UNIVERSITY OF TORINO

Ph.D. in Modeling and Data Science

XXXV cycle

Final dissertation

UNIVERSITÀ
DI TORINO

Multimorbidity Analysis and Covid-19 Severity Prediction using Machine Learning and Evolutionary Algorithms

Supervisors:
Giuseppe Costa,
Fulvio Ricceri,
Mario Giacobini

Candidate: Dayana Benny

ACADEMIC YEAR 2022/2023

Summary

Multimorbidity holds paramount importance in public health, representing a multidimensional state where multiple pre-existing medical conditions coexist and interact. This condition has been linked to an elevated risk of COVID-19. Those with multimorbidity who succumb to COVID-19 experienced a substantial increased risk of negative outcomes. The post-pandemic period also sees an acceleration of frailty. Therefore, it is imperative to incorporate existing multimorbidity details into epidemiological risk assessments. Handling clinical data with medical history poses significant challenges, notably the data's sparsity due to the rarity of multimorbidity conditions and the intricate enumeration of combinatorial multimorbidity features, which introduces a combinatorial explosion issue.

In this research, based on the health administrative data of the piedmont region, each patient profile in the dataset is depicted as a binary vector, where each feature denotes the presence or absence of a specific multimorbidity condition. In the first project of this research, by clustering the sparse medical data, newly engineered features are generated as a bin of features, and they are combined with the prevalent features for COVID-19 severity predictive modeling. In the second project, a sparsity-addressing Evolutionary Machine Learning model for analyzing pre-existing multimorbidity in COVID-19 hospitalized patients using their medical history is proposed. This research attempt to discover the optimal set of multimorbidity feature combinations that are highly associated with COVID-19 severity.

This research distinguishes the severity of COVID-19 on infected people who have multiple medical conditions alongside their demographic characteristics, age, and sex. Contrary to misconceptions, the concept of multimorbidity analysis of COVID-19 patient is outdated, this research introduces a groundbreaking tool designed to analyze the intricate interactions among diverse chronic health conditions and their collective impact that could be useful in situations analogous to recent health crises.

Acknowledgements

I began my PhD journey around a month after the initial outbreak of the Covid-19 pandemic in Northern Italy, a period marked by a multitude of obstacles and unpredictabilities. Nonetheless, I persevered and reached the finish line! I would like to express my profound gratitude to all the individuals who offered wonderful support and provided guidance throughout this transformative journey, enabling me to overcome challenges.

First and foremost, my profound gratitude extends to my PhD supervisors: Prof. Giuseppe Costa, Prof. Fulvio Ricceri, and Prof. Mario Giacobini. Their consistent support, guidance, and encouragement proved invaluable throughout the entire journey. I am deeply indebted for the immense contributions they made to my personal and academic growth.

In addition to my supervisors, I am profoundly grateful to my ever-supportive PhD coordinator, Prof. Laura Sacerdote, whose unwavering support served as a continuous wellspring of motivation. Special appreciation is reserved for Dr. Roberto Gnavi, Alberto Catalano for their valuable collaboration on my research, and my exceptional colleagues and peers at the Unit of Epidemiology, ASL TO3 and Modeling & Data Science Doctoral School.

To my better half Lippin, and my son Edward, whose unwavering belief in my abilities and unwavering support were instrumental in my achievements, I offer my deepest thanks. To my Mom Princy, Pappa Benny (in heaven), and all my family members, along with all those who imparted their wisdom to me, I express my heartfelt gratitude. Last but not least, I dedicate this PhD thesis to the Almighty Jesus; thank you for everything!

Contents

List of Tables

List of Figures

Chapter 1

Introduction

1.1 Background

Coronavirus Disease 19 (COVID-19) is classified as a highly infectious disease, posing a severe threat to vulnerable populations, and thus, it is a critical public health concern and a significant epidemiological situation worldwide. On February 21, 2020, the first Italian case of COVID-19 was diagnosed in the Lombardy region. The virus rapidly spread throughout the country, resulting in a nationwide lockdown and overwhelming the healthcare system. Italy was one of the countries that suffered the most from the COVID-19 pandemic, with Piedmont, a region in the northwest of Italy, being one of the areas with a large number of instances in the first wave.

In the case of people infected with Severe Acute Respiratory Syndrome Coronavirus 2 (SARS-CoV-2), multimorbidity can increase the severity of the infection [1, 2]. Multimorbidity refers to the presence of multiple coexisting medical conditions in a patient, which interact with each other and can result in a complex and multidimensional health condition [3]. It has been established at population level that interactions between diseases can increase the severity of the medical condition and make the treatment of other diseases in the combination complex [4, 5]. Therefore, it is important to identify specific disease combinations that could have an impact on the severity of COVID-19 among individuals with multimorbidity.

Multimorbidity is consistently linked to a lower health-related quality of life in mid-life [6, 7]. Additionally, there is evidence suggesting that women have a higher likelihood of developing multimorbidity compared to their male counterparts [8]. Moreover, having multiple health problems at the same time has been found to make healthcare more expensive and

create difficulties for healthcare systems in terms of resource allocation and providing appropriate care [9].

Moreover, multimorbidity can worsen the effects of long COVID in several ways [10, 11], when multimorbidity is present, additional symptoms related to other chronic conditions can compound the overall symptom burden, making it more challenging for individuals with long COVID to manage and recover from their illness. Research studies have indicated that individuals with multimorbidity have been adopting various precautionary behaviors during the pandemic [12, 13]. This is reflected in the restrictive guidelines recommended by authorities to control transmission [14]. Furthermore, studies have found that females are more likely to adopt protective measures compared to males [13]. The difference in precautionary behaviors based on gender underlines the importance of considering various demographic factors in the development of public health interventions during a pandemic.

It is important to note that having one or more of these chronic health conditions does not necessarily mean that an individual will develop severe COVID-19, but it does increase the risk and that different diseases could act in a different way on COVID-19 outcomes. Therefore, identifying specific disease combinations and studying the interactions between different chronic health conditions is relevant when analyzing the severity of COVID-19 among individuals with multimorbidity. This can help healthcare professionals to identify those at highest risk of severe complications and provide appropriate prevention, care, and treatment.

1.2 Challenges in Dealing with Multimorbidity Clinical Data

Studying multimorbidity using traditional methods can be labor-intensive and requires identifying high-dimensional combinatorial features, especially when dealing with data coming from health administrative registers. Furthermore, there is no universally accepted list of medical conditions to elucidate the state of multimorbidity. To address these challenges, efforts must be put forward to identify low-dimensional representations of multimorbidity features for effective prediction of outcome. High-order input features make Machine Learning models more prone to overfitting, and identifying meaningful high-order combinatorial features requires extensive effort from experts with domain knowledge.

Rare features such as diseases and drugs with low occurrence rates in the data can pose significant challenges for both statistical and Machine Learning analyses. This is because their lower prevalence in the data can result in sparsity, which may lead to poor predictions. The sparse data generated by the absence of medical conditions in a cohort is illustrated in Figure 1.1

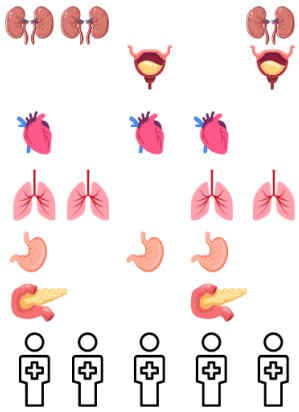


Figure 1.1: Absence of medical conditions in a cohort - sparse data

The challenges of sparsity in predictive modeling include (i) biased and unreliable results, (ii) imbalances in data due to under representation by certain medical conditions, (iii) hampering of the model's ability to effectively learn from the data, (iv) noise and difficulty in pattern detection.

1.3 State of the Art: Multimorbidity Analysis

Traditionally, research in multimorbidity relies on counting the total number of chronic conditions instead of considering the difference in individual experiences and the effects of various combinations of diseases [15]. Count-based multimorbidity measures have been used for emergency hospitalization prediction [16]. Patients with the same number of chronic conditions may have vastly different experiences. For instance, two patients with three chronic conditions each might have completely different symptoms, treatment regimens, and levels of disability. Counting conditions alone does not account for these individual variations. In [15], authors suggest to undertake the task of identifying and condensing prevailing multimorbidity indices used to assess multimorbidity beyond mere disease counts. In a recent literature [17], researchers state that these indices might offer limited clinical practicality. Because, even though an index score can anticipate particular

outcomes within a population or aid in mitigating confounding variables in various research, it typically does not help in patient-level management or offer guidance for interventions. Moreover, Indices for measuring multimorbidity fail to adequately consider the intricate connections between various chronic health conditions [17]. In this context, Machine learning techniques can play a crucial role in identifying and characterizing these intricate multimorbidity patterns, offering valuable insights for personalized patient care and treatment strategies [17, 18].

Frequent combinations of medical conditions have been reported to describe multimorbidity patterns [19, 20]. According to [19], in the context of the COVID-19 pandemic, there are various combinations of multiple health conditions that people experience. Some of these combinations have been consistently common since the start of the pandemic, while others occur less frequently but tend to appear when there are more cases of COVID-19. Moreover, when a specific combination of health conditions is of low prevalence, it can make clinical management more complicated since they often include Orphan diseases [19].

Previous works have also investigated multimorbidity combinations through latent class [21], cluster-based [22], network-based [23], factor-based [24].

A latent class is a concept used in statistical modeling and analysis to represent unobservable or hidden groupings or categories within a dataset [25]. These groupings are inferred based on patterns or associations in observed data. In [21], latent class analysis was used to group patients based on similar combinations of long-term health conditions, capturing complex interactions between these conditions while maintaining specificity. The study then analyzed patient characteristics and treatment patterns among these multimorbidity phenotype clusters, using statistical tests. A limitation of using latent class analysis (LCA) in the context of grouping multimorbidity is that it may not be well-suited for scenarios where the underlying health conditions exhibit continuous or overlapping characteristics, as LCA assumes categorical, mutually exclusive, and exhaustive subgroups [26], which may not fully capture the complexity of certain multimorbidity patterns that can involve overlapping conditions or varying degrees of severity.

In [22], the sum of squared errors determines the number of clusters and the ratio of within-cluster variance to between-cluster variance. Later, the individuals are put into various clusters based on the composition of their health conditions. A limitation of the described study is that the method

assigns individuals to clusters based solely on the majority composition of their health conditions within a cluster, which might not fully capture the complexity of individual health profiles, potentially overlooking important nuances in multimorbidity patterns.

In [23], authors identified complex network of disease associations using network graphs. However, a potential limitation is that the network analysis often relies on predefined relationships or connections between variables, which may not capture more complex and non-linear relationships that Machine Learning algorithms can uncover.

In [24], authors used Exploratory Factor Analysis (EFA), a statistical a statistical method used to uncover latent patterns or relationships among variables in a dataset without making prior assumptions about the structure of those relationships. This method can identify interactions among variois medical conditions. But, EFA relies on the assumption that observed variables are linear combinations of underlying latent factors [27] and doesn't have predictive capabilities, which can be a drawback when the goal is to make predictions or decisions based on data patterns, a task more commonly addressed by Machine Learning techniques.

As many studies employ various techniques to discern patterns and relationships among multimorbidity, we can see that Machine Learning holds distinct advantages over these approaches. However, it is worth noting that research specifically addressing sparse multimorbidity data, where there are limited instances or occurrences of comorbid conditions, is currently lacking, and this presents a gap in the literature. Therefore, while Machine Learning offers significant potential for multimorbidity analysis, addressing the unique challenges posed by sparse data remains an area that requires further exploration and investigation within the field.

Some works that use Machine Learning to investigate multimorbidity patterns address sparsity in the dataset by either removing the sparsity-generating features [28], merging the categories of features after performing one-hot encoding [29], or clustering the rare features [30]. However, while these methods may reduce sparsity, they may also lead to the loss of important information and hinder the meaningful interpretation of multimorbidity features.

From literature it is evident that combinations of less prevalent medical conditions are markedly associated with worse outcomes and amplify the risks associated with individual conditions [31].

1.4 Importance of Proposed Method for Multimorbidity Research

With the growing prevalence of electronic health records and other large datasets, there is a growing need for efficient and effective methods to analyze and understand multimorbidity. By leveraging Machine Learning algorithms and other advanced computational techniques, researchers can gain deeper insights into the underlying mechanisms and risk factors associated with multimorbidity, which can ultimately inform more effective prevention and treatment strategies.

Multimorbidity is typically associated with deficient health-related quality of life in mid-life, and the likelihood of developing multimorbidity in women is elevated. In some studies, epidemiological data reveals no visible sex-based discrepancy in disease severity, suggesting that the progression of the virus is comparably favorable in both women and men, and there is a similarity in the age at which the rate of SARS-CoV-2 infection peaks for both male and female [32], [33]. However, the specific comorbidities that increase the risk of severe COVID-19 outcomes can differ significantly between men and women [34]. Also, according to literature, women appear to be relatively less susceptible to SARS-CoV-2 infection than men [35]. This underscores the need for a refined understanding of gender-specific factors influencing susceptibility and outcomes in the context of the COVID-19 pandemic. While existing literature provides valuable insights, there is a distinct lack of in-depth investigation specifically focusing on women [32]. To comprehensively address this gap in knowledge, it is imperative to advocate for targeted research works dedicated to understanding the unique aspects of women's vulnerability or protection against COVID-19.

The first project of this research address the issue of data sparsity in non-prevalent features by clustering the binary data of various rare medical conditions in a cohort of middle-aged women. This study aims to enhance understanding of how multimorbidity affects COVID-19 severity by clustering rare medical conditions and combining them with prevalent features for predictive modeling.

In this project, clustering is performed on less prevalent features and put such features into various Bins to enhance the interpretability of our data. By strategically grouping less common features into Bins and integrating them with prevalent ones, this research aim to capture a comprehensive picture of multimorbidity among women in midlife.

To group the multimorbidity features into various bins, a matrix is reconstructed based on the cluster structures. The clustering process involves two levels: feature level and data level. Feature level is performed to assign features into different clusters which are the Bins and data level clustering is performed where patients' records are grouped into clusters based on the features within each Bin before predictive modeling.

Constructing clusters of multimorbidity and interpreting the outcomes at the patient level aids in identifying, in case of future patients, which cluster value of a Bin contribute to whether a group of patients will be hospitalized or not due to COVID-19. Furthermore, in this study, identifying the most predictive feature or a Bin that includes less prevalent features helps in revealing the underlying combination of multimorbidity that predicts the severity of COVID-19 among the studied cohort. The insights gained can guide the development of targeted interventions and improved management strategies for individuals with multiple health conditions.

In second project, an Evolutionary Algorithm with deep learning-based feature scoring is used and it is a powerful approach for analyzing multimorbidity data [36]. The application of evolutionary model might be better not because of its higher predictive performance alone but because it handles sparsity more effectively. This could manifest in better identification of key features, more stable predictions, or better performance in certain subgroups of the data [37].

Also, a logistic model can uncover complex multimorbidity patterns [38]. However, while linear models offer high interpretability, they may fall short in sparse datasets where feature selection is key [39]. Here, evolutionary algorithms, particularly Genetic Algorithms, excel by efficiently navigating complex feature interactions and identifying optimal feature subsets, a task challenging for linear models in sparse data scenarios [40].

This method involves several steps to identify the most relevant features for predicting the target variable while minimizing the number of features used. The dataset is preprocessed by generating various subsets or bins of the multimorbidity features using a feature binning approach. This step reduces sparsity in the data and allows for more efficient feature scoring. Next, deep learning is used to score the features within each subset based on their importance for predicting the target variable. The output of this step is a feature score for each feature within each subset. An Evolutionary Algorithm is then applied to select the best subset of features based on their scores. The algorithm generates a population of candidate feature subsets

and iteratively improves the population through selection, crossover, and mutation operations [41]. The fitness of each candidate solution is evaluated using a fitness function that incorporates the deep learning-based feature scores of each subset or bin of features.

The output of the Evolutionary Algorithm is a subset of features that are most relevant for predicting the target variable. These features can be used for further analysis, such as building a predictive model or identifying the underlying associations of the multimorbidity patterns. In summary, an Evolutionary Algorithm with deep learning-based feature scoring provides a powerful approach for analyzing multimorbidity data by identifying the most relevant features for predicting the target variable. This approach can lead to better model performance, faster training times, and improved interpretability in complex datasets with multimorbidity features [42].

1.5 Goal of this Research

1.5.1 Research question

What are the multimorbidity predictors of severe COVID-19 outcomes (specifically, as proxy of a more severe COVID-19 outcome), considering the sparsity challenges posed by rare features in the data and the optimal set of morbidity feature combinations that are highly associated with COVID-19 severity.

1.5.2 Scope of the research

The research aims to mitigate the challenges while dealing with sparse multimorbidity data and develop effective models to predict severe COVID-19 outcomes and to identify the specific combinations of medical conditions that are most strongly associated with severe COVID-19 outcomes.

Rare diseases, characterized by their lower prevalence in the population, are a diverse group of conditions affecting only a small fraction of individuals [31]. A recent study have shown a clear link between rare diseases and negative inpatient outcomes, suggesting these patients may require individualized care protocols [31]. Combining less prevalent conditions with other morbidity features might create complex interactions that amplify or diminish the impact on the outcome. The impact of less prevalent medical conditions on outcomes is a complex and under-researched area [43].

The 2001 article suggests that future research on comorbidity consequences should focus on specific combinations of diseases [44]. But apparently, there is still limited knowledge regarding the outcomes when morbidity conditions are considered together [45].

In this study, it is demonstrated how the innovative tool used in this research has the potential to revolutionize traditional risk assessment approaches. By incorporating intricate combinations of diseases, the tool aims to enhance the accuracy of predicting severe outcomes for individuals who have multiple chronic conditions. Through its adaptable design, it ensures applicability even in evolving scenarios of different communicable diseases, underscoring its continued relevance. This study focuses on investigating the complexities of disease interactions, showcasing how this aforementioned tool could reshape risk assessment for similar contexts.

Chapter 2

Methodology

2.1 Multimorbidity Dataset

Data for the multimorbidity analyses were gathered from the Piedmont Longitudinal Study (PLS), which is a health-administrative cohort composed of anonymous records linked at the individual level from various social, health, and administrative databases [2]. Furthermore, since February 2020, the PLS has been enhanced by the regional COVID-19 platform that collects COVID-19 infection data. From these databases, we used registers for: (i) hospital discharges, (ii) drug prescriptions data and the (iii) COVID-19 hospitalizations of the individuals diagnosed with a SARS-CoV-2 infection for the first time between February 22, 2020 and May 31, 2020. We retrieved the 5-year medical history of COVID-19-positive patients from these datasets. The extracted data consists of 12,793 people aged 45 to 74 years who tested positive for the first time for the SARS-COV-2 infection. Table 5.1 depicts the descriptive statistics of the dataset utilized.

This research focused on individuals who were aged 45 to 74 years, eliminating the potential influence of both younger (people aged less than 45) and elderly subjects (people aged 75 and more) on the results. Because, there are clinical differences between younger and elderly COVID-19 patients [46] and growing older is linked to a rising prevalence of multiple health conditions [46]. Additionally, the non utilization of the patient profiles of people aged 75 and more allowed us to eliminate any bias associated with patients residing in nursing homes.

As the study was incorporated into the National Statistical Plan, no ethical approvals or permits were required and the database used for the analyses included anonymized data only.

2.2 Construction of the Exposure Variables

In this longitudinal cohort study, the patients with their presence and absence of various multimorbidity in the past 5 years (2015 - 2019) are compared for a particular outcome (hospitalization due to COVID-19). It's important to note that the study population exclusively consisted of COVID-19 patients, and the primary focus was on assessing hospitalization as the outcome. Multimorbidity has been defined using records from hospital discharges' and drug prescriptions' registers. In the case of hospital discharges and drug prescriptions datasets, there are multiple entries for a single COVID-19 infected patient. The drug prescriptions dataset consists of around 1 million records; the hospital discharges dataset consists of around 19,000 entries. From the drug prescriptions dataset, the Anatomical Therapeutic Chemical classification system (ATC) code is used. All distinct ATC codes up to the 4th level (the first 5 digits of the ATC codes) are considered in this study. One-hot encoding is carried out for converting categorical codes into different feature columns with 0 or 1 values (absence and presence of drugs in the history of prescriptions for each patient). Similarly, from the hospital discharge data, the 9th International Classification of Diseases-Clinical Modification (ICD9-CM) code [47] (as a diagnosis code of disease) is used and one-hot encoding is performed. After these transformations, only the drug codes and diagnosis codes that comply with this condition are kept: "at least 100 patients with this code in the COVID-19-positive patients' data". Thus, 194 features are derived from drug codes (112) and diagnosis codes (82) as multimorbidity features from the whole data, where the presence and absence of those features are marked as 0 and 1, respectively. The other two features are age and sex, where sex is coded as 1 for female and 0 for male. Later, this pre-processed data is divided into four datasets based on age and sex. The dataset transformation steps are illustrated in Figure 2.1.

The study conducted separate analyses for four cohorts, which are subsets of the original data stratified by sex and age. A combined analysis can be informative only if the research is primarily interested in the overall effect of the main variable, regardless of subgroups. Also, combining data might be acceptable if the effect modifier has a small impact on the relationship.However, based on existing literature [2, 48], age and sex are recognized as effect modifiers, making separate analyses more appropriate.

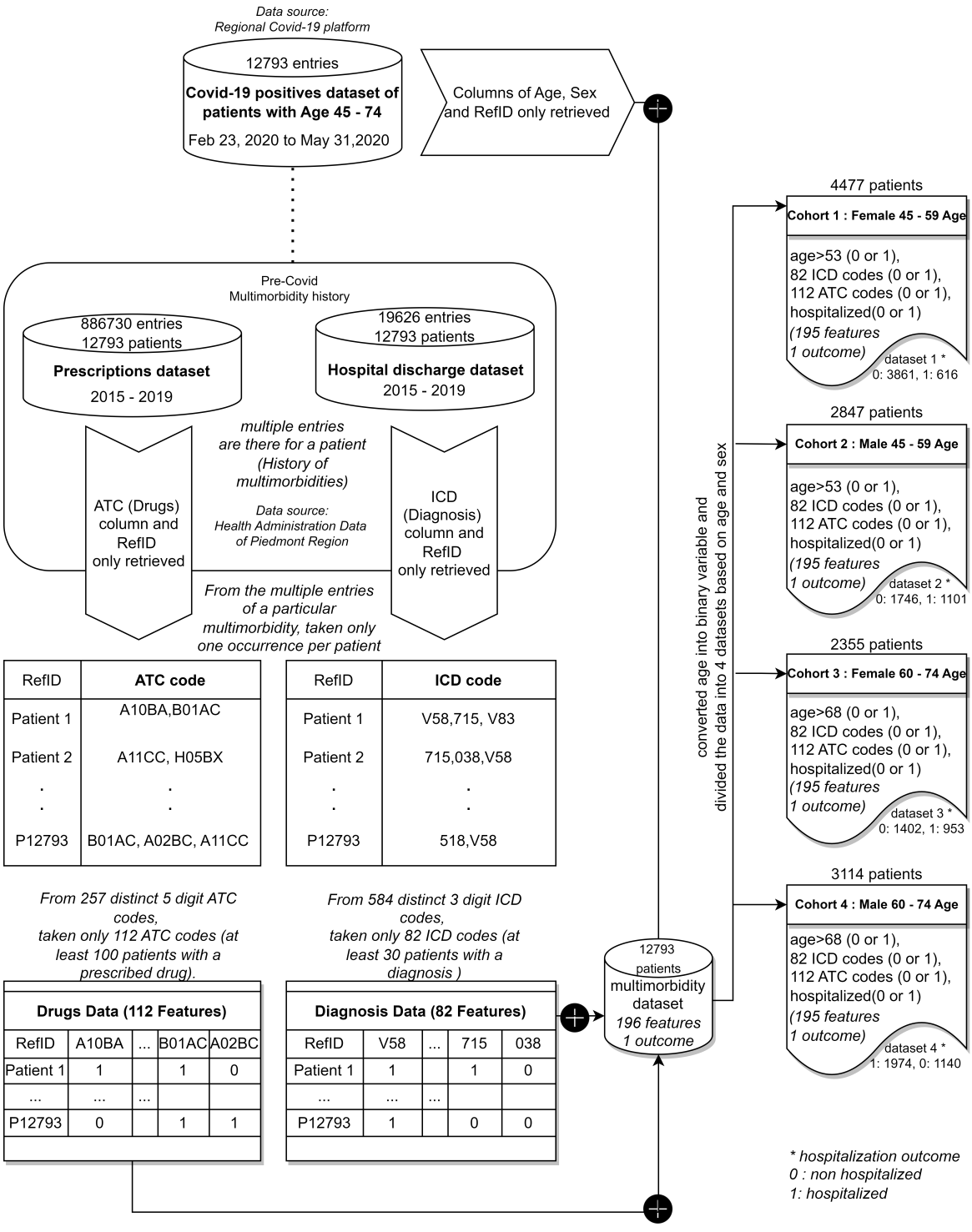


Figure 2.1: Dataset transformation steps

Subsets of various cohorts are obtained by considering the study population who fall within the age criteria of "aged 45 – 59 years" and "aged 60 – 74 years". This sub-division is performed since people over 60 can be considered as part of the older population.

The various cohorts in this research are : cohort 1 - Female COVID-19 patients aged 45 to 59, cohort 2 - Male COVID-19 patients aged 45 to 59, cohort 3 - Female COVID-19 patients aged 60 to 74 and cohort 4 - Male COVID-19 patients aged 60 to 74.

In the datasets for the middle-aged cohorts (cohort 1 and cohort 2), the age feature is converted into a binary variable where 1: age>53, 0: age≤53. The age values are taken from the 2020 COVID-19 data, and the age of 53 is used as a threshold to divide the younger population into two subgroups (45-53, 54-59). Similarly, the elderly population is also divided into two subgroups (60-68, 69-74), where the age feature is converted into a binary variable where 1: age>68, 0: age≤68. All four cohort datasets are considered as separate binary classification problems. The input variables, which include multimorbidity history and age, and the outcome variable, which denotes whether a patient was hospitalized due to COVID-19 or not, are represented as binary values.

In this study, the presence and absence of prescribed drugs and diagnosed diseases including patient age, and sex were considered as the multimorbidity features. However, due to the rarity of many medical conditions in the study population, the resulting dataset becomes sparse when encoding absence as zero values.

2.3 Data Imbalance Rectification

A major challenge when handling clinical data is predicting rare events, which can lead to an imbalance problem when the target feature has more observations in a class than in other classes [49]. Therefore, it is essential to treat the imbalanced raw data properly to avoid bias towards a particular class. Oversampling, a data imbalance rectification process offers the advantage of retaining all information from the original training set, as it preserves all members from both the minority and majority classes [50]. In a recent study that compares performance of various data resampling methods on imbalanced medical data [51], the findings indicate that utilizing class imbalance techniques can contribute to the diagnosis of lung cancer. The study suggests that among various imbalanced learning methods, the

oversampling technique demonstrates superior performance.

SMOTE works by connecting existing data points, but when features are binary, creating intermediate values through interpolation is not meaningful [52]. This can result in synthetic data points that may not accurately reflect the original data distribution. There are variations for SMOTE that specifically handles nominal features like binary data by randomly picking a neighbor instead of interpolating. SMOTE-NC is an extension of the original SMOTE algorithm designed to handle datasets with both nominal and continuous features [53]. However, for SMOTE-NC to function properly, the dataset must contain at least one continuous attribute [54]. In this research not only all the datasets are unbalanced, but also all the data points are binary and there is no continuous valued attribute. Moreover, the performance assessment of different interpolation methods confirmed previous findings regarding the limited effectiveness of established SMOTE-based variations.

Also, oversampling introduces the drawback of significantly increasing the training set size; undersampling, on the other hand, outperforms oversampling in time and memory complexity when resampling time is not considered [55]. In a study [56] addressing class imbalance in cardiovascular data, the well-known SMOTE oversampling [53] technique is utilized, accompanied by exploration of under-sampling methods. Also, they used an undersampling technique, and experimental results demonstrate its superior performance compared to existing methods.

The findings from a recent study show that Cost-Sensitive Learning (CSL) is effective in predicting imbalanced medical data [57]. Instead of artificially creating balanced class distributions through sampling techniques, CSL addresses the imbalanced class issue by using cost matrices that specify the costs linked to misclassification for each class [57]. This cost matrix is used during the training process to adjust the model's behavior by assigning costs or penalties to different types of classification errors, influencing the learning algorithm to prioritize certain outcomes over others. Sparse and imbalanced binary datasets may exhibit unique characteristics that are not well-represented in the training set. CSL might struggle to generalize to unseen sparse patterns, impacting the model's performance on new data. In cost-sensitive learning, features might be used to dynamically adjust the misclassification costs for different instances [58]. Also, adaptive cost adjustment may introduce additional computational complexity [58], particularly if the algorithm needs to iteratively update costs during the

training process. This can impact the efficiency of the learning algorithm, especially when using in conjunction with evolutionary approach.

Class weighting allows the algorithm to retain all instances in the dataset, including those from the majority class [59]. In binary data, class weighting often assigns higher weights to the minority class, which can inadvertently push the model to solely prioritize its prediction. Also, studies show that if the class imbalance is moderate, under-sampling can be advantageous [60].

All the datasets used in this study are unbalanced, and resampling is proposed. To achieve this, randomly class-balanced sample data is taken from the unbalanced original dataset, followed by a statistical hypothesis test called the one proportion z-test. This test is carried out to draw an analogy between the proportion of the sampled population and the population in raw data. This test ensures the representativeness of randomly balanced sample data and the original cohort dataset, avoiding any potential bias.

The steps performed to obtain an unbiased balanced dataset with significant features are:

- Extract all minority and majority samples attributed to the outcome value from the original cohort dataset.
- Randomly select samples belonging to the majority class such that they are equal in number to the minority class to obtain a balanced dataset.
- Calculate the prevalence of each feature in the randomly selected samples and the original population.
- Conduct one proportion z-test on all non-zero variables to determine whether the frequency distribution of a feature in the resampled data is representative of the same feature in the original cohort dataset, using a significance level of .05.
- Evaluate the obtained one proportion z-test statistic and P values to support the significance of the conclusion of the test and eliminate non similar features. Thus the features for which there is statistical evidence of a significant difference in proportions between the original and resampled datasets are eliminated from the sampled data.

The rationale for conducting a one-proportion z-test in this context is to assess whether the frequency distribution of a specific feature in the resampled data is statistically representative of the same feature in the original cohort dataset. This test helps determine whether any observed

differences in the proportions of the feature between the resampled data and the original data are statistically significant. By using a significance level of 0.05 (commonly chosen in hypothesis testing), the test aims to identify features in the resampled data that deviate significantly from the original dataset, indicating that they may not be representative. Features with non-significant differences would likely be retained, while those with significant differences would be considered for elimination, as they may not accurately reflect the original data's characteristics. Thus, in this research, the one-proportion z-test is employed as a statistical tool to ensure the validity of the resampled data by identifying and potentially excluding features that do not maintain the expected distribution. The statistics before and after rectification is tabulated in Appendices B.1, B.2, B.3, and B.4.

2.4 Model Development

2.4.1 Machine Learning algorithms

To choose a best model, we examine the performance of various supervised Machine Learning algorithms.

The labeled health records allow for the use of supervised learning, and the binary classification method is used to classify the multimorbidity profile of a patient. Deep learning and other Machine Learning algorithms are applied to all cohort datasets as shown in Figure 2.2. The results are compared using a scoring grid with average cross-validated scores.

2.4.2 SHAP analysis for interpretation

The SHapley Additive exPlanations (SHAP) [74] values are used to explain the contribution of individual features in predicting hospitalization outcomes of the cohort.

SHAP values provide a way to quantify the contribution of each feature to the prediction of hospitalization. This can help identify which multimorbidity features have the most significant impact on the likelihood of hospitalization due to COVID-19. SHAP values offer an intuitive way to explain the predictions of a Machine Learning model. For healthcare professionals and policymakers, it's essential to understand why a model predicts a certain outcome. SHAP values provide a clear breakdown of how each feature influences the prediction, making it easier to trust and act upon

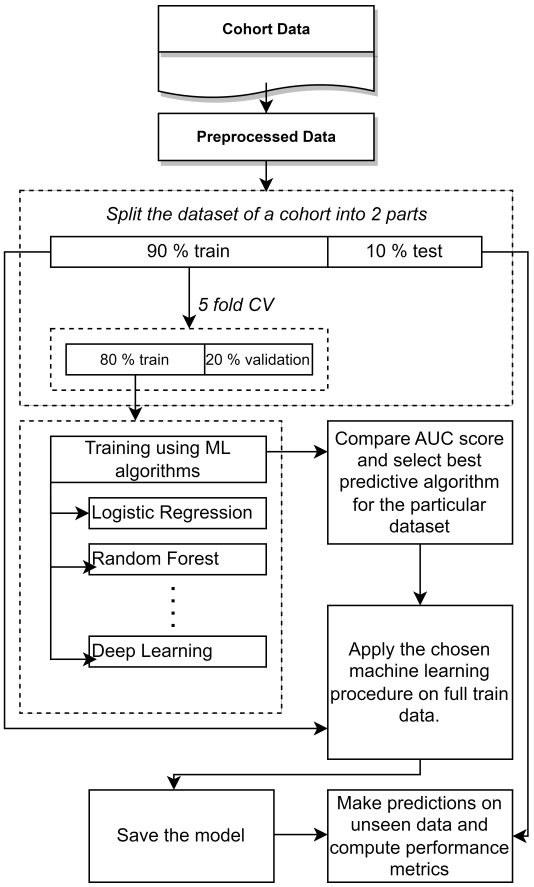


Figure 2.2: Selecting best Machine Learning model for each cohort dataset

the model's recommendations. Moreover, by analyzing SHAP values, individuals who are at higher risk of hospitalization due to COVID-19 based on their multimorbidity features can be identified. This can inform targeted interventions, such as vaccination prioritization, remote monitoring, or preventive measures, for individuals with specific comorbidities.

The SHAP values of all features are plotted and positioned on the y-axis based on their impact on the model outcome. The SHAP beeswarm plots are used to investigate the distribution of the influence each feature has on the outcome of the model, with features having higher importance positioned at the top of the graph. A data point for a feature corresponds to a single patient, with the position of the data point (SHAP value) on the X-axis representing the effect of that feature on the outcome of the model for that particular patient. If the multimorbidity is present (feature value is 1: red) for a patient accounting for a more positive SHAP value, it indicates that the presence of that feature acts as a risk factor for hospitalization.

Similarly, if the multimorbidity is present (feature value is 1: red) for a patient having a more negative SHAP value, it indicates that the presence of that feature acts as a protective factor against hospitalization risk for that patient.

2.5 Projects

We have undertaken two projects aimed at identifying predictors of severe COVID-19 outcomes related to multimorbidity. These projects address the challenge of dealing with sparse data, particularly rare features, and focus on determining the optimal combinations of morbidity features strongly linked to the severity of COVID-19.

- Project 1: Multimorbidity in Middle-aged Women and COVID-19: Binary Data Clustering for Unsupervised Binning of Rare Multimorbidity Features and Predictive Modeling.
- Project 2: Evolutionary Machine Learning Based Multimorbidity Analysis In COVID-19 Hospitalized Patients: A Longitudinal Study Using Health-administrative Data of a Region in the North-West of Italy

Table 2.1: Machine Learning Algorithm Descriptions

Algorithm	Underlying Principle	Strengths	Weaknesses
Logistic Regression (LR) [61]	Linear model for binary classification.	Simplicity and interpretability.	May underperform with complex relationships.
CatBoost Classifier [62]	Gradient boosting with category-aware features.	Handles categorical data well, robust to overfitting.	Can be computationally intensive.
Gradient Boosting Classifier [63]	Ensemble method that builds trees sequentially.	High predictive accuracy, handles complex relationships.	Prone to overfitting, longer training times.
AdaBoost Classifier [64]	Ensemble method that assigns weights to data points.	Good at correcting misclassifications, works well with weak learners.	Sensitive to noisy data and outliers.
Linear Discriminant Analysis (LDA) [65]	Dimensionality reduction and classification based on linear combinations.	Effective in high-dimensional data, reduces multicollinearity.	Assumes Gaussian distributions.
Random Forest Classifier [66]	Ensemble of decision trees with bootstrapped samples and feature randomness.	Excellent at handling complex data, resistant to overfitting.	Less interpretable than individual trees.
Naive Bayes [67]	Probabilistic classifier based on Bayes' theorem with independence assumptions.	Simple, computationally efficient, good for text classification.	Assumes feature independence, may not capture complex relationships.
LightGBM [68]	Gradient boosting framework with histogram-based learning.	High-speed training and good accuracy, handles large datasets.	Prone to overfitting with small datasets.
Extra Tree Classifier [69]	Ensemble of decision trees with random feature splits.	Low computational cost, robust to noise.	Less interpretable than other tree-based models.
Extreme Gradient Boosting (XGBoost) [70]	Gradient boosting with optimized tree algorithms.	High performance, strong regularization, and feature selection.	Can be sensitive to hyperparameters.
Decision Tree Classifier	Hierarchical structure of binary decisions.	Simple to understand, interpretable.	Prone to overfitting, not suitable for complex relationships.
K-Nearest Neighbors Classifier [71]	Classifies data points based on their neighbors.	Non-parametric, simple to implement.	Sensitive to the choice of k, computationally expensive for large datasets.
Quadratic Discriminant Analysis [65]	Extension of LDA allowing for different covariance matrices.	More flexible than LDA when covariance structures differ.	Requires more data, computationally expensive.
Support Vector Machine (SVM) Linear Kernel [72]	Creates a linear decision boundary with maximal margin.	Effective for high-dimensional data, works well with small to medium-sized datasets.	Limited ability to capture complex non-linear patterns.
Ridge Classifier [73]	Linear model with L2 regularization.	Reduces multicollinearity, robust to noisy data.	Less interpretable than logistic regression.

Chapter 3

Project 1: Binary Data Clustering for Unsupervised Binning

3.1 Clustering Patients with Multimorbidity

This project, specifically focuses on clustering binary data related to various medical conditions in middle-aged women (cohort 1). Cluster analysis is a valuable statistical technique for grouping objects based on their similarity in terms of indicator variables or features, and can be applied to identify clinically significant multimorbid groupings of medical conditions [75]. By using cluster analysis, researchers can learn important information about how different medical conditions are related and occur together. This helps them understand the complex connections between diseases and to develop personalized ways of treatment. It is also evident from the existing studies that clustering methodology can be applied to identify patient subgroups with similar disease profiles or symptom patterns [76]. Furthermore, it also can be utilized for identifying patient subgroups with distinct healthcare utilization trends and identifying risk factors associated with adverse outcomes [77]. In a multimorbidity study [78], the authors utilized K-means non-hierarchical cluster analysis to identify patterns of multimorbidity. Similarly, another study [79] focused on stratifying a population of high-risk multimorbid patients by using cluster analysis for risk stratification and identifying distinct characteristics of each cluster. These findings emphasize the significance of healthcare reform in addressing the unique needs of different patient clusters. By tailoring interventions and care strategies based

on these identified clusters, healthcare providers can effectively address the diverse challenges associated with multimorbidity. Self-Organizing Feature Maps (SOFMs) have been widely employed in various clustering applications, including tasks like handwritten digit recognition [80]. In another study [81], the authors employed SOFMs to identify clusters of patients based on their healthcare data. However, SOFMs are not commonly used for clustering multimorbidity patterns, as these patterns typically involve clinical and demographic data rather than image data. Instead, other clustering approaches such as k-means, hierarchical clustering, and latent class analysis are more commonly employed for multimorbidity clustering.

3.2 Clustering Rare Features

This study focus on clustering rare features, which are medical conditions that are not commonly observed in patient data. We grouped multimorbidity features into bins using a matrix based on cluster structures. This process involves two levels of clustering: the feature level and the data level, without making assumptions about the number of feature clusters. Once the features associated with each cluster are identified, they are mapped to corresponding bins. The unsupervised bins are then merged with prevalent features to create a new engineered feature matrix. The performance of models using clustered data is compared to models without clustered data, and the importance of the features is investigated, leading to the interpretation of the models.

3.2.1 Unsupervised feature binning

To group the multimorbidity features into various bins, a matrix is reconstructed based on the cluster structures. The clustering process involves two levels: feature level and data level, as shown in Figure 3.1.

At the feature level clustering, the Binary Matrix Decomposition (BMD) algorithm [82] is used to assign features into different clusters without bootstrapping on labeled train data. The clustering method makes no presumptions regarding the number of feature clusters. After identifying the features associated with each cluster, each feature is mapped to its corresponding bin. Features that are not considered rare (i.e., present in at least 20% of the data) are not mapped to any bin and are used as they are. Only the rare features are mapped to their corresponding cluster, forming the Cluster

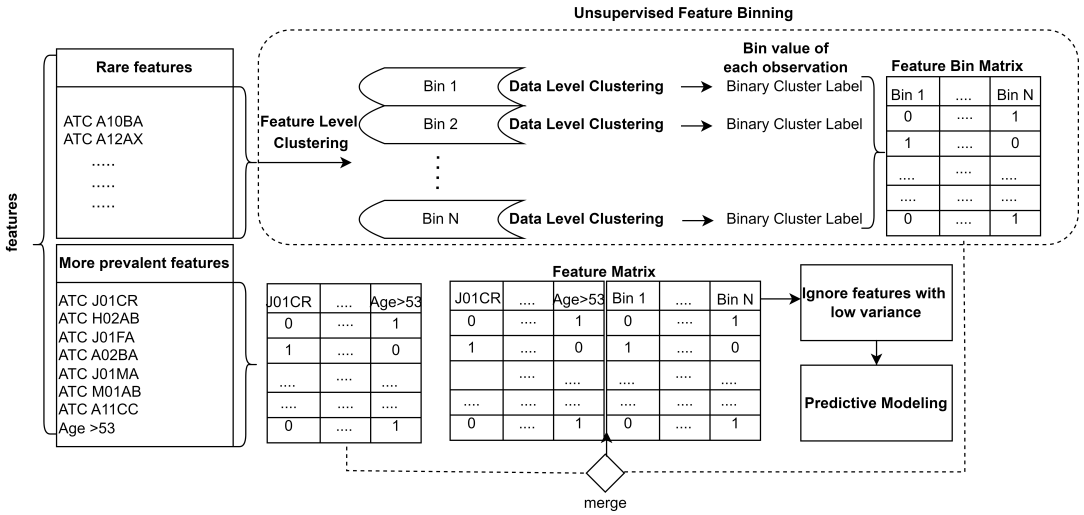


Figure 3.1: Feature level and data level clustering is performed before predictive modeling

Map.

Using the Cluster Map, the features within each cluster are represented as a Feature Bin Matrix (FBM). The training FBM consists of the features in the corresponding cluster, along with the feature values for all patients in the training dataset (without the class label). The unsupervised learning [83] is performed on the training FBM using the same BMD algorithm, iteratively for each cluster in the Cluster Map. The resulting values for each cluster are obtained. The trained model is then used to predict the cluster labels for the test FBM.

The unsupervised bins engineered from the FBMs are merged with the prevalent features (with the features excluded from the Cluster Map) to form a new engineered Feature Matrix (FM). This process is carried out separately for the training and test sets, resulting in the train FM and test FM, respectively. During the data level clustering, both datasets are handled separately without the class label to prevent data leaks. The entire procedure is illustrated in Figure 3.2.

3.2.2 Predictive modeling

To assess the performance of different Machine Learning algorithms in predicting hospital admission due to Covid-19, we utilized the train and test FM datasets. Since the data is labeled, we employed a supervised learning approach on this engineered dataset. The trained binary classification

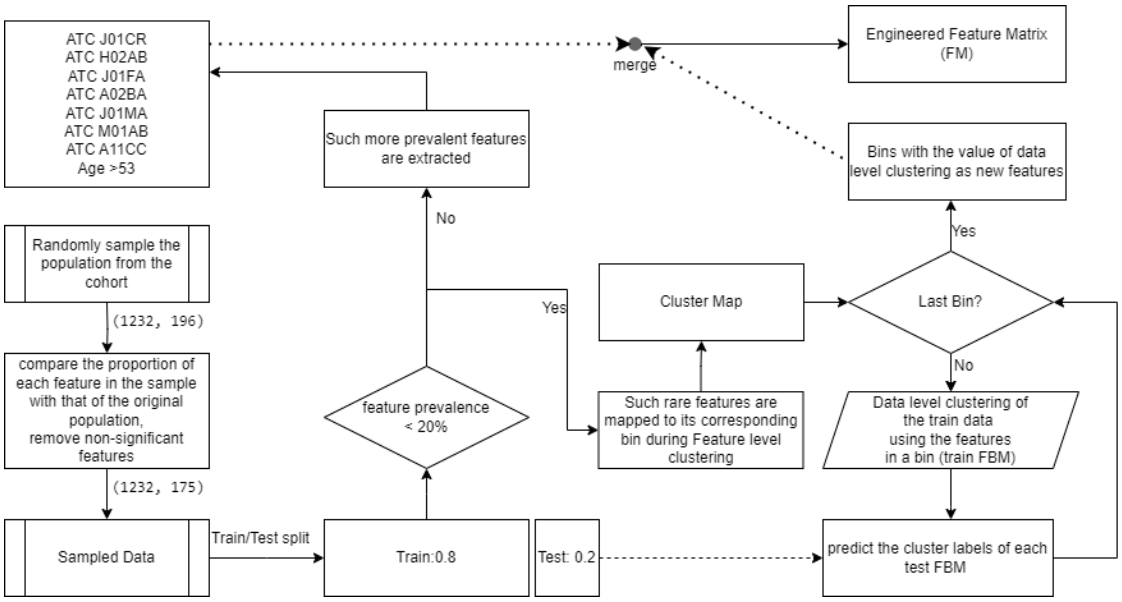


Figure 3.2: Unsupervised feature binning of rare features and generation of the Feature Matrix using new engineered features and other features

model was then applied to the holdout data to classify patients into one of two classes: whether or not hospital admission is required, taking into account their multimorbidity history.

Following the creation of the train and test FM datasets with the newly engineered features, we analyzed the variance of each feature. We trained the train FM using various Machine Learning algorithms available in the Pycaret package [84], employing 5-fold cross-validation.

Due to the sparsity of the data and the skewed distribution of value levels (0 or 1), certain levels may dominate others, resulting in insufficient variation to generate informative features. Therefore, during the Machine Learning-based predictive modeling, such non-informative features can be disregarded. The criteria for ignoring low-variance features [85] are as follows:

$$\frac{\text{number of unique values in a feature}}{\text{sample size}} < 10\%$$

and

$$\frac{\text{number of most prevalent value}}{\text{number of second most prevalent value}} > 20$$

The best-performing model is selected by examining the mean area under the curve (AUC) score of each Machine Learning model. Later, the best

model is evaluated using the test FM and the performance scores are reported.

Chapter 4

Project 2: Evolutionary Machine Learning for Feature Engineering

4.1 An Evolutionary Approach for Discovering Frequent Associated Bins

Firstly, Machine Learning algorithms are compared for selecting the optimal classification algorithm for the evolutionary approach. Employing 5-fold cross-validation, each model's performance is evaluated. Subsequently, the proposed Evolutionary Algorithm utilized a deep learning classifier to generate prediction-based fitness scores for identifying multimorbidity combinations associated with COVID-19 hospitalization. The selected models are interpreted using SHAP values to understand the relationship between the multimorbidity features of the various cohort data and hospitalization outcome. The proposed method also generates a feature-engineered dataset consisting of a user-specified number of outcome-associated combinations or bins of multimorbidity. Finally, the best-performing bins has been analyzed to discover the frequency of various multimorbidity patterns in all cohorts.

4.1.1 Deep Learning with sparse data

The sparse healthcare dataset of this research contains rare medical conditions and drugs which pose a challenge for statistical and Machine Learning analyses due to their lower prevalence [37]. To overcome this issue, the study uses sequential deep learning with Adagrad - Adaptive Gradient Algorithm,

an optimization algorithm that is well-suited for handling sparse data [86]. By an optimization method systematic adjusting of a model's parameters to minimize or maximize a specified objective function can be achieved. Thus, model's performance on a given task can be improved. In sparse datasets, many features have zero or near-zero values, and traditional optimization algorithms may struggle to adapt their learning rates properly. Adagrad's adaptive learning rates help overcome this issue. Adagrad's adaptive scaling of the learning rate eliminates the need for manual tuning, and it is more robust than stochastic gradient descent. Additionally, the study employs the early stopping functionality to improve the model's performance. By using early stopping, the training is halted when performance on a validation set starts to degrade.

Dropout is a regularization technique used during training in neural networks to reduce overfitting by randomly deactivating a portion of neurons during both forward and backward passes, the network is compelled to learn features that are more robust and generalizable [87]. In all deep learning models of this research, Dropout as a regularization technique is used to minimize overfitting while training [88] and also introduced a dropout layer with 20% dropout after the first and second layers in the sequential model. Since this research deals with a binary classification problem, the default loss function is used for such a problem, which is Binary Cross Entropy Loss [89]. Binary Cross Entropy Loss, also known as log loss, is a loss function commonly used for binary classification problems in Machine Learning [90]. It measures the dissimilarity between the predicted probability distribution and the actual binary labels, penalizing the model more as its predictions deviate from the true labels, providing a gradient signal that guides the model towards better classification performance [91].

4.1.2 Feature selection for discovering the optimal set of multimorbidity features

Feature selection as a pre-processing method eliminates irrelevant and redundant information and aid in dimensionality reduction [92]. Three methods of feature selection are: filter-based, embedded, and wrapper-based methods [93]. The filter-based method generates models with reduced predictive performance compared to the other two methods. The embedded method performs optimum feature subset search while constructing the model, while the wrapper method selects the best feature subset using the performance of the classifier used. This study uses a wrapper method that

uses deep learning as a classifier algorithm and an Evolutionary Algorithm as a search strategy to generate feature subsets (bins). The best-performing bin is estimated using AUC and selected as the optimal subset of the multimorbidity features that are highly associated with Covid-19 hospitalization.

4.1.3 Evolutionary Algorithms

Evolutionary Algorithms encompass a group of optimization techniques that draw inspiration from the biological evolution proces [94]. They are used to solve complex optimization and search problems. In this algorithms, a population of potential solutions or individuals (in this research it can be represented as feature groups with more association to the outcome) evolves over generations. Each generation undergoes a selection process where individuals (feature groups) are assessed by considering their fitness, which quantifies their problem-solving capability. Individuals demonstrating higher fitness are likelier to be chosen for reproduction.[95].

Genetic Algorithms

Genetic Algorithms are a specific subset of Evolutionary Algorithms [96]. These algorithms are a type of optimization algorithm inspired by the process of natural selection. They involve creating a population of individuals, which represent potential solutions, and subsequently, evolving this population over many generations through various operations. Reproduction involves creating new individuals (offspring) through operations like mutation, crossover (recombination), and selection. These processes introduce genetic diversity into the population and promote the evolution of better solutions [97]. Over multiple generations, the algorithm seeks to improve the population's overall fitness and, consequently, find optimal or near-optimal solutions to the problem [98].

4.1.4 Evolutionary Machine Learning

The use of Evolutionary Algorithms is a promising approach for extracting a reduced set of meaningful rare associations that are accurate, especially for problems such as sparse data, epistatic association with features, and high dimensional representations of features. Evolutionary Machine Learning is a hybrid method that uses evolutionary computation to overcome obstacles in various Machine Learning tasks [94]. Compared to traditional algorithms

that rely on exhaustive search-based techniques, Evolutionary Algorithms offer a more robust solution.

These are important points to consider when performing feature engineering using Evolutionary Algorithms: (i) just because a feature is not prevalent does not mean it is irrelevant, as it could still have a strong association with the outcome, (ii) dealing with sparsity in the data is a challenge for many Machine Learning methods, especially when it comes to features with near-zero variance and (iii) considering combinations of features may be more predictive than just looking at isolated features, which highlights the importance of searching for feature interactions.

This study used a Genetic Algorithm to create an optimized feature matrix. Initially, the features are randomly grouped into bins, and a feature matrix is created for each bin. The bins are regrouped using a Genetic Algorithm and a wrapper-based method to interact with a classifier. The study adopts the elitism principle to preserve the best-performing bins and save them as a checkpoint. The final feature matrix is the engineered matrix evolved after all iterations, and it can be utilized for tackling the problems of data sparsity and including the perspective of interactions among various multimorbidity features. The proposed evolutionary approach is an Evolutionary Algorithm-based wrapper method and it is illustrated in Figure 4.1. It is a modified version of an Evolutionary Algorithm called Relevant Association Rare-variant-bin Evolver [37]. The idiosyncrasies that differentiate our proposed method from the existing one in terms of the following: (i) In the section of evolutionary approach, our method uses a prediction-based method with train and test methods. Nonetheless, the existing instance count-based prediction method uses all available data without any model fitting. Furthermore, this research used this instance count-based prediction method only for the last evolutionary cycle to calculate the scores of the final bins. The Pseudocode for calculating final bin scores using prediction-based feature scoring method is given in Appendix 1. However, this research has adopted sparsity addressing contribution of the existing method by means of the summation of the values of the features in a group of final bins for generating a value for that feature group. (ii) In this study, it is implemented not only a feature learning algorithm but also a deep learning technique with an Adagrad optimizer for predicting the outcome in each iteration of the evolutionary cycle, enabling the Evolutionary Algorithm to converge to solutions (multimorbidity feature combinations) that generalize well regardless of the feature selection for a single model. While executing the Genetic

Algorithm, the scores generated from the deep learning model are used for genetic operations, where this prediction score is used as the fitness score to evaluate the performance of multimorbidity combinations (feature matrices) evolved in each cycle. (iii) Existing method discovers new feature combinations and then encode them as features. But, in this study, in addition to this, the discovered feature combinations are further analyzed to estimate the frequency of occurrence of a particular feature in best-performing feature combinations rather than using them for a single model. From the outcome-associated multimorbidity combinations, the most prevalent multimorbidity combinations are also extracted to identify the multimorbidity pattern among COVID-19 patients using Apriori algorithm.

4.1.5 Frequent multimorbidity features

Most prevalent multimorbidity combinations are extracted to identify the multimorbidity pattern among COVID-19 patients using the apriori algorithm. Apriori algorithm is applied to the dataset, which contains various combinations of multimorbidity features obtained from the Evolutionary Algorithm. The support measure was used to determine how common a feature combination is in the feature matrix, with the rows representing the various feature groups in the final bins. To avoid analyzing irrelevant feature combinations, only the most common multimorbidity feature combinations in the final bins were analyzed. The frequent combinations of the features were then analyzed using a threshold value of 0.5 for minimum support (smin) to obtain the frequent itemsets.

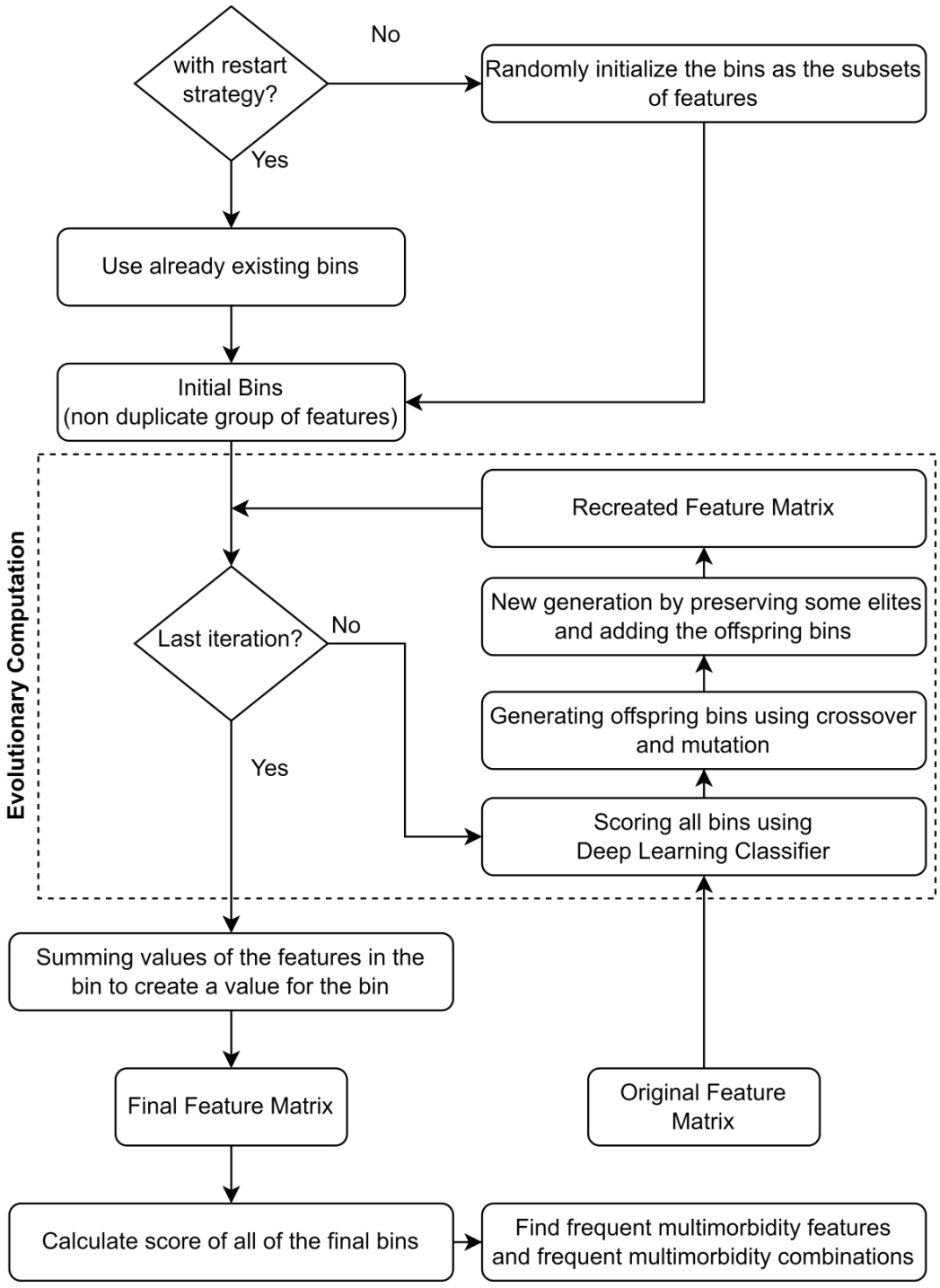


Figure 4.1: Illustration of the evolutionary approach carried out in this study.

Chapter 5

Results

5.1 COVID-19 population

Table 5.1 provides the summary of the characteristics of the COVID-19 population and the distribution of hospitalized and non-hospitalized patients represented as counts and percentages.

Table 5.1: Characteristics and distribution of the COVID-19 population

Age Groups	**45-53**	**54-59**	**60-68**	**69-74**
Overall Count	N = 7324		N = 5469	
	4179(57.1%)	3145(42.9%)	3296(60%)	2173(40%)
Female	4477(61.1%)		2355 (43.1%)	
Male	2847(38.9%)		3114 (56.9%)	
Mean age	52.3 (SD 4.18)		67 (SD 4.55)	
Age groups	**Hospitalized**		**Non-Hospitalized**	
	Male	**Female**	**Male**	**Female**
	N=1717		**N=5607**	
45 - 59	1101	616	1746	3861
45 - 53	825 (48%)		3352(60%)	
	522	303	1031	2323
54 - 59	892(52%)		2253(40%)	
	579	313	715	1538
	N=2927		**N=2542**	
60 - 74	1974	953	1140	1402
60 - 68	1582 (54%)		1711 (67%)	
	1073	512	740	971
69 - 74	1342(46%)		831(33%)	
	901	441	400	431

5.2 One Proportion z-test Results

The one proportion z-test is performed on all features and the results of the one proportion z-test between randomly taken samples and original cohort datasets are presented in Appendices B.1, B.2, B.3, and B.4.

5.3 Machine Learning model performance comparison

The performance evaluation of Machine Learning models as 5-fold Cross Validation of all cohorts is tabulated in Table 5.2, 5.3, 5.4 and 5.5.

Table 5.2: Performance of Machine Learning models: Cohort 1

Model	Acc[1]	AUC[2]	Recall	Prec.[3]	F1	TT[4]
Logistic Regression	0.7175	0.7634	0.6275	0.7588	0.6863	6.260
CatBoost Classifier	0.7193	0.7602	0.6037	0.7773	0.6792	5.214
Gradient Boosting Classifier	0.7148	0.7583	0.5963	0.7735	0.6725	0.286
Ada Boost Classifier	0.7039	0.7512	0.6294	0.7340	0.6770	0.132
Naive Bayes	0.6489	0.7502	0.3266	0.8968	0.4772	0.012
Random Forest Classifier	0.6878	0.7262	0.5945	0.7237	0.6524	0.246
Light Gradient Boosting Machine	0.6742	0.7241	0.5853	0.7044	0.6387	0.204
Extreme Gradient Boosting	0.6751	0.7216	0.5780	0.7093	0.6358	0.948
Extra Trees Classifier	0.6814	0.7168	0.5963	0.7089	0.6473	0.248
Linear Discriminant Analysis	0.6787	0.7061	0.5229	0.7510	0.6148	0.042
Decision Tree Classifier	0.6291	0.6166	0.5413	0.6459	0.5882	0.016
K Neighbors Classifier	0.5858	0.6072	0.2385	0.7465	0.3611	0.060
Quadratic Discriminant Analysis	0.5272	0.5329	0.8606	0.5750	0.6255	0.052
SVM - Linear Kernel	0.6444	0.0000	0.6257	0.7345	0.6213	0.040
Ridge Classifier	0.6931	0.0000	0.5450	0.7668	0.6349	0.012

[1]Acc : Accuracy Score obtained by the corresponding Machine Learning model

[2]AUC: Area under the ROC Curve

[3]Prec: Precision score.

[4]TT : Time taken in seconds

Table 5.3: Performance of Machine Learning models: Cohort 2

Model	Acc	AUC	Recall	Prec.	F1	TT (Sec)
CatBoost Classifier	0.6295	0.6699	0.5296	0.6644	0.5886	6.564
Gradient Boosting Classifier	0.6179	0.6683	0.4854	0.6638	0.5602	0.550
Logistic Regression	0.6315	0.6669	0.5417	0.6631	0.5962	6.582
Ada Boost Classifier	0.6204	0.6614	0.5226	0.6525	0.5797	0.312
Extreme Gradient Boosting	0.6204	0.6518	0.5176	0.6554	0.5779	1.558
Linear Discriminant Analysis	0.6204	0.6509	0.4955	0.6637	0.5673	0.134
Light Gradient Boosting Machine	0.6164	0.6443	0.5347	0.6417	0.5831	0.452
Naive Bayes	0.5810	0.6396	0.2281	0.7854	0.3536	0.050
Random Forest Classifier	0.6073	0.6366	0.5286	0.6313	0.5752	0.468
Extra Trees Classifier	0.6048	0.6212	0.5387	0.6238	0.5779	0.510
K Neighbors Classifier	0.5507	0.5640	0.3035	0.6045	0.4033	0.282
Decision Tree Classifier	0.5543	0.5310	0.4492	0.5735	0.5031	0.080
Quadratic Discriminant Analysis	0.5078	0.5065	0.8211	0.5455	0.5805	0.130
SVM - Linear Kernel	0.5876	0.0000	0.4492	0.7409	0.4866	0.098
Ridge Classifier	0.6285	0.0000	0.5116	0.6707	0.5803	0.044

Table 5.4: Performance of Machine Learning models: Cohort 3

Model	Acc	AUC	Recall	Prec.	F1	TT (Sec)
Gradient Boosting Classifier	0.6035	0.6569	0.5157	0.6292	0.5659	0.432
CatBoost Classifier	0.6064	0.6541	0.5343	0.6296	0.5769	7.624
Random Forest Classifier	0.6157	0.6520	0.5912	0.6236	0.6065	0.352
Extra Trees Classifier	0.5994	0.6512	0.5982	0.6020	0.5992	0.408
Naive Bayes	0.5761	0.6420	0.3171	0.6633	0.4221	0.050
Ada Boost Classifier	0.5924	0.6294	0.5145	0.6134	0.5591	0.220
Logistic Regression	0.5901	0.6290	0.5238	0.6076	0.5617	6.366
Light Gradient Boosting Machine	0.5948	0.6288	0.5529	0.6052	0.5773	0.324
Extreme Gradient Boosting	0.5819	0.6252	0.5366	0.5921	0.5627	1.302
Linear Discriminant Analysis	0.5866	0.6060	0.5006	0.6082	0.5483	0.124
K Neighbors Classifier	0.5522	0.5800	0.2939	0.6114	0.3961	0.240
Quadratic Discriminant Analysis	0.5545	0.5760	0.6497	0.5946	0.5655	0.088
Decision Tree Classifier	0.5271	0.5245	0.5227	0.5286	0.5252	0.080
SVM - Linear Kernel	0.5656	0.0000	0.5804	0.6149	0.5429	0.090
Ridge Classifier	0.5895	0.0000	0.5075	0.6109	0.5534	0.040

Table 5.5: Performance of Machine Learning models: Cohort 4

Model	Acc	AUC	Recall	Prec.	F1	TT (Sec)
Gradient Boosting Classifier	0.5717	0.6032	0.5520	0.5741	0.5618	0.482
Ada Boost Classifier	0.5668	0.6004	0.5267	0.5726	0.5475	0.188
CatBoost Classifier	0.5604	0.5992	0.5549	0.5614	0.5577	6.952
Logistic Regression	0.5721	0.5984	0.5423	0.5779	0.5590	6.468
Linear Discriminant Analysis	0.5629	0.5918	0.5286	0.5682	0.5474	0.124
Random Forest Classifier	0.5585	0.5845	0.6142	0.5536	0.5820	0.436
Naive Bayes	0.5429	0.5796	0.7805	0.5383	0.6286	0.048
Light Gradient Boosting Machine	0.5546	0.5787	0.5471	0.5565	0.5512	0.394
Extra Trees Classifier	0.5565	0.5784	0.6161	0.5515	0.5818	0.412
Extreme Gradient Boosting	0.5546	0.5739	0.5287	0.5583	0.5424	1.656
Decision Tree Classifier	0.5429	0.5361	0.5549	0.5434	0.5488	0.112
K Neighbors Classifier	0.5209	0.5333	0.4422	0.5262	0.4798	0.286
Quadratic Discriminant Analysis	0.5000	0.5002	0.4999	0.5117	0.4015	0.136
SVM - Linear Kernel	0.5317	0.0000	0.3484	0.5745	0.3405	0.112
Ridge Classifier	0.5692	0.0000	0.5316	0.5759	0.5524	0.114

5.4 Results - Project 1

5.4.1 Cluster Map

After applying feature-level clustering to the training data, a Cluster Map is generated. In this Cluster Map, rare features are clustered and assigned to their respective bins, resulting in 13 feature clusters. The bin values for each observation are calculated by determining the cluster label of the corresponding features in that bin. Table 5.6 illustrates the resulting 11 bins after excluding features with low variance.

5.4.2 Analysing performance score for model selection

To select the best model from various Machine Learning algorithms, the AUC score of each Machine Learning model is compared after executing a 5-fold cross-validation. During cross-validation using the train data with all 174 features of cohort 1 data, the best performance was obtained by LR (accuracy 0.72, AUC 0.76, F1-score 0.69), CatBoost Classifier (accuracy 0.72, AUC 0.76, F1-score 0.68), and Gradient Boosting Classifier (accuracy 0.72, AUC 0.76, F1-score 0.67).

Later, using the features which are reduced by clustering technique and ignoring the features with low variance, performance is analysed. During

cross-validation using the train data with only 17 features, the best performance was obtained by LR (accuracy 0.7, AUC 0.74, F1-score 0.68), LDA (accuracy 0.7, AUC 0.74, F1-score 0.66) and Ada Boost Classifier (accuracy 0.7, AUC 0.73, F1-score 0.67). The 5-fold cross-validation scores of each Machine Learning model are tabulated in Table 5.7.

5.4.3 Model performance evaluation

After analyzing the cross-validation results, the top three models are selected based on their performance. To assess the predictive ability of these Machine Learning algorithms on the reduced data without sparsity, the selected models are utilized to predict the outcome of Covid-19 hospital admission using the test Feature Matrix (FM).

The performance metrics of the selected models on the test FM (holdout data) are as follows: LR (accuracy 0.72, AUC 0.77, F1-score 0.69), LDA (accuracy 0.7, AUC 0.77, F1-score 0.67) and Ada Boost (accuracy 0.7, AUC 0.77, F1-score 0.68). For a comprehensive overview, please refer to Table 5.8 for the complete set of results.

5.4.4 Feature importance

Feature importance refers to the scores assigned to input features, which indicate their relative significance in making predictions. These scores provide insights into the importance of each feature in the data and the model. Feature importance helps not only in explaining the influential features but also in understanding the data and model better.

Feature importance score from the model coefficients

In linear algorithms such as LR and LDA, the predictions are calculated as a weighted sum of the observations, with the coefficients determined by the algorithm. In this context, negative coefficients indicate that as the value of a feature increases, the severity due to Covid-19 is predicted to decrease, suggesting no hospital admission. The features with negative coefficients in both LR and LDA algorithms are bin 2, bin 3, bin 4, bin 7, bin 10, J01CR, J01FA, and Age >53. On the other hand, features with positive coefficients have a positive association with the severity outcome. A higher negative coefficient indicates a stronger negative association between the input feature and the severity outcome. For example, if the value of

a cluster or feature is 1, it suggests that most patients belonging to that cluster or feature category have a lower chance of severe Covid-19 outcomes, and vice versa. Conversely, in the case of a positive coefficient, if the cluster or feature value is 1, it indicates an increased likelihood of severe Covid-19 outcomes, and vice versa.

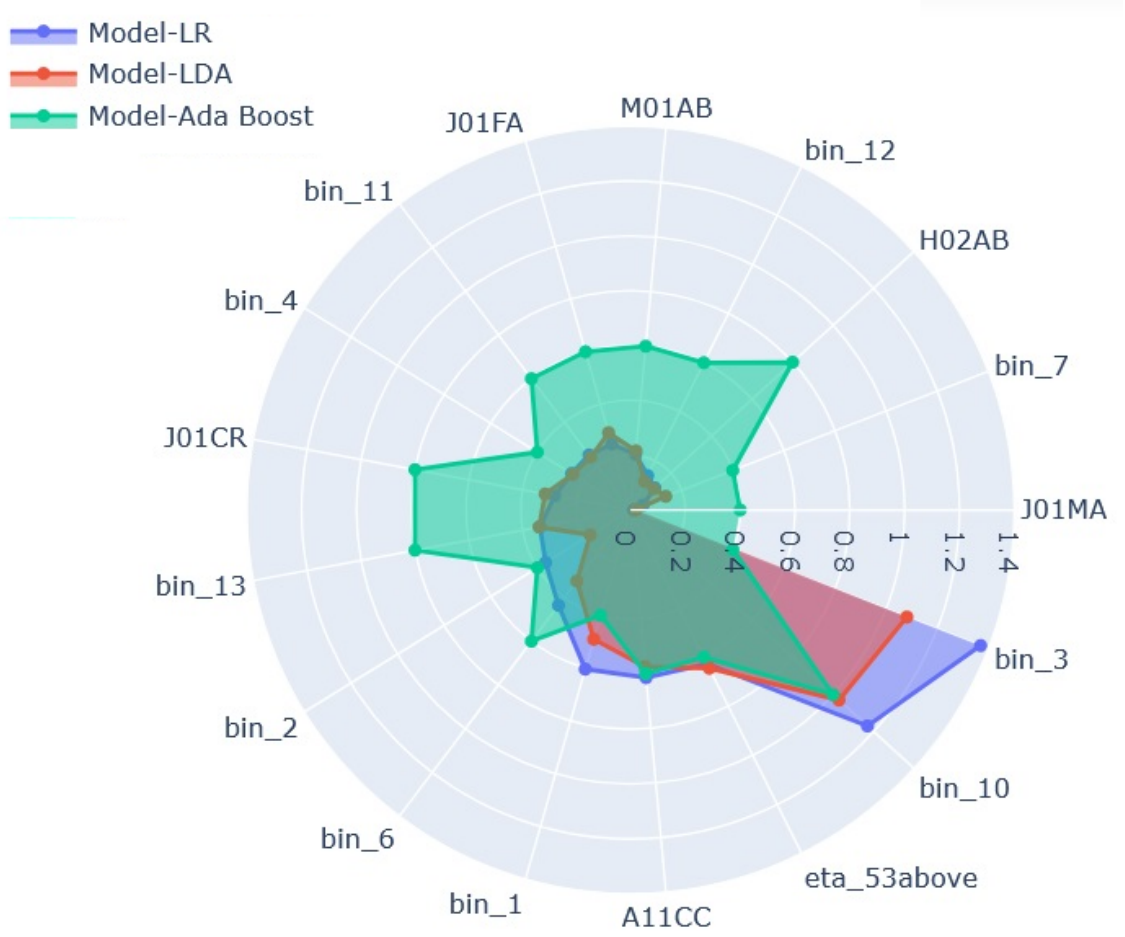


Figure 5.1: Feature importance scores from LR, LDA, and Ada Boost Models

The impurity-based feature importance

In the Ada Boost algorithm, the feature scores are determined using the Gini importance [99]. This score is calculated for each decision tree based on how much a single feature split improves the model's performance, and it is normalized by the number of observations accounted for by that feature.

To analyze the feature importance of all three models (LR, LDA, and Ada Boost), the feature importance values are aggregated and visualized the results in Figure 5.1. In the case of linear models (LR and LDA), the feature importance is represented by the absolute values of the coefficients. For the Ada Boost Classifier, the feature importance values are scaled and presented in the visualization.

5.4.5 Interpretation of the model

We used SHAP to interpret the most impactful features that our models utilize [74] in determining the status of the hospitalization. The SHAP heatmaps for the linear models depicted in Figure 5.2 and Figure 5.3 are based on the 20% test samples (X-axis). The sorted global feature importance is represented by the Y-axis and the bar plot (right-hand side). The magnitude of SHAP values of each observation (each patient) is represented by colors. The blue color for a feature denotes, in that patient profile, that particular feature has a value of 0 and this feature contributed to or impacted the prediction of the severity either positively or negatively. The topmost graph, f(x) represents the model predictions of each patient's multimorbidity profile.

In the LR heatmap of SHAP values, while examining the f(x), the 0th patient observation number possesses a higher prediction. So, it is predicted that the patient is admitted to the hospital, and the features in cluster "bin 10" contribute more positively to the Covid-19 severity of that particular patient. Similarly, we can interpret the results of other patients for all the features using this visualization.

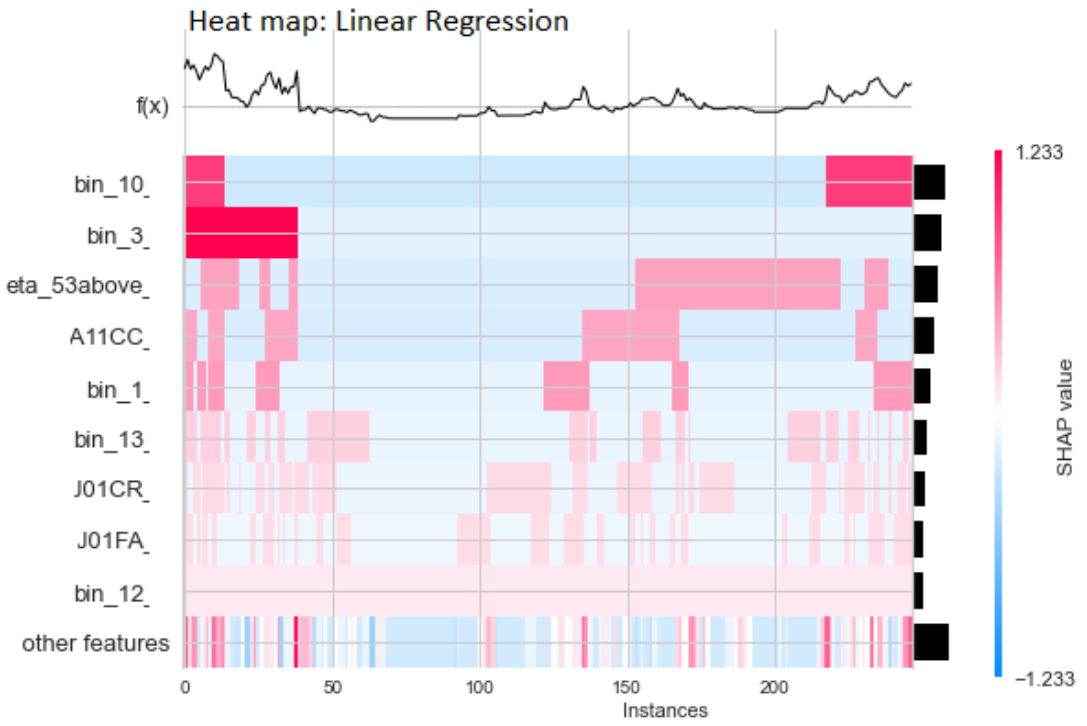


Figure 5.2: Heatmap matrix and global importance of features - LR

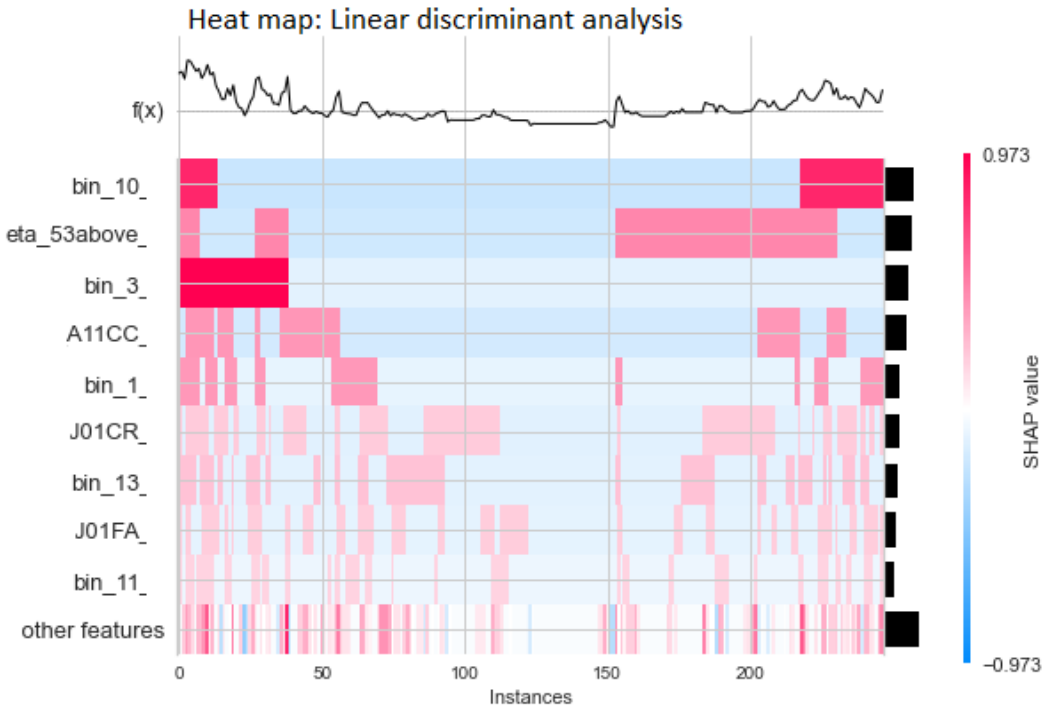


Figure 5.3: Heatmap matrix and global importance of features - LDA

Table 5.6: Cluster Map: Rare features are clustered and mapped to their corresponding cluster (Bins) after feature level clustering

Bin 1	Bin 2	Bin 3		Bin 4	Bin 5		Bin 6		Bin 7		Bin 8	Bin 9	Bin 10	Bin 11
ATC	ATC	ATC	ICD	ATC	ATC	ICD	ATC	ICD	ATC	ICD	ATC	ATC	ATC	ATC
A10BA	A07EC	A03FA	338	G03AA	C01BC	041	A07EA	038	B01AA	278	A02BX	A02AD	JO1DD	J01CA
A12AX	B02AA	A05AA	574	G03CA	C01BD	162	A10AB	174	B05BB	295	B03BB	A07AA		J01XX
B01AB	C03EA	A12AA	727	G03DB	C01DA	211	A10BB	218	C09BX	427	C03CA	B03AA		N06AB
B01AC	C09BA	A12BA	V58	J05AB	C02AC	241	C03AA	296	M05BA	455	C08CA	C07AB		R03BA
C09DA	J01AA	B03BA			S01EE	250	C03DA	301	N04AA	553	C09AA	H03AA		
M01AX	J01DC	C02CA		**ICD**		298	C07AG	454	N05AA	618	C09CA	J02AC		
R03AC	J01EE	C03BA		621		354	C07BB	473	N05AD	626	C10AA	M01AH		
R06AE	M04AA	C07AA				410	C10BA	518	N05AH	717	N02AA	N02AX		
	N06AA	C09BB				428	L01BA	592	S01EC	726	N02CC	N02BE		
	P01AB	C10AB				434	N01BB	735		812	N03AX			
	R03AL	C10AX				437	N02AJ	V54		996	N06AX			
		D05AX				438	V43	N03AF		998	R03AK			
		N02AB				440	V53	P01BA		V64	R06AX			
		N02BA				470	V56	R03BB						
		N03AE				482	V57	R03DA						
		N03AG				486		S01ED						

Table 5.7: Score of the Machine Learning models obtained during 5-fold Cross Validation using reduced features

	Model	Acc	AUC	Recall	Prec.	F1	TT
LR	Logistic Regression	0.7015	0.7376	0.6186	0.7436	0.6752	2.410
LDA	Linear Discriminant Analysis	0.7025	0.7370	0.5781	0.7712	0.6605	0.008
Ada Boost	Ada Boost Classifier	0.6964	0.7347	0.6248	0.7315	0.6737	0.030
NB	Naive Bayes	0.6843	0.7305	0.5823	0.7345	0.6492	0.006
RF	Random Forest Classifier	0.6772	0.7301	0.6267	0.6980	0.6601	0.196
CatBoost	CatBoost Classifier	0.6853	0.7272	0.5800	0.7398	0.6490	0.674
XGBoost	Extreme Gradient Boosting	0.6761	0.7184	0.5900	0.7159	0.6451	0.402
QDA	Quadratic Discriminant Analysis	0.6772	0.7171	0.5701	0.7267	0.6387	0.008
ET	Extra Trees Classifier	0.6690	0.7155	0.6064	0.6947	0.6469	0.178
GBC	Gradient Boosting Classifier	0.6914	0.7147	0.5761	0.7507	0.6516	0.028
LightGBM	Light Gradient Boosting Machine	0.6843	0.7146	0.5962	0.7260	0.6541	0.258
KNN	K Neighbors Classifier	0.6569	0.7058	0.5537	0.7001	0.6162	0.422
DT	Decision Tree Classifier	0.6548	0.6522	0.5618	0.6956	0.6201	0.006
Dummy	Dummy Classifier	0.4975	0.5000	0.4000	0.1990	0.2658	0.006
SVM	SVM - Linear Kernel	0.5513	0.0000	0.9091	0.5393	0.6700	0.010
Ridge	Ridge Classifier	0.7025	0.0000	0.5781	0.7712	0.6605	0.006

Table 5.8: Performance Evaluation of the selected Machine Learning models using Holdout data

Model	Acc	AUC	Recall	Prec.	F1
LR	0.72	0.77	0.63	0.76	0.69
LDA	0.70	0.77	0.59	0.76	0.67
AdaBoost	0.70	0.77	0.65	0.72	0.68

5.5 Results - Project 2

5.5.1 Performance evaluation of Deep Learning model

Table 5.9 depicts the evaluation of the performance of the deep learning model used in all four cohorts.

Table 5.9: Performance evaluation of Deep Learning model

	AUC score 5-fold CV	**Train AUC score**	**Test AUC score**	**Acc**	**Prec.**	**Recall**	**F1**
Cohort 1	77% (SD 1.87%)	82% Loss: .28	80% Loss: .29	76%	85%	63%	72%
Cohort 2	68% (SD 1.94%)	71% Loss: .30	67% Loss: .32	62%	62%	61%	62%
Cohort 3	67% (SD 1.87%)	74% Loss: .31	69% Loss: .32	67%	70%	60%	65%
Cohort 4	61% (SD 2.44%)	65% Loss: .34	62% Loss: .34	63%	62%	68%	65%

For each cohort, as shown in Figure 5.4, two line-plots are obtained while validating the efficiency of the model using cross-validation. The topmost plot depicts the Binary Cross Entropy Loss for the epochs for the train dataset and validation dataset, and the bottommost one presents the classification performance (AUC score) over epochs.

In the case of Cohort 1, it is visible that the problem is learned by the model quite well and quickly, attaining an AUC score of 82% in the train dataset and 80% in the test dataset. The obtained scores are nearly equivalent, indicating that the model is apparently neither over-fitting nor under-fitting. The plot of Cross Entropy Loss depicts that the model has converged. Moreover, on either dataset, the loss is admissible and the classification performance plot also indicates the convergence. The model performance and converging manner advocate that Cross Entropy Loss is appropriate for the neural network to learn this problem. In the case of Cohort 2, the model obtained performance scores of 71% and 67% for train and test datasets respectively with reasonable loss. The difference between these scores is very less suggesting the model satisfactorily learned the problem. In the case of Cohort 3, the model returned 74% as a train score and 69%

as the AUC test score. It is visible that there is not much improvement after 30 epochs, so Early stopping can be introduced while model training to circumvent the problem of overfitting and for the validation loss to stay the course. In Cohort 4 scenario, the loss plot appears to be converged well, albeit at a lesser classification performance than the models of other cohorts.

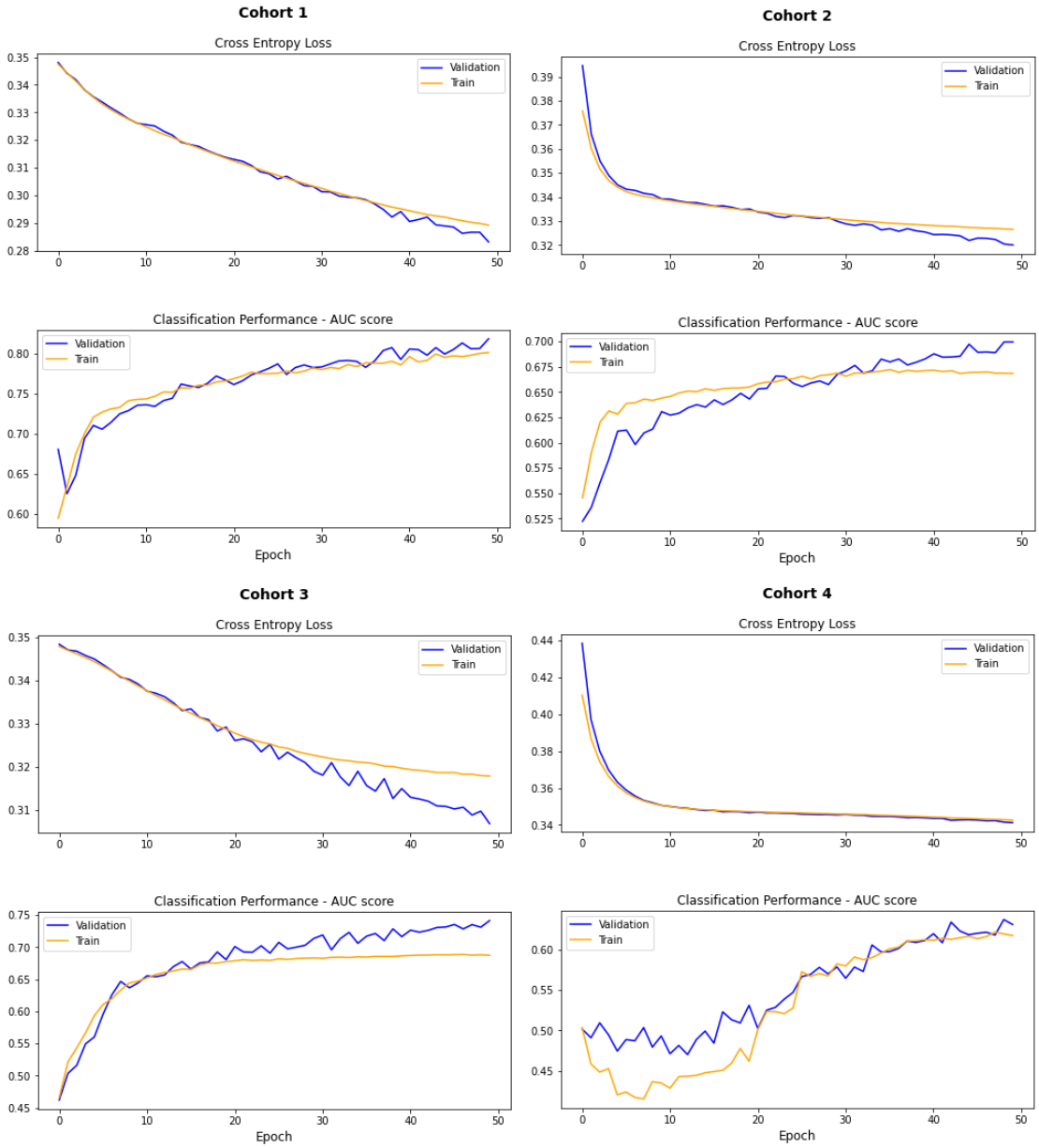


Figure 5.4: Model Loss Plot and AUC Score over Epochs

5.5.2 Influence of individual features on COVID-19 Hospitalization

SHAP beeswarm plots are illustrated to depict the impact of all features on COVID-19 hospitalization for all four models in Figure 5.5.

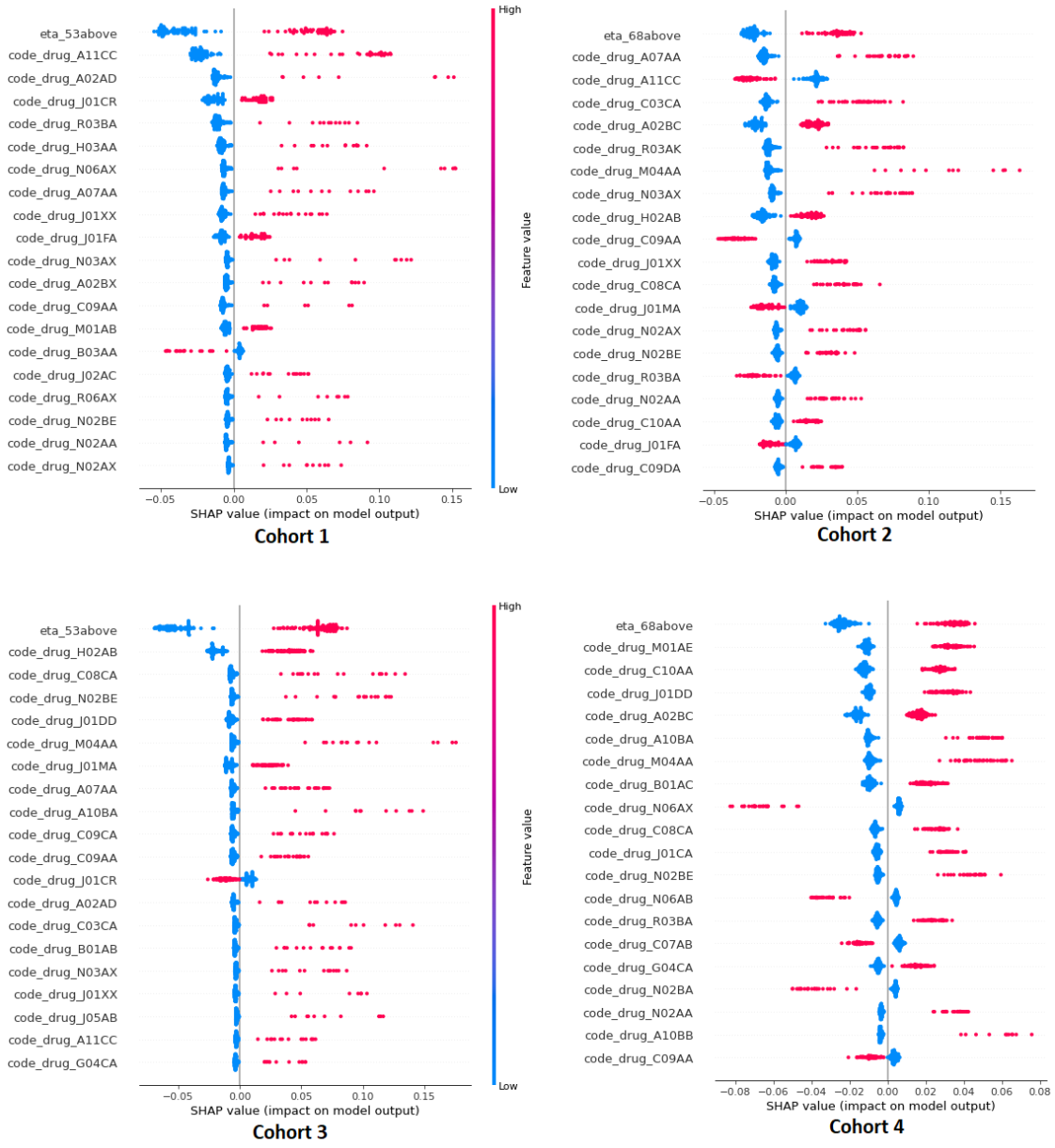


Figure 5.5: SHAP beeswarm plots - impact of features on COVID-19 hospitalization

5.5.3 Most Prevalent Multimorbidity Features in Evolved Bins

The accuracy score of evolutionarily obtained final bins are calculated. The highest accuracy obtained for Cohort 1 using evolutionary approach for finding outcome-associated best subsets of features is 71.43% (95% CI 67.31-67.97) using 64 features, for Cohort 2 is 63% (95 % CI 59.43-59.75) using 69 features, for Cohort 3 is 62.38% (95 % CI 59.84-60.09) using 53 features and for Cohort 4 is 58% (95% CI 55.42-55.63) using 61 features. These results are compared with the accuracy score of the deep learning model that use all features is illustrated in Figure 5.6. It represents maximum classification accuracy achieved by a bin vs number of features in that bin using evolutionary approach (left side) and the accuracy score achieved exclusively by the Deep Learning model (right side) with all the available features in the Cohort.

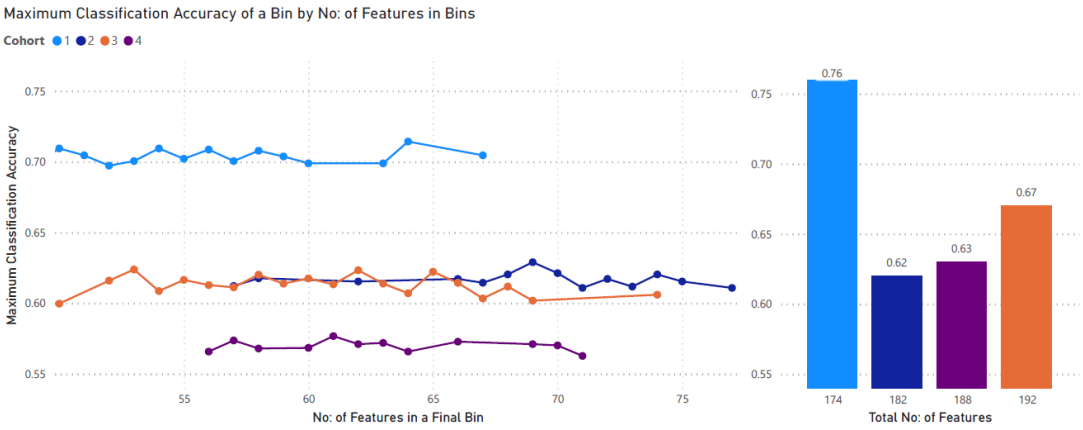


Figure 5.6: Final Bin's maximum classification accuracy VS No: of features

Frequently occurred morbidity features are the variables Age>53, R03BA (glucocorticoid inhalants), N03AX (other antiepileptics) in Cohort 1, A10BA (biguanide or metformin), N02BE (anilides) in Cohort 2, N02AX (other opioids), M04AA (preparations inhibiting uric acid production) in Cohort 3 and G04CA (Alpha-adrenoreceptor antagonists) in Cohort 4.

Table 5.10 shows the multimorbidity features that occurred more frequently in the final bins dataset of all cohorts with a minimum support (smin) measure of 0.6 along with the prevalence of the features in the sampled dataset. All other features with their statistics in case of each cohort are also given in Appendices C.1, C.2, C.3, and C.4.

Table 5.10: Frequently occurred morbidity features in the evolutionarily obtained final bins dataset

Cohort	Features	Description	P value	Support	Prevalence
1					
Age	Age >53		<.001	0.84	41.15
ATC#					
	R03BA	Glucocorticoids	<.001	0.85	15.5
	N03AX	Other antiepileptics	<.001	0.82	5.6
	R06AX	Other antihistamines for systemic use	<.001	0.79	6.74
	J01XX	Other antibacterials	<.001	0.78	14.2
	C03CA	Sulfonamides, plain	<.001	0.76	5.19
	N02AX	Other opioids	<.001	0.74	6.9
	A11CC	Vitamin D and analogues	<.001	0.73	23.05
	C09CA	Angiotensin II receptor blockers (ARBs), plain	<.001	0.69	5.44
	J01CA	Penicillins with extended spectrum	<.001	0.66	14.12
	J01EE	Combinations of sulfonamides and trimethoprim, incl. derivatives	.03	0.61	2.44
ICD#					
	298	Other nonorganic psychoses	.16	0.68	0.16
	411	Other acute and subacute forms of ischemic heart disease	.32	0.62	0.08
2					
ATC#					
	A10BA	Biguanides	<.001	0.86	4.31
	N02BE	Anilides	<.001	0.79	6.4

Continued on next page

Table 5.10: Frequently occurred morbidity features in the evolutionarily obtained final bins dataset (Continued)

Cohort	Features	Description	P value	Support	Prevalence
	J05AB	Nucleosides and nucleotides excl. reverse transcriptase inhibitors	<.001	0.76	2.91
	C03CA	Sulfonamides, plain	<.001	0.76	4.09
	M04AA	Preparations inhibiting uric acid production	<.001	0.74	5.13
	C09CA	Angiotensin II receptor blockers (ARBs), plain	<.001	0.71	8.4
	C02CA	Alpha-adrenoreceptor antagonists	<.001	0.65	3.22
	C08CA	Dihydropyridine derivatives	<.001	0.65	7.4
	J02AC	Triazole and tetrazole derivatives	.03	0.64	6.18
	N06AB	Selective serotonin reuptake inhibitors	.03	0.63	8.58
	S01EE	Prostaglandin analogues	.07	0.62	0.68
	N03AG	Fatty acid derivatives	.08	0.61	2.5
	M01AB	Acetic acid derivatives and related substances	.001	0.6	18.21
	N03AE	Benzodiazepine derivatives	.17	0.6	1.54
ICD#					
	V64	Surgical or other procedure not carried out because of contraindication	1.0	0.64	0.18
	V54	Other orthopedic aftercare	.26	0.64	0.32
	188	Malignant neoplasm of bladder	.32	0.63	0.18
	735	Acquired deformities of toe	1.0	0.6	0.18
	454	Varicose veins of lower extremities	.83	0.6	1.04

Continued on next page

Table 5.10: Frequently occurred morbidity features in the evolutionarily obtained final bins dataset (Continued)

Cohort	Features	Description	P value	Support	Prevalence
	820	Fracture of neck of femur	.32	0.6	0.05
3					
ATC#					
	N02AX	Other opioids	<.001	0.84	12.96
	M04AA	Preparations inhibiting uric acid production	<.001	0.82	8.5
	C03EA	Low-ceiling diuretics and potassium-sparing agents	<.001	0.76	5.35
	A02BA	H2-receptor antagonists	.004	0.75	4.04
	B01AB	Heparin group	<.001	0.73	12.59
	N03AX	Other antiepileptics	<.001	0.7	11.7
	N02AA	Natural opium alkaloids	<.001	0.68	13.9
	J05AB	Nucleosides and nucleotides excl. reverse transcriptase inhibitors	.008	0.65	5.77
	A12AA	Calcium	.11	0.62	4.67
	C07BB	Beta blocking agents, selective, and thiazides	.16	0.62	2.62
	B03BB	Folic acid and derivatives	<.001	0.61	9.23
	R03AC	Selective beta-2-adrenoreceptor agonists	.005	0.6	7.19
ICD#					
	295	Schizophrenic disorders	.03	0.68	0.73
	813	Fracture of radius and ulna	.62	0.68	0.84
4					
ATC#					
	G04CA	Alpha-adrenoreceptor antagonists	.02	0.8	25.75

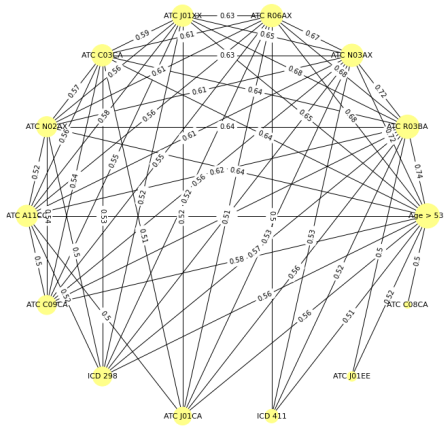
Continued on next page

Table 5.10: Frequently occurred morbidity features in the evolutionarily obtained final bins dataset (Continued)

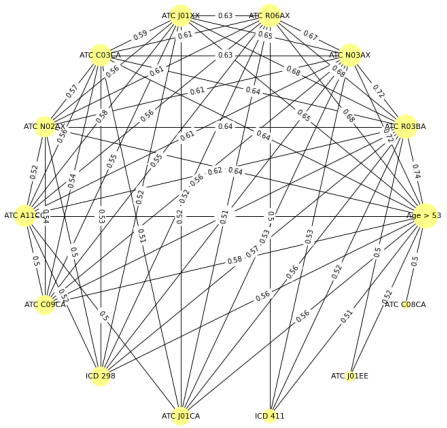
Cohort	Features	Description	P value	Support	Prevalence
	J01CA	Penicillins with extended spectrum	.008	0.73	14.47
	C09DA	Angiotensin II receptor blockers (ARBs) and diuretics	.07	0.66	13.11
	C09AA	ACE inhibitors, plain	.03	0.66	26.32
	B01AA	Vitamin K antagonists	.001	0.64	4.61
	C03CA	Sulfonamides, plain	.002	0.62	16.49
ICD#					
	995	Certain adverse effects not elsewhere classified	1.0	0.61	0.44

The graph illustrated in Figure 5.7 represents the combinations obtained by analyzing all of the two variable combinations with smin = 0.5. The results for all combinations are provided in Appendices D.1, D.2, D.3 and D.4.

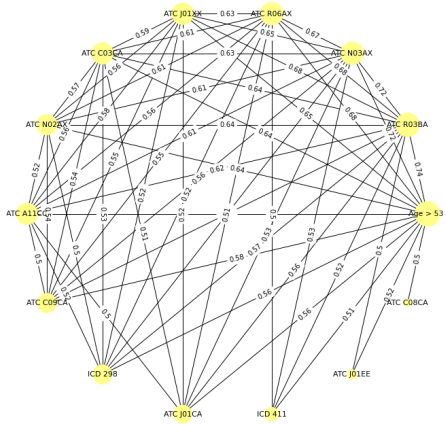
We can see that some multimorbidity features appear in most of the outcome-associated bins. Moreover, some acts as common frequent features in the final bins of various cohorts. In Table 5.11, the features and combinations that are frequently appeared in the final bins dataset when configured the support (s) between 0.7–1.0 are tabulated and they are graphically presented in Figure 5.8.



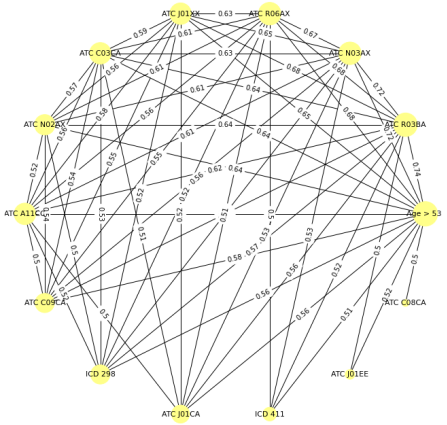
(a) Cohort 1



(b) Cohort 2



(c) Cohort 3



(d) Cohort 4

Figure 5.7: Frequent outcome-associated multimorbidity feature combinations (two variable combinations with smin = 0.5)

Table 5.11: Frequently appeared features and combinations in the final bins dataset when configured the support (s) between 0.7–1.0

Support	Length of the combination	Frequent Features	Cohort
0.85	1	ATC R03BA	cohort 1
0.84	1	age >53	cohort 1
0.82	1	ATC N03AX	cohort 1
0.79	1	ATC R06AX	cohort 1
0.78	1	ATC J01XX	cohort 1
0.76	1	ATC C03CA	cohort 1
0.74	1	ATC N02AX	cohort 1
0.74	2	age >53, ATC R03BA	cohort 1
0.73	1	ATC A11CC	cohort 1
0.72	2	ATC N03AX, ATC R03BA	cohort 1
0.72	2	age >53, ATC N03AX	cohort 1
0.86	1	ATC A10BA	cohort 2
0.79	1	ATC N02BE	cohort 2
0.76	1	ATC C03CA	cohort 2
0.76	1	ATC J05AB	cohort 2
0.74	1	ATC M04AA	cohort 2
0.71	1	ATC C09CA	cohort 2
0.84	1	ATC N02AX	cohort 3
0.82	1	ATC M04AA	cohort 3
0.76	1	ATC C03EA	cohort 3
0.75	1	ATC A02BA	cohort 3
0.73	1	ATC B01AB	cohort 3
0.71	2	ATC M04AA, ATC N02AX	cohort 3
0.7	1	ATC N03AX	cohort 3
0.8	1	ATC G04CA	cohort 4
0.73	1	ATC J01CA	cohort 4

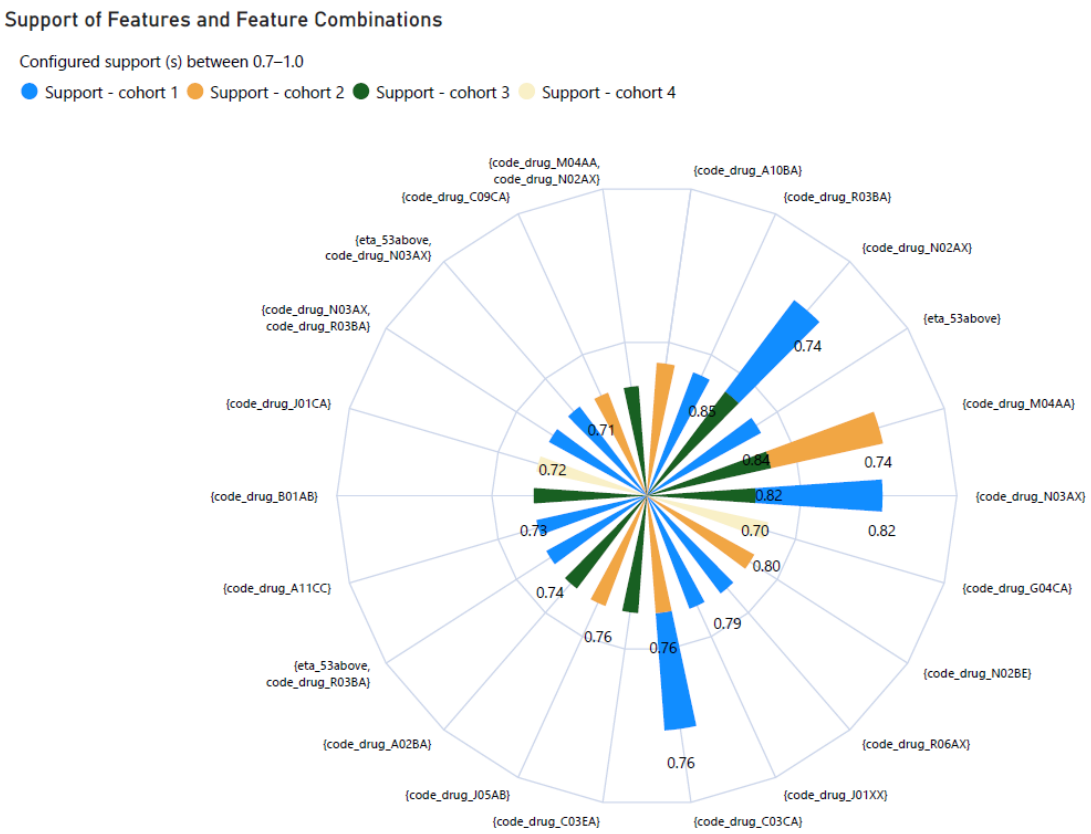


Figure 5.8: Illustration of the features and combinations that are frequently appeared in the final bins dataset when configured the support (s) between 0.7–1.0 as radar chart with features presented in more than one cohort is stacked.

Chapter 6

Discussion

6.1 Principal Observations

The primary findings of the study highlight prevalent morbidity patterns within the evolved dataset. These patterns, characterized by specific ATC codes and ICD codes, demonstrate significant associations with the hospitalization outcome, particularly among specific demographic groups. This analysis not only contributes insights in the context of COVID-19 but also offers potential for broader applications. By repurposing data initially collected for administrative purposes, this innovative approach holds promise for multimorbidity analysis in the realm of public health. This underscores the adaptability and versatility of the methodology, with the potential to extract valuable insights from existing datasets, thereby informing effective public health strategies and interventions.

The use of the method to address data sparsity in medical data and improve the understanding of the factors associated with the impact of infectious diseases on health outcomes in a population with multimorbidity is significant. In the first project of the research, the methodology adopted maps data from higher-dimensional spaces into lower-dimensional ones. This approach lead to a loss of information, but it can also yield advantages in Machine Learning by mitigating sparsity issues, ultimately enhancing predictive capabilities without much computational requirement as Evolutionary Algorithms.

In the second project of the research, it not only underscores the effectiveness of Evolutionary Machine Learning but also opens up promising avenues for future exploration in the realm of managing multimorbid conditions.

Moreover, the utilization of a novel Evolutionary Machine Learning approach demonstrates that even with rare events, meaningful results can be obtained. The evolutionary mining of prevalent morbidity patterns that are more associated with the outcome showcases the potential of this method to derive valuable insights, which can be particularly impactful in situations involving sparsity of data due to rare events.

In this research, prevalent morbidity patterns from the evolved dataset are identified. Notably, multimorbidity features like higher age with specific ATC codes (N03AX, R03BA) were common in outcome-related bins, especially in middle-aged females. Also, while analyzing SHAP values, it is found that in the case of Cohort 1, the inhaled corticosteroid medication used for Asthma (R03BA) has a very high positive impact on the analyzed hospitalized outcome. This supports the Open SAFELY study, which highlighted asthma as a noteworthy risk factor for mortality in individuals with COVID-19 and indicated that patients using inhaled corticosteroids face the highest risk in this context [100].

ATC NO3AX group comprises of other antiepileptics. This group include medications utilized for bipolar disease, epilepsy treatment, migraine management as well as schizophrenia in some cases. Individuals afflicted by a severe mental illness exhibited a slightly elevated risk of experiencing severe clinical outcomes due to COVID-19 compared to individuals not affected by prior mental health conditions [101]. Moreover, there have been reports of a link between the use of antiepileptics and the occurrence of vitamin D deficiency [102]. In our study, multimorbidity associated with presence of A11CC (Vitamin D and analogues) in history makes middle aged female more vulnerable for hospitalization. On the other hand, for elderly female, presence of this feature is associated with smaller SHAP values. So, presence of this drug in the history makes such patients not to be hospitalized and thus act as protective.

In a multimorbidity study of hospitalized COVID-19 patients [103], the ATC group most closely linked with prolonged hospital stays is M04AA, preparations inhibiting uric acid production. In our study also, in case of elderly females, M04AA and NO3AX combinations were prominent. M04AA also appeared frequently in middle-aged males, while G04CA (Alpha-adrenoreceptor antagonists), used for benign prostatic hypertrophy, was prominent among elderly males. Research indicates that male COVID-19 cohorts experience more unfavorable clinical outcomes compared to females [104, 105]. Notably, while cancer patients face heightened susceptibility to SARS-CoV-2

infection, individuals with prostate cancer receiving androgen-deprivation therapies seem to possess some level of protection against the infection [105].

6.2 Strength and Limitations

While there is a connection between ICD and ATC codes, the factors such as the nature of drugs and their diverse uses introduce complexity [106]. This research is carried out under the assumption ATC and ICD codes are dependent and they are not collinear. Collinearity occurs when two features are highly correlated and provide redundant information. In this case, dependence implies that there is a connection between the two code systems, but it doesn't mean that one can be perfectly predicted from the other. Also, a drug can be prescribed for different diseases, leading to different ICD codes. For instance, an antiacid might be used for conditions beyond just one specific disease. This diversity in usage adds complexity to the relationship between drugs and ICD codes.

The dimensionally reduced data with newly engineered features can be used for the predictive modeling and the removal of data sparsity by the proposed unsupervised binning of the rare features offered a low dimensional feature matrix for the predictive modeling.

It is important to note that the absence of a detailed clustering validity analysis leaves a potential gap in fully understanding the robustness of these clusters. However, the binary nature of the data, where diseases are represented as present or absent for each patient, inherently limits the complexity of the clustering process and the clusters formed through BMD are likely to reflect clear patterns without significant noise. BMD is specifically designed to capture underlying patterns or latent features in binary data. By decomposing the binary patient-disease matrix, BMD inherently identifies clusters that represent meaningful associations between patients and diseases. Moreover, previous studies in similar domains have demonstrated the effectiveness of BMD in extracting meaningful clusters from binary data [82].

The predictive ability of the new sparsity-free feature matrix and the original sparse data is compared and found that with a very low number of features itself, the model achieves nearly equal prediction performance. Also, the predictive utility of the new feature matrix by interpreting the feature importance and impact of the new features in the Machine Learning model is checked.

All Evolutionary Algorithms possess a bias towards the best-performing choices that are available. So, even though Evolutionary Algorithms are stochastic, these biases lead these algorithms to perform better. Also, for identifying the best performing group of features, each evolutionary cycle involves bin fitness evaluation and genetic operations for generating an elite population. In this study, the Evolutionary Algorithm is not just used for feature selection in a sparse data. On the other hand, in each evolutionary cycle, the epistatic association between features are indirectly assessed by using the strategy of binned multimorbidity features. These group of features are scored based on the ability of a Deep Learning Classifier to predict the outcome and the features in the bins are regrouped after each evolutionary cycle.

Many works that use Machine Learning for investigating multimorbidity patterns only use the removal of sparsity-generating features from the dataset to deal with the sparse datasets. In some cases, they merge the categories of features to avoid sparsity. However, these methods lead to more information loss and vague interpretation of multimorbidity features. Rather than just concluding the analysis based on a simple sequential Deep Learning model, all the evolved bins are aggregated and obtained a new dataset contain these final bins. By analyzing the evolutionarily evolved bins, the frequent multimorbidity features and combinations are obtained.

It is computationally very expensive to analyze all possible combinations of multimorbidity features in a dataset, and many irrelevant feature combinations need not be analyzed further. An Evolutionary Algorithm has already been applied to obtain meaningful combinations, including less prevalent features in those combinations. As a result, only the most common multimorbidity features in the top bins were further analyzed.

6.3 Future Perspective

This research has the potential to enhance the delivery of personalized medicine and patient-centered care by tailoring treatments and interventions to the specific multimorbid profiles of individuals. This could lead to improved patient outcomes, reduced healthcare costs, and an overall enhancement in the quality of healthcare services.

By clustering and creating new features, the method adopted in project 1 could provide a more detailed understanding of multimorbidity patterns and the associations between different diseases in the context of sparse binary

medical data. Continuing to refine and expand the clustering and feature generation methods can lead to even more nuanced insights into how different medical conditions co-occur and influence one another. This could result in the identification of previously unrecognized disease clusters or risk factors. Moreover, building on the method's capabilities, researchers could explore its potential in predicting long-term health outcomes for patients with specific multimorbidity profiles. This could aid in early intervention and preventive measures.

Continued research can concentrate on refining and optimizing the Evolutionary Machine Learning algorithm employed in Project 2, tailoring it specifically for the analysis of multimorbidity features. These algorithms could be designed to further decrease computational costs while preserving and improving their capacity to identify meaningful feature combinations.

Improving the understanding of the factors associated with the severity of COVID-19 in this population could have important implications for public health policies, and for the assessment of patients particularly vulnerable to the disease. The method has the potential to lead to better healthcare outcomes and inform public health policies related to COVID-19 and other similar public health contexts.

Chapter 7

Conclusions

When combined with other multimorbidity features, it is identified that even less prevalent medical conditions show association with the outcome. The discovery of hidden interconnections between the different multimorbidity features can offer a new research pathway for the study of multidimensional medical conditions in combination.

In the future, further investigation into these intricate associations can lead to a deeper understanding of disease interactions and the development of more effective treatment strategies tailored to patients with complex multimorbid profiles.

Mapping higher-dimensional data to a low-dimensional space using the clustering technique adopted in this research can result in information loss, but reducing sparsity can be beneficial for Machine Learning modeling due to improved predictive ability.

In this research, the issue of data sparsity in electronic health records is addressed and created a model that incorporates both prevalent and rare medical conditions, leading to more accurate and effective predictive modeling. Looking ahead, this approach has the potential to revolutionize healthcare by enabling the development of predictive models that can assist clinicians in making informed decisions and ultimately improving patient outcomes. The identification of complex associations between multimorbidity and the severity of COVID-19 has shed light on crucial areas for future research. This includes in-depth studies on long COVID, which remains a challenging and poorly understood aspect of the pandemic. Furthermore, intervention efforts can be more precisely targeted based on the insights gained from these complex associations, potentially leading to more effective public health strategies.

By employing an innovative Evolutionary Machine Learning approach, prevalent morbidity patterns associated with hospitalization risk is uncovered, particularly among specific age and gender cohorts. The findings of this research underscore the adaptability of this methodology, showcasing its potential to derive meaningful results even from situations involving rare events. Moreover, the repurposing of administrative data for multimorbidity analysis presents an innovative avenue for public health research. This opens up possibilities for leveraging existing data sources to gain a better understanding of health trends and disparities. As public health challenges continue to evolve, this approach can inform effective strategies and interventions, enhancing the overall well-being of communities. As a future direction, this research can be enhanced by integrating patient stratification based on their healthcare requirements. This entails grouping patient data to identify cohorts with similar healthcare utilization patterns. This approach will aid in identifying patient subgroups with distinct clinical profiles, enabling the design of targeted interventions and personalized care plans. By refining patient stratification methods, we can optimize healthcare delivery and improve patient outcomes, ultimately advancing the field of healthcare and public health research.

Chapter 8

Submitted Articles

1. Dayana Benny, Mario Giacobini, Alberto Catalano, Giuseppe Costa, Roberto Gnavi, Fulvio Ricceri. Evolutionary Machine Learning Based Multimorbidity Analysis In COVID-19 Hospitalized Patients: A Longitudinal Study Using Health-administrative Data of a Region in the North-West of Italy. JMIR Public Health and Surveillance (submitted on 31 agust 2023, Submission under second round of peer review).
2. Dayana Benny, Mario Giacobini, Giuseppe Costa, Roberto Gnavi, Fulvio Ricceri. Multimorbidity in Middle-aged Women and COVID-19: Binary Data Clustering for Unsupervised Binning of Rare Multimorbidity Features and Predictive Modeling. BMC Medical Research Methodology (Submitted on: 27 May 2023, Submission passed technical check 30 June 2023, Submission under second round of peer review).

Bibliography

[1] Andrew Clark, Mark Jit, Charlotte Warren-Gash, Bruce Guthrie, Harry HX Wang, Stewart W Mercer, Colin Sanderson, Martin McKee, Christopher Troeger, Kanyin L Ong, et al. Global, regional, and national estimates of the population at increased risk of severe covid-19 due to underlying health conditions in 2020: a modelling study. *The Lancet Global Health*, 8(8):e1003–e1017, 2020.

[2] Alberto Catalano, Lucia Dansero, Winston Gilcrease, Alessandra Macciotta, Carlo Saugo, Luca Manfredi, Roberto Gnavi, Elena Strippoli, Nicolás Zengarini, Valeria Caramello, et al. Multimorbidity and sars-cov-2–related outcomes: Analysis of a cohort of italian patients. *JMIR Public Health and Surveillance*, 9(1):e41404, 2023.

[3] Helga Radner, Kazuki Yoshida, Josef S Smolen, and Daniel H Solomon. Multimorbidity and rheumatic conditions—enhancing the concept of comorbidity. *Nature Reviews Rheumatology*, 10(4):252–256, 2014.

[4] Finn Breinholt Larsen, Marie Hauge Pedersen, Karina Friis, Charlotte Glümer, and Mathias Lasgaard. A latent class analysis of multimorbidity and the relationship to socio-demographic factors and health-related quality of life. a national population-based study of 162,283 danish adults. *PloS one*, 12(1):e0169426, 2017.

[5] Yutao Zhang, Robert Chen, Jie Tang, Walter F Stewart, and Jimeng Sun. Leap: learning to prescribe effective and safe treatment combinations for multimorbidity. In *proceedings of the 23rd ACM SIGKDD international conference on knowledge Discovery and data Mining*, pages 1315–1324, 2017.

[6] Christopher Harrison, Martin Fortin, Marjan van den Akker, Frances Mair, Amaia Calderon-Larranaga, Fiona Boland, Emma Wallace,

Bhautesh Jani, and Susan Smith. Comorbidity versus multimorbidity: Why it matters, 2021.

[7] Jeeva Kanesarajah, Michael Waller, Jennifer A Whitty, and Gita D Mishra. Multimorbidity and quality of life at mid-life: a systematic review of general population studies. *Maturitas*, 109:53–62, 2018.

[8] Masoomeh Alimohammadian, Azam Majidi, Mehdi Yaseri, Batoul Ahmadi, Farhad Islami, Mohammad Derakhshan, Alireza Delavari, Mohammad Amani, Akbar Feyz-Sani, Hossein Poustchi, et al. Multimorbidity as an important issue among women: results of a gender difference investigation in a large population-based cross-sectional study in west asia. *BMJ open*, 7(5):e013548, 2017.

[9] Steven M McPhail. Multimorbidity in chronic disease: impact on health care resources and costs. *Risk management and healthcare policy*, pages 143–156, 2016.

[10] Simone Turner, M Asad Khan, David Putrino, Ashley Woodcock, Douglas B Kell, and Etheresia Pretorius. Long covid: pathophysiological factors and abnormalities of coagulation. *Trends in Endocrinology & Metabolism*, 34(6):321–344, 2023.

[11] Clark D Russell, Nazir I Lone, and J Kenneth Baillie. Comorbidities, multimorbidity and covid-19. *Nature Medicine*, 29(2):334–343, 2023.

[12] Regi Jose, Meghana Narendran, Anil Bindu, Nazeema Beevi, L Manju, and PV Benny. Public perception and preparedness for the pandemic covid 19: a health belief model approach. *Clinical epidemiology and global health*, 9:41–46, 2021.

[13] Alice Delerue Matos, Andreia Fonseca de Paiva, Cláudia Cunha, and Gina Voss. Precautionary behaviours of individuals with multimorbidity during the covid-19 pandemic. *European journal of ageing*, pages 1–9, 2022.

[14] Dayana Benny, Silvia Castro, Omer Mujahid, and Olga Lugovska Abrosimova. Contact tracing for covid-19 in ukraine: insights from a case study in the region of chernivtsi. *Regional Academy on United Nations*, 2021.

[15] Lucy E Stirland, Laura González-Saavedra, Donncha S Mullin, Craig W Ritchie, Graciela Muniz-Terrera, and Tom C Russ. Measuring multimorbidity beyond counting diseases: systematic review of community and population studies and guide to index choice. *Bmj*, 368, 2020.

[16] Vijaya Sundararajan, Toni Henderson, Catherine Perry, Amanda Muggivan, Hude Quan, and William A Ghali. New icd-10 version of the charlson comorbidity index predicted in-hospital mortality. *Journal of clinical epidemiology*, 57(12):1288–1294, 2004.

[17] Bryant R England, Yangyuna Yang, Punyasha Roul, Christian Haas, Lotfollah Najjar, Harlan Sayles, Fang Yu, Brian C Sauer, Joshua F Baker, Fenglong Xie, et al. Identification of multimorbidity patterns in rheumatoid arthritis through machine learning. *Arthritis care & research*, 75(2):220–230, 2023.

[18] Javier Alvarez-Galvez and Esteban Vegas-Lozano. Discovery and classification of complex multimorbidity patterns: unravelling chronicity networks and their social profiles. *Scientific Reports*, 12(1):20004, 2022.

[19] Julián A Fernández-Niño, John A Guerra-Gómez, and Alvaro J Idrovo. Multimorbidity patterns among covid-19 deaths: proposal for the construction of etiological models. *Revista Panamericana de Salud Pública*, 44, 2020.

[20] Hendrik Van den Bussche, Daniela Koller, Tina Kolonko, Heike Hansen, Karl Wegscheider, Gerd Glaeske, Eike-Christin von Leitner, Ingmar Schäfer, and Gerhard Schön. Which chronic diseases and disease combinations are specific to multimorbidity in the elderly? results of a claims data based cross-sectional study in germany. *BMC public health*, 11(1):1–9, 2011.

[21] Marlous Hall, Tatendashe B Dondo, Andrew T Yan, Mamas A Mamas, Adam D Timmis, John E Deanfield, Tomas Jernberg, Harry Hemingway, Keith AA Fox, and Chris P Gale. Multimorbidity and survival for patients with acute myocardial infarction in england and wales: Latent class analysis of a nationwide population-based cohort. *PLoS medicine*, 15(3):e1002501, 2018.

[22] Alessandra Bisquera, Martin Gulliford, Hiten Dodhia, Lesedi Ledwaba-Chapman, Stevo Durbaba, Marina Soley-Bori, Julia Fox-Rushby, Mark Ashworth, and Yanzhong Wang. Identifying longitudinal clusters of multimorbidity in an urban setting: a population-based cross-sectional study. *The Lancet Regional Health–Europe*, 3, 2021.

[23] Belinda Hernández, Richard B Reilly, and Rose Anne Kenny. Investigation of multimorbidity and prevalent disease combinations in older irish adults using network analysis and association rules. *Scientific reports*, 9(1):14567, 2019.

[24] Alexandra Prados-Torres, Beatriz Poblador-Plou, Amaia Calderón-Larrañaga, Luis Andrés Gimeno-Feliu, Francisca González-Rubio, Antonio Poncel-Falcó, Antoni Sicras-Mainar, and José Tomás Alcalá-Nalvaiz. Multimorbidity patterns in primary care: interactions among chronic diseases using factor analysis. *PloS one*, 7(2):e32190, 2012.

[25] Kayvan Aflaki, Simone Vigod, and Joel G Ray. Part i: A friendly introduction to latent class analysis. *Journal of Clinical Epidemiology*, 147:168–170, 2022.

[26] Stephanie T Lanza, Bethany C Bray, and Linda M Collins. An introduction to latent class and latent transition analysis. *Handbook of Psychology, Second Edition*, 2, 2012.

[27] An Gie Yong, Sean Pearce, et al. A beginner's guide to factor analysis: Focusing on exploratory factor analysis. *Tutorials in quantitative methods for psychology*, 9(2):79–94, 2013.

[28] Guiying Dong, Zi-Chao Zhang, Jianfeng Feng, and Xing-Ming Zhao. Morbidgcn: prediction of multimorbidity with a graph convolutional network based on integration of population phenotypes and disease network. *Briefings in Bioinformatics*, 23(4):bbac255, 2022.

[29] Maede S Nouri, Daniel J Lizotte, Kamran Sedig, and Sheikh S Abdullah. Visemure: A visual analytics system for making sense of multimorbidity using electronic medical record data. *Data*, 6(8):85, 2021.

[30] Charles Gadd, Krishnarajah Nirantharakumar, and Christopher Yau. mmvae: multimorbidity clustering using relaxed bernoulli beta variational autoencoders. In *Machine Learning for Health*, pages 88–102. PMLR, 2022.

[31] Reka Maria Blazsik, Patrick Emanuel Beeler, Karol Tarcak, Marcus Cheetham, Viktor von Wyl, and Holger Dressel. Impact of single and combined rare diseases on adult inpatient outcomes: a retrospective, cross-sectional study of a large inpatient population. *Orphanet Journal of Rare Diseases*, 16:1–8, 2021.

[32] R Craig Stillwell. Exclusion of women from covid-19 studies harms women's health and slows our response to pandemics. *Biology of sex Differences*, 13(1):27, 2022.

[33] Derek M Griffith, Garima Sharma, Christopher S Holliday, Okechuku K Enyia, Matthew Valliere, Andrea R Semlow, Elizabeth C Stewart, and Roger Scott Blumenthal. Men and covid-19: a biopsychosocial approach to understanding sex differences in mortality and recommendations for practice and policy interventions. *Preventing chronic disease*, 17:E63, 2020.

[34] Janice L Atkins, Jane AH Masoli, Joao Delgado, Luke C Pilling, Chia-Ling Kuo, George A Kuchel, and David Melzer. Preexisting comorbidities predicting covid-19 and mortality in the uk biobank community cohort. *The Journals of Gerontology: Series A*, 75(11):2224–2230, 2020.

[35] Rosario Pivonello, Renata S Auriemma, Claudia Pivonello, Andrea M Isidori, Giovanni Corona, Annamaria Colao, and Robert P Millar. Sex disparities in covid-19 severity and outcome: are men weaker or women stronger? *Neuroendocrinology*, 111(11):1066–1085, 2021.

[36] Faouzi Marzouki and Omar Bouattane. Deep learning based model for automatic multimorbidity pattern prognosis. In *2023 3rd International Conference on Innovative Research in Applied Science, Engineering and Technology (IRASET)*, pages 1–6. IEEE, 2023.

[37] Satvik Dasariraju and Ryan J Urbanowicz. Rare: evolutionary feature engineering for rare-variant bin discovery. In *Proceedings of the Genetic and Evolutionary Computation Conference Companion*, pages 1335–1343, 2021.

[38] Yiming Chen, Lei Shi, Xiao Zheng, Juan Yang, Yaqing Xue, Shujuan Xiao, Benli Xue, Jiachi Zhang, Xinru Li, Huang Lin, et al. Patterns and determinants of multimorbidity in older adults: Study in

health-ecological perspective. *International Journal of Environmental Research and Public Health*, 19(24):16756, 2022.

[39] Dong Ling Tong and Robert Mintram. Genetic algorithm-neural network (gann): a study of neural network activation functions and depth of genetic algorithm search applied to feature selection. *International Journal of Machine Learning and Cybernetics*, 1:75–87, 2010.

[40] Seyedmostafa Sheikhalishahi, Anirban Bhattacharyya, Leo Anthony Celi, and Venet Osmani. An interpretable deep learning model for time-series electronic health records: Case study of delirium prediction in critical care. *Artificial Intelligence in Medicine*, 144:102659, 2023.

[41] Ritam Guha, Manosij Ghosh, Souvik Kapri, Sushant Shaw, Shyok Mutsuddi, Vikrant Bhateja, and Ram Sarkar. Deluge based genetic algorithm for feature selection. *Evolutionary intelligence*, 14:357–367, 2021.

[42] R Vaishali, R Sasikala, S Ramasubbareddy, S Remya, and Sravani Nalluri. Genetic algorithm based feature selection and moe fuzzy classification algorithm on pima indians diabetes dataset. In *2017 international conference on computing networking and informatics (ICCNI)*, pages 1–5. IEEE, 2017.

[43] Marjan van den Akker, Frank Buntinx, Sjef Roos, and J André Knottnerus. Problems in determining occurrence rates of multimorbidity. *Journal of clinical epidemiology*, 54(7):675–679, 2001.

[44] Ronald Gijsen, Nancy Hoeymans, François G Schellevis, Dirk Ruwaard, William A Satariano, and Geertrudis AM van den Bos. Causes and consequences of comorbidity: a review. *Journal of clinical epidemiology*, 54(7):661–674, 2001.

[45] Mark Q Thompson, Solomon Yu, Graeme R Tucker, Robert J Adams, Matteo Cesari, Olga Theou, and Renuka Visvanathan. Frailty and sarcopenia in combination are more predictive of mortality than either condition alone. *Maturitas*, 144:102–107, 2021.

[46] Keqiang Wan, Chang Su, Lingxi Kong, Juan Liao, Wenguang Tian, and Hua Luo. Clinical characteristics of covid-19 in young patients differ from middle-aged and elderly patients. *Archives of Medical Science: AMS*, 18(3):704, 2022.

[47] CDC. Centers for disease control and prevention, Nov 2021. Available at https://www.cdc.gov/nchs/icd/icd9cm.htm.

[48] Ashraf El-Metwally, Faris Fatani, Nouf Binhowaimel, Badr F Al Khateeb, Hanan M Al Kadri, Awad Alshahrani, Aljohrah I Aldubikhi, Mona I Bin Amer, Abdulrahman Almuflih, and Abdulaziz S Alangari. Effect modification by age and gender in the correlation between diabetes mellitus, hypertension, and obesity. *Journal of Primary Care & Community Health*, 14:21501319231220234, 2023.

[49] Neus Llop Torrent, Giorgio Visani, and Enrico Bagli. Psd2 explainable ai model for credit scoring. *arXiv preprint arXiv:2011.10367*, 2020.

[50] María Pérez-Ortiz, Pedro Antonio Gutiérrez, Peter Tino, and César Hervás-Martínez. Oversampling the minority class in the feature space. *IEEE transactions on neural networks and learning systems*, 27(9):1947–1961, 2015.

[51] Matloob Khushi, Kamran Shaukat, Talha Mahboob Alam, Ibrahim A Hameed, Shahadat Uddin, Suhuai Luo, Xiaoyan Yang, and Maranatha Consuelo Reyes. A comparative performance analysis of data resampling methods on imbalance medical data. *IEEE Access*, 9:109960–109975, 2021.

[52] Sebastián Maldonado, Julio López, and Carla Vairetti. An alternative smote oversampling strategy for high-dimensional datasets. *Applied Soft Computing*, 76:380–389, 2019.

[53] Nitesh V Chawla, Kevin W Bowyer, Lawrence O Hall, and W Philip Kegelmeyer. Smote: synthetic minority over-sampling technique. *Journal of artificial intelligence research*, 16:321–357, 2002.

[54] Mimi Mukherjee and Matloob Khushi. Smote-enc: A novel smote-based method to generate synthetic data for nominal and continuous features. *Applied System Innovation*, 4(1):18, 2021.

[55] Alexander Yun-chung Liu. The effect of oversampling and undersampling on classifying imbalanced text datasets. 2004.

[56] M Mostafizur Rahman and Darryl N Davis. Addressing the class imbalance problem in medical datasets. *International Journal of Machine Learning and Computing*, 3(2):224, 2013.

[57] Ibomoiye Domor Mienye and Yanxia Sun. Performance analysis of cost-sensitive learning methods with application to imbalanced medical data. *Informatics in Medicine Unlocked*, 25:100690, 2021.

[58] Vasileios Iosifidis, Symeon Papadopoulos, Bodo Rosenhahn, and Eirini Ntoutsi. Adacc: cumulative cost-sensitive boosting for imbalanced classification. *Knowledge and Information Systems*, 65(2):789–826, 2023.

[59] Mohamed Bekkar and Taklit Akrouf Alitouche. Imbalanced data learning approaches review. *International Journal of Data Mining & Knowledge Management Process*, 3(4):15, 2013.

[60] Joseph Prusa, Taghi M Khoshgoftaar, David J Dittman, and Amri Napolitano. Using random undersampling to alleviate class imbalance on tweet sentiment data. In *2015 IEEE international conference on information reuse and integration*, pages 197–202. IEEE, 2015.

[61] David W Hosmer Jr, Stanley Lemeshow, and Rodney X Sturdivant. *Applied logistic regression*, volume 398. John Wiley & Sons, 2013.

[62] Liudmila Prokhorenkova, Gleb Gusev, Aleksandr Vorobev, Anna Veronika Dorogush, and Andrey Gulin. Catboost: unbiased boosting with categorical features. *Advances in neural information processing systems*, 31, 2018.

[63] Jerome H Friedman. Greedy function approximation: a gradient boosting machine. *Annals of statistics*, pages 1189–1232, 2001.

[64] Yoav Freund, Robert E Schapire, et al. Experiments with a new boosting algorithm. In *icml*, volume 96, pages 148–156. Citeseer, 1996.

[65] Ronald A Fisher. The use of multiple measurements in taxonomic problems. *Annals of eugenics*, 7(2):179–188, 1936.

[66] Leo Breiman. Random forests. *Machine learning*, 45:5–32, 2001.

[67] Thomas Bayes. Lii. an essay towards solving a problem in the doctrine of chances. by the late rev. mr. bayes, frs communicated by mr. price, in a letter to john canton, amfr s. *Philosophical transactions of the Royal Society of London*, (53):370–418, 1763.

[68] Guolin Ke, Qi Meng, Thomas Finley, Taifeng Wang, Wei Chen, Weidong Ma, and Qiwei Ye. Lightgbm: A highly efficient gradient boosting decision tree. *Advances in Neural Information Processing Systems*, 30:3146–3154, 2017.

[69] Pierre Geurts, Damien Ernst, and Louis Wehenkel. Extremely randomized trees. *Machine Learning*, 63:3–42, 2006.

[70] Tianqi Chen and Carlos Guestrin. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 785–794. ACM, 2016.

[71] Trevor Hastie, Robert Tibshirani, Jerome H Friedman, and Jerome H Friedman. *The elements of statistical learning: data mining, inference, and prediction*, volume 2. Springer, 2009.

[72] Corinna Cortes and Vladimir Vapnik. Support-vector networks. *Machine learning*, 20:273–297, 1995.

[73] Arthur E Hoerl and Robert W Kennard. Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, 12(1):55–67, 1970.

[74] Scott M Lundberg and Su-In Lee. A unified approach to interpreting model predictions. *Advances in neural information processing systems*, 30, 2017.

[75] John E Cornell, Jacqueline A Pugh, John W Williams Jr, Lewis Kazis, Austin FS Lee, Michael L Parchman, John Zeber, Thomas Pederson, Kelly A Montgomery, and Polly Hitchcock Noël. Multimorbidity clusters: clustering binary data from multimorbidity clusters: clustering binary data from a large administrative medical database. *Applied multivariate research*, 12(3):163–182, 2008.

[76] Michael A Ghebre, Mona Bafadhel, Dhananjay Desai, Suzanne E Cohen, Paul Newbold, Laura Rapley, Jo Woods, Paul Rugman, Ian D Pavord, Chris Newby, et al. Biological clustering supports both "dutch" and "british" hypotheses of asthma and chronic obstructive pulmonary disease. *Journal of Allergy and Clinical Immunology*, 135(1):63–72, 2015.

[77] Sheryl Hui-Xian Ng, Nabilah Rahman, Ian Yi Han Ang, Srinath Sridharan, Sravan Ramachandran, Debby D Wang, Chuen Seng Tan, Sue-Anne Toh, and Xin Quan Tan. Characterization of high healthcare utilizer groups using administrative data from an electronic medical record database. *BMC health services research*, 19(1):1–14, 2019.

[78] Concepción Violán, Albert Roso-Llorach, Quintí Foguet-Boreu, Marina Guisado-Clavero, Mariona Pons-Vigués, Enriqueta Pujol-Ribera, and Jose M Valderas. Multimorbidity patterns with k-means nonhierarchical cluster analysis. *BMC family practice*, 19:1–11, 2018.

[79] Pablo E Bretos-Azcona, Eduardo Sánchez-Iriso, and Juan M Cabasés Hita. Tailoring integrated care services for high-risk patients with multiple chronic conditions: a risk stratification approach using cluster analysis. *BMC Health Services Research*, 20:1–9, 2020.

[80] Dayana Benny, Kumary R Soumya, and K Nageswara Rao. New dynamic self-organizing feature maps for the classification of extracted feature vectors of characters. In *2015 International Conference on Robotics, Automation, Control and Embedded Systems (RACE)*, pages 1–3. IEEE, 2015.

[81] Mohammed A Khalilia, Mihail Popescu, and James Keller. Patient stratification based on activity of daily living score using relational self-organizing maps. In *2014 IEEE Symposium on Computational Intelligence in Healthcare and e-health (CICARE)*, pages 112–116. IEEE, 2014.

[82] Tao Li. A general model for clustering binary data. In *Proceedings of the eleventh ACM SIGKDD international conference on Knowledge discovery in data mining*, pages 188–197, 2005.

[83] Dayana Benny and Kumary R Soumya. New local adaptive thresholding and dynamic self-organizing feature map techniques for handwritten character recognizer. In *2015 International Conference on Circuits, Power and Computing Technologies [ICCPCT-2015]*, pages 1–4. IEEE, 2015.

[84] Moez Ali. Pycaret: An open source, low-code machine learning library in python. *PyCaret version*, 2, 2020.

[85] Pycaret. Feature selection - docs. Accessed 30 January 2023.

[86] Malka N Halgamuge, Eshan Daminda, and Ampalavanapillai Nirmalathas. Best optimizer selection for predicting bushfire occurrences using deep learning. *Natural Hazards*, 103(1):845–860, 2020.

[87] Alex Labach, Hojjat Salehinejad, and Shahrokh Valaee. Survey of dropout methods for deep neural networks. *arXiv preprint arXiv:1904.13310*, 2019.

[88] Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Dropout: a simple way to prevent neural networks from overfitting. *The journal of machine learning research*, 15(1):1929–1958, 2014.

[89] Liu Liu and Hairong Qi. Learning effective binary descriptors via cross entropy. In *2017 IEEE winter conference on applications of computer vision (WACV)*, pages 1251–1258. IEEE, 2017.

[90] Samanyou Garg. Group emotion recognition using machine learning. *arXiv preprint arXiv:1905.01118*, 2019.

[91] Hong-Jian Xue, Xinyu Dai, Jianbing Zhang, Shujian Huang, and Jiajun Chen. Deep matrix factorization models for recommender systems. In *IJCAI*, volume 17, pages 3203–3209. Melbourne, Australia, 2017.

[92] Rizgar Zebari, Adnan Abdulazeez, Diyar Zeebaree, Dilovan Zebari, and Jwan Saeed. A comprehensive review of dimensionality reduction techniques for feature selection and feature extraction. *Journal of Applied Science and Technology Trends*, 1(2):56–70, 2020.

[93] Randall Wald, Taghi M Khoshgoftaar, Amri Napolitano, and Chris Sumner. Using twitter content to predict psychopathy. In *2012 11th International Conference on Machine Learning and Applications*, volume 2, pages 394–401. IEEE, 2012.

[94] Jun Zhang, Zhi-hui Zhan, Ying Lin, Ni Chen, Yue-jiao Gong, Jing-hui Zhong, Henry SH Chung, Yun Li, and Yu-hui Shi. Evolutionary computation meets machine learning: A survey. *IEEE Computational Intelligence Magazine*, 6(4):68–75, 2011.

[95] Emad Elbeltagi, Tarek Hegazy, and Donald Grierson. Comparison among five evolutionary-based optimization algorithms. *Advanced engineering informatics*, 19(1):43–53, 2005.

[96] Pradnya A Vikhar. Evolutionary algorithms: A critical review and its future prospects. In *2016 International conference on global trends in signal processing, information computing and communication (ICGTSPICC)*, pages 261–265. IEEE, 2016.

[97] Sourabh Katoch, Sumit Singh Chauhan, and Vijay Kumar. A review on genetic algorithm: past, present, and future. *Multimedia tools and applications*, 80:8091–8126, 2021.

[98] Manoj Kumar, Dr Mohammad Husain, Naveen Upreti, and Deepti Gupta. Genetic algorithm: Review and application. *Available at SSRN 3529843*, 2010.

[99] Sangram Patil, Aum Patil, and Vikas M Phalle. Life prediction of bearing by using adaboost regressor. In *Proceedings of TRIBOINDIA-2018 An International Conference on Tribology*, 2018.

[100] Elizabeth J Williamson, Alex J Walker, Krishnan Bhaskaran, Seb Bacon, Chris Bates, Caroline E Morton, Helen J Curtis, Amir Mehrkar, David Evans, Peter Inglesby, et al. Factors associated with covid-19-related death using opensafely. *Nature*, 584(7821):430–436, 2020.

[101] Seung Won Lee, Jee Myung Yang, Sung Yong Moon, In Kyung Yoo, Eun Kyo Ha, So Young Kim, Un Min Park, Sejin Choi, Sang-Hyuk Lee, Yong Min Ahn, et al. Association between mental illness and covid-19 susceptibility and clinical outcomes in south korea: a nationwide cohort study. *The Lancet Psychiatry*, 7(12):1025–1031, 2020.

[102] Sergej Nadalin, Hrvoje Jakovac, Vjekoslav Peitl, Dalibor Karlović, and Alena Buretić-Tomljanović. Dysregulated inflammation may predispose patients with serious mental illnesses to severe covid-19. *Molecular Medicine Reports*, 24(2):1–9, 2021.

[103] Mohammad-Reza Malekpour, Mohsen Abbasi-Kangevari, Ali Shojaee, Sahar Saeedi Moghaddam, Seyyed-Hadi Ghamari, Mohammad-Mahdi Rashidi, Alireza Namazi Shabestari, Mohammad Effatpanah, Mohammadmehdi Nasehi, Mehdi Rezaei, et al. Effect of the chronic medication use on outcome measures of hospitalized covid-19 patients: Evidence from big data. *Frontiers in public health*, 11:1061307, 2023.

[104] Jonathan D Strope, Cindy H Chau, and William D Figg. Are sex discordant outcomes in covid-19 related to sex hormones? In *Seminars in oncology*, volume 47, pages 335–340. Elsevier, 2020.

[105] Monica Montopoli, Sara Zumerle, Roberto Vettor, Massimo Rugge, Manuel Zorzi, Carlo V Catapano, GM Carbone, Andrea Cavalli, Francesco Pagano, Eugenio Ragazzi, et al. Androgen-deprivation therapies for prostate cancer and risk of infection by sars-cov-2: a population-based study (n= 4532). *Annals of Oncology*, 31(8):1040–1045, 2020.

[106] Irene López-Rodríguez, Cesár F Reyes-Manzano, Ariel Guzmán-Vargas, and Lev Guzmán-Vargas. The complex structure of the pharmacological drug–disease network. *Entropy*, 23(9):1139, 2021.

APPENDICES

Appendix A

Pseudocode for the feature score calculation on final bins

Algorithm 1: Pseudo code for calculating feature scores using prediction based method

```
Data: feature matrix
Result: best score of each feature combination
for each feature combination do
    get all the values in the feature combination;
    for each value in the feature combination as threshold do
        for each patient do
            if value of feature combination>threshold then
                put the patient in above threshold instances;
            else
                put the patient in under threshold instances;

        predicted as hospitalized=above threshold instances;
        predicted as non-hospitalized=under threshold instances;

        TP=actual hospitalized & predicted as hospitalized;
        TN=actual non-hospitalized & predicted as non-hospitalized;
        FP=actual non-hospitalized & predicted as hospitalized;
        FN=actual hospitalized & predicted as non-hospitalized;

        //calculate score as accuracy;
        score=count(TP)+count(TN)/(count(TP)+count(TN)+count(FP)+count(FN));
    find max score and corresponding threshold;
return max scores as best scores of each feature combination;
```

Appendix B

One Proportion z-test Results

If P value <= 0.05, the features are eliminated from the sampled dataset.

Table B.1: One Proportion z-test Results - Cohort 1

Features	**z score**	**P value**	**Prevalence of the feature in raw data (%)**	**Prevalence of the feature in sampled data (%)**
ATC J01CR	-1.92	.06	47.2	44.48
ATC H02AB	-1.68	.09	43.67	41.31
Age >53	-0.14	.89	41.34	41.15
ATC J01FA	-1.25	.21	36.77	35.06
ATC A02BC	-2.57	.01	32.39	29.06
ATC J01MA	-1.78	.07	29.28	27.03
ATC M01AB	-1.78	.07	26.11	23.94
ATC A11CC	-1.34	.18	24.66	23.05
ATC M01AE	-2.65	.008	21.44	18.51
ATC J01DD	-0.13	.9	21.33	21.19
ATC R03BA	-1.04	.3	16.57	15.5
ATC J01XX	-1.48	.14	15.68	14.2
ATC J01CA	-0.76	.45	14.88	14.12
ATC N06AB	-0.77	.44	13.8	13.07
ATC B03AA	-0.33	.74	11.99	11.69
ATC H03AA	-1.1	.27	11.77	10.8
ATC J02AC	-0.41	.68	11.08	10.71
ATC A02AD	-1.66	.10	10.97	9.58
ATC N02BE	-1.47	.14	10.72	9.5
ATC C07AB	-0.57	.57	10.14	9.66

Continued on next page

Table B.1: One Proportion z-test Results - Cohort 1 (Continued)

Features	z score	P value	Prevalence of the feature in raw data (%)	Prevalence of the feature in sampled data (%)
ATC A07AA	-0.29	.77	9.98	9.74
ATC R03AK	-1.3	.19	9.56	8.52
ATC M01AH	-0.4	.69	9.34	9.01
ATC C09AA	-1.73	.08	9.02	7.71
ATC A02BX	-0.8	.43	8.06	7.47
ATC N02AX	-1.06	.29	7.66	6.9
ATC N02AA	-0.59	.55	7.33	6.9
ATC R06AX	-0.45	.65	7.06	6.74
ATC B03BB	-0.74	.46	6.68	6.17
ATC N06AX	-0.02	.99	6.34	6.33
ATC C08CA	-0.71	.48	6.32	5.84
ATC N03AX	-0.96	.34	6.23	5.6
ATC C10AA	-1.64	.10	5.96	4.95
ATC C09CA	-0.54	.59	5.79	5.44
ATC B01AB	-0.76	.44	5.76	5.28
ATC N02CC	-0.59	.56	5.65	5.28
ATC R06AE	-1.6	.11	5.14	4.22
ATC R03AC	-1.03	.3	5.07	4.46
ATC A12AX	-0.63	.53	4.67	4.3
ATC C03CA	0.87	.39	4.65	5.19
ATC J05AB	-1.01	.31	4.62	4.06
ATC M01AX	-1.24	.22	4.58	3.9
ATC B01AC	-0.18	.85	4.24	4.14
ATC C09DA	-1.02	.31	4.11	3.57
ATC M01AC	-2.63	.009	3.89	2.68
ATC G03DB	-0.91	.36	3.71	3.25
ATC B02AA	-0.13	.9	3.64	3.57
ATC C09BA	-0.12	.91	3.31	3.25
ATC G03CA	-1.12	.26	3.28	2.76
ATC A10BA	-0.07	.94	3.28	3.25
ATC N06AA	-1.69	.09	3.08	2.35
ATC A02BA	0.59	.55	2.95	3.25
ATC P01AB	-0.21	.83	2.86	2.76
ICD 621	0.92	.36	2.26	2.68

Continued on next page

Table B.1: One Proportion z-test Results - Cohort 1 (Continued)

Features	z score	P value	Prevalence of the feature in raw data (%)	Prevalence of the feature in sampled data (%)
ATC J01EE	0.46	.65	2.23	2.44
ATC G03AA	0.51	.61	2.21	2.44
ATC J01AA	0.38	.7	2.19	2.35
ATC A03FA	-1.01	.31	2.17	1.79
ATC C03EA	0.54	.59	2.12	2.35
ATC A07EC	-0.06	.95	2.05	2.03
ATC D05AX	0.51	.61	2.05	2.27
ATC C09BB	-1.21	.23	1.97	1.54
ICD 218	-1.56	.12	1.9	1.38
ATC J01DC	0.99	.32	1.85	2.27
ICD 727	-0.12	.9	1.83	1.79
ATC A05AA	0.21	.84	1.79	1.87
ICD 574	-0.63	.53	1.68	1.46
ATC N03AG	-0.12	.9	1.59	1.54
ATC M04AA	0.56	.57	1.5	1.7
ATC A12AA	-0.82	.41	1.47	1.22
ATC C02CA	-0.54	.59	1.47	1.3
ATC C07AA	0.69	.49	1.45	1.7
ATC R03DC	-2.24	.02	1.38	0.81
ATC B03BA	-0.2	.84	1.36	1.3
ATC R03AL	0.99	.32	1.34	1.7
ICD 454	0.19	.85	1.32	1.38
ATC S01ED	-0.75	.45	1.18	0.97
ICD 735	-0.44	.66	1.18	1.06
ATC C07BB	-0.29	.77	1.14	1.06
ICD V58	1.41	.16	1.12	1.62
ATC C03DA	-0.75	.45	1.09	0.89
ATC C03BA	1.14	.26	1.07	1.46
ATC P01BA	-0.93	.35	1.05	0.81
ATC N02BA	0.36	.72	1.03	1.14
ICD 174	0.17	.86	1.01	1.06
ATC R03BB	-1.46	.15	0.98	0.65
ATC C10AX	-0.25	.8	0.96	0.89
ATC N03AF	-0.49	.62	0.94	0.81

Continued on next page

Table B.1: One Proportion z-test Results - Cohort 1 (Continued)

Features	z score	P value	Prevalence of the feature in raw data (%)	Prevalence of the feature in sampled data (%)
ATC M05BA	-1.16	.24	0.92	0.65
ATC A12BA	0.48	.63	0.92	1.06
ATC N03AE	0.08	.94	0.87	0.89
ICD 278	0.08	.94	0.87	0.89
ATC L01BA	0.61	.54	0.8	0.97
ICD 717	0.33	.74	0.8	0.89
ATC A07EA	0.33	.74	0.8	0.89
ATC C09BX	-1.49	.14	0.78	0.49
ATC C09DB	-2.68	.007	0.76	0.32
ICD 338	-0.29	.77	0.71	0.65
ICD 626	-1.04	.3	0.69	0.49
ATC A10AB	-0.19	.85	0.69	0.65
ICD 726	-0.19	.85	0.69	0.65
ATC A10BB	0.25	.8	0.67	0.73
ATC N01BB	0.91	.36	0.65	0.89
ATC N05AD	0.34	.73	0.65	0.73
ICD 715	-1.99	.05	0.65	0.32
ATC C10AB	0.34	.73	0.65	0.73
ATC N02AB	0.73	.47	0.63	0.81
ATC N02AJ	-0.7	.49	0.63	0.49
ATC S01EE	-1.72	.09	0.6	0.32
ATC C03AA	-0.47	.64	0.58	0.49
ATC S01EC	-0.06	.95	0.58	0.57
ICD 473	-0.25	.8	0.54	0.49
ICD 455	-1.3	.19	0.54	0.32
ATC N04AA	-0.6	.55	0.51	0.41
ATC C10BA	0.89	.37	0.51	0.73
ICD 553	-0.47	.64	0.49	0.41
ICD V54	0.09	.93	0.47	0.49
ATC C03EB	-0.35	.73	0.47	0.41
ICD 296	1.08	.28	0.47	0.73
ATC N03AA	-2.48	.01	0.45	0.16
ATC C07AG	0.57	.57	0.45	0.57
ATC B01AA	0.89	.38	0.45	0.65

Continued on next page

Table B.1: One Proportion z-test Results - Cohort 1 (Continued)

Features	z score	P value	Prevalence of the feature in raw data (%)	Prevalence of the feature in sampled data (%)
ATC N05AA	-0.75	.45	0.45	0.32
ICD V64	0.67	.5	0.42	0.57
ATC N05AH	0.43	.67	0.4	0.49
ATC R03DA	0.02	.98	0.4	0.41
ICD 592	0.43	.67	0.4	0.49
ATC B05BB	-0.97	.33	0.38	0.24
ICD 241	-1.51	.13	0.34	0.16
ICD 996	-0.65	.51	0.34	0.24
ICD 618	0.39	.7	0.34	0.41
ATC C01BC	-0.49	.62	0.31	0.24
ICD 038	1.47	.14	0.31	0.65
ICD 812	0.88	.38	0.31	0.49
ICD 354	-1.12	.26	0.29	0.16
ATC C02AC	0.64	.52	0.29	0.41
ICD 478	-0.33	.74	0.29	0.24
ICD V53	-0.17	.86	0.27	0.24
ICD 301	0.35	.73	0.27	0.32
ICD 780	0.35	.73	0.27	0.32
ICD 295	-0.02	.99	0.25	0.24
ICD 298	-0.53	.59	0.22	0.16
ICD 518	1.33	.18	0.22	0.49
ATC C01DA	-0.53	.59	0.22	0.16
ICD 434	-1.75	.08	0.22	0.08
ICD 562	-0.53	.59	0.22	0.16
ICD 998	0.14	.89	0.22	0.24
ICD 599	-0.34	.74	0.2	0.16
ICD 585	1.13	.26	0.2	0.41
ICD V43	-1.48	.14	0.2	0.08
ICD 427	0.3	.76	0.2	0.24
ICD 722	0.3	.76	0.2	0.24
ICD 786	-1.2	.23	0.18	0.08
ICD 820	-0.14	.89	0.18	0.16
ICD 550	-0.14	.89	0.18	0.16
ICD V56	1.25	.21	0.18	0.41

Continued on next page

Table B.1: One Proportion z-test Results - Cohort 1 (Continued)

Features	z score	P value	Prevalence of the feature in raw data (%)	Prevalence of the feature in sampled data (%)
ICD 410	-0.14	.89	0.18	0.16
ICD 470	0.46	.64	0.18	0.24
ICD V57	0.46	.64	0.18	0.24
ICD 560	-0.93	.35	0.16	0.08
ICD 482	0.78	.44	0.13	0.24
ICD 813	0.25	.8	0.13	0.16
ICD 486	0.44	.66	0.11	0.16
ICD 041	0.44	.66	0.11	0.16
ICD 211	0.44	.66	0.11	0.16
ICD 162	-0.1	.92	0.09	0.08
ICD 437	-0.1	.92	0.09	0.08
ICD 424	1.1	.27	0.09	0.24
ICD 995	0.64	.52	0.09	0.16
ICD V71	-0.1	.92	0.09	0.08
ICD 571	-0.1	.92	0.09	0.08
ICD 728	0.17	.86	0.07	0.08
ICD 428	0.17	.86	0.07	0.08
ATC C01BD	1.26	.21	0.07	0.24
ICD 188	0.17	.86	0.07	0.08
ICD 250	0.83	.41	0.07	0.16
ICD 438	0.45	.65	0.04	0.08
ICD 366	0.45	.65	0.04	0.08
ICD 411	0.45	.65	0.04	0.08
ICD 440	0.73	.47	0.02	0.08
ICD 153	0.73	.47	0.02	0.08

Removed features: ATC A02BC, ATC M01AE, ATC M01AC, ATC R03DC, ATC C09DB, ICD 715, ATC N03AA

Table B.2: One Proportion z-test Results - Cohort 2

Features	z score	P value	Prevalence of the feature in raw data (%)	Prevalence of the feature in sampled data (%)
Age >53	-0.72	.47	45.45	44.69

Continued on next page

Table B.2: One Proportion z-test Results - Cohort 2 (Continued)

Features	z score	P value	Prevalence of the feature in raw data (%)	Prevalence of the feature in sampled data (%)
ATC J01CR	-1.07	.28	41.45	40.33
ATC H02AB	-1.25	.21	34.77	33.51
ATC J01FA	-0.76	.45	29.43	28.7
ATC J01MA	-1.76	.08	27.43	25.79
ATC A02BC	-3.06	.002	26.91	24.11
ATC M01AB	-1.73	.08	19.63	18.21
ATC J01DD	-1.34	.18	16.89	15.85
ATC M01AE	-0.78	.43	16.23	15.62
ATC R03BA	-1.33	.18	13.94	12.99
ATC C09AA	-1.25	.21	12.86	11.99
ATC J01CA	-0.67	.51	11.94	11.49
ATC C10AA	-2.93	.003	10.92	9.13
ATC A07AA	-1.61	.11	10.78	9.76
ATC C07AB	-2.24	.02	10.75	9.36
ATC C09CA	-0.82	.41	8.89	8.4
ATC N06AB	-0.51	.61	8.89	8.58
ATC C08CA	-1.46	.14	8.22	7.4
ATC R03AK	-0.82	.41	8.18	7.72
ATC G04CA	-1.96	.05	8.01	6.95
ATC B01AC	-2.68	.007	7.55	6.18
ATC A02AD	-1.61	.11	7.48	6.63
ATC A11CC	-0.74	.46	7.02	6.63
ATC N02BE	-0.59	.56	6.71	6.4
ATC J02AC	-0.9	.37	6.64	6.18
ATC A02BX	-0.69	.49	6.25	5.9
ATC C09DA	-1.16	.25	5.87	5.31
ATC M04AA	-1.26	.21	5.73	5.13
ATC B01AB	-0.52	.6	5.66	5.4
ATC N03AX	-0.97	.33	5.44	5
ATC M01AH	-1.45	.15	5.23	4.59
ATC N02AA	-1.26	.21	5.2	4.63
ATC R06AX	-0.66	.51	5.16	4.86
ATC A10BA	-1.8	.07	5.09	4.31
ATC N02AX	-1.07	.29	4.78	4.31

Continued on next page

Table B.2: One Proportion z-test Results - Cohort 2 (Continued)

Features	z score	P value	Prevalence of the feature in raw data (%)	Prevalence of the feature in sampled data (%)
ATC N06AX	-1.22	.22	4.6	4.09
ATC C03CA	-1.05	.29	4.53	4.09
ATC B03BB	-1.31	.19	4.5	3.95
ATC J01XX	-0.69	.49	4.43	4.13
ATC R03AC	-0.11	.91	4.18	4.13
ATC C09BA	-0.81	.42	4.14	3.81
ATC M01AX	-1.57	.12	3.86	3.27
ATC C02CA	-0.95	.34	3.58	3.22
ATC C09BB	-0.39	.7	3.23	3.09
ATC B03AA	-1.12	.26	3.16	2.77
ATC R06AE	-1.12	.26	3.16	2.77
ATC J05AB	-0.52	.61	3.09	2.91
ATC J01EE	-0.85	.39	3.02	2.72
ATC H03AA	-1.15	.25	2.88	2.5
ATC N03AG	-0.94	.35	2.81	2.5
ATC C10AX	-0.8	.43	2.81	2.54
ATC A07EC	-0.59	.56	2.74	2.54
ATC D05AX	-0.59	.56	2.6	2.41
ATC M01AC	-0.23	.82	2.53	2.45
ATC A07EA	-1.22	.22	2.21	1.86
ATC C10AB	-0.25	.8	2.07	2
ATC N02BA	-1.43	.15	2.07	1.68
ICD 550	0.62	.53	2.07	2.27
ATC A10AB	-0.57	.57	1.93	1.77
ATC C09DB	-0.74	.46	1.93	1.73
ATC A02BA	-1.15	.25	1.9	1.59
ATC N03AE	-0.94	.35	1.79	1.54
ATC N02CC	0.25	.81	1.79	1.86
ATC A10BB	-1.29	.2	1.58	1.27
ATC P01AB	-0.14	.89	1.58	1.54
ATC C03DA	-1	.32	1.51	1.27
ICD 717	-0.13	.9	1.44	1.41
ATC S01ED	-0.22	.83	1.37	1.32
ATC J01AA	-0.46	.64	1.33	1.23

Continued on next page

Table B.2: One Proportion z-test Results - Cohort 2 (Continued)

Features	z score	P value	Prevalence of the feature in raw data (%)	Prevalence of the feature in sampled data (%)
ATC B03BA	-2.11	.03	1.33	0.91
ATC A05AA	0.03	.98	1.26	1.27
ATC R03AL	-0.36	.72	1.26	1.18
ATC N05AD	-0.36	.72	1.26	1.18
ATC N03AF	-0.21	.83	1.23	1.18
ATC N05AH	-0.06	.95	1.19	1.18
ATC G04CB	-0.76	.45	1.16	1
ATC N04AA	-0.53	.6	1.16	1.04
ATC N06AA	-0.53	.6	1.16	1.04
ATC J01DC	-0.31	.75	1.16	1.09
ICD 574	-0.89	.37	1.09	0.91
ICD V58	-0.2	.84	1.09	1.04
ICD 727	-0.42	.67	1.09	1
ATC A12AX	-0.2	.84	1.09	1.04
ATC C01DA	-1.41	.16	1.09	0.82
ATC B01AA	-0.97	.33	1.05	0.86
ATC A03FA	-0.26	.8	1.05	1
ATC R03BB	-0.97	.33	1.05	0.86
ICD 592	-0.48	.63	1.05	0.95
ATC C09BX	-0.31	.75	1.02	0.95
ATC N05AA	-0.37	.71	0.98	0.91
ATC B02AA	-1.13	.26	0.98	0.77
ICD 454	0.28	.78	0.98	1.04
ATC R03DC	-0.95	.34	0.95	0.77
ICD 410	-1.64	.10	0.91	0.64
ATC A12AA	0.15	.88	0.88	0.91
ATC C03BA	-0.57	.57	0.88	0.77
ICD 553	-0.13	.89	0.84	0.82
ATC A12BA	0.1	.92	0.84	0.86
ATC C10BA	-0.92	.36	0.84	0.68
ATC C07AG	-0.19	.85	0.81	0.77
ATC C03EA	-0.45	.65	0.81	0.73
ATC C07BB	-0.72	.47	0.81	0.68
ATC N03AA	-0.81	.42	0.77	0.64

Continued on next page

Table B.2: One Proportion z-test Results - Cohort 2 (Continued)

Features	z score	P value	Prevalence of the feature in raw data (%)	Prevalence of the feature in sampled data (%)
ATC S01EE	-0.52	.6	0.77	0.68
ICD 715	0.42	.68	0.74	0.82
ICD 038	-0.78	.44	0.67	0.54
ATC C02AC	-0.26	.8	0.63	0.59
ATC R03DA	-0.88	.38	0.63	0.5
ATC B05BB	-0.88	.38	0.63	0.5
ICD 518	-0.26	.8	0.63	0.59
ICD 455	-0.11	.91	0.56	0.54
ICD 473	-0.11	.91	0.56	0.54
ATC N02AB	-0.75	.45	0.56	0.45
ICD 413	-0.11	.91	0.56	0.54
ICD 470	-0.42	.68	0.56	0.5
ATC L01BA	-0.75	.45	0.56	0.45
ATC C07AA	-0.51	.61	0.53	0.45
ATC N02AJ	-0.18	.86	0.53	0.5
ICD V53	0.12	.91	0.53	0.54
ATC C03AA	-1.28	.2	0.53	0.36
ICD 600	0.05	.96	0.49	0.5
ICD 301	-1.45	.15	0.49	0.32
ATC C03EB	0.34	.73	0.49	0.54
ATC S01EC	-0.02	.99	0.46	0.45
ATC C01BC	-1.16	.25	0.46	0.32
ICD 585	0.29	.78	0.46	0.5
ATC N01BB	0.52	.6	0.42	0.5
ICD 298	0.23	.82	0.42	0.45
ICD 214	0.52	.6	0.42	0.5
ICD 278	0.23	.82	0.42	0.45
ICD 996	0.16	.87	0.39	0.41
ICD 995	-1.03	.31	0.39	0.27
ICD 427	-0.57	.57	0.39	0.32
ICD 722	-0.18	.86	0.39	0.36
ICD 434	-1.03	.31	0.39	0.27
ICD 482	0.09	.93	0.35	0.36
ICD 726	-0.28	.78	0.35	0.32

Continued on next page

Table B.2: One Proportion z-test Results - Cohort 2 (Continued)

Features	z score	P value	Prevalence of the feature in raw data (%)	Prevalence of the feature in sampled data (%)
ICD 295	0.42	.67	0.35	0.41
ICD V43	0.09	.93	0.35	0.36
ICD 414	-0.71	.48	0.35	0.27
ICD 438	-1.22	.22	0.35	0.23
ICD 571	-0.71	.48	0.35	0.27
ICD 478	-0.88	.38	0.32	0.23
ICD 786	-0.39	.69	0.32	0.27
ICD V54	0.01	.99	0.32	0.32
ICD 431	-0.39	.69	0.32	0.27
ICD 486	-0.39	.69	0.32	0.27
ATC C01BD	-1.09	.27	0.28	0.18
ICD 560	-0.08	.94	0.28	0.27
ICD 296	-0.53	.59	0.28	0.23
ICD 411	-1.84	.07	0.28	0.14
ICD 728	-0.08	.94	0.28	0.27
ICD 415	-1.84	.07	0.28	0.14
ICD 338	-0.08	.94	0.28	0.27
ICD 428	0.31	.76	0.28	0.32
ICD V64	-1.09	.27	0.28	0.18
ICD V57	-1.84	.07	0.28	0.14
ICD V71	-0.19	.85	0.25	0.23
ICD 211	-0.19	.85	0.25	0.23
ICD 780	-0.19	.85	0.25	0.23
ICD 424	-0.32	.75	0.21	0.18
ICD V56	0.56	.58	0.21	0.27
ICD 241	0.16	.87	0.21	0.23
ICD 173	0.16	.87	0.21	0.23
ATC P01BA	-0.95	.34	0.21	0.14
ICD 812	0.07	.95	0.18	0.18
ICD 250	0.07	.95	0.18	0.18
ICD 041	-0.5	.62	0.18	0.14
ICD 437	-0.5	.62	0.18	0.14
ICD 998	0.51	.61	0.18	0.23
ICD 354	-0.5	.62	0.18	0.14

Continued on next page

Table B.2: One Proportion z-test Results - Cohort 2 (Continued)

Features	z score	P value	Prevalence of the feature in raw data (%)	Prevalence of the feature in sampled data (%)
ICD 440	-0.05	.96	0.14	0.14
ICD 366	-0.05	.96	0.14	0.14
ICD 735	0.45	.65	0.14	0.18
ICD 188	0.45	.65	0.14	0.18
ICD 562	0.45	.65	0.14	0.18
ICD 813	-0.05	.96	0.14	0.14
ICD 153	-0.23	.82	0.11	0.09
ICD 185	-0.23	.82	0.11	0.09
ICD 584	0.39	.69	0.11	0.14
ATC M05BA	-0.23	.82	0.11	0.09
ICD 599	-0.23	.82	0.11	0.09
ICD 820	-1.32	.19	0.11	0.05
ICD 276	-0.55	.58	0.07	0.05
ICD 162	-0.55	.58	0.07	0.05
ATC G03DB	0.23	.82	0.04	0.05
ICD 331	0.23	.82	0.04	0.05

Removed features: ATC A02BC, ATC C10AA, ATC C07AB, ATC B01AC, ATC B03BA

Table B.3: One Proportion z-test Results - Cohort 3

Features	z score	P value	Prevalence of the feature in raw data (%)	Prevalence of the feature in sampled data (%)
ATC J01CR	0.16	.87	54.01	54.2
ATC A02BC	-0.13	.89	49.47	49.32
ATC H02AB	-0.09	.93	47.69	47.59
ATC A11CC	0.13	.9	45.18	45.33
ATC J01MA	-0.19	.85	41.4	41.19
ATC J01FA	-0.32	.75	38.6	38.25
Age >68	1.42	.15	37.03	38.61
ATC M01AB	-0.05	.96	33.16	33.11
ATC M01AE	1.1	.27	30.62	31.79
ATC J01DD	-0.01	.99	29.13	29.12
ATC C10AA	0.04	.97	26.24	26.29

Continued on next page

Table B.3: One Proportion z-test Results - Cohort 3 (Continued)

Features	z score	P value	Prevalence of the feature in raw data (%)	Prevalence of the feature in sampled data (%)
ATC C07AB	0.18	.86	24.59	24.76
ATC R03BA	-0.91	.36	23.27	22.4
ATC N06AB	0.18	.85	22.12	22.3
ATC J01XX	0.03	.98	21.49	21.51
ATC B01AC	-0.16	.87	21.19	21.04
ATC C03CA	-0.1	.92	20.08	19.99
ATC C09AA	0.09	.93	19.49	19.57
ATC M01AH	0.54	.59	17.83	18.31
ATC A07AA	0.12	.91	17.58	17.68
ATC J01CA	-0.63	.53	17.54	17
ATC H03AA	-0.73	.47	17.2	16.58
ATC N02BE	0.22	.83	17.07	17.26
ATC C08CA	0.43	.67	15.8	16.16
ATC A02AD	-0.69	.49	15.63	15.06
ATC R03AK	-0.19	.85	15.37	15.22
ATC C09CA	0.48	.63	15.03	15.42
ATC C09DA	0.72	.47	14.31	14.9
ATC N02AA	-0.03	.98	13.93	13.9
ATC N06AX	0.1	.92	13.04	13.12
ATC N02AX	0.23	.82	12.78	12.96
ATC A12AX	-0.77	.44	12.7	12.12
ATC A02BX	-0.74	.46	12.57	12.01
ATC B01AB	0.53	.59	12.19	12.59
ATC N03AX	-0.14	.89	11.8	11.7
ATC A10BA	0.18	.86	11.25	11.39
ATC B03BB	-1.06	.29	9.94	9.23
ATC M04AA	-0.26	.8	8.66	8.5
ATC M01AX	-0.26	.8	8.66	8.5
ATC C09BA	-0.12	.9	8.58	8.5
ATC J02AC	0.54	.59	8.15	8.5
ATC B03AA	-0.73	.47	7.94	7.5
ATC R03AC	-0.05	.96	7.22	7.19
ATC M01AC	0.51	.61	6.84	7.14
ATC J05AB	-0.64	.52	6.11	5.77

Continued on next page

Table B.3: One Proportion z-test Results - Cohort 3 (Continued)

Features	z score	P value	Prevalence of the feature in raw data (%)	Prevalence of the feature in sampled data (%)
ATC R06AE	0.05	.96	5.9	5.93
ATC N02BA	0.05	.96	5.9	5.93
ATC R06AX	-0.67	.5	5.86	5.51
ATC G03CA	-0.18	.86	5.39	5.3
ATC C02CA	0.5	.62	5.35	5.61
ATC M05BA	0.08	.93	5.31	5.35
ATC R03BB	0.07	.94	5.05	5.09
ATC J01EE	0.16	.88	5.01	5.09
ATC C03EA	0.91	.36	4.88	5.35
ATC A12AA	-0.09	.93	4.71	4.67
ATC C03DA	0.06	.95	4.59	4.62
ATC A07EC	0.04	.96	4.54	4.56
ATC A02BA	-0.83	.4	4.42	4.04
ATC A10AB	0.01	.99	4.25	4.25
ATC C09BB	0.81	.42	4.08	4.46
ATC C10AX	-0.01	.99	3.99	3.99
ATC A12BA	-0.03	.97	3.95	3.93
ATC N06AA	0.3	.77	3.91	4.04
ATC A10BB	0	1.0	3.78	3.78
ATC S01ED	0.56	.58	3.74	3.99
ATC D05AX	-0.82	.41	3.69	3.36
ATC A05AA	0	1.0	3.57	3.57
ATC B01AA	0.38	.7	3.35	3.52
ATC A03FA	-0.15	.88	3.31	3.25
ATC N03AG	0.36	.72	3.31	3.46
ICD 715	0.09	.93	3.27	3.31
ATC C03EB	0.42	.67	3.18	3.36
ATC B03BA	-0.88	.38	3.06	2.73
ATC N05AD	-0.45	.65	3.06	2.89
ATC C03BA	0.71	.48	3.01	3.31
ATC R03AL	0.22	.83	2.8	2.89
ATC N02CC	-0.09	.93	2.76	2.73
ICD V58	-0.4	.69	2.72	2.57
ATC C07BB	-0.03	.98	2.63	2.62

Continued on next page

Table B.3: One Proportion z-test Results - Cohort 3 (Continued)

Features	z score	P value	Prevalence of the feature in raw data (%)	Prevalence of the feature in sampled data (%)
ATC P01AB	-0.86	.39	2.55	2.26
ATC C10BA	-0.23	.82	2.55	2.47
ATC C01DA	0.18	.86	2.51	2.57
ATC J01DC	0.46	.64	2.51	2.68
ATC N02AJ	0.32	.75	2.51	2.62
ICD 518	0.07	.94	2.34	2.36
ATC C07AA	-0.27	.79	2.29	2.2
ATC N02AB	0.61	.54	2.25	2.47
ATC N03AE	0.11	.91	2.17	2.2
ATC P01BA	-0.27	.79	2.08	1.99
ATC C09DB	0.64	.52	2.04	2.26
ATC B05BB	-0.64	.52	1.87	1.68
ICD V43	0.23	.82	1.87	1.94
ATC L01BA	-0.1	.92	1.87	1.84
ICD 574	0.17	.86	1.78	1.84
ICD 727	0.34	.74	1.78	1.89
ATC S01EE	0.14	.89	1.74	1.78
ATC R03DA	-0.4	.69	1.74	1.63
ATC C10AB	0.45	.65	1.7	1.84
ATC S01EC	0.45	.65	1.7	1.84
ICD 038	0.45	.65	1.7	1.84
ATC N03AA	-0.29	.77	1.66	1.57
ATC B02AA	-0.33	.74	1.61	1.52
ATC N03AF	-0.52	.6	1.61	1.47
ATC R03DC	-1	.32	1.57	1.31
ATC N04AA	0.19	.85	1.57	1.63
ATC N05AA	0.19	.85	1.57	1.63
ATC C07AG	-0.37	.71	1.57	1.47
ATC N05AH	0.34	.74	1.53	1.63
ATC C01BC	-0.67	.5	1.49	1.31
ATC J01AA	0.65	.51	1.49	1.68
ICD 338	0.09	.93	1.44	1.47
ICD 621	0.09	.93	1.44	1.47
ATC N01BB	-0.3	.76	1.44	1.36

Continued on next page

Table B.3: One Proportion z-test Results - Cohort 3 (Continued)

Features	z score	P value	Prevalence of the feature in raw data (%)	Prevalence of the feature in sampled data (%)
ATC C03AA	0.94	.35	1.4	1.68
ICD 296	-0.34	.73	1.4	1.31
ICD 820	-0.61	.54	1.36	1.21
ICD 996	-0.22	.82	1.32	1.26
ICD 410	-0.06	.95	1.27	1.26
ATC C02AC	-0.06	.95	1.27	1.26
ICD 174	-0.06	.95	1.27	1.26
ICD 735	0.14	.88	1.27	1.31
ICD 454	-0.14	.89	1.19	1.15
ICD 428	0.07	.94	1.19	1.21
ICD 427	0.03	.97	1.15	1.15
ATC C09BX	0.21	.84	1.1	1.15
ICD 618	0.21	.84	1.1	1.15
ICD 278	0.55	.58	1.02	1.15
ICD 585	-0.15	.88	0.98	0.94
ATC C01BD	0.09	.93	0.98	1
ATC A07EA	0.49	.62	0.93	1.05
ICD 486	-0.15	.88	0.76	0.73
ICD 295	-0.15	.88	0.76	0.73
ICD 366	0.32	.75	0.72	0.79
ICD 331	0.32	.75	0.72	0.79
ICD 482	0.32	.75	0.72	0.79
ICD 813	0.56	.57	0.72	0.84
ICD V56	-0.34	.73	0.64	0.58
ICD V53	0.24	.81	0.64	0.68
ICD 717	0.24	.81	0.64	0.68
ICD 812	0.19	.85	0.59	0.63
ICD 424	0.14	.88	0.55	0.58
ICD 411	-0.17	.87	0.55	0.52
ICD 995	0.14	.88	0.55	0.58
ICD 413	-1.33	.18	0.55	0.37
ICD 250	0.14	.88	0.55	0.58
ICD 786	-0.61	.54	0.51	0.42
ICD 241	0.09	.93	0.51	0.52

Continued on next page

Table B.3: One Proportion z-test Results - Cohort 3 (Continued)

Features	z score	P value	Prevalence of the feature in raw data (%)	Prevalence of the feature in sampled data (%)
ICD 301	0.39	.7	0.51	0.58
ICD 414	0.09	.93	0.51	0.52
ICD 415	0.09	.93	0.51	0.52
ICD 434	-1.03	.3	0.51	0.37
ICD 728	-0.24	.81	0.51	0.47
ICD 550	0.03	.97	0.47	0.47
ICD 562	0.63	.53	0.47	0.58
ICD 437	0.63	.53	0.47	0.58
ICD 592	0.35	.73	0.47	0.52
ICD V57	0.03	.97	0.47	0.47
ICD V64	-0.32	.75	0.47	0.42
ICD 438	-0.03	.97	0.42	0.42
ICD V71	0.3	.76	0.42	0.47
ICD 571	-0.41	.68	0.42	0.37
ICD 726	0.3	.76	0.42	0.47
ICD 153	0.25	.8	0.38	0.42
ICD 211	0.25	.8	0.38	0.42
ICD 041	-1.02	.31	0.38	0.26
ICD 560	0.25	.8	0.38	0.42
ICD 173	0.25	.8	0.38	0.42
ICD 276	-0.11	.91	0.38	0.37
ICD V54	0.25	.8	0.38	0.42
ICD 599	-0.11	.91	0.38	0.37
ICD 780	-0.53	.6	0.38	0.31
ICD 218	-0.53	.6	0.38	0.31
ICD 478	-0.11	.91	0.38	0.37
ICD 188	-0.19	.85	0.34	0.31
ICD 584	0.54	.59	0.34	0.42
ICD 214	-0.3	.77	0.3	0.26
ICD 431	-0.83	.4	0.3	0.21
ICD 998	0.06	.95	0.25	0.26
ICD 455	-1.07	.28	0.25	0.16
ICD 553	0.06	.95	0.25	0.26
ICD 433	0.47	.64	0.25	0.31

Continued on next page

Table B.3: One Proportion z-test Results - Cohort 3 (Continued)

Features	z score	P value	Prevalence of the feature in raw data (%)	Prevalence of the feature in sampled data (%)
ICD 440	-1.07	.28	0.25	0.16
ICD 722	0.06	.95	0.25	0.26
ICD 298	0.43	.67	0.21	0.26
ICD 473	-0.02	.98	0.21	0.21
ICD 162	0.38	.7	0.17	0.21
ICD 470	0.38	.7	0.17	0.21
ICD 626	0.33	.74	0.13	0.16
ICD 354	-1.43	.15	0.13	0.05
ATC G04CA	0.27	.79	0.08	0.1
ATC G03AA	-0.62	.54	0.08	0.05
ATC G03DB	0.19	.85	0.04	0.05

No features are removed from the sampled data.

Table B.4: One Proportion z-test Results - Cohort 4

Features	z score	P value	Prevalence of the feature in raw data (%)	Prevalence of the feature in sampled data (%)
ATC A02BC	-0.31	.76	49.71	49.39
ATC J01CR	0.07	.95	48.39	48.46
Age >68	-1.22	.22	41.78	40.53
ATC J01MA	0.44	.66	40.08	40.53
ATC H02AB	-0.11	.91	36.99	36.89
ATC J01FA	0.58	.56	32.85	33.42
ATC C10AA	-0.42	.67	31.73	31.32
ATC B01AC	-0.16	.87	31.34	31.18
ATC C07AB	-0.91	.36	28.84	27.98
ATC M01AB	-0.11	.92	27.07	26.97
ATC C09AA	-0.54	.59	26.81	26.32
ATC G04CA	-0.82	.41	26.49	25.75
ATC M01AE	0.82	.41	24.82	25.57
ATC J01DD	-0.09	.93	24.25	24.17
ATC C08CA	-0.66	.51	21.71	21.14
ATC C09CA	0.01	.99	17.53	17.54

Continued on next page

Table B.4: One Proportion z-test Results - Cohort 4 (Continued)

Features	z score	P value	Prevalence of the feature in raw data (%)	Prevalence of the feature in sampled data (%)
ATC R03BA	0.67	.5	17.05	17.59
ATC C03CA	-0.51	.61	16.89	16.49
ATC A10BA	-0.49	.62	16.7	16.32
ATC M04AA	-0.22	.83	16.44	16.27
ATC A07AA	-0.17	.86	16.18	16.05
ATC J01CA	0.29	.77	14.26	14.47
ATC A11CC	-1.08	.28	13.97	13.2
ATC C09DA	-0.71	.48	13.62	13.11
ATC R03AK	0.26	.8	12.85	13.03
ATC N06AB	0.28	.78	12.04	12.24
ATC N02BE	-0.14	.89	11.14	11.05
ATC B01AB	1.21	.23	10.98	11.8
ATC M01AH	0.39	.69	10.53	10.79
ATC N03AX	-0.44	.66	10.28	10
ATC C09BA	-0.27	.78	10.08	9.91
ATC G04CB	-0.55	.58	9.99	9.65
ATC N02AA	-1.33	.18	9.92	9.12
ATC N02AX	-0.03	.97	9.54	9.52
ATC A02AD	0.1	.92	9.28	9.34
ATC B03BB	-0.44	.66	9.12	8.86
ATC N02BA	0.65	.52	8.99	9.39
ATC A02BX	-0.27	.79	8.8	8.64
ATC N06AX	0.38	.7	8.29	8.51
ATC C02CA	-0.57	.57	8.12	7.81
ATC M01AX	-0.07	.94	7.8	7.76
ATC B03AA	0.11	.91	6.87	6.93
ATC C10AX	-0.38	.7	6.78	6.58
ATC J01XX	-0.15	.88	6.74	6.67
ATC A10AB	-0.56	.57	6.65	6.36
ATC J02AC	0.29	.77	6.17	6.32
ATC C09BB	0.44	.66	6.01	6.23
ATC C03DA	-0.4	.69	5.94	5.75
ATC A10BB	0.07	.95	5.84	5.88
ATC R03AC	0.28	.78	5.78	5.92

Continued on next page

Table B.4: One Proportion z-test Results - Cohort 4 (Continued)

Features	z score	P value	Prevalence of the feature in raw data (%)	Prevalence of the feature in sampled data (%)
ATC R03BB	-1.14	.26	5.43	4.91
ATC J01EE	-0.6	.55	5.14	4.87
ATC M01AC	-1.03	.3	5.01	4.56
ATC H03AA	-0.63	.53	4.66	4.39
ATC J05AB	-0.38	.71	4.59	4.43
ATC C01DA	-1.16	.25	4.56	4.08
ATC S01ED	0.59	.56	4.43	4.69
ATC B01AA	0.54	.59	4.37	4.61
ATC R06AX	-0.43	.66	4.3	4.12
ATC D05AX	0.86	.39	4.27	4.65
ICD 550	-0.08	.93	3.98	3.95
ATC A07EA	-0.27	.79	3.79	3.68
ATC A02BA	-0.53	.6	3.76	3.55
ATC R06AE	-1.2	.23	3.6	3.16
ATC A07EC	0.25	.81	3.5	3.6
ATC C10BA	0.38	.7	3.4	3.55
ATC C09DB	-0.58	.56	3.37	3.16
ATC N03AG	-0.17	.87	3.31	3.25
ATC C10AB	-0.99	.32	3.24	2.89
ATC R03AL	0.05	.96	3.05	3.07
ATC A12BA	-0.11	.91	2.89	2.85
ICD 600	-0.28	.78	2.86	2.76
ICD 518	-0.56	.57	2.73	2.54
ATC C09BX	-0.89	.37	2.7	2.41
ATC C03EA	0.41	.68	2.67	2.81
ATC A12AX	0.09	.93	2.6	2.63
ICD 410	0.35	.73	2.6	2.72
ATC A05AA	0.31	.75	2.57	2.68
ATC S01EE	0.44	.66	2.57	2.72
ATC C03BA	0.02	.98	2.54	2.54
ATC B03BA	0.75	.45	2.5	2.76
ICD 715	0.69	.49	2.44	2.68
ATC C07AG	-0.27	.79	2.41	2.32
ICD V58	-0.31	.76	2.38	2.28

Continued on next page

Table B.4: One Proportion z-test Results - Cohort 4 (Continued)

Features	z score	P value	Prevalence of the feature in raw data (%)	Prevalence of the feature in sampled data (%)
ATC C03EB	-0.84	.4	2.31	2.06
ATC B02AA	0.54	.59	2.28	2.46
ICD 413	0.24	.81	2.25	2.32
ATC C07AA	0.61	.54	2.22	2.41
ATC C01BC	0.07	.95	2.22	2.24
ATC N06AA	0.03	.98	2.18	2.19
ATC C01BD	-0.01	.99	2.15	2.15
ICD 427	-0.82	.41	2.12	1.89
ATC N05AD	0.24	.81	2.12	2.19
ATC N03AE	0.48	.63	2.09	2.24
ATC J01DC	-0.17	.87	2.02	1.97
ICD 414	0.34	.73	1.96	2.06
ICD 038	-0.51	.61	1.89	1.75
ATC J01AA	-0.19	.85	1.89	1.84
ATC P01AB	0.12	.9	1.89	1.93
ATC A03FA	-0.39	.69	1.86	1.75
ICD 727	0.31	.76	1.8	1.89
ICD 428	-0.72	.47	1.77	1.58
ICD 574	0.53	.59	1.73	1.89
ATC N03AF	-0.96	.34	1.73	1.49
ATC C07BB	0.23	.82	1.73	1.8
ATC A12AA	-0.25	.8	1.73	1.67
ATC S01EC	0.19	.85	1.7	1.75
ICD 434	0.19	.85	1.7	1.75
ATC N02AJ	-1.35	.18	1.64	1.32
ATC C02AC	-0.69	.49	1.57	1.4
ATC N02AB	-0.2	.84	1.54	1.49
ICD 411	0.43	.67	1.51	1.62
ICD 996	-0.87	.38	1.48	1.27
ATC N05AA	-0.22	.83	1.41	1.36
ICD 188	-0.09	.93	1.38	1.36
ATC R03DA	-0.66	.51	1.38	1.23
ICD V43	-0.09	.93	1.38	1.36
ATC B05BB	-0.33	.74	1.35	1.27

Continued on next page

Table B.4: One Proportion z-test Results - Cohort 4 (Continued)

Features	z score	P value	Prevalence of the feature in raw data (%)	Prevalence of the feature in sampled data (%)
ATC N03AA	-0.86	.39	1.28	1.1
ICD 440	0.27	.79	1.25	1.32
ATC N05AH	-0.78	.43	1.22	1.05
ATC N04AA	-0.86	.39	1.19	1.01
ICD 185	-1.53	.13	1.12	0.83
ATC L01BA	0.77	.44	1.09	1.27
ATC N02CC	0.22	.83	1.09	1.14
ICD V56	-0.53	.59	1.03	0.92
ICD 486	-0.61	.54	1	0.88
ICD 585	-0.85	.39	1	0.83
ICD 438	0.83	.4	1	1.18
ICD V57	-0.44	.66	0.96	0.88
ATC C03AA	0.22	.83	0.96	1.01
ICD 717	-0.11	.91	0.9	0.88
ICD 295	-0.35	.73	0.9	0.83
ICD 338	-0.42	.68	0.87	0.79
ICD 482	0.48	.63	0.87	0.96
ATC N01BB	-1.24	.22	0.87	0.66
ICD V53	0.22	.83	0.83	0.88
ATC R03DC	-0.58	.56	0.8	0.7
ICD 820	0.16	.87	0.8	0.83
ICD 296	-0.14	.89	0.77	0.75
ICD 780	0.33	.74	0.77	0.83
ICD 433	0.71	.48	0.74	0.88
ICD 431	-0.21	.83	0.74	0.7
ICD 331	0.04	.97	0.74	0.75
ICD 250	-1.19	.23	0.71	0.53
ICD 584	-0.37	.71	0.67	0.61
ICD 162	-1.32	.19	0.67	0.48
ICD 153	0.16	.88	0.67	0.7
ICD 276	-0.1	.92	0.67	0.66
ICD 599	-0.17	.86	0.64	0.61
ICD 592	-0.77	.44	0.64	0.53
ATC P01BA	0.34	.73	0.64	0.7

Continued on next page

Table B.4: One Proportion z-test Results - Cohort 4 (Continued)

Features	z score	P value	Prevalence of the feature in raw data (%)	Prevalence of the feature in sampled data (%)
ICD 553	0.34	.73	0.64	0.7
ICD 728	-0.46	.65	0.64	0.57
ICD 722	0.34	.73	0.64	0.7
ICD 454	0.28	.78	0.61	0.66
ICD 571	0.75	.45	0.61	0.75
ICD 366	0.02	.98	0.61	0.61
ICD V64	0.28	.78	0.61	0.66
ICD 473	0.15	.88	0.55	0.57
ICD 726	-1.32	.19	0.51	0.35
ICD 786	0.36	.72	0.51	0.57
ICD V71	0.29	.77	0.48	0.53
ICD 455	0.56	.57	0.48	0.57
ICD 424	0.76	.44	0.45	0.57
ICD 278	-0.8	.43	0.45	0.35
ICD 560	-0.08	.94	0.45	0.44
ICD 995	0.15	.88	0.42	0.44
ICD 173	0.45	.65	0.42	0.48
ICD 437	0.15	.88	0.42	0.44
ICD 562	0.38	.7	0.39	0.44
ICD 041	0.07	.94	0.39	0.39
ATC M05BA	0.32	.75	0.35	0.39
ICD 211	-0.02	.98	0.35	0.35
ICD 298	-0.02	.98	0.35	0.35
ICD 354	-0.4	.69	0.35	0.31
ICD 998	-0.12	.9	0.32	0.31
ICD 478	0.24	.81	0.32	0.35
ICD 415	0.81	.42	0.29	0.39
ICD 214	0.5	.62	0.29	0.35
ICD 812	0.43	.67	0.26	0.31
ICD 813	-0.56	.57	0.22	0.18
ICD 735	-0.06	.96	0.22	0.22
ICD 470	-0.06	.96	0.22	0.22
ICD V54	-0.2	.84	0.19	0.18
ICD 301	0.17	.87	0.16	0.18

Continued on next page

Table B.4: One Proportion z-test Results - Cohort 4 (Continued)

Features	**z score**	**P value**	**Prevalence of the feature in raw data (%)**	**Prevalence of the feature in sampled data (%)**
ICD 241	-0.14	.89	0.1	0.09
ICD 174	0.27	.79	0.03	0.04

No features are removed from the sampled data.

Appendix C

Outcome Association of the Feature and the Support

Sup[1]: Support - the Occurrence of the Feature in the Evolutionarily Obtained Final Bin Dataset.

Prev[2]: Prevalance - the Occurrence of the Feature in the Dataset.

Table C.1: Outcome Association of the Feature and the Support - Cohort 1

Features	P value	**Sup[1]**	**Prev[2]**	Features	P value	Sup	Prev
ATC R03BA	<.001	0.85	15.5	**ATC C10BA**	.003	0.28	0.73
Age >53	<.001	0.84	41.15	**ATC R03AC**	<.001	0.28	4.46
ATC N03AX	<.001	0.82	5.6	**ATC N05AH**	.01	0.28	0.49
ATC R06AX	<.001	0.79	6.74	**ICD 454**	.23	0.27	1.38
ATC J01XX	<.001	0.78	14.2	**ATC R03AL**	<.001	0.26	1.7
ATC C03CA	<.001	0.76	5.19	**ICD 996**	.56	0.26	0.24
ATC N02AX	<.001	0.74	6.9	**ATC S01EC**	.71	0.26	0.57
ATC A11CC	<.001	0.73	23.05	**ICD 455**	1.0	0.25	0.32
ATC C09CA	<.001	0.69	5.44	**ATC C01DA**	.16	0.24	0.16
ICD 298	.16	0.68	0.16	**ICD 820**	.16	0.24	0.16
ATC J01CA	<.001	0.66	14.12	**ICD 438**	.32	0.24	0.08
ICD 411	.32	0.62	0.08	**ATC C03EB**	.65	0.24	0.41
ATC J01EE	.03	0.61	2.44	**ICD 211**	.16	0.24	0.16

Features	P value	**Sup[1]**	**Prev[2]**
ATC A02BX	<.001	0.57	7.47
ATC C08CA	<.001	0.57	5.84
ICD 550	.16	0.56	0.16
ATC B01AC	<.001	0.56	4.14
ATC D05AX	.13	0.55	2.27
ATC C07BB	.17	0.54	1.06
ATC A07EC	.003	0.54	2.03
ICD 618	.18	0.52	0.41
ICD 592	.41	0.51	0.49
ATC C07AA	<.001	0.5	1.7
ATC R06AE	<.001	0.5	4.22
ICD V54	.10	0.5	0.49
ATC B03BA	.05	0.5	1.3
ATC P01AB	<.001	0.49	2.76
ATC N03AE	<.001	0.48	0.89
ATC N06AB	<.001	0.48	13.07
ICD 278	.007	0.48	0.89
ICD 998	.08	0.47	0.24
ICD 574	.005	0.47	1.46
ATC B05BB	.08	0.47	0.24
ATC C09DA	<.001	0.46	3.57
ICD 038	.005	0.46	0.65
ICD V53	.08	0.46	0.24
ATC A03FA	.003	0.46	1.79
ATC C03DA	<.001	0.46	0.89
ICD 301	.05	0.44	0.32
ATC J01DC	.13	0.44	2.27
ATC J01AA	.09	0.44	2.35
ATC N04AA	.65	0.42	0.41
ATC N02AB	.002	0.42	0.81
ICD 482	.08	0.42	0.24
ICD 434	.32	0.42	0.08
ATC M05BA	1.0	0.42	0.65
ATC P01BA	.06	0.42	0.81
ICD 188	.32	0.4	0.08
ATC C09BB	<.001	0.4	1.54
ICD 218	.81	0.4	1.38
ICD 250	.16	0.4	0.16
ATC N05AD	.32	0.4	0.73
ICD 473	.41	0.4	0.49

Features	P value	Sup	Prev
ATC R03BB	.005	0.24	0.65
ICD 727	.39	0.24	1.79
ATC C09AA	<.001	0.24	7.71
ICD V43	.32	0.24	0.08
ATC N03AG	.003	0.23	1.54
ATC A12BA	<.001	0.23	1.06
ICD 162	.32	0.23	0.08
ICD V64	.71	0.23	0.57
ATC A07EA	.13	0.22	0.89
ICD 241	.16	0.22	0.16
ATC R03DA	.03	0.22	0.41
ATC N01BB	.37	0.22	0.89
ICD 786	.32	0.22	0.08
ATC A10BA	<.001	0.21	3.25
ATC B01AB	<.001	0.21	5.28
ICD 437	.32	0.2	0.08
ICD 995	1.0	0.2	0.16
ICD 366	.32	0.2	0.08
ICD 410	.16	0.2	0.16
ICD V58	<.001	0.2	1.62
ICD 585	.03	0.2	0.41
ICD 478	.56	0.2	0.24
ICD 626	1.0	0.2	0.49
ATC G03DB	.75	0.2	3.25
ATC C01BC	.56	0.19	0.24
ICD 518	.01	0.19	0.49
ATC C02CA	<.001	0.18	1.3
ATC B01AA	.005	0.18	0.65
ATC G03CA	.17	0.18	2.76
ICD 174	.17	0.17	1.06
ICD 354	.16	0.17	0.16
ATC N06AX	<.001	0.16	6.33
ATC N02CC	.03	0.16	5.28
ATC C10AX	.03	0.15	0.89
ATC S01EE	.32	0.14	0.32
ICD 424	.08	0.14	0.24
ATC B03BB	<.001	0.14	6.17
ATC C03AA	.10	0.14	0.49
ATC A05AA	.02	0.13	1.87
ATC C07AB	<.001	0.13	9.66

Features	P value	**Sup[1]**	**Prev[2]**
ICD 571	.32	0.4	0.08
ATC N05AA	.05	0.38	0.32
ICD 812	.10	0.38	0.49
ICD 153	.32	0.38	0.08
ICD 553	.18	0.38	0.41
ICD 470	.56	0.38	0.24
ATC C01BD	.08	0.38	0.24
ATC A02BA	<.001	0.37	3.25
ATC A10AB	.005	0.37	0.65
ATC J05AB	.78	0.36	4.06
ATC S01ED	.02	0.36	0.97
ATC N02AJ	.01	0.36	0.49
ICD 780	.05	0.36	0.32
ATC N02BA	.001	0.36	1.14
ATC C03EA	<.001	0.36	2.35
ATC N03AF	.06	0.35	0.81
ATC L01BA	<.001	0.35	0.97
ATC C03BA	<.001	0.34	1.46
ICD 427	.56	0.34	0.24
ATC C10AB	.02	0.34	0.73
ATC C07AG	.008	0.34	0.57
ICD V71	.32	0.34	0.08
ATC N06AA	.005	0.34	2.35
ICD 296	.003	0.34	0.73
ICD 562	1.0	0.32	0.16
ICD 599	.16	0.32	0.16
ICD 717	.76	0.32	0.89
ICD 041	1.0	0.32	0.16
ATC C09BX	.10	0.32	0.49
ICD 560	.32	0.32	0.08
ICD 295	.56	0.31	0.24
ATC A12AX	<.001	0.3	4.3
ATC A10BB	.003	0.3	0.73
ATC C02AC	.03	0.3	0.41
ICD 428	.32	0.3	0.08
ICD 440	.32	0.3	0.08
ICD 735	.78	0.3	1.06
ICD 722	.08	0.3	0.24
ICD V56	.03	0.3	0.41
ATC M04AA	<.001	0.3	1.7
ICD V57	.08	0.3	0.24
ICD 726	1.0	0.29	0.65
ATC B03AA	.13	0.13	11.69
ATC B02AA	.03	0.12	3.57
ATC A12AA	<.001	0.12	1.22
ATC C09BA	<.001	0.12	3.25
ATC G03AA	.03	0.12	2.44
ICD 338	.03	0.12	0.65
ICD 728	.32	0.12	0.08
ATC M01AH	<.001	0.11	9.01
ATC J02AC	<.001	0.1	10.71
ATC J01MA	<.001	0.1	27.03
ATC M01AB	<.001	0.1	23.94
ICD 486	.16	0.09	0.16
ATC N02AA	<.001	0.09	6.9
ATC H03AA	<.001	0.09	10.8
ATC A07AA	<.001	0.08	9.74
ATC C10AA	<.001	0.08	4.95
ICD 621	.12	0.08	2.68
ATC N02BE	<.001	0.07	9.5
ATC M01AX	.009	0.06	3.9
ATC J01FA	<.001	0.06	35.06
ATC H02AB	<.001	0.06	41.31
ATC R03AK	<.001	0.05	8.52
ATC J01CR	<.001	0.04	44.48

Table C.2: Outcome Association of the Feature and the Support - Cohort 2

Features	P value	Sup	Prev	Features	P value	Sup	Prev
ATC A10BA	<.001	0.86	4.31	**ATC N02AJ**	.03	0.38	0.5
ATC N02BE	<.001	0.79	6.4	**ATC R03BB**	<.001	0.38	0.86
ATC J05AB	<.001	0.76	2.91	**ATC C09DB**	.33	0.38	1.73
ATC C03CA	<.001	0.76	4.09	**ATC H03AA**	.14	0.38	2.5
ATC M04AA	<.001	0.74	5.13	**ICD V71**	.65	0.38	0.23
ATC C09CA	<.001	0.71	8.4	**ICD 366**	.56	0.37	0.14
ATC C02CA	<.001	0.65	3.22	**ICD 295**	.32	0.37	0.41
ATC C08CA	<.001	0.65	7.4	**ATC A10BB**	<.001	0.37	1.27
ICD V64	1.0	0.64	0.18	**ICD V58**	<.001	0.37	1.04
ICD V54	.26	0.64	0.32	**ATC C10AB**	.07	0.36	2
ATC J02AC	.03	0.64	6.18	**ICD 599**	.16	0.35	0.09
ICD 188	.32	0.63	0.18	**ICD 998**	.18	0.35	0.23
ATC N06AB	.03	0.63	8.58	**ATC N05AA**	.65	0.34	0.91
ATC S01EE	.07	0.62	0.68	**ATC R03DC**	.002	0.34	0.77
ATC N03AG	.08	0.61	2.5	**ATC C10BA**	.44	0.34	0.68
ATC M01AB	.001	0.6	18.21	**ICD 354**	.56	0.34	0.14
ICD 735	1.0	0.6	0.18	**ATC S01EC**	.06	0.33	0.45
ICD 454	.83	0.6	1.04	**ATC J01XX**	<.001	0.33	4.13
ATC N03AE	.17	0.6	1.54	**ICD 553**	<.001	0.33	0.82
ICD 820	.32	0.6	0.05	**ICD 331**	.32	0.32	0.05
ATC B01AA	.01	0.59	0.86	**ATC A12AA**	<.001	0.32	0.91
ICD 211	.65	0.56	0.23	**ICD 786**	.10	0.32	0.27
ATC P01AB	<.001	0.56	1.54	**ICD 486**	.10	0.32	0.27
ICD 574	.65	0.56	0.91	**ATC N01BB**	.13	0.32	0.5
ATC A07EC	.003	0.56	2.54	**ATC N02AB**	.01	0.32	0.45
ATC C09BX	.05	0.56	0.95	**ICD 241**	.65	0.31	0.23
ICD 482	.005	0.55	0.36	**ICD 338**	.10	0.3	0.27
ATC B03AA	<.001	0.55	2.77	**ATC C09AA**	<.001	0.3	11.99
ICD V56	.01	0.55	0.27	**ICD 437**	.08	0.3	0.14
ATC M01AC	.01	0.55	2.45	**ICD 434**	.01	0.3	0.27
ICD 550	1.0	0.55	2.27	**ICD 153**	.16	0.3	0.09
ATC G04CB	<.001	0.54	1	**ATC M01AX**	.06	0.3	3.27
ICD 427	.71	0.54	0.32	**ATC N05AD**	.24	0.3	1.18
ICD 571	.10	0.53	0.27	**ATC R03DA**	.37	0.29	0.5
ICD V53	.25	0.52	0.54	**ICD 455**	.56	0.29	0.54
ICD 428	.008	0.52	0.32	**ICD 301**	.71	0.29	0.32
ICD 813	.08	0.52	0.14	**ICD 041**	.56	0.29	0.14

Features	P value	Sup	Prev	Features	P value	Sup	Prev
ICD 276	.32	0.51	0.05	**ICD 431**	.10	0.29	0.27
ATC S01ED	.005	0.51	1.32	**ATC D05AX**	.89	0.29	2.41
ATC N02AX	.003	0.51	4.31	**ATC A02AD**	<.001	0.28	6.63
ATC R03AL	<.001	0.5	1.18	**ICD 411**	.08	0.28	0.14
ATC N05AH	.43	0.5	1.18	**ATC C07AA**	.06	0.28	0.45
ICD 278	1.0	0.5	0.45	**ATC B05BB**	.76	0.28	0.5
ICD 996	.02	0.5	0.41	**ATC N06AA**	.3	0.26	1.04
ICD 440	.56	0.5	0.14	**ATC A10AB**	<.001	0.26	1.77
ICD 727	.2	0.5	1	**ATC A02BA**	.06	0.26	1.59
ICD 410	<.001	0.48	0.64	**ATC C09DA**	<.001	0.26	5.31
ATC N02CC	.88	0.48	1.86	**ICD 717**	.59	0.26	1.41
ATC C07AG	.008	0.48	0.77	**ATC A02BX**	.003	0.25	5.9
ICD 995	.41	0.48	0.27	**ATC C09BA**	.38	0.25	3.81
ICD 518	.01	0.47	0.59	**ATC B01AB**	<.001	0.24	5.4
ATC J01EE	<.001	0.47	2.72	**ICD 424**	.05	0.24	0.18
ATC B02AA	.03	0.46	0.77	**ICD 562**	.05	0.24	0.18
ATC C07BB	.44	0.46	0.68	**ICD 722**	.48	0.24	0.36
ICD 812	.32	0.46	0.18	**ATC A12BA**	.003	0.24	0.86
ICD V57	.56	0.46	0.14	**ATC C10AX**	.003	0.24	2.54
ICD 038	.02	0.45	0.54	**ATC N02AA**	.03	0.24	4.63
ICD 600	.76	0.45	0.5	**ATC N06AX**	.83	0.22	4.09
ATC C03AA	.03	0.45	0.36	**ICD 298**	.53	0.22	0.45
ATC N04AA	.3	0.44	1.04	**ATC C03EA**	.01	0.22	0.73
ATC B03BB	.007	0.44	3.95	**ICD 592**	.83	0.22	0.95
ATC C01DA	<.001	0.44	0.82	**ICD 473**	.56	0.21	0.54
ATC C03BA	.81	0.44	0.77	**ICD 173**	.18	0.21	0.23
ATC C01BC	.71	0.44	0.32	**ATC C01BD**	.05	0.2	0.18
ICD 478	.65	0.43	0.23	**ATC A07EA**	.88	0.2	1.86
ATC C09BB	.23	0.43	3.09	**ICD 728**	.41	0.2	0.27
ATC M05BA	.16	0.43	0.09	**ICD 438**	.18	0.2	0.23
ICD V43	1.0	0.43	0.36	**ICD 185**	1.0	0.2	0.09
ICD 470	.03	0.42	0.5	**ATC A07AA**	<.001	0.2	9.76
ATC A12AX	.53	0.42	1.04	**ATC J01AA**	.34	0.19	1.23
ICD 250	.32	0.42	0.18	**ATC N03AF**	.69	0.18	1.18
ICD 715	.35	0.42	0.82	**ATC R03AC**	.005	0.16	4.13
ICD 415	.56	0.42	0.14	**ATC A03FA**	.2	0.14	1
ICD 780	.65	0.42	0.23	**ATC J01FA**	.01	0.14	28.7
ATC L01BA	.53	0.42	0.45	**ATC J01CA**	.007	0.12	11.49
ATC J01DC	.68	0.42	1.09	**ATC R03BA**	.003	0.11	12.99
ICD 726	.71	0.41	0.32	**ATC R06AE**	.52	0.11	2.77

Features	P value	Sup	Prev	Features	P value	Sup	Prev
ICD 414	.01	0.41	0.27	**Age >53**	<.001	0.1	44.69
ATC P01BA	.08	0.4	0.14	**ATC A11CC**	<.001	0.1	6.63
ATC G03DB	.32	0.4	0.05	**ATC R06AX**	.02	0.1	4.86
ATC C03DA	<.001	0.4	1.27	**ATC J01CR**	.001	0.1	40.33
ATC H02AB	<.001	0.4	33.51	**ATC M01AH**	.001	0.09	4.59
ICD 585	<.001	0.4	0.5	**ATC M01AE**	<.001	0.09	15.62
ATC C03EB	.004	0.4	0.54	**ATC N03AX**	<.001	0.07	5
ATC A05AA	1.0	0.4	1.27	**ATC R03AK**	<.001	0.06	7.72
ATC N03AA	.11	0.4	0.64	**ATC J01DD**	<.001	0.06	15.85
ATC N02BA	<.001	0.4	1.68	**ATC G04CA**	<.001	0.06	6.95
ICD 162	.32	0.39	0.05				
ICD 296	.18	0.39	0.23				
ICD 584	.56	0.39	0.14				
ICD 413	<.001	0.39	0.54				
ATC C02AC	.002	0.39	0.59				

Table C.3: Outcome Association of the Feature and the Support - Cohort 3

Features	P value	Sup	Prev	Features	P value	Sup	Prev
ATC N02AX	<.001	0.84	12.96	**ATC A11CC**	.12	0.29	45.33
ATC M04AA	<.001	0.82	8.5	**ATC C09AA**	.11	0.28	19.57
ATC C03EA	<.001	0.76	5.35	**ATC M01AH**	.05	0.28	18.31
ATC A02BA	.004	0.75	4.04	**ICD 410**	<.001	0.28	1.26
ATC B01AB	<.001	0.73	12.59	**ICD 998**	.18	0.28	0.26
ATC N03AX	<.001	0.7	11.7	**ATC G03DB**	.32	0.28	0.05
ICD 295	.03	0.68	0.73	**ATC S01ED**	.82	0.27	3.99
ICD 813	.62	0.68	0.84	**ATC J02AC**	.002	0.27	8.5
ATC N02AA	<.001	0.68	13.9	**ICD V58**	.007	0.26	2.57
ATC J05AB	.008	0.65	5.77	**ATC N01BB**	.02	0.26	1.36
ATC A12AA	.11	0.62	4.67	**ICD 560**	.16	0.26	0.42
ATC C07BB	.16	0.62	2.62	**ATC N03AG**	1.0	0.26	3.46
ATC B03BB	<.001	0.61	9.23	**ATC C10AB**	.13	0.26	1.84
ATC R03AC	.005	0.6	7.19	**ATC S01EE**	.09	0.26	1.78
ICD 427	.03	0.58	1.15	**ATC C03DA**	<.001	0.26	4.62
ICD 413	.71	0.56	0.37	**ICD 995**	.03	0.26	0.58
ICD V53	.002	0.56	0.68	**ATC C10BA**	.002	0.25	2.47

Features	P value	Sup	Prev	Features	P value	Sup	Prev
ATC A12BA	.008	0.56	3.93	**ICD V56**	.13	0.25	0.58
ICD 518	.002	0.56	2.36	**ATC M05BA**	.07	0.24	5.35
ICD 574	.24	0.55	1.84	**ICD 278**	.09	0.24	1.15
ATC J01DC	.07	0.55	2.68	**ICD 250**	.03	0.24	0.58
ATC N06AA	.14	0.55	4.04	**ICD 218**	.10	0.24	0.31
ATC A05AA	.008	0.54	3.57	**ATC J01EE**	<.001	0.24	5.09
ATC G03AA	.32	0.53	0.05	**ICD 431**	1.0	0.23	0.21
ATC C01DA	<.001	0.52	2.57	**ATC A07EA**	.03	0.23	1.05
ICD 618	1.0	0.5	1.15	**ATC N06AX**	.005	0.23	13.12
ICD 727	.74	0.5	1.89	**ATC N02CC**	.17	0.22	2.73
ICD 434	.71	0.5	0.37	**ATC C09DB**	.09	0.22	2.26
ICD 553	.65	0.5	0.26	**ATC C03BA**	.17	0.22	3.31
ATC G04CA	1.0	0.49	0.1	**ICD 996**	.10	0.21	1.26
ICD 354	.32	0.48	0.05	**ICD 162**	1.0	0.21	0.21
ICD 241	.06	0.48	0.52	**ICD 041**	.18	0.2	0.26
ICD 038	<.001	0.48	1.84	**ATC C02AC**	<.001	0.2	1.26
ICD 455	.56	0.47	0.16	**ICD 786**	1.0	0.2	0.42
ATC N02AJ	.005	0.47	2.62	**ATC M01AB**	.10	0.2	33.11
ATC R06AX	.14	0.47	5.51	**ICD 584**	.03	0.2	0.42
ICD 478	.71	0.46	0.37	**ICD 440**	.08	0.19	0.16
ICD 599	.71	0.46	0.37	**ICD 473**	.32	0.19	0.21
ICD 433	.41	0.46	0.31	**ATC H03AA**	.26	0.19	16.58
ATC M01AE	.009	0.46	31.79	**ICD 735**	.55	0.18	1.31
ICD 728	.10	0.44	0.47	**ICD 415**	.53	0.18	0.52
ICD 717	.41	0.44	0.68	**ATC C03AA**	.01	0.18	1.68
ICD 437	.13	0.43	0.58	**ICD 411**	.01	0.18	0.52
ATC D05AX	.32	0.42	3.36	**ATC R03BA**	.06	0.18	22.4
ATC R03BB	<.001	0.42	5.09	**ATC C02CA**	<.001	0.18	5.61
ICD 188	.41	0.42	0.31	**ATC C09BB**	.23	0.17	4.46
ATC B05BB	.72	0.42	1.68	**ICD 592**	.53	0.16	0.52
ICD 482	.2	0.42	0.79	**ATC C08CA**	<.001	0.16	16.16
ICD 331	.8	0.42	0.79	**ICD V64**	.16	0.16	0.42
ATC B02AA	.19	0.41	1.52	**ICD 715**	.17	0.16	3.31
ATC R03DA	.007	0.41	1.63	**ATC C09BX**	.09	0.16	1.15
ATC A07EC	.004	0.4	4.56	**ATC J01CR**	.01	0.16	54.2
ATC C09DA	<.001	0.4	14.9	**ATC N04AA**	.86	0.16	1.63
ATC N02BE	<.001	0.4	17.26	**ATC C03EB**	<.001	0.16	3.36
ATC N06AB	.002	0.39	22.3	**ATC R06AE**	.01	0.16	5.93
ICD 296	.84	0.39	1.31	**ATC M01AX**	.04	0.16	8.5
ATC N03AA	1.0	0.39	1.57	**ATC N05AD**	.5	0.16	2.89

Features	P value	Sup	Prev
ICD 153	.48	0.39	0.42
ICD 550	.74	0.39	0.47
ICD 438	1.0	0.39	0.42
ICD 470	.32	0.39	0.21
ICD 428	.02	0.39	1.21
ICD 214	.65	0.38	0.26
ATC B03BA	<.001	0.38	2.73
ATC C07AG	<.001	0.38	1.47
ATC J01AA	.03	0.38	1.68
ATC C01BC	.55	0.38	1.31
ICD V43	.14	0.37	1.94
ATC N03AF	.26	0.37	1.47
ATC S01EC	.61	0.37	1.84
ATC M01AC	.3	0.37	7.14
ATC N05AH	.86	0.36	1.63
ICD 812	.25	0.36	0.63
ATC P01AB	.29	0.36	2.26
ICD 780	.10	0.36	0.31
ICD 298	.18	0.36	0.26
ICD 486	.03	0.35	0.73
ICD V71	.10	0.35	0.47
ATC R03DC	.07	0.34	1.31
ATC C07AA	1.0	0.34	2.2
ICD 626	.56	0.34	0.16
ATC N03AE	.54	0.34	2.2
ICD 820	.06	0.33	1.21
ICD V57	.74	0.32	0.47
ICD 424	.13	0.32	0.58
ICD 174	.41	0.32	1.26
ATC A12AX	.01	0.32	12.12
ICD V54	.16	0.32	0.42
ICD 338	.13	0.32	1.47
ICD 366	.2	0.32	0.79
ATC G03CA	.77	0.32	5.3
ICD 621	.26	0.3	1.47
ATC N02AB	.03	0.3	2.47
ATC C01BD	.25	0.3	1
ICD 571	.06	0.3	0.37
ICD 211	.03	0.3	0.42
ICD 414	.06	0.29	0.52
ICD 722	.18	0.15	0.26
ATC J01DD	.02	0.14	29.12
ATC C09BA	.16	0.14	8.5
ATC C09CA	.005	0.14	15.42
ATC B01AC	<.001	0.14	21.04
ICD 562	.76	0.14	0.58
ICD 726	.74	0.14	0.47
ATC L01BA	.13	0.14	1.84
ICD 173	.16	0.13	0.42
ATC A03FA	.31	0.12	3.25
ATC N05AA	.86	0.12	1.63
ATC A10BB	<.001	0.12	3.78
ATC B01AA	<.001	0.12	3.52
ATC J01CA	.003	0.12	17
ATC C10AX	.001	0.1	3.99
ATC A02AD	.09	0.1	15.06
ATC A02BC	<.001	0.1	49.32
ATC A10AB	<.001	0.1	4.25
ATC R03AL	.04	0.1	2.89
ICD 276	.26	0.09	0.37
ATC C07AB	<.001	0.09	24.76
ATC C03CA	<.001	0.08	19.99
ATC A02BX	.003	0.08	12.01
ATC J01FA	.02	0.08	38.25
ICD 454	1.0	0.08	1.15
ATC A10BA	<.001	0.07	11.39
ATC R03AK	<.001	0.07	15.22
ATC A07AA	<.001	0.06	17.68
Age >68	<.001	0.06	38.61
ATC B03AA	.002	0.06	7.5
ATC J01XX	.002	0.06	21.51
ATC N02BA	.006	0.06	5.93
ATC C10AA	<.001	0.05	26.29
ATC J01MA	<.001	0.04	41.19
ATC H02AB	.002	0.03	47.59

Table C.4: Outcome Association of the Feature and the Support - Cohort 4

Features	P value	Sup	Prev	Features	P value	Sup	Prev
ATC G04CA	.02	0.8	25.75	**ATC A12AX**	.8	0.32	2.63
ATC J01CA	.008	0.73	14.47	**ATC C01DA**	<.001	0.32	4.08
ATC C09DA	.07	0.66	13.11	**ATC N02CC**	.69	0.32	1.14
ATC C09AA	.03	0.66	26.32	**ATC N03AA**	<.001	0.32	1.1
ATC B01AA	.001	0.64	4.61	**ICD 996**	.19	0.32	1.27
ATC C03CA	.002	0.62	16.49	**ICD V53**	1.0	0.32	0.88
ICD 995	1.0	0.61	0.44	**ICD 338**	.64	0.32	0.79
ATC N04AA	<.001	0.59	1.01	**ATC C10BA**	.44	0.31	3.55
ICD 153	.13	0.57	0.7	**ICD 174**	.32	0.31	0.04
ATC J02AC	.02	0.57	6.32	**ICD 427**	.88	0.31	1.89
ICD 437	.53	0.56	0.44	**ICD 301**	.32	0.3	0.18
ATC C09BX	.22	0.56	2.41	**ATC A02BX**	.18	0.3	8.64
ICD 250	1.0	0.55	0.53	**ICD 331**	.002	0.3	0.75
ICD 298	.16	0.55	0.35	**ICD 727**	.29	0.3	1.89
ATC N05AH	.004	0.55	1.05	**ATC J01FA**	.15	0.29	33.42
ATC A10BA	<.001	0.54	16.32	**ATC C03EA**	.32	0.28	2.81
ICD 415	.32	0.52	0.39	**ATC A10AB**	.08	0.28	6.36
ATC C07AG	.02	0.52	2.32	**ATC N03AE**	.48	0.28	2.24
ICD 786	.41	0.5	0.57	**ICD 998**	.71	0.28	0.31
ICD 424	.17	0.5	0.57	**ICD 562**	.53	0.28	0.44
ATC J05AB	.007	0.5	4.43	**ATC C09DB**	.35	0.28	3.16
ATC C03EB	.47	0.5	2.06	**ATC R03BB**	.19	0.28	4.91
ICD 780	.25	0.5	0.83	**ICD 553**	.62	0.28	0.7
ICD 592	.08	0.5	0.53	**ICD 414**	.03	0.28	2.06
ICD 185	.82	0.49	0.83	**ATC B03BB**	.89	0.28	8.86
ICD 722	.13	0.48	0.7	**ATC B01AB**	.36	0.28	11.8
ATC C03DA	.001	0.48	5.75	**ICD 599**	.29	0.28	0.61
ATC H03AA	.55	0.48	4.39	**ATC A07AA**	.14	0.28	16.05
ICD 428	.10	0.48	1.58	**ATC N03AX**	.09	0.27	10
ICD 473	.41	0.47	0.57	**ATC A10BB**	<.001	0.27	5.88
ATC N05AA	<.001	0.47	1.36	**ICD 214**	1.0	0.27	0.35
ATC A03FA	1.0	0.47	1.75	**ATC A07EA**	.38	0.26	3.68
ICD 518	.79	0.46	2.54	**ICD 820**	.11	0.26	0.83
ATC S01EE	.2	0.46	2.72	**ICD 211**	1.0	0.26	0.35
ICD 486	1.0	0.46	0.88	**ATC N02BA**	.41	0.26	9.39
ICD 366	.11	0.46	0.61	**ATC R03AL**	.09	0.26	3.07
ATC S01EC	.75	0.45	1.75	**ICD V54**	.05	0.26	0.18

Features	P value	Sup	Prev	Features	P value	Sup	Prev
ATC A05AA	.16	0.45	2.68	**ATC N03AG**	.005	0.26	3.25
ICD 585	.82	0.45	0.83	**ATC N02AB**	.49	0.24	1.49
ATC C09CA	.009	0.45	17.54	**ICD 717**	.65	0.24	0.88
ICD V71	.004	0.45	0.53	**ICD V57**	1.0	0.24	0.88
ATC D05AX	.03	0.44	4.65	**ATC C07AA**	.69	0.24	2.41
ATC M01AX	.15	0.44	7.76	**ATC R06AX**	.68	0.24	4.12
ATC B02AA	.29	0.44	2.46	**ATC N01BB**	.44	0.24	0.66
ATC R03DC	.62	0.44	0.7	**ATC N02AX**	.07	0.24	9.52
ATC A12AA	.02	0.43	1.67	**ATC M05BA**	.32	0.24	0.39
ICD 413	.04	0.42	2.32	**ATC P01AB**	.76	0.24	1.93
ATC G04CB	.18	0.42	9.65	**ICD 434**	.53	0.22	1.75
ATC L01BA	.005	0.42	1.27	**ICD 241**	.16	0.22	0.09
ICD 454	.8	0.42	0.66	**ATC N02BE**	.004	0.22	11.05
ICD 296	.23	0.41	0.75	**ICD 438**	.18	0.22	1.18
ICD 162	.13	0.41	0.48	**ATC M01AC**	.02	0.22	4.56
ATC C03BA	.12	0.41	2.54	**ATC N03AF**	.09	0.22	1.49
ICD 574	.88	0.41	1.89	**ATC N06AA**	1.0	0.21	2.19
ATC N02AA	.006	0.4	9.12	**ATC J01DC**	.10	0.21	1.97
ATC A07EC	.27	0.4	3.6	**ICD V58**	.58	0.2	2.28
ICD 715	.37	0.4	2.68	**ATC C01BD**	.003	0.2	2.15
ICD 726	.48	0.4	0.35	**ATC B03BA**	.38	0.2	2.76
ICD 295	.11	0.4	0.83	**ICD 173**	.76	0.2	0.48
ATC J01EE	.11	0.4	4.87	**ATC N06AX**	.01	0.19	8.51
ATC P01BA	.05	0.4	0.7	**ATC B05BB**	.19	0.19	1.27
ICD 431	.62	0.4	0.7	**ICD 278**	1.0	0.18	0.35
ATC C02CA	.45	0.4	7.81	**ATC A02AD**	.15	0.18	9.34
ATC R03AK	.15	0.4	13.03	**ATC C07AB**	.08	0.18	27.98
ICD 188	.86	0.4	1.36	**ICD 410**	.04	0.18	2.72
ICD 038	.11	0.39	1.75	**ATC A12BA**	.39	0.17	2.85
ICD 550	.4	0.38	3.95	**ATC R03AC**	.10	0.16	5.92
ICD 433	.18	0.38	0.88	**ATC B03AA**	.63	0.16	6.93
ICD 470	.18	0.38	0.22	**ICD 600**	.53	0.16	2.76
ICD 584	.11	0.38	0.61	**ATC R06AE**	.81	0.16	3.16
ICD 411	.25	0.38	1.62	**ICD 354**	.71	0.15	0.31
ICD 478	.16	0.38	0.35	**ATC J01XX**	.33	0.14	6.67
ICD 455	.78	0.38	0.57	**ATC C09BB**	.04	0.14	6.23
ATC N05AD	.05	0.37	2.19	**ATC B01AC**	<.001	0.14	31.18
ICD V56	.05	0.37	0.92	**ATC C10AX**	<.001	0.12	6.58
ICD 571	.008	0.36	0.75	**ATC C08CA**	.004	0.12	21.14
ATC S01ED	.5	0.36	4.69	**ATC M01AH**	.37	0.12	10.79

Features	P value	Sup	Prev
ATC A02BA	.44	0.36	3.55
ATC C07BB	.88	0.36	1.8
ICD 440	.72	0.36	1.32
ATC R03DA	.45	0.36	1.23
ICD V43	.11	0.36	1.36
ATC C10AB	.003	0.36	2.89
ATC J01AA	.22	0.36	1.84
ICD 560	.53	0.35	0.44
ICD 482	.67	0.35	0.96
ICD 813	.32	0.34	0.18
ICD 735	.65	0.34	0.22
ICD 041	.32	0.34	0.39
ICD 812	.71	0.34	0.31
ATC C01BC	.21	0.34	2.24
ICD 728	.78	0.34	0.57
ATC C03AA	.14	0.34	1.01
ATC N06AB	.17	0.34	12.24
ICD V64	.8	0.33	0.66
check			

Features	P value	Sup	Prev
ATC J01DD	.004	0.12	24.17
ICD 276	.8	0.12	0.66
ATC M01AB	.14	0.12	26.97
ATC J01MA	.07	0.12	40.53
ATC N02AJ	.72	0.1	1.32
ATC J01CR	.18	0.09	48.46
ATC C10AA	<.001	0.09	31.32
ATC C09BA	.02	0.08	9.91
ATC H02AB	.01	0.08	36.89
ATC M04AA	<.001	0.08	16.27
ATC M01AE	.001	0.06	25.57
ATC R03BA	.003	0.06	17.59
ATC A02BC	<.001	0.05	49.39
Age >68	<.001	0.04	40.53
ATC A11CC	.03	0.03	13.2

Appendix D

Most Prevalent Multimorbidity Feature Combinations in Evolved Bins

check

Table D.1: Most Prevalent Multimorbidity Feature Combinations in Evolved Bins - Cohort 1

Support	**Length**	**Frequent Item Sets**
0.85	1	{'code_drug_R03BA'}
0.84	1	{'eta_53above'}
0.82	1	{'code_drug_N03AX'}
0.79	1	{'code_drug_R06AX'}
0.78	1	{'code_drug_J01XX'}
0.76	1	{'code_drug_C03CA'}
0.74	1	{'code_drug_N02AX'}
0.73	1	{'code_drug_A11CC'}
0.69	1	{'code_drug_C09CA'}
0.68	1	{'code_diag_298'}
0.66	1	{'code_drug_J01CA'}
0.62	1	{'code_diag_411'}
0.61	1	{'code_drug_J01EE'}
0.57	1	{'code_drug_C08CA'}
0.57	1	{'code_drug_A02BX'}
0.56	1	{'code_drug_B01AC'}
0.56	1	{'code_diag_550'}
0.55	1	{'code_drug_D05AX'}

Continued on next page

Table D.1: Most Prevalent Multimorbidity Feature Combinations in Evolved Bins - Cohort 1 (Continued)

Support	Length	Frequent Item Sets
0.54	1	{'code_drug_C07BB'}
0.54	1	{'code_drug_A07EC'}
0.52	1	{'code_diag_618'}
0.51	1	{'code_diag_592'}
0.5	1	{'code_drug_C07AA'}
0.5	1	{'code_drug_R06AE'}
0.74	2	{'eta_53above', 'code_drug_R03BA'}
0.72	2	{'eta_53above', 'code_drug_N03AX'}
0.72	2	{'code_drug_N03AX', 'code_drug_R03BA'}
0.68	2	{'code_drug_R06AX', 'code_drug_R03BA'}
0.68	2	{'code_drug_R06AX', 'eta_53above'}
0.68	2	{'code_drug_R03BA', 'code_drug_J01XX'}
0.67	2	{'code_drug_R06AX', 'code_drug_N03AX'}
0.65	2	{'code_drug_N03AX', 'code_drug_J01XX'}
0.65	2	{'eta_53above', 'code_drug_J01XX'}
0.64	2	{'eta_53above', 'code_drug_N02AX'}
0.64	2	{'code_drug_C03CA', 'eta_53above'}
0.64	2	{'code_drug_R03BA', 'code_drug_N02AX'}
0.64	2	{'code_drug_C03CA', 'code_drug_R03BA'}
0.63	2	{'code_drug_C03CA', 'code_drug_N03AX'}
0.63	2	{'code_drug_R06AX', 'code_drug_J01XX'}
0.62	2	{'code_drug_A11CC', 'code_drug_R03BA'}
0.62	2	{'eta_53above', 'code_drug_A11CC'}
0.61	2	{'code_drug_R06AX', 'code_drug_C03CA'}
0.61	2	{'code_drug_N03AX', 'code_drug_N02AX'}
0.61	2	{'code_drug_R06AX', 'code_drug_N02AX'}
0.61	2	{'code_drug_A11CC', 'code_drug_N03AX'}
0.59	2	{'code_drug_C09CA', 'code_drug_R03BA'}
0.59	2	{'code_drug_C03CA', 'code_drug_J01XX'}
0.58	2	{'code_drug_A11CC', 'code_drug_J01XX'}
0.58	2	{'code_drug_C09CA', 'eta_53above'}
0.57	2	{'code_drug_C03CA', 'code_drug_N02AX'}
0.57	2	{'code_drug_R03BA', 'code_diag_298'}
0.57	2	{'code_drug_N03AX', 'code_diag_298'}
0.56	2	{'code_drug_R06AX', 'code_drug_A11CC'}

Continued on next page

Table D.1: Most Prevalent Multimorbidity Feature Combinations in Evolved Bins - Cohort 1 (Continued)

Support	**Length**	**Frequent Item Sets**
0.56	2	{'code_drug_R03BA', 'code_drug_J01CA'}
0.56	2	{'code_drug_C03CA', 'code_drug_A11CC'}
0.56	2	{'eta_53above', 'code_drug_J01CA'}
0.56	2	{'eta_53above', 'code_diag_298'}
0.56	2	{'code_drug_N02AX', 'code_drug_J01XX'}
0.56	2	{'code_drug_C09CA', 'code_drug_N03AX'}
0.55	2	{'code_drug_C09CA', 'code_drug_J01XX'}
0.55	2	{'code_drug_R06AX', 'code_drug_C09CA'}
0.54	2	{'code_drug_C09CA', 'code_drug_N02AX'}
0.54	2	{'code_drug_C03CA', 'code_drug_C09CA'}
0.53	2	{'code_drug_C03CA', 'code_diag_298'}
0.53	2	{'code_diag_411', 'code_drug_N03AX'}
0.53	2	{'code_drug_N03AX', 'code_drug_J01CA'}
0.52	2	{'code_drug_R06AX', 'code_diag_298'}
0.52	2	{'code_drug_A11CC', 'code_diag_298'}
0.52	2	{'code_drug_A11CC', 'code_drug_N02AX'}
0.52	2	{'code_diag_411', 'code_drug_R03BA'}
0.52	2	{'code_drug_J01CA', 'code_drug_J01XX'}
0.52	2	{'code_diag_298', 'code_drug_J01XX'}
0.52	2	{'eta_53above', 'code_drug_J01EE'}
0.51	2	{'code_diag_411', 'eta_53above'}
0.51	2	{'code_drug_R06AX', 'code_drug_J01CA'}
0.51	2	{'code_drug_C03CA', 'code_drug_J01CA'}
0.5	2	{'code_drug_A11CC', 'code_drug_J01CA'}
0.5	2	{'eta_53above', 'code_drug_C08CA'}
0.5	2	{'code_diag_411', 'code_drug_R06AX'}
0.5	2	{'code_drug_R03BA', 'code_drug_J01EE'}
0.5	2	{'code_diag_298', 'code_drug_N02AX'}
0.5	2	{'code_drug_C09CA', 'code_drug_A11CC'}
0.63	3	{'eta_53above', 'code_drug_R03BA', 'code_drug_N03AX'}
0.59	3	{'eta_53above', 'code_drug_R03BA', 'code_drug_J01XX'}
0.59	3	{'code_drug_R06AX', 'eta_53above', 'code_drug_R03BA'}
0.58	3	{'code_drug_R06AX', 'eta_53above', 'code_drug_N03AX'}
0.58	3	{'code_drug_N03AX', 'code_drug_R03BA', 'code_drug_J01XX'}
0.58	3	{'code_drug_R06AX', 'code_drug_N03AX', 'code_drug_R03BA'}

Continued on next page

Table D.1: Most Prevalent Multimorbidity Feature Combinations in Evolved Bins - Cohort 1 (Continued)

Support	Length	Frequent Item Sets
0.57	3	{'eta_53above', 'code_drug_R03BA', 'code_drug_N02AX'}
0.56	3	{'code_drug_R06AX', 'code_drug_R03BA', 'code_drug_J01XX'}
0.56	3	{'eta_53above', 'code_drug_N03AX', 'code_drug_J01XX'}
0.55	3	{'code_drug_C03CA', 'code_drug_N03AX', 'code_drug_R03BA'}
0.55	3	{'code_drug_C03CA', 'eta_53above', 'code_drug_R03BA'}
0.55	3	{'code_drug_R06AX', 'eta_53above', 'code_drug_J01XX'}
0.55	3	{'code_drug_R06AX', 'eta_53above', 'code_drug_N02AX'}
0.55	3	{'eta_53above', 'code_drug_N03AX', 'code_drug_N02AX'}
0.54	3	{'code_drug_R06AX', 'code_drug_R03BA', 'code_drug_N02AX'}
0.54	3	{'eta_53above', 'code_drug_R03BA', 'code_drug_A11CC'}
0.54	3	{'eta_53above', 'code_drug_N03AX', 'code_drug_A11CC'}
0.54	3	{'code_drug_N03AX', 'code_drug_R03BA', 'code_drug_N02AX'}
0.54	3	{'code_drug_R06AX', 'code_drug_N03AX', 'code_drug_J01XX'}
0.54	3	{'code_drug_C03CA', 'eta_53above', 'code_drug_N03AX'}
0.52	3	{'code_drug_C03CA', 'code_drug_R03BA', 'code_drug_J01XX'}
0.52	3	{'code_drug_N03AX', 'code_drug_R03BA', 'code_drug_A11CC'}
0.52	3	{'code_drug_R03BA', 'code_drug_N02AX', 'code_drug_J01XX'}
0.52	3	{'code_drug_A11CC', 'code_drug_R03BA', 'code_drug_J01XX'}
0.52	3	{'code_drug_R06AX', 'eta_53above', 'code_drug_C03CA'}
0.52	3	{'code_drug_R06AX', 'code_drug_N03AX', 'code_drug_C03CA'}
0.52	3	{'code_drug_R06AX', 'code_drug_R03BA', 'code_drug_C03CA'}
0.52	3	{'eta_53above', 'code_drug_R03BA', 'code_drug_C09CA'}
0.51	3	{'code_drug_R06AX', 'code_drug_N03AX', 'code_drug_N02AX'}
0.5	3	{'code_drug_C09CA', 'code_drug_N03AX', 'eta_53above'}
0.5	3	{'code_drug_C09CA', 'code_drug_R03BA', 'code_drug_J01XX'}

Table D.2: Most Prevalent Multimorbidity Feature Combinations in Evolved Bins - Cohort 2

Support	Length	Frequent Item Sets
0.86	1	{'code_drug_A10BA'}
0.79	1	{'code_drug_N02BE'}
0.76	1	{'code_drug_C03CA'}
0.76	1	{'code_drug_J05AB'}
0.74	1	{'code_drug_M04AA'}
0.71	1	{'code_drug_C09CA'}

Continued on next page

Table D.2: Most Prevalent Multimorbidity Feature Combinations in Evolved Bins - Cohort 2 (Continued)

Support	**Length**	**Frequent Item Sets**
0.65	1	{'code_drug_C02CA'}
0.65	1	{'code_drug_C08CA'}
0.64	1	{'code_diag_V54'}
0.64	1	{'code_diag_V64'}
0.64	1	{'code_drug_J02AC'}
0.63	1	{'code_drug_N06AB'}
0.63	1	{'code_diag_188'}
0.62	1	{'code_drug_S01EE'}
0.61	1	{'code_drug_N03AG'}
0.6	1	{'code_diag_454'}
0.6	1	{'code_drug_N03AE'}
0.6	1	{'code_diag_820'}
0.6	1	{'code_drug_M01AB'}
0.6	1	{'code_diag_735'}
0.59	1	{'code_drug_B01AA'}
0.56	1	{'code_diag_211'}
0.56	1	{'code_diag_574'}
0.56	1	{'code_drug_C09BX'}
0.56	1	{'code_drug_A07EC'}
0.56	1	{'code_drug_P01AB'}
0.55	1	{'code_drug_B03AA'}
0.55	1	{'code_drug_M01AC'}
0.55	1	{'code_diag_482'}
0.55	1	{'code_diag_550'}
0.55	1	{'code_diag_V56'}
0.54	1	{'code_drug_G04CB'}
0.54	1	{'code_diag_427'}
0.53	1	{'code_diag_571'}
0.52	1	{'code_diag_813'}
0.52	1	{'code_diag_V53'}
0.52	1	{'code_diag_428'}
0.51	1	{'code_drug_S01ED'}
0.51	1	{'code_drug_N02AX'}
0.51	1	{'code_diag_276'}
0.5	1	{'code_drug_N05AH'}

Continued on next page

Table D.2: Most Prevalent Multimorbidity Feature Combinations in Evolved Bins - Cohort 2 (Continued)

Support	**Length**	**Frequent Item Sets**
0.5	1	{'code_diag_996'}
0.5	1	{'code_diag_278'}
0.5	1	{'code_drug_R03AL'}
0.68	2	{'code_drug_A10BA', 'code_drug_N02BE'}
0.64	2	{'code_drug_A10BA', 'code_drug_M04AA'}
0.64	2	{'code_drug_A10BA', 'code_drug_J05AB'}
0.64	2	{'code_drug_C03CA', 'code_drug_A10BA'}
0.62	2	{'code_drug_C09CA', 'code_drug_A10BA'}
0.61	2	{'code_drug_J05AB', 'code_drug_N02BE'}
0.6	2	{'code_drug_C03CA', 'code_drug_N02BE'}
0.59	2	{'code_drug_A10BA', 'code_drug_C02CA'}
0.58	2	{'code_drug_M04AA', 'code_drug_N02BE'}
0.57	2	{'code_drug_C03CA', 'code_drug_J05AB'}
0.57	2	{'code_drug_A10BA', 'code_drug_C08CA'}
0.57	2	{'code_drug_C03CA', 'code_drug_M04AA'}
0.57	2	{'code_drug_C09CA', 'code_drug_N02BE'}
0.57	2	{'code_drug_A10BA', 'code_drug_J02AC'}
0.56	2	{'code_drug_C09CA', 'code_drug_J05AB'}
0.55	2	{'code_drug_A10BA', 'code_drug_N06AB'}
0.55	2	{'code_diag_V54', 'code_drug_A10BA'}
0.55	2	{'code_drug_A10BA', 'code_diag_V64'}
0.55	2	{'code_drug_C09CA', 'code_drug_M04AA'}
0.55	2	{'code_drug_J05AB', 'code_drug_M04AA'}
0.54	2	{'code_drug_A10BA', 'code_drug_N03AE'}
0.54	2	{'code_drug_A10BA', 'code_drug_M01AB'}
0.53	2	{'code_drug_A10BA', 'code_diag_188'}
0.52	2	{'code_drug_N02BE', 'code_drug_J02AC'}
0.52	2	{'code_drug_J05AB', 'code_drug_C08CA'}
0.52	2	{'code_drug_N02BE', 'code_drug_C08CA'}
0.52	2	{'code_diag_V54', 'code_drug_N02BE'}
0.52	2	{'code_drug_N02BE', 'code_diag_V64'}
0.52	2	{'code_drug_N02BE', 'code_diag_188'}
0.52	2	{'code_drug_C03CA', 'code_diag_V64'}
0.52	2	{'code_drug_C09CA', 'code_drug_C03CA'}
0.52	2	{'code_drug_A10BA', 'code_diag_735'}

Continued on next page

Table D.2: Most Prevalent Multimorbidity Feature Combinations in Evolved Bins - Cohort 2 (Continued)

Support	**Length**	**Frequent Item Sets**
0.52	2	{'code_drug_A10BA', 'code_diag_454'}
0.52	2	{'code_drug_A10BA', 'code_drug_N03AG'}
0.52	2	{'code_drug_A10BA', 'code_drug_S01EE'}
0.52	2	{'code_drug_A10BA', 'code_drug_B01AA'}
0.51	2	{'code_drug_M04AA', 'code_diag_188'}
0.51	2	{'code_drug_P01AB', 'code_drug_A10BA'}
0.51	2	{'code_drug_A10BA', 'code_diag_820'}
0.5	2	{'code_drug_C03CA', 'code_drug_J02AC'}
0.5	2	{'code_drug_N02BE', 'code_diag_454'}
0.5	2	{'code_drug_N02BE', 'code_diag_820'}
0.5	2	{'code_drug_N06AB', 'code_drug_M04AA'}
0.5	2	{'code_drug_C02CA', 'code_drug_N02BE'}
0.52	3	{'code_drug_A10BA', 'code_drug_N02BE', 'code_drug_J05AB'}
0.51	3	{'code_drug_C03CA', 'code_drug_A10BA', 'code_drug_N02BE'}

Table D.3: Most Prevalent Multimorbidity Feature Combinations in Evolved Bins - Cohort 3

Support	**Length**	**Frequent Item Sets**
0.84	1	{'code_drug_N02AX'}
0.82	1	{'code_drug_M04AA'}
0.76	1	{'code_drug_C03EA'}
0.75	1	{'code_drug_A02BA'}
0.73	1	{'code_drug_B01AB'}
0.7	1	{'code_drug_N03AX'}
0.68	1	{'code_diag_813'}
0.68	1	{'code_diag_295'}
0.68	1	{'code_drug_N02AA'}
0.65	1	{'code_drug_J05AB'}
0.62	1	{'code_drug_C07BB'}
0.62	1	{'code_drug_A12AA'}
0.61	1	{'code_drug_B03BB'}
0.6	1	{'code_drug_R03AC'}
0.58	1	{'code_diag_427'}
0.56	1	{'code_diag_V53'}
0.56	1	{'code_diag_518'}

Continued on next page

Table D.3: Most Prevalent Multimorbidity Feature Combinations in Evolved Bins - Cohort 3 (Continued)

Support	**Length**	**Frequent Item Sets**
0.56	1	{'code_drug_A12BA'}
0.56	1	{'code_diag_413'}
0.55	1	{'code_drug_J01DC'}
0.55	1	{'code_diag_574'}
0.55	1	{'code_drug_N06AA'}
0.54	1	{'code_drug_A05AA'}
0.53	1	{'code_drug_G03AA'}
0.52	1	{'code_drug_C01DA'}
0.5	1	{'code_diag_618'}
0.5	1	{'code_diag_553'}
0.5	1	{'code_diag_434'}
0.5	1	{'code_diag_727'}
0.71	2	{'code_drug_M04AA', 'code_drug_N02AX'}
0.64	2	{'code_drug_N02AX', 'code_drug_A02BA'}
0.64	2	{'code_drug_M04AA', 'code_drug_A02BA'}
0.63	2	{'code_drug_N02AX', 'code_drug_B01AB'}
0.62	2	{'code_drug_C03EA', 'code_drug_N02AX'}
0.61	2	{'code_drug_M04AA', 'code_drug_B01AB'}
0.61	2	{'code_drug_C03EA', 'code_drug_M04AA'}
0.6	2	{'code_drug_N02AX', 'code_drug_N03AX'}
0.59	2	{'code_drug_N02AX', 'code_diag_813'}
0.58	2	{'code_drug_M04AA', 'code_drug_N03AX'}
0.58	2	{'code_drug_M04AA', 'code_diag_813'}
0.57	2	{'code_drug_N02AA', 'code_drug_N02AX'}
0.57	2	{'code_drug_N02AX', 'code_diag_295'}
0.57	2	{'code_drug_N02AA', 'code_drug_M04AA'}
0.56	2	{'code_drug_C03EA', 'code_drug_A02BA'}
0.56	2	{'code_drug_M04AA', 'code_diag_295'}
0.55	2	{'code_drug_N02AX', 'code_drug_C07BB'}
0.55	2	{'code_drug_A02BA', 'code_drug_B01AB'}
0.55	2	{'code_drug_A02BA', 'code_drug_N03AX'}
0.55	2	{'code_drug_C03EA', 'code_drug_B01AB'}
0.54	2	{'code_drug_C03EA', 'code_drug_N03AX'}
0.54	2	{'code_drug_M04AA', 'code_drug_J05AB'}
0.54	2	{'code_drug_N02AA', 'code_drug_C03EA'}

Continued on next page

Table D.3: Most Prevalent Multimorbidity Feature Combinations in Evolved Bins - Cohort 3 (Continued)

Support	**Length**	**Frequent Item Sets**
0.54	2	{'code_drug_A02BA', 'code_diag_813'}
0.54	2	{'code_drug_J05AB', 'code_drug_N02AX'}
0.52	2	{'code_drug_N02AX', 'code_drug_R03AC'}
0.52	2	{'code_drug_C03EA', 'code_drug_J05AB'}
0.52	2	{'code_diag_813', 'code_drug_N03AX'}
0.52	2	{'code_diag_813', 'code_drug_B01AB'}
0.52	2	{'code_drug_A02BA', 'code_diag_295'}
0.52	2	{'code_drug_M04AA', 'code_drug_A12AA'}
0.52	2	{'code_drug_C03EA', 'code_diag_295'}
0.52	2	{'code_drug_C03EA', 'code_diag_813'}
0.51	2	{'code_drug_N02AX', 'code_drug_A12AA'}
0.5	2	{'code_drug_N02AA', 'code_drug_A02BA'}
0.5	2	{'code_drug_B01AB', 'code_drug_N03AX'}
0.5	2	{'code_drug_B03BB', 'code_drug_N02AX'}
0.5	2	{'code_drug_M04AA', 'code_drug_R03AC'}
0.5	2	{'code_drug_M04AA', 'code_drug_B03BB'}
0.5	2	{'code_drug_M04AA', 'code_drug_C07BB'}
0.56	3	{'code_drug_M04AA', 'code_drug_N02AX', 'code_drug_A02BA'}
0.54	3	{'code_drug_M04AA', 'code_drug_N02AX', 'code_drug_B01AB'}
0.52	3	{'code_drug_M04AA', 'code_drug_N02AX', 'code_drug_N03AX'}
0.52	3	{'code_drug_C03EA', 'code_drug_M04AA', 'code_drug_N02AX'}
0.52	3	{'code_drug_M04AA', 'code_drug_N02AX', 'code_diag_813'}

Table D.4: Most Prevalent Multimorbidity Feature Combinations in Evolved Bins - Cohort 4

Support	**Length**	**Frequent Item Sets**
0.8	1	{'code_drug_G04CA'}
0.73	1	{'code_drug_J01CA'}
0.66	1	{'code_drug_C09AA'}
0.66	1	{'code_drug_C09DA'}
0.64	1	{'code_drug_B01AA'}
0.62	1	{'code_drug_C03CA'}
0.61	1	{'code_diag_995'}
0.59	1	{'code_drug_N04AA'}
0.57	1	{'code_diag_153'}

Continued on next page

Table D.4: Most Prevalent Multimorbidity Feature Combinations in Evolved Bins - Cohort 4 (Continued)

Support	**Length**	**Frequent Item Sets**
0.57	1	{'code_drug_J02AC'}
0.56	1	{'code_diag_437'}
0.56	1	{'code_drug_C09BX'}
0.55	1	{'code_diag_298'}
0.55	1	{'code_drug_N05AH'}
0.55	1	{'code_diag_250'}
0.54	1	{'code_drug_A10BA'}
0.52	1	{'code_diag_415'}
0.52	1	{'code_drug_C07AG'}
0.5	1	{'code_diag_424'}
0.5	1	{'code_diag_786'}
0.5	1	{'code_drug_C03EB'}
0.5	1	{'code_diag_780'}
0.57	2	{'code_drug_G04CA', 'code_drug_J01CA'}
0.57	2	{'code_drug_C09AA', 'code_drug_G04CA'}
0.53	2	{'code_drug_G04CA', 'code_drug_C09DA'}
0.52	2	{'code_diag_995', 'code_drug_G04CA'}
0.52	2	{'code_drug_C09AA', 'code_drug_J01CA'}
0.51	2	{'code_drug_B01AA', 'code_drug_G04CA'}
0.51	2	{'code_drug_C09DA', 'code_drug_J01CA'}
0.5	2	{'code_drug_C03CA', 'code_drug_G04CA'}
0.5	2	{'code_drug_B01AA', 'code_drug_J01CA'}