Acute alcohol abuse/intoxication
P17	Tobacco abuse
P18	Drug abuse
P19	Substance abuse
P19.01	Soft drug abuse/addiction
P19.02	Hard drug abuse/addiction
P20	Memory/concentration/orientation disorders
P21	Attention-deficit/hyperactivity disorder

Continues on the next page ...

P22	Other child behavior concerns
P23	Other adolescent behavior concerns
P24	Specific learning difficulty
P24.01	Dyslexia
P24.02	Specific language/speech development disorder
P24.03	Motor Development Disorder
P25	Adult life stage problem
P27	Fear of mental illness
P28	Functional limitation/disability mental illness
P29	Other psychological symptoms/complaints
P70	Dementia
P70.01	Alzheimer's disease
P70.02	Vascular dementia
P71	Other organic psychosis(s)
P71.04	Delirium [ex.P15.02]
P72	Schizophrenia
P73	Affective psychosis
P73.02	Bipolar disorder
P74	Anxiety disorder/anxiety state
P74.01	Panic attacks/disorder
P74.02	Generalized anxiety disorder
P75	Somatization disorder
P76	Depression
P76.01	Postpartum depression
P76.02	Dysthymia
P77	Suicide attempt
P77.01	Suicide attempt
P77.02	Suicide
P78	Overvoltage
P79	Other neurosis
P79.01	Phobia
P79.02	Obsessive-compulsive disorder
P80	Personality/character disorder
P80.01	Borderline personality disorder
P80.02	Gambling addiction
P85	Mental retardation/intellectual disability
P98	Other/unspecified psychosis(s)
P99	Other mental disorders
P99.01	Autism/autism spectrum
P99.02	Adjustment disorder
U04	Urinary incontinence [ex. P12]
U04.01	Stress incontinence
U04.02	Urge incontinence
U04.03	Mixed incontinence

- **Z ICPC codes**

Table 13: 'Z' ICPC codes

ICPC	Concern
Z01	Poverty/financial problem
Z02	Food/water problem
Z03	Housing/neighborhood problem
Z03.01	Inadequate housing
Z03.02	Homeless
Z03.03	Neighborhood dispute/noise nuisance
Z04	Problem with social/cultural background
Z04.01	Discrimination against race/religion/sex
Z04.02	Problem due to illegal stay
Z04.03	Loneliness
Z04.04	Problem with retirement
Z04.05	Empty nest syndrome
Z05	Problem with work situation
Z05.01	Impending dismissal
Z05.02	Labor dispute/problem with colleagues
Z05.03	High workload
Z05.04	Exposure to noise at work
Z05.05	Exposure to toxic substances at work
Z06	Problem with unemployment
Z07	Problem with training
Z07.01	Illiteracy
Z07.02	Falling behind at school
Z07.03	Failed exam
Z08	Problem with social insurance/welfare
Z08.01	Problem with sickness benefit/WAO
Z08.02	Problem with social assistance
Z09	Problem with justice/police
Z09.01	Imprisonment/imprisonment
Z09.02	Custody/alimony issue
Z10	Problem of accessibility/availability of healthcare
Z10.01	Awaiting treatment
Z10.02	Waiting at the location of a care/nursing home
Z11	Problem with being sick
Z12	Relationship problem with partner
Z12.01	Separation from spouse
Z12.02	Abuse/sexual abuse by partner
Z13	Problem with partner's behavior
Z13.01	Problem with partner's addiction
Z13.02	Problem with partner aggression
Z13.03	Problem with partner's infidelity
Z14	Problem with partner's illness
Z15	Loss/death of partner
Z16	Relationship problem with child
Z16.01	Maltreatment/sexual abuse of child

Continues on the next page ...

Z16.02	Neglect of child
Z16.03	Raising a child outside the family context
Z18	Problem with child's illness
Z19	Loss/death of child
Z20	Relationship problem with parents/family
Z21	Problem with parental/family behavior
Z21.01	Problem with addiction of parents/family
Z21.02	Problem with parental/family aggression
Z22	Problem with illness of parents/family
Z23	Loss/death of parents/family
Z24	Relationship problem with friends
Z25	Problem due to violence
Z25.01	Problem resulting from sexual assault/rape
Z27	Fear of having a social problem
Z28	Social disability/disability
Z29	Other social problem
Z29.01	Burnout
Z29.02	Lack of free time/relaxation
Z29.03	Dependence on others

A.2 ATC codes

The ATC codes corresponding to opioid prescriptions in given in table 14.
Abbreviations used:

- Unit
  1. g = gram
  2. mg = milligram
- Route of Administration
  1. N = nasal
  2. O = oral
  3. P = parenteral
  4. R = rectal
  5. SL = sublingual/buccal/oromucosal
  6. TD = transdermal

Table 14: ATC codes

ATC	Medication
N01AH01	fentanyl
N01AH02	alfentanil

Continues on the next page ...

N01AH03	sufentanil
N01AH04	phenoperidine
NA01AH05	anileridine
NA01AH06	remifentanil
N01AH51	fentanyl, combinations
N02AA01	morphine (0.1g O/ 30mg P/ 30mg R)
N02AA02	opium
N02AA03	hydromorphone (20mg O/ 4mg P/ 4mg R)
N02AA04	nicomorphine (30mg O/ 30mg P/ 30mg R)
N02AA05	oxycodone (75mg O/ 30mg P)
N02AA08	dihydrocodeine (0.15g O)
N02AA10	papaveretum
N02AA11	oxymorphone
N02AA51	morphine, combinations
N02AA53	hydromorphone and naloxone
N02AA55	oxycodone and naloxone (75mg O)
N02AA56	oxycodone and naltrexone
N02AA58	dihydrocodeine, combinations
N02AA59	codeine, combinations excl. psycholeptics
N02AA79	codeine, combinations with psycholeptics
N02AB01	ketobemidone (50mg O/ 50mg P)
N02AB02	pethidine (0.4g O/ 0.4g P/ 0.4g R)
N02AB03	fentanyl (0.6mg N/ 0.6mg SL/ 1.2mg TD)
N02AB52	pethidine, combinations excl. psycholeptics
N02AB72	pethidine, combinations with psycholeptics
N02AC01	dextromoramide (20mg O/ 20mg P/ 40mg R)
N02AC03	piritramide (45mg P)
N02AC04	dextropropoxyphene (0.2g O/ 0.3g O)
N02AC05	bezitramide (15mg O)
N02AC52	methadone, combinations excl. psycholeptics
N02AC54	dextropropoxyphene, combinations excl. psycholeptics
N02AC74	dextropropoxyphene, combinations with psycholeptics
N02AD01	pentazocine (0.2g O, 0.2g P)
N02AD02	phenazocine (3mg P)
N02AD51	pentazocine and naloxone
N02AE01	buprenorphine (1.2mg P/ 1.2mg SL/ 1.2mg TD)
N02AF01	butorphanol (12mg P)
N02AF02	nalbuphine (80mg P)
N02AG01	morphine and antispasmodics
N02AG02	ketobemidone and antispasmodics
N02AG03	pethidine and antispasmodics
N02AG04	hydromorphone and antispasmodics
N02AJ01	dihydrocodeine and paracetamol
N02AJ02	dihydrocodeine and acetylsalicylic acid
N02AJ03	dihydrocodeine and other non-opioid analgesics
N02AJ06	codeine and paracetamol

Continues on the next page ...

N02AJ07	codeine and acetylsalicylic acid
N02AJ08	codeine and ibuprofen
N02AJ09	codeine and other non-opioid analgesics
N02AJ13	tramadol and paracetamol
N02AJ14	tramadol and dexketoprofen
N02AJ15	tramadol and other non-opioid analgesics
N02AJ16	tramadol and celecoxib
N02AJ17	oxycodone and paracetamol
N02AJ18	oxycodone and acetylsalicylic acid
N02AJ19	oxycodone and ibuprofen
N02AJ22	hydrocodone and paracetamol
N02AJ23	hydrocodone and ibuprofen
N02AX01	tilidine (0.2g O/ 0.2g P)
N02AX02	tramadol (0.3g O/ 0.3g P/ 0.3g R)
N02AX03	dezocine
N02AX05	meptazinol (1.2g O/ 1.2g P)
N02AX06	tapentadol (0.4g O)
N02AX07	oliceridine
N02AX51	tilidine and naloxone

B Algorithms

Algorithm 1 Apriori Algorithm for discovering frequent itemsets[44]

Input:

1. D, a dataset of transactions;
2. *min_sup*, the minimum support count threshold

Output: L, frequent itemsets in D

Method:

```
L_1 = frequent_1_itemsets(D)
for k = 2; L_{k-1} ≠ ∅; k++ do
    C_k = APRIORI_GEN(L_{k-1})
    for each transaction t ∈ D do                    ▷ scan D for counts
        C_t = subset(C_k, t)                         ▷ get the subsets of t that are candidates
        for each candidate c ∈ C_t do
            c.count++
        end for
    end for
    L_k = {c ∈ C_k | c.count ≥ min_sup}
end for
return L = ∪_k L_k

procedure APRIORI_GEN(L_{k-1})
    for each itemset l_1 ∈ L_{k-1} do
        for each itemset l_2 ∈ L_{k-1} do
            if (l_1[1] = l_2[1]) ∧ (l_1[2] = l_2[2]) ∧ ... ∧ (l_1[k-2] = l_2[k-2]) ∧ (l_1[k-1] < l_2[k-1])
then
                c = l_1 ⋈ l_2                        ▷ Join step
                if HAS_INFREQUENT_SUBSET(c, L_{k-1}) then
                    delete c
                else
                    add c to C_k
                end if
            end if
        end for
    end for
    return C_k
end procedure

procedure HAS_INFREQUENT_SUBSET(c, L_{k-1})
    for each (k-1)-subset s of c do
        if s ∉ L_{k-1} then
            return TRUE
        end if
    end for
    return FALSE
end procedure
```

Algorithm 2 FP-growth Algorithm for discovering frequent itemsets[44]

Input:

1. D, a dataset of transactions;
2. *min_sup*, the minimum support count threshold

Output: The complete set of frequent patterns
Method:

1. The FP-tree is constructed in the following steps:
   (a) Scan the transaction database D once. Collect F, the set of frequent items, and their support counts. Sort F in support count descending order as L, the list of frequent items.
   (b) Create the root of an FP-tree, and label it as "null." For each transaction *Trans* in D do the following. Select and sort the frequent items in *Trans* according to the order of L. Let the sorted frequent item list in *Trans* be $[p|P]$, where p is the first element and P is the remaining list. Call INSERT_TREE($[p|P], T$), which is performed as follows. If T has a child N such that $N.itemname = p.itemname$, then increment N's count by 1; else create a new node N, and let its count be 1, its parent link be linked to T, and its node-link to the nodes with the same itemname via the node-link structure. If P is nonempty, call INSERT_TREE(P, N) recursively.
2. The FP-tree is mined by calling FP_GROWTH($FP\_tree, null$), which is implemented as follows.

procedure FP_GROWTH($Tree, \alpha$)
  if $Tree$ contains a single path P **then**
    for each combination (denoted as β) of the nodes in the path P **do**
      generate pattern $\beta \cup \alpha$ with *support_count* = *minimum support count of nodes in* β
    end for
  else
    for each a_i in the header of *Tree* **do**
      generate pattern $\beta = a_i \cup \alpha$ with $support\_count = a_i.support_count$
      construct β's conditional pattern base and then β's conditional FP_tree $Tree_\beta$
      if $Tree_\beta \neq \emptyset$ **then**
        call FP_GROWTH($Tree_\beta, \beta$)
      end if
    end for
  end if
end procedure