A comparative analysis of existing comorbidity indices with disease clusters generated by using stochastic block models

Mrinmoy Sonowal

4th Year Project Report
Artificial Intelligence and Computer Science
School of Informatics
University of Edinburgh

2023

Abstract

In his thesis report we use inferential community detection, in particular nested or hierarchical stochastic block models to find disease clusters in the Beth-Israel patient dataset to compare and contrast performance and clusters with existing indices like the Elixhauser index with 30 comorbidity featueres. We found that on a bipartite model, the sbm model generated 153 clusters that was more performant than the Elixhauser index, while the association network of only Elixhauser morbidities gave a similar performance to the Elixhauser index.

Research Ethics Approval

This project was planned in accordance with the Informatics Research Ethics policy. It did not involve any aspects that required approval from the Informatics Research Ethics committee.

Declaration

I declare that this thesis was composed by myself, that the work contained herein is my own except where explicitly stated otherwise in the text, and that this work has not been submitted for any other degree or professional qualification except as specified.

(*Mrinmoy Sonowal*)

Acknowledgements

I would like to firstly dedicate this and salute all frontline workers, especially the ones in NHS intensive care and A&E, for their tireless dedication and service.

I would like to thank Valerio Restocchi (my supervisor), Valerio's PhD students - Jorge and Guillermo, for their support and help throughout this project. I would also like to extend my thanks to Jacques Fleuriot (my second marker), whose guidance and understanding, helped to bring it back on track and for their extensive.

Lastly, I thank God, this country and my parents, for enabling this enriching opportunity, not only in the last year to make this dissertation possible but throughout the 4 years in the University of Edinburgh. Special thanks to my friends, who have been with me through this and have helped greatly mentally and emotionally. They have been important pillars in helping me bring this project to a conclusion.

Table of Contents

Chapter 1

Introduction

Comorbidity classifications are systems used to categorize and identify the presence of multiple chronic diseases or conditions in a patient. These classifications are important for understanding the complex interactions between different conditions and how they may impact patient outcomes. In this project we are going to be focusing on creating a better set of clusters with higher predictive power than the Elixhauser index as well as the Charlson index[1]. Predictive power would be measured as the ability to predict mortality of in-patient intensive care unit (ICU). Most comparisons would be made to the Elixhauser index, while both indices are important, we believe a comparison to one is sufficient to prove the hypothesis.

A comorbidity is defined in Elixhauser 1998 as:

> a clinical condition that exists before a patient's admission to the hospital, is not related to the principal reason for the hospitalization, and is likely to be a significant factor influencing mortality and resource use in the hospital

We shall use the above definition of a comorbidity across this paper, although this paper attempts to focus only on mortality.

The motivation for this research is to compare a pure data-driven method of identifying comorbidities to medically-driven approaches. While medically-driven approaches rely on clinical expertise and judgment, data-driven methods can potentially identify patterns and relationships in the data and work alongside clinical judgement.

Furthermore the increased risk of multi morbidity on in-patient mortality rates as well as resource usages, means that it is pressing to find clusters that are better able to predict mortality, resource usage etc. [2]

In this research, we are using community detection, specifically stochastic block modeling (SBM), rather than traditional machine learning clustering methods. This is mainly because the data used in this study is sparse, and SBM has been shown to be effective in identifying communities in sparse networks [3][4].

We will see that SBMs have been used in a wide variety of tasks to perform community detection, however this project presents a novel technique in finding disease clusters

by using hierarchical stochastic block modelling. We would then proceed to use these clusters found by the different models, as features, for a simple logistic regressor and measure performance based on the ROC-AUC score.

Chapter 2

Background

This chapter aims to introduce the key concepts, both medical and computational, to better understand the motivations, prior work and the literature required in understanding this project.

2.1 Medical literature

A preliminary understanding of the medical literature is required for the understanding of the problem tackled by this research. The following sub sections seek to introduce the various concepts and methods that are currently used in the field and that we will be using in this dissertation project. We will also be discussing existing methods and literature on predicting in-patient mortality.

2.1.1 The International Classification of Diseases (ICD)

The ICD-9th revision or ICD-9, is the World Health Organisation's (WHO) method to encode diseases, diagnosis and procedures[5]. It is arranged in a tabular format with each disease or procedure assigned a numerical value. There are also ICD-9s that are start with letters, namely V codes and E codes that represent non-disease factors and external causes respectively. These numerical codes are grouped together in a hierarchy. For example, 001-139 represents infectious diseases, in this 020 represents plague, whereas 0200 represents Bubonic plague. In this research, we are interested only in the first 3 digits as consulted with ICU doctors. For ICD9s that start with letters we take the first 4 digits.

2.1.2 Elixhauser Index

The Elixhauser Comorbidity Index [1] is a widely used tool for characterizing co-morbidities and to predict hospital charges, length of stay, and in-hospital mortality. It is a classification system that includes 30 different co-morbid conditions, grouped into categories such as cardiovascular, metabolic, hypertension etc. The conditions are defined using specific International Classification of Diseases (ICD) codes. Each

co-morbidity is dichotomous, i.e., assigned a value of 1 or 0 (present or absent).The Elixhauser index is widely used in research and administrative purpose, it's also used as a measure of risk adjustment for hospital morbidity and mortality rates.[6]The data was collected from all adult, non-maternal inpatients from 438 acute care hospitals in California in 1992.

The Elixhauser index has been modified to create a Elixhauser score, which is in Van-Walraven et al[7], where they add up the the dichotomous values to get the Elixhauser score, this has been shown to have the same discriminative power as using the entire table. Furthermore the paper identified that 21 of the 30 groups were independent to mortality.

In this paper we treat these co-morbidities as the communities/clusters, although this isn't entirely accurate as Elixhauser comorbidities weren't only measured as a cluster of ICD-9 disease codes. However, from experiments outlined below we can prove that for the intention to predict mortality these codes can be treated as clusters.

2.1.2.1 Using Elixhauser index to predict mortality

In Liu et al [8],a data set of 123,641 adult inpatient admissions from a large hospital system in New York City was collected from the year 2013-14. After training a binary logistic regression model as well as a zero-truncated binomial regression model. They achieved a C-statistic 0.80 (95% CI [0.80–0.81]) for the VW Elixhauser index [7].

However, in Ladha et al [9], we see that they only achieved, 0.66 (95% CI 0.65 to 0.66) for the VW Elixhauser index, for critically ill icu patients before 365 days in the hospital. The ROC-AUC improved after 365 days and the scientists also observed that adding basic demographic indicators did not improve performance.

The papers above proved to be important baselines to follow for this dissertation.

2.1.3 Charlson Index

Mary Charlson et al introduced the Charlson index to predict in patient mortality within 1 year of admission[10]. They originally developed 16 comorbidities each with a score that was weighted on the ability to predict mortality, scores ranged from 1 to 6. The scores were then summed to form the patient's comorbidity score. 559 patients were studied to develop the index and a further 685 individuals were tested on a 10 year period for the index's ability to predict mortality. It was found that a total comorbidity score of 3 had a 59% chance of mortality. Note that the Charlson index always considers demographics when accounting for the comorbidity score, however as we will see and from prior research, we find that demographics does not show major improvements.

Future studies have proven that the Elixhauser index is marginally better than the Charlson index. [6] Although studies such as Ladha et al [9] which compares the Deyo-charlson [11] index and VW Elixhauser index, prove otherwise.

2.1.4 MIMIC-III dataset

Medical Information Mart in Critical Care -III is an anonymised dataset of medical records from the Beth Israel Deaconess Medical Center in Boston, Massachusetts[12]. This dataset has been used in a in a wide variety of both medical and AI tasks, such as predicting sepsis, in-hospitality mortality etc. [13][14].

Following the work of Restocchi et al, we use the cleaned version of the dataset, which consists only of the first admission of patients into the ICU, where all patients are above the age of 16[15]. In this paper the dataset was used to create a bipartite network of patient nodes and comorbities (secondary ICD-9) to find patient clusters, including latent patient clusters.

2.2 Existing mathematical and computer science tools

In this section we introduce the existing mathematics and computer science tools that aids in achieving the goal of this dissertation. In particular we shall introduce graphs and the role of graphs in the project and proceed to explain the problem of community detection using stochastic block models.

2.2.1 Graphs

This section is to elucidate the graphs that we would be primarily working with. While some of the information may seem trivial, it would give a better understanding of the data and models we aim to work with in this dissertation. A graph defined as G=$\{$V,E$\}$ where V is set of all vertices and E are edges in a graph, these vertices and edges can be represented in a adjacency matrix as well as a weighted matrix. In this project we are primarily concerned with bipartite and multigraphs.

Since the solving of Seven Bridges of Königsberg problem by Euler in 1736, graphs have been extensively used as a cross disciplinary tool and in particular, have been of keen interest to mathematicians and computer scientist alike to solve and model various problems.

While graphs aren't as frequently used to represent patient data, its usages in the medicine is increasing [16]. Recently it has been used to perform community detection for Covid-19 [17] as well as to map electronic health records of patients to better represent them as temporal graphs to improve diagnosis performance. [18] The framework introduced in the latter paper was also used for predictive tasks, such as, predicting onset of heart failure.

2.2.2 Community detection

Community detection [19] or clustering is the task of finding clusters that have a higher density of occurrence without groups than outside of groups. In the context of this thesis we aim to find clusters of diseases and use these clusters as features to predict mortality. These clusters would be identified from a dataset of ICU patient and their respective diseases.

There are several ways to perform clustering, however for this research we are primarily interested in hierarchical clustering and link community detection these concepts would be described in detail in the next section on stochastic block model.

2.2.3 Stochastic block model

SBM is a probabilistic graphical model that represents a network as a collection of blocks or communities, with the probability of an edge between two nodes being dependent on the block membership of the nodes[3]. While a basic model would require the knowledge of the number of blocks, using a nested block model would find the optimal number of blocks. We hence use the nested block model structure to find our clusters.

These models are very flexible and have been used in a wide variety of tasks including topic modelling [20], as well as to study social networks as described in Tang et al[21] and modelling protein-protein interaction [22].

Combined with the ability of graphs to represent a wide variety of tasks and SBM's ability to find latent interactions through community detection, we too aim to find structures in our data that would enable clustering of diseases to form a different set of commodities.

We use SBMs over other methods as stochastic block models are scalable as they scale to several nodes and inherently capture the community structures of a given graph which are often found in real networks. [3]. Furthermore as stochastic block models are generative they can indicate the intricate processes by which a network is created and therefore we can identify structures and sub structures inherent to a given network.

2.2.4 Markov chain monte carlo for SBM

We use graph tools [23] implementation of the Merge-split Markov chain Monte Carlo (mcmw) for community detection [24]. This algorithm allows us to reduce the entropy of our graph clusters by repeated merging and splitting and therefore arrive at a better parameter set for the stochastic block model. Originally the mcmw algorithm involved moving nodes from one block to another, however this was computationally very expensive and as the algorithm swept through low probability states, it became increasing difficult to generate new clusters by moving some nodes out of old blocks as well to merge clusters. Therefore in merge-split variant this is achieved through taking larger sets of nodes together and moving them to either new groups or some other older group.

Markov chain monte carlo is a versatile algorithm used for several tasks, such as image segmentation [25]. As the solution space for image segmentation is complex, mcmw is able to find better solutions by iteratively labelling different regions of an image and updating these labels based on a posterior. Likewise, mcmw is also used in natural language processing tasks, such as story generation [26].

Chapter 3

Experimental Design

3.1 Data

As described in 2.1.4 and following Restocchi et al, a data set comprising of 38417 patients data consisting of ham_id (patient id), 30 Elixhauser columns and, one sepsis and mortality column represented as 1 or 0 to indicate the presence or absence of a morbidity, sepsis and mortality respectively. The data set also consists of one-hot encoded demographic data as well primary and secondary diagnoses (represented as ICD-9s). The secondary ICD-9s were present in one column, where the ICD-9s were comma separated.

Throughout this project only this dataset would be used, other datasets such as one hot encoded patient and their respective secondary ICD-9 (size: 38717, 1167) and the Elixhauser one-hot encoded secondary ICD-9s, are all generated using the above data. This second Elixhauser dataset is not the above morbidity dataset (size: 38717,30) but consists of the ICD-9 codes used to create the original Elixhauser index.

This dataset consists of 1169 unique secondary ICD-9 or morbidity diagnosis, we find that it follows an inverse graph, where only the first 343 diseases had a count of more than 100 and only the first 4 diseases had a count of more than 10000. This may prove tricky as our dataset is imbalanced however both stochastic block models and logistic regression works well with imbalanced data.

While extensive cleaning of data wasn't required, the code base ensures that there are no repeated secondary ICD-9s, this had to be done as were considering only the first 3 (or 4) digits. Secondly primary ICD-9 codes and sepsis icd9 codes were removed as the project aims to predict mortality *purely* from secondary ICD-9s, i.e., the morbidites of a patient. In the following section we discuss how we further cleaned the dataset.

3.1.1 Cleaning Data

In this project we were primarily concerned with only the first 3 digits of the ICD-9 codes provided and if it were to start with an alphabet we considered the first 4 characters. Therefore when removing primary ICD-9s from the secondary ICD-9 list

we considered the first 3 (or 4) digits of the primary ICD-9 and the secondary ICD-9. Likewise when removing sepsis ICD-9 the first 3 (or 4) digits of all diseases that contained the word sepsis or septic (but not aseptic) was compared with the secondary ICD-9 set.

3.1.1.1 Finding the sepsis ICD-9s

In the search function of the ICD-cms module in python [27], 'search('n/a')' returns the root of the tree data structure. As the module was arranged in this tree format, a basic depth first search was executed to find all the nodes in this tree that contained the words sepsis or septic but not asepctic and the respective ICD-9 code was added to a set. This set was then used as the list of sepsis codes throughout the project.

3.2 Back to Graphs

In this section we reintroduce the concept of graphs more formally and adapt it to our use case. We are using the following graphs in this dissertation project:

3.2.1 Simple Graph

We define this as any graph where nodes, representing 3 (or 4) digit ICD-9s, are connected by and edge if nodes exist together with a certain association metric. When this graph is weighted, in the context of a stochastic block model, it is an association network. (see: 3.1)

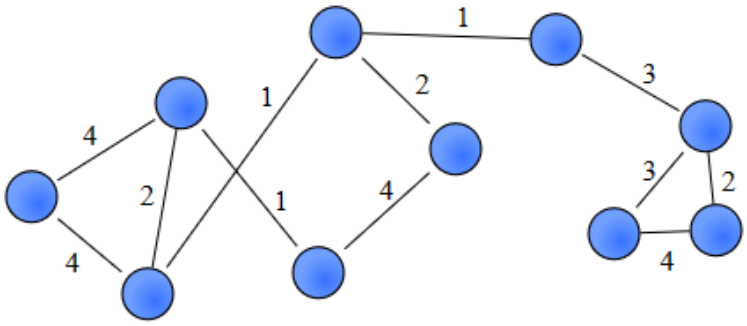

Figure 3.1: Simple weighted graph

3.2.2 Bipartite graphs

A bipartite graph G={V,E}, where $\forall v1 \in V1$ *and* $\forall v2 \in V2$, and $V1, V2 \subset V$. Furthermore, if $e \in E$, and e connects v1 and v2 then $v1 \notin V2$ and $v2 \notin V1$.

In experiment a bipartite graph is used to represent the patients on one side and their respective disease node on the other. This graph is unweighted, which means that if a patient has a disease an edge is formed with no weight in consideration.

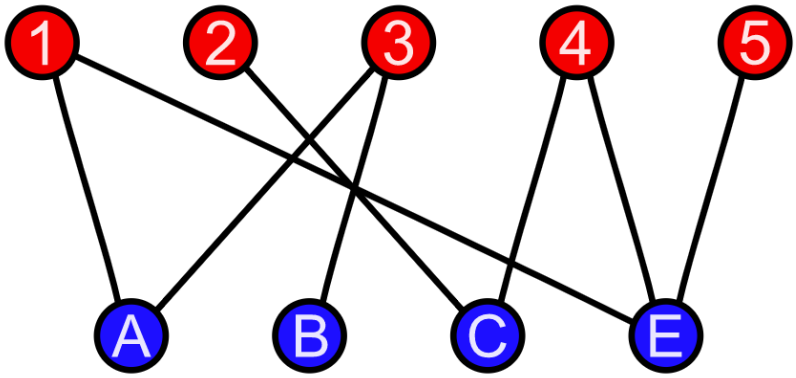

Figure 3.2: Basic bipartite graph

3.3 Stochastic block model

An SBM represents the relationships in a graph by clustering nodes that are 'closer' together in order to reduce the minimum description length (MDL). A minimum description length is the number of bits required to encode the graph, in this case it would be to find the number of blocks and node memberships that would best reduce MDL. In the context of this project it would be find out disease clusters which when defined as a comorbidity is able to encode all the comorbid/secondary ICD-9s. To put simply, we aim to find unique clusters whose members are best able to define that cluster. For example, a diabetes cluster should include the different diabetes ICD-9s, however as we shall later find, this is not always true given the data. [4]

3.3.1 Nested Block model

A nested block model [28] is required for this dissertation project as the intention is to find comorbidity clusters where the number of clusters are unknown. A nested block model is created by generating a stochastic block model with one large block and then further subdividing each block into smaller blocks. The aim of the algorithm is find the block structure that incurs the lowest cost. The cost is the probability of a node being in a block, where edges in between nodes in a block have a higher probability of occurrence than a node in one block being connected to another.

For this project we use the **sbmtm** model as defined in sbmtm.py based on the work of Gerlach et al. [29][20]

In 3.3 The blue lines represent the overall nested block structures and the frontier blocks (blocks on the edges represent by a blue square) are level 0 blocks and each level higher has a smaller number of blocks until we reach the 4th level (the central square) which is the highest level block that includes all the ICD-9s. Each small square then has several nodes represented as coloured circle, the colours represent block membership at level 0 and we can see that each node is connected to another node. These nodes are the disease nodes.

Each block thus contains *distinct* ICD-9 codes, where each ICD-9 code has a probability

Figure 3.3: Nested graph of disease nodes after running nested SBM

of being in the cluster, these probabilities are summed to one for each block.

3.3.2 Clusters as features

In this dissertation we primarily work with two kinds of graphs, the association and the bipartite graph, the following section explains how these clusters were extracted.

3.3.2.1 Getting node cluster membership

The sbmtm module works with bipartite networks and consists of the get_topic_dict function which returns a dictionary of cluster-ids as keys and a list of tuples of nodes and its respective probability within the class, as its values. For the association network, we wrote the get_clusters function in the CooccurrenceGraph.py script, which uses the get_bs function under the NestedBlockModel class in graph tools.

The get_bs function returns a state-membership list, where the index represents the node, ie, it corresponds to the same index as the list of vertices and the list item represents the cluster number. With this a topic_dict of our own was created where the keys again represented cluster id and the values represented the list of nodes in the cluster.

In the following sections we explain the 3 kinds of features used to predict mortality.

3.3.2.2 Boolean features

These features are generated by checking if a patient has any one of the diseases from the cluster, if so a value of 1 is assigned otherwise it is 0. In order to check if a patient has one of the diseases, the patient's secondary ICD-9s are processed by taking the first 3 digits (or 4), and a set is created from these ICD-9s, then using a set intersection function with the set of ICD-9s in the clusters we check if the intersection has 1 or more elements.

3.3.2.3 Weighted features

Continuing from 3.3.2.2 instead of checking if there is an element in the intersection of the cluster and patient ICD-9, we take the number of ICD-9 in the intersection and normalise by the number of ICD-9s in a cluster. Put simply we take the ratio of diseases a patient has from the cluster to the total number of diseases in a cluster.

3.3.2.4 Node-Cluster weighted feature

After performing running the hSBM using the sbmtm module, we get the probability of the node in a cluster and the sum of probabilities in a cluster sum up to one. Hence these weights are calculated simply by adding the probabilities of the ICD-9 codes found for the patient.

3.4 Classification models

As the primary task of this project is to find better clusters and therefore comorbidity than the existing indices. A simple, working metric is sufficient to compare and contrast between the existing models and the proposed hSBM cluster models. For this we use Logistic regression is which defined below.

3.4.1 Logistic regression

This is a type of linear statistical model that analyses and models the relationship between a dependant binary variable (in our case mortality) with some independent variables (cluster features, Elixhauser comorbidity and one hot encoded ICD-9 patient features).

A logistic regression works by adjusting the coefficients of a linear expression based on some loss to fit the model to the data. These coefficients can be either positive or negative, whereby, a positive coefficient indicates that the presence of that feature is likely to output true for the dependant variable while a negative coefficient will tend to output a false.

The output of the linear expression when passed through a sigmoid function outputs true or false.

A regressor is said to have been overfit if it performs exceedingly well on the training set but not on the test set. This is a result of the model adjusting its coefficients to best

minimise loss on the training set data.

A model is underfit if its output is largely random and is unable to predict either train set or test set data accurately.

The following gives the sigmoid probability of y-1, given x, using a threshold value a decision boundary can be constructed whereby the model can predict if given x it should return true or false.

$$p(y=1|x) = \frac{1}{1+e^{-(\beta_0+\beta_1 x_1+\beta_2 x_2+\cdots+\beta_n x_n)}}$$

where β are the respective coefficients of the linear model and x_1 to x_n are the features.

In this dissertation we will use the scikit-learn's sklearn module in python to implement our logistic regressor [30].

3.4.2 GridSearchCV

This module from sklearn works by running various parameters and 'searching' the different models for the performance, it then returns the best provided parameters. For training our logistic regressors, we used a 5 fold cross validation using grid search cv and provided a balanced and un-balanced class-weight. This module helped us find the best solver to be liblinear and class weight to be balanced, these parameters were maintained throughout the various experiments for consistency.

3.5 Metrics

This section describes the various metrics used in this project for different tasks. We first define the roc metric and its associate metrics like the confusion matrix and we will then proceed in this section to describe other metrics, such as, confidence intervals which are used to indicate statistical significance. Relative risk, association beyond chance and fisher exacts were used when generating the association network.

3.5.1 ROC-AUC Metric

We shall measure the performance of our models using the ROC-AUC score however to calculate this we first need the following:

3.5.1.1 Confusion matrix

		Actual Class	
		Positive	Negative
Predicted Class	Positive	TP	FP
	Negative	FN	TN

Table 3.1: Confusion Matrix

A confusion matrix is a square matrix consisting the number of true positives (TP), false positives (FP), true negatives (TN) and false negatives (FN). A true positive is when the model outputs true when the real label is true and a true negative is for a false label. A false positive is when the model predicts true when in reality the label is false. Lastly, a false negative is when the model predicts that a the dependant variable is false when it was actually true.

3.5.1.2 Specificity, Sensitivity and False positive rate

Using the above definitions of TP, FP, FN and TN.

$$TPR = Sensitivity = \frac{TP}{TP+FN}$$

Sensitivity or the true positive rate is the probability of the model to correctly identify true results.

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

Specificity or the true negative rate, is the probability of the model to correctly identify negative results.

$$FPR = 1 - Specificity = \frac{FP}{(FP+TN)}$$

False Positive Rate (FPR) measures the actual negative classes that are wrongly labelled by the model.

3.5.1.3 ROC-AUC

Receiver Operating Characteristic - Area under curve, which is a curve that plots sensitivity against specificity. This ratio is crucial for binary classification and is far superior to accuracy, in particular for sparse data. This ratio's values range from 0.5 to 1.0, where 0.5 represents the inability to discriminate and 1.0 represents perfect discrimination. Every 0.1 is considered to be a different class of classifier, where 0.5-0.6 is bad, 0.6-0.7 is okay, 0.7-0.8 is good and so on.

$$ROC\_AUC = \frac{TPR}{FPR}$$

3.5.2 Confidence intervals

Confidence intervals are calculated to test if the population means are indeed different than the sample means of a given distribution. In our case we plot confidence intervals on the distribution of bootstrapped roc-auc values for a given model. If the intervals overlap then we reject the hypothesis that a model is better or worse than the other and are indeed performing at the same level. It is also important to note that as stated before roc-auc scores are classified every 0.1

		Disease j	
		Absent	Present
Disease i	Absent	$n_{0,0}$	$n_{0,1}$
	Present	$n_{1,0}$	$n_{1,1}$

Table 3.2: Contingency matrix for diseases A and B.

3.5.3 Contingency table

A contingency table represents the multi variant frequency distribution of two variables (in our case any two ICD-9 codes or morbidities).

Furthermore n_i represents to the total number of patients who had disease i and be calculated as $n_i = n_{1,0} + n_{1,1}$ Likewise $n_j = n_{0,1} + n_{1,1}$.

3.5.4 Association beyond chance

The joint probability of conditions ($P_{i,j}$) is expressed as:

$$P_{i,j} = \frac{\sigma_i \sigma_j + \sigma_i \sigma_j a_{i,j}}{1 + \sigma_i \sigma_j a_{i,j}}$$

Here σ represents the appearance due to factors not affecting other conditions and $a_{i,j}$ **represents the association between two diseases i and j**.

Prevalence of a condition P_i is expressed as:

$$P_i = \frac{\sigma_i + \sigma_i \sigma_j a_{i,j}}{1 + \sigma_i \sigma_j a_{i,j}}$$

Here the smoothing (1 +) represents the contribution independent of associations and $\sigma_i \sigma_j a_{i,j}$ represents contributions of associations.

ABC measures the prevalence of two diseases based on common factors This metric is useful as it does not overestimate for diseases that appear a lot in the dataset.

Negative abc values represent associations that aren't likely to occur together and positive values represent those are likely to occur together beyond chance.

3.5.4.1 Scaling negative abc

We used the following formula to bring negative abcs to the positive scale

$$abc_{pos} = \frac{-1}{1 + abc_{neg}}$$

where abc_{neg} is the negative abc score, ie, $abc_{neg} < 0$

3.5.5 Relative risk

Using n_i and n_j as defined in 3.5.3, where n_i and n_j are total counts of disease i and j. Using this, Relative risk is calculated as:

$$RR_{i,j} = \frac{C_{i,j}}{n_i n_j}$$

where $C_{i,j}$ is the co-occurrence of diseases i and j.

Relative risk also measures the chances of two variables occurring together, a relative risk greater than 1 represents a positive correlation whereas less than 1 represents no correlation. However relative risks fails to capture underlying reasons and can be overestimated by skewed datasets.

3.5.6 Fisher exact

Fisher exact is a statistical significance test, which given a contingency table yields the p value for the results, we used this to filter out those pairs of diseases who p values were less than 0.05. This is to ensure that the measures outlined above are statistically significant and helps filter out cases where there isn't enough data, i.e., certain disease pairs may not have enough data to provide a good rr or abc and the fisher exact helps us with this.

Chapter 4

Results and Discussions

In this chapter using the models, methods and data explained in chapter 3, we first tested the baseline using basic Elixhauser comorbidity features as well as Charlson and Elixhauser with demographic features. We then proceeded to run a bipartite model for all ICD-9s and Elixhauser ICD-9s, and also experimented with removing certain clusters and taking only the best 30 clusters. Lastly instead of using a bipartite network we used an association network with the abc and rr metric.

We shall highlight the experiments and the different results we achieved from them in the following sections:

4.1 Experiment 1: Elixhauser comorbidity as features

We used the one-hot encoded Elixhauser comorbidity table from the dataset (size: 38717,30) to train a logistic regressor to predict mortality. Using a logisitic regression model with balanced weights set to true we got the baseline metric. The balanced weights parameter helps a logitic regressor train for spare datasets such as ours. It is important to know that this baseline is based on our dataset and higher scores have been achieved by different researchers as outlined in the background section 2.1.2.1. From fig. 4.2, we can see that that the Elixhauser features were all sub 0.7 roc-auc, bootstrapping 200 times roc_auc score of the Elixhauser features was found to be (0.628, 0.655, 95% CI) on the test set.

This meant that the predictive power of the Elixhauser index was okay. Charlson too had a very similar predictive power [0.637-0.664, 95% CI]. Lastly adding demographic data to the Elixhauser index did improve performance [0.657-0.683, 95% CI].

We know from prior research that the Elixhauser index uses diseases that are indicative of mortality as per medical literature and data. In the upcoming sections we shall use the same set of diseases to run a stochastic block model to find disease clusters (see figs: 4.3 and 4.6).

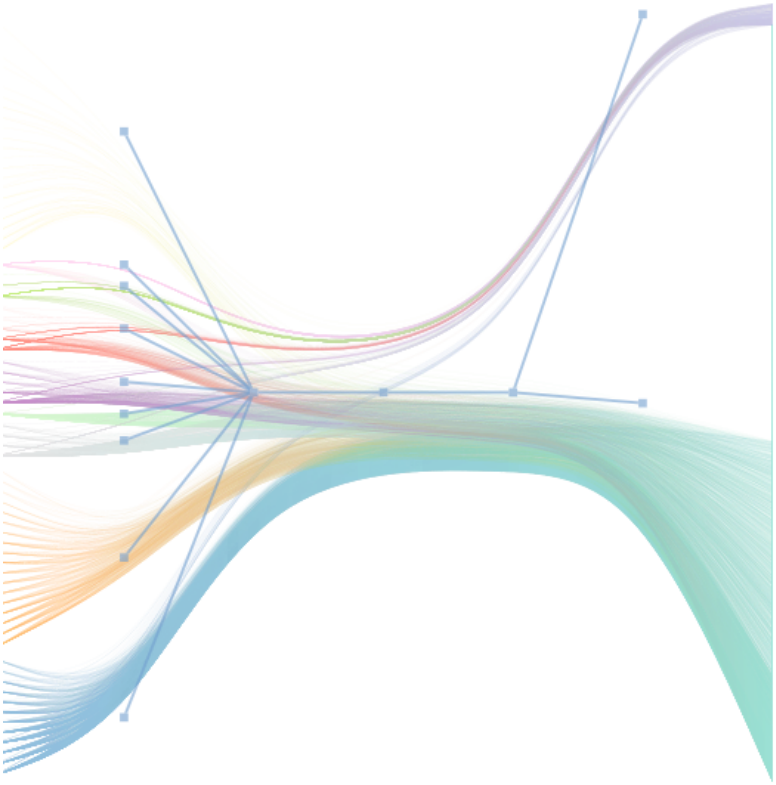

Figure 4.1: Bipartite graph of Elixhauser ICD-9s (left) to patient nodes (right) after running hSBM

4.2 Experiment 2: Bipartite network

We used sbmtm, a topic modelling tool, to generate a bipartite network of patients on one side and diseases on the other. The stochastic block model was then run for 10 epochs followed by the mcmw sweep algorithm for 1000 iterations. We ran this for two of the graphs as outlined below, one for the Elixhauser ICD-9s and another graph with all the secondary ICD-9 codes baring sepsis and the respective primary ICD-9 of the patient. The stochastic block models generated a topic dictionary that contained node membership to their respective clusters, where clusters were key. The mcmw helped the stochastic block model refine the graph it generated and running for a 1000 iterations meant that while we may never reach the best clustering for a graph we can be close enough.

The next experiment outlines the use of a bipartite network in generating clusters for Elixhauser diseases.

4.3 Experiment 2.1: Bipartite network - Elixhauser ICD-9

As a proof of concept, we wanted to prove that the stochastic block model is able to find clusters from the Elixhauser ICD-9 set and that it is able to predict mortality

with a similar predictive power as the Elixhauser index itself. For this the ICD-9s from the secondary ICD-9 column in the patient data were taken and a matrix of just Elixhauser ICD-9s was generated, where each row represented a patient. However, after running the hSBM model, only 9 clusters were found (see fig: 4.1), whereas the original Elixhauser index had 30. In the next 3 sections we will explain the results retrieved from the various feature types.

4.3.1 Boolean features

From fig 4.2 we see that the bipartite model performed sub par to the existing baseline (Elixhauser index), while this is not ideal, it may still prove useful to ICU doctors who make minute by minute decisions and therefore would only have to go through 9 comorbidities.

4.3.2 Weighted features

Taking the fraction of diseases a patient had from a cluster yielded a roc-auc score of [0.618-0.628, 95%CI]. While this seems like a an improvement from using boolean features, as the intervals overlap, it is not statistically significant on the population mean. This means that the both weighted features and boolean features may have similar predictive power given the true population data.

4.3.3 Node-cluster weights as features

Lastly we took the sum of weights of the diseases present. This yielded a roc-auc score of [0.593-0.612, 95%CI].

From the 3 models above we can see that while it may be useful to have only 9 clusters, we haven't yet achieved the goal of finding better predictive features. However even these sbm models were able to discriminate between different types of diseases for instance, alcohol related issues existed in its own cluster while depression was set into its own cluster. For certain diseases like liver diseases and heart diseases, as they constituted the larger number of patients, descriptive clusters were not found but instead were placed in several clusters. However the most informative cluster that predicted mortality was that of the cluster that included other diseases, this was possibly because brain conditions, anemia etc. while dangerous, were also infrequent as compared to heart conditions and hence was able to better predict mortality. It is important to note that a roc-auc of around 0.6 is sub optimal as it means that the logistic regression model was only somewhat able to find the patient's mortality.

4.4 Experiment 2.2: Bipartite network - All secondary ICD-9 excluding sepsis and primary diagnosis

Similar to the above bipartite model this model was run with all the ICD-9 codes excluding sepsis and the respective primary diagnoses. 153 clusters were found after running the stochastic block model over the bipartite graph and this by far was the most

performant model as it yielded the best roc-auc score (see fig: 4.5). Before proceeding to fitting the stochastic block model a table of patient and their one hot encoded ICD-9s was generated (size: 38417,1169) after removing sepsis, a simple logistic regression model was trained using this dataset. We got a training roc-auc of [0.814-0.820, 95% CI] and [0.777-0.800, 95% CI], this can be considered a baseline for these ICD-9 codes as it is the model with the highest dimensionality and hence should be able to capture the data better, this is particularly true as we do not see major overfitting.

4.4.1 Boolean features

On this model we achieved a roc-auc of [(0.786-0.795, 95% CI] on the train set and [0.769-0.794, 95% CI] on the test set, a well fit model with a high performance as the features are only boolean, a patient with any of the diseases in the cluster is enough for the patient to have that cluster.

4.4.2 Weighted features

We then proceeded to train with the weights, which like before captured the fraction of diseases a patient has in the cluster. To our surprise its predictive power was no more different than the boolean matrix. With a roc-auc score of [0.785-0.792, 95% CI] on the training sets and [0.768-0.792, 95% CI]. This clearly shows that **the model is agnostic to the number of diseases a patient has in a given cluster**.

4.4.3 Node-cluster weights as features

Lastly, we hypothesised that perhaps some features were very informant of their clusters and hence their probability in the cluster could be used as weights for a given cluster. We also achieved [(0.788-0.796, 95%CI] on the train set and [0.773-0.795, 95% CI], this further proves that just boolean features, ie, the presence or absence of one of the diseases in a cluster is therefore indicative of the cluster. These clusters are hence similar to the Elixhauser co morbidities as they encompasses a wide number secondary ICD-9 while being features themselves.

4.4.4 Further analysis

It is important to note that this predictive power could be entirely due to the larger number of features, as stated earlier increasing the number of features is better able to fit the data until it reaches an inflexion point where it overfits. However from 4.4, we can see that we haven't yet reached the number of features required to overfit. Therefore a sub-experiment was carried out whereby only the best 30 features were taken from the boolean model, this enabled us to better compare with the Elixhauser model which also has 30 feature columns.

While this model did drop in performance as compared to the previous 3 models, it still performed better overall than the Elixhauser index (from: 4.5). This further solidified our hypothesis that stochastic block models, being able to understand latent structures in the graph (and hence the data), are able to produce better comorbidity clusters than

existing indices. As these clusters only need one secondary ICD-9 to be identified as true, it would also help doctors save crucial time treating their ICU patients.

We also tested the model using level 1 clusters, as our block model is nested. This however gave extremely sub par results as at this level there were only 10 clusters (see fig. 4.3). We can therefore conclude with a certain degree of confidence that in order to create performant features that a stochastic block model is able to identify and cluster together secondary ICD-9 (morbidities) in a meaningful manner.

From A.1, we see that our model found succinct disease clusters where diseases were similar to the diseases found in the Elixhauser index - such as liver chirrosis, heart diseases and kidney disease. This validates the model as it shows that it aligns with medically curated index but at the same time, with addition of some other ICD-9s and improved clustering we were able to get better results.

Lastly, we stress on the boolean model having the same performance as the weighted ones as this means that the clusters are predictive of mortality instead of the respective disease. If they were to have different performances we would have to study the individual diseases and therefore their contribution to the weight of a feature vector. This would make the task of predicting mortality harder and the clusters may not be that significant.

As we conclude the experiments using bipartite network the next section aims to explain how we used association networks to cluster diseases.

4.5 Experiment 3: Association network

Association networks are graphs where an edge between two nodes exists if they satisfy some correlation or association metric. We initially experimented with the relative risk metric (see: 3.5.5), i.e., we used relative risk (3.5.5) two nodes as weight of the edge and using the fisher exact metric filtered out those edges that were below $p < 0.05$. We found that while relative risk gives a better predictive power this was largely due to its inability to cluster the diseases into meaningful clusters, therefore giving us more features to fit to the data. Relative risk's main drawback is that its effected by imbalanced datasets such as ours and overestimates on larger groups, for smaller groups the risk becomes more random. Therefore we proceed to using the association beyond chance (abc) metric, to add edges to the graph. The abc metric is also a probability metric where negative values between two variables means that two variables are unlikely to occur together while a positive value signifies they are likely to occur together.

4.6 Experiment 3.1: Association network - Elixhauser ICD-9s non nested

Using the Elixhauser ICD-9 set, we built an association network using the abc metric, as we knew we were using Elixhauser features we first wanted to see if it could find

features that could be clustered in 30 clusters. For this we used minimize_blockmodel_dl from the graph tool library and supplied the number of blocks as 30. While we can indicate what the number of block should be, this function doesn't guarantee that it will find those many clusters. Hence the model was only able to find 14 clusters.

4.6.1 Boolean features

We found that while the results were not as good as the Elixhauser index, the Elixhauser index was still not significantly better than the clusters found by hSBM. Furthermore the number of clusters found were 14 (see fig: 4.6) and hence difficult to compare with Elixhauser which has 30 feature columns. As our model only produced 14 clusters it was expected to perform poorly compared to the baseline Elixhauser model but a model with a similar performance as the baseline would also be very helpful to ICU doctors as they would have to go through 14 instead of 30 comorbidities.

It is also interesting to note that on the train set the baseline does better (see fig: 4.7) however it is on the test set that our model matches the baseline performance. This is due to both the models have a tighter confidence interval on the train set than the test set.

4.6.2 Weighted features

While the performance was slightly improved on the mean, as the confidence intervals overlapped this improvement was not statistically significant. This ones again proves that **knowing if a patient has a disease in a cluster is sufficient to predict mortality**, and hence the clusters are largely independent.

The above experiments may not be as performant as the bipartite network however with 14 clusters and using the Elixhauser ICD-9 set these clusters indicate that our stochastic block model is to able to find relations between diseases and is able to cluster them together that produce a feature set that works similar to the existing Elixhauser index.

4.7 Experiment 3.2: Association network - all secondary IC9-s except sepsis and primary diagnosis

Now that we have been able to establish that the association network using abc metric is able to perform almost as well as the baseline, we proceeded to doing the same for the whole set of ICD-9s. However this is where we first ran into a major road block. We considered edges with non-zero abc, which means that the connected nodes were either to be placed in different clusters (if negative weight) and same clusters if positive. However when running the hsbm model it was unable to converge. Due to limited time and availability of computing resources we had to find an alternative, so we used the adjusted abc metric which brings negative abc to the positive scale. This was able to run and provided us with 152 clusters. 4.8

4.7.1 Boolean features

With an AUC of (0.579-0.607 95% CI) on the test set, we achieved sub par results, while the 153 clusters maybe similar in number to the bipartite network the clusters themselves were different and many a times did not make sense, very large clusters constituting of several diseases were clustered and we suspect that this could have impacted performance.

4.7.2 Weighted features

The weighted features also achieved similar results to the boolean features and were not statistically significant, which again shows that a stochastic block model generates clusters that are independent of how many diseases or which diseases were found.

We plotted these results in fig:4.9 which demonstrates that only the rr model achieved a much higher result albeit as a result of improper clustering. We suspect that as we used the modified abc metric we couldn't achieve the necessary convergence required, while a nested block state was acquired and the mcmw algorithm was run over it, the latent structures to capture the inter connectedness of the diseases and hence formulate clusters indicative of mortality, could not be acquired using this model.

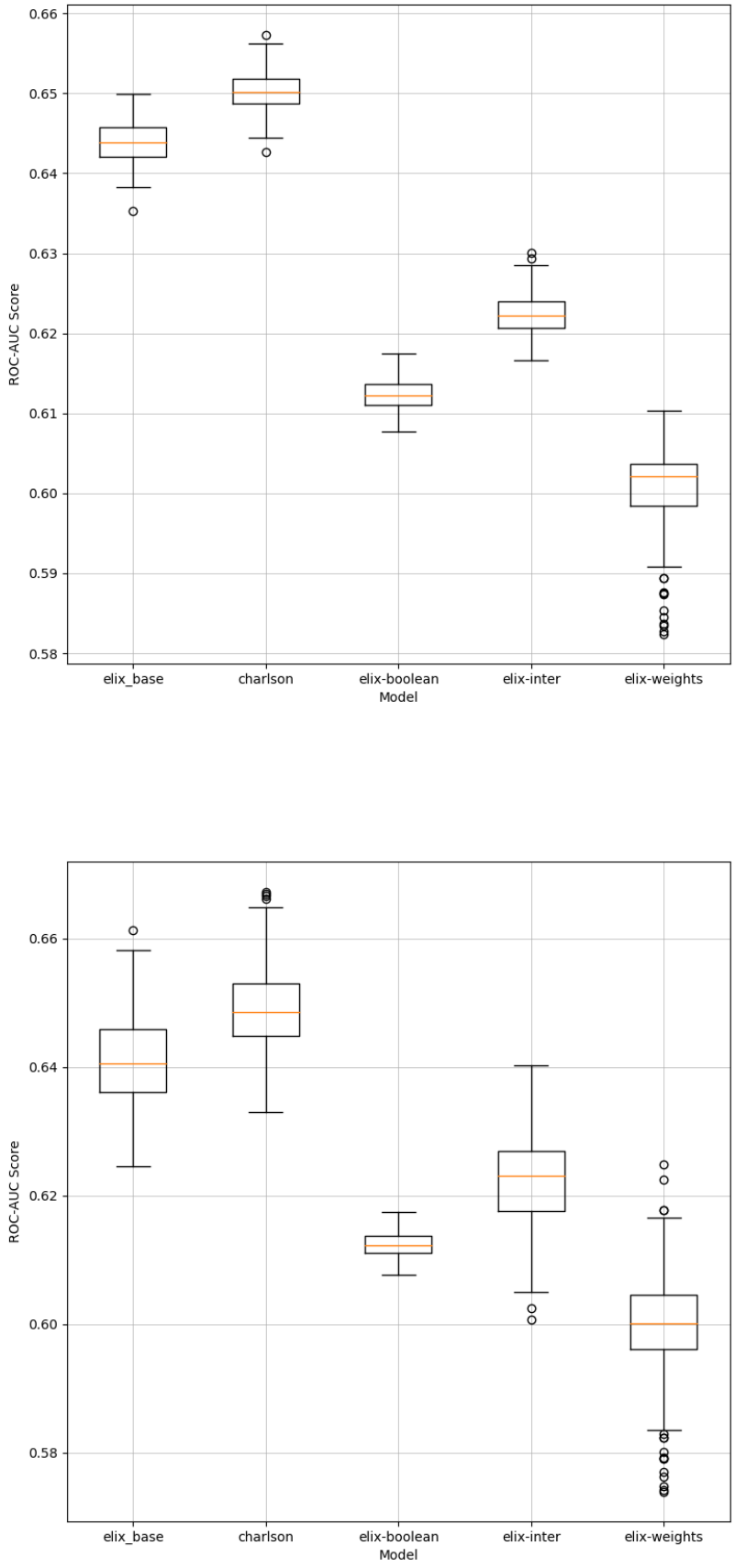

Figure 4.2: Box and whiskers of Elixhauser ICD-9 hSBM bipartite - Train(top) and test(bottom)

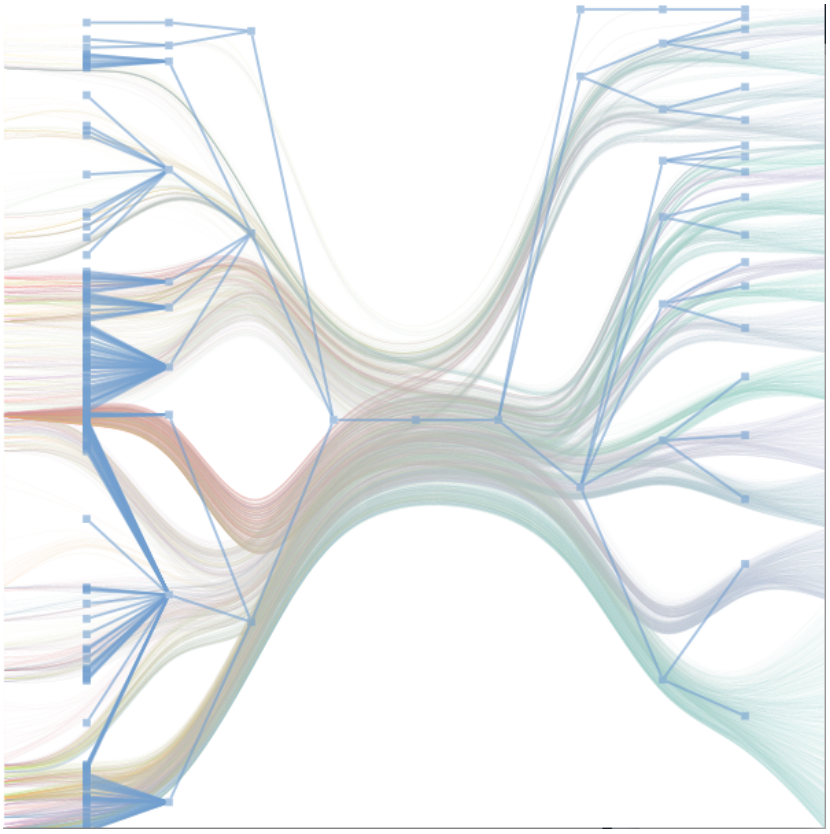

Figure 4.3: Bipartite graph of Elixhauser diseases (left) to patient nodes (right) after running hSBM

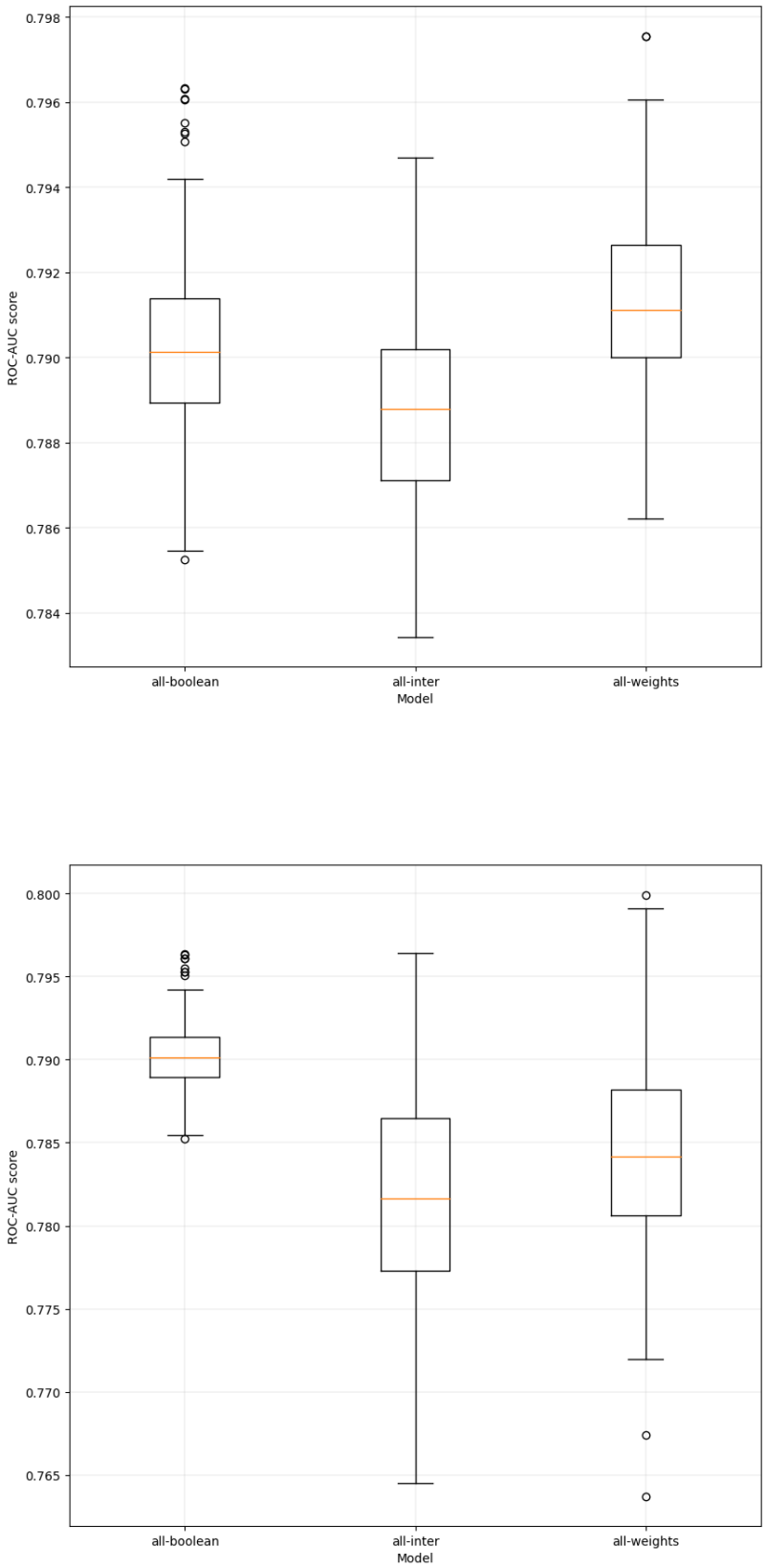

Figure 4.4: Box and whiskers of hSBM bipartite - Train(top) and test(bottom)

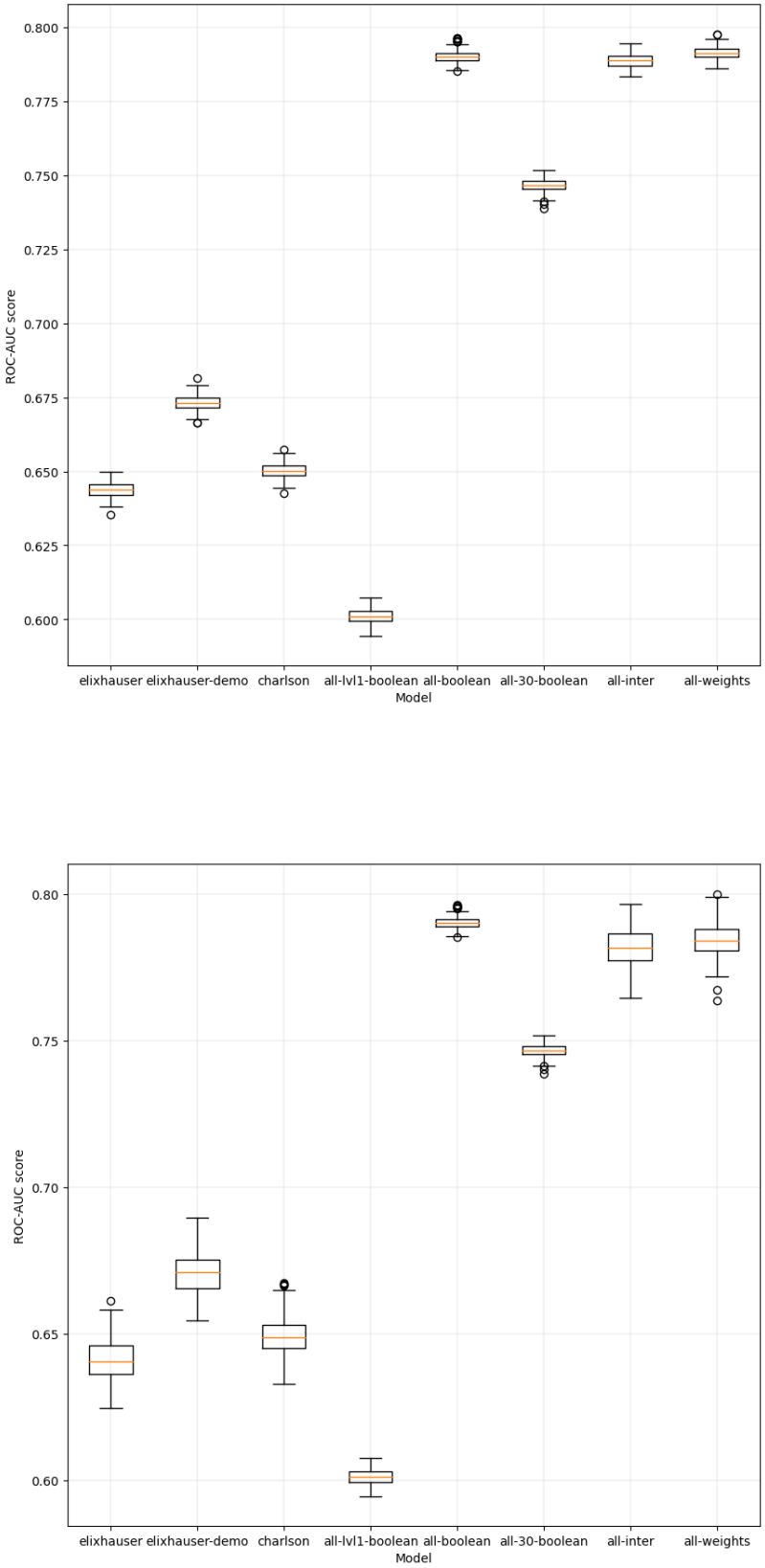

Figure 4.5: Box and whiskers of the existing comorbidity models (Elixhauser and Charlson) and hSBM bipartite - Train (top) and test(bottom)

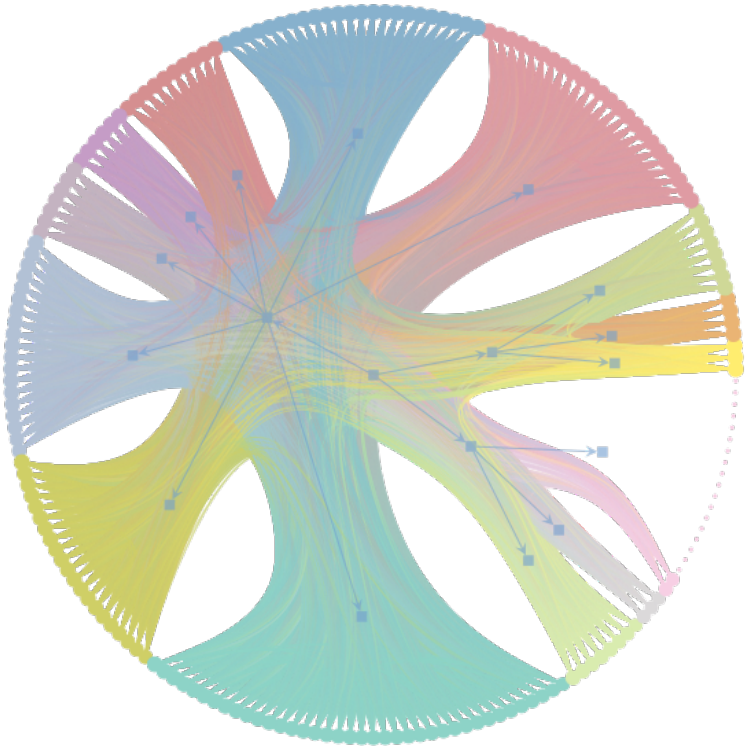

Figure 4.6: Association network of Elixhauser ICD-9 demonstrating 14 clusters at level 0

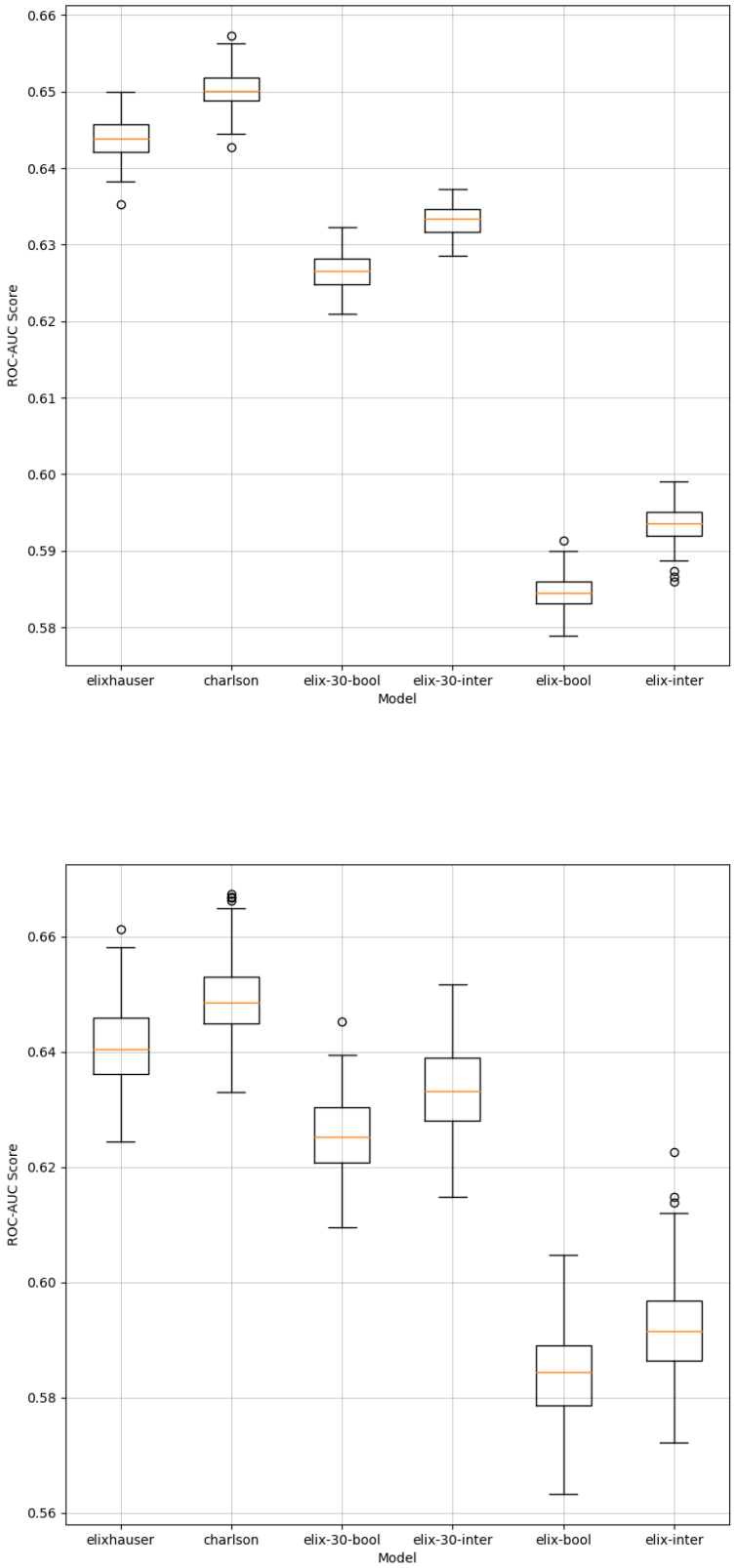

Figure 4.7: AUC box and whiskers of association network, association network using weighted cluster features and basic Elixhauser comorbidity features - Train (top) & Test (bottom)

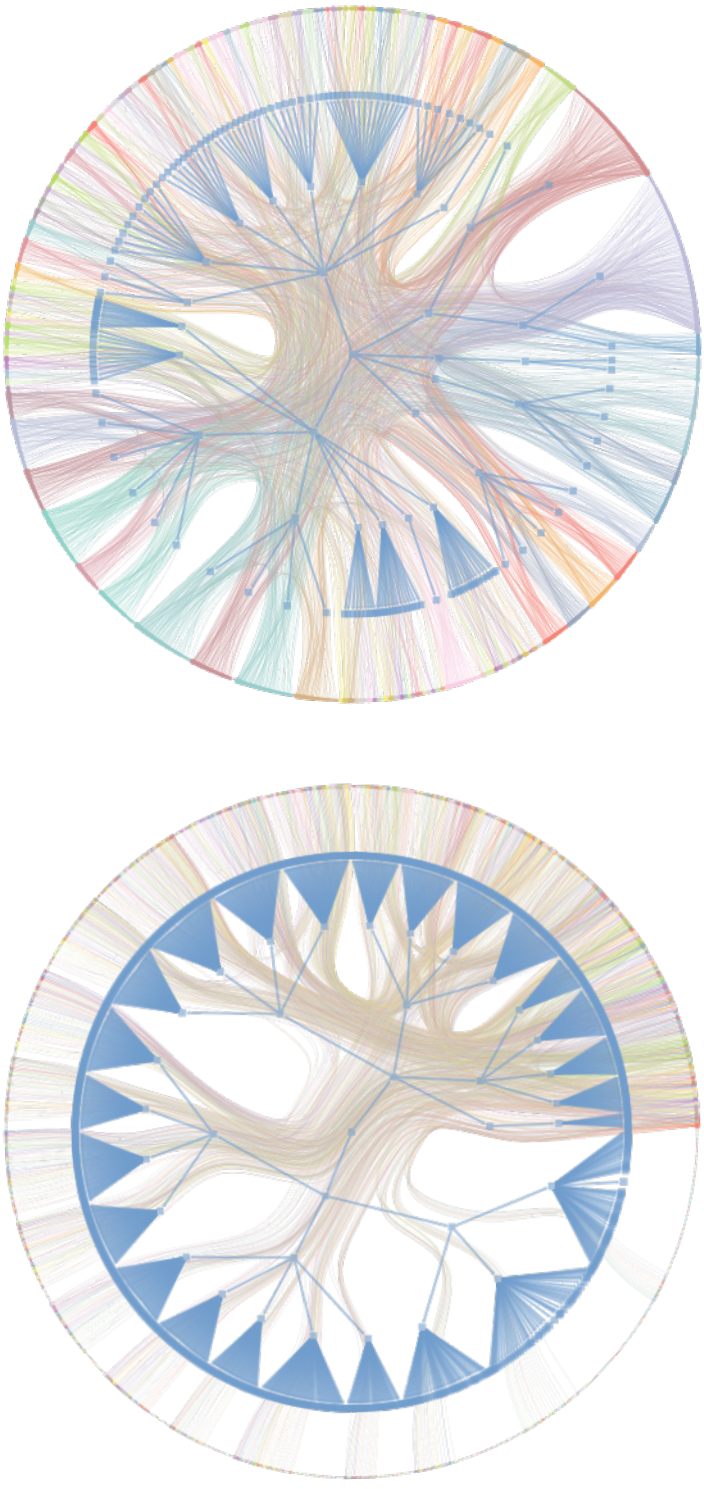

Figure 4.8: Association networks comprising of all the secondary ICD-9s - abc metric(top) rr metric(bottom) - sub sample of 2000 nodes

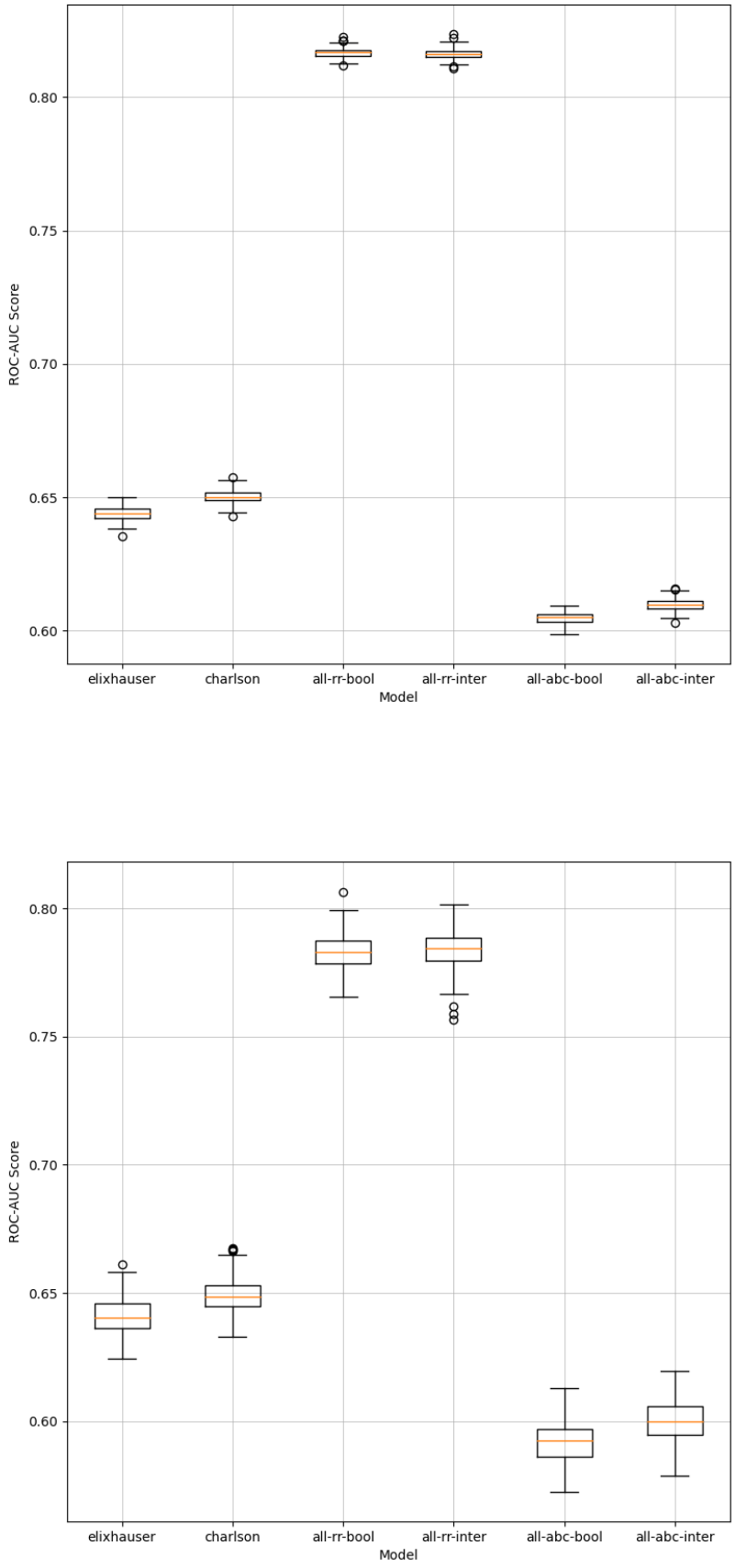

Figure 4.9: Box and whiskers of the existing comorbidity models (Elixhauser and Charlson) and hSBM association - Train (top) and test(bottom)

Chapter 5

Conclusions

The primary objective of finding clusters indicative of mortality was achieved as the confidence interval ([(0.786-0.795, 95% CI]) within which the sample results of the bipartite network lie far beyond the confidence intervals of the Charlson and Elixhauser indices [0.639-0.648, 95% CI]. We also saw that even while taking the 30 most predictive features, we still achieved a higher score [0.734-0.757, 95% CI] showing that the model is well fit and given the same dimensionality of data, the cluster-features were better at predicting mortality. Lastly, we got to see that the clusters themselves were indicative of mortality irrespective of how many diseases a patient had in that cluster and the individual disease itself, this proved interesting as in the real setting ICU doctors would only have to check if a patient had one of the diseases in the cluster and can mark the presence of a cluster in a patient with only one disease saving crucial time.

In the next section we will discuss the limitations of this project and then suggest improvements and future work.

5.1 Limitations

While this project was successful at achieving better results a clearer, medically verified understanding of the clusters are required. We used the Beth Israel dataset of 38717 patients, however as we purely used only this dataset to train, validate and test it is entirely possible our model is overfit for this hospital. This would mean that our model may very well be useless for a different population group, especially one who lives outside of the developed world. At best our model only predicts based on morbidities provided that were recorded. Plus, we have seen from other papers that the Elixhauser index has shown better prediction powers, in particular for other datasets.

The other limitation of this project is that while we were able to prove improvements using a bipartite model of patient to their diseases, this model doesn't fully capture the respective prevalence of diseases, while it is able to cluster better than the association network, we believe that a full proof association model that captures relative prevalence of diseases is important.

This project can undergo rigorous testing, in particular, all the metrics can be unit tested

and the code can be tested for mathematical consistency. While assertions and basic flags and checks have been added to the code base, extensive testing will help fix unseen bugs

Lastly, as discussed earlier this project is still in need of expert opinion on the clusters and the respective diseases that were found, while liver and kidney diseases were proven by earlier papers to be indicative of mortality in ICU patients, the focus of this project is on all the clusters and extracting information to better predict mortality.

5.2 Future work

We believe that this work should be reproduced for other datasets, especially in other parts of the world with a different morbidity profile. This would enable us to better test if the model has been overfit to our dataset and if it can truly generalise on any given morbidity profiles. We could approach this in two ways:

5.2.1 Train a large varied dataset

We could test on a larger dataset, similar to VW Elixhauser, although with the added task of compiling different datasets from different hospitals, this would also require greater computing power as more unique secondary ICD-9s could be found. However this would better identify intricate relations among diseases on a more global scale and therefore we can better identify the disease clusters that best predict mortality.

5.2.2 Train on different datasets

Similar to the Beth Israel dataset, other datasets can be acquired and the steps we have outlined in this paper could be reproduced individually for the datasets. If we see a marked difference in performance we would have to study why that dataset was different from the other datasets. This approach would help in finding intricate differences among patient data from different places and hence if there is a difference in performance latent details about the dataset could be identified.

We are unsure as to why the abc metric on the larger dataset was unable to converge, we did observe that the block model did run but was simply unable to finish execution, which could mean that we were not running it long enough. We believe that in the future a model with the simple abc metric considering all non zero values should be run for longer.

Bibliography

[1] A. Elixhauser, C. Steiner, D. R. Harris, and R. M. Coffey, "Comorbidity measures for use with administrative data," *Medical care*, vol. 36, no. 1, pp. 8–27, 1998. [Online]. Available: https://pubmed.ncbi.nlm.nih.gov/9431328/

[2] Y. V. Chudasama, A. K. Khunti, and M. J. Davies, "INTEGRATED CARE Clustering of comorbidities," *Future Healthcare Journal*, vol. 8, no. 2, pp. 224–233, 2021.

[3] P. W. Holland, K. B. Laskey, and S. Leinhardt, "Stochastic blockmodels: First steps," *Social Networks*, vol. 5, no. 2, pp. 109–137, 6 1983.

[4] C. Lee and D. J. Wilkinson, "A review of stochastic block models and extensions for graph clustering," *Applied Network Science 2019 4:1*, vol. 4, no. 1, pp. 1–50, 12 2019. [Online]. Available: https://appliednetsci.springeropen.com/articles/10.1007/s41109-019-0232-2

[5] "ICD - ICD-9-CM - International Classification of Diseases, Ninth Revision, Clinical Modification." [Online]. Available: https://www.cdc.gov/nchs/icd/icd9cm.htm

[6] M. E. Menendez, V. Neuhaus, C. N. Van Dijk, and D. Ring, "The Elixhauser comorbidity method outperforms the Charlson index in predicting inpatient death after orthopaedic surgery," *Clinical Orthopaedics and Related Research*, vol. 472, no. 9, pp. 2878–2886, 2014. [Online]. Available: https://pubmed.ncbi.nlm.nih.gov/24867450/

[7] C. Van Walraven, P. C. Austin, A. Jennings, H. Quan, and A. J. Forster, "A modification of the Elixhauser comorbidity measures into a point system for hospital death using administrative data," *Medical care*, vol. 47, no. 6, pp. 626–633, 6 2009. [Online]. Available: https://pubmed.ncbi.nlm.nih.gov/19433995/

[8] J. Liu, E. Larson, P. Zachariah, D. Caplan, B. . Senior, and J. Shang, "Comparison of Measures to Predict Mortality and Length of Stay in Hospitalized Patients."

[9] K. S. Ladha, K. Zhao, S. A. Quraishi, T. Kurth, M. Eikermann, H. M. Kaafarani, E. N. Klein, R. Seethala, and J. Lee, "The Deyo-Charlson and Elixhauser-van Walraven Comorbidity Indices as predictors of mortality in critically ill patients," *BMJ Open*, vol. 5, no. 9, p. e008990, 9 2015. [Online]. Available: https://bmjopen.bmj.com/content/5/9/e008990 https://bmjopen.bmj.com/content/5/9/e008990.abstract

[10] M. E. Charlson, P. Pompei, K. L. Ales, and C. R. MacKenzie, "A new method of classifying prognostic comorbidity in longitudinal studies: development and validation," *Journal of chronic diseases*, vol. 40, no. 5, pp. 373–383, 1987. [Online]. Available: https://pubmed.ncbi.nlm.nih.gov/3558716/

[11] R. A. Deyo, D. C. Cherkin, and M. A. Ciol, "Adapting a clinical comorbidity index for use with ICD-9-CM administrative databases," *Journal of clinical epidemiology*, vol. 45, no. 6, pp. 613–619, 1992. [Online]. Available: https://pubmed.ncbi.nlm.nih.gov/1607900/

[12] A. E. Johnson, T. J. Pollard, L. Shen, L. W. H. Lehman, M. Feng, M. Ghassemi, B. Moody, P. Szolovits, L. Anthony Celi, and R. G. Mark, "MIMIC-III, a freely accessible critical care database," *Scientific Data 2016 3:1*, vol. 3, no. 1, pp. 1–9, 5 2016. [Online]. Available: https://www.nature.com/articles/sdata201635

[13] N. Ding, C. Guo, C. Li, Y. Zhou, and X. Chai, "An Artificial Neural Networks Model for Early Predicting In-Hospital Mortality in Acute Pancreatitis in MIMIC-III," *BioMed Research International*, vol. 2021, 2021.

[14] M. Scherpf, F. Gräßer, H. Malberg, and S. Zaunseder, "Predicting sepsis with a recurrent neural network using the MIMIC III database," *Computers in Biology and Medicine*, vol. 113, p. 103395, 10 2019.

[15] V. Restocchi, J. G. Villegas, and J. D. Fleuriot, "Multimorbidity profiles and stochastic block modeling improve ICU patient clustering," *Proceedings - 22nd IEEE/ACM International Symposium on Cluster, Cloud and Internet Computing, CCGrid 2022*, pp. 925–932, 2022.

[16] J. Schrodt, A. Dudchenko, P. Knaup-Gregori, and M. Ganzinger, "Graph-Representation of Patient Data: a Systematic Literature Review," *Journal of Medical Systems*, vol. 44, no. 4, 4 2020. [Online]. Available: /pmc/articles/PMC7067737/ /pmc/articles/PMC7067737/?report=abstract https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7067737/

[17] T. Choudhury, R. Arunachalam, A. Khanna, E. Jasinska, V. Bolshev, V. Panchenko, and Z. Leonowicz, "A Social Network Analysis Approach to COVID-19 Community Detection Techniques," *International Journal of Environmental Research and Public Health*, vol. 19, no. 7, 4 2022. [Online]. Available: /pmc/articles/PMC8997780/ /pmc/articles/PMC8997780/?report=abstract https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8997780/

[18] C. Liu, F. Wang, J. Hu, and H. Xiong, "Temporal phenotyping from longitudinal Electronic Health Records: A graph based framework," *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, vol. 2015-August, pp. 705–714, 8 2015. [Online]. Available: https://dl.acm.org/doi/10.1145/2783258.2783352

[19] M. Rosvall, J.-C. Delvenne, M. T. Schaub, and R. Lambiotte, "Different approaches to community detection," *Advances in Network Clustering and Blockmodeling*, pp. 105–119, 12 2017. [Online]. Available: http://arxiv.org/abs/1712.06468 http://dx.doi.org/10.1002/9781119483298.ch4

[20] M. Gerlach, T. P. Peixoto, and E. G. Altmann, "A network approach to topic models," *Science Advances*, vol. 4, no. 7, 7 2018. [Online]. Available: https://www.science.org/doi/10.1126/sciadv.aaq1360

[21] X. Tang and C. C. Yang, "Detecting Social Media Hidden Communities Using Dynamic Stochastic Blockmodel with Temporal Dirichlet Process," *ACM Transactions on Intelligent Systems and Technology (TIST)*, vol. 5, no. 2, 4 2014. [Online]. Available: https://dl.acm.org/doi/10.1145/2517085

[22] E. M. Airoldi, D. M. Blei, S. E. Fienberg, and E. P. Xing, "Mixed Membership Stochastic Block Models for Relational Data with Application to Protein-Protein Interactions."

[23] Tiago P. Peixoto, "Graph tools."

[24] T. P. Peixoto, "Merge-split Markov chain Monte Carlo for community detection," *Physical Review E*, vol. 102, no. 1, 3 2020. [Online]. Available: https://arxiv.org/abs/2003.07070v4

[25] Z. Tu and S. C. Zhu, "Image segmentation by data-driven Markov Chain Monte Carlo," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 24, no. 5, pp. 657–673, 2002.

[26] B. Harrison, C. Purdy, and M. O. Riedl, "Toward Automated Story Generation with Markov Chain Monte Carlo Methods and Deep Neural Networks," *Proceedings of the AAAI Conference on Artificial Intelligence and Interactive Digital Entertainment*, vol. 13, no. 2, pp. 191–197, 6 2017. [Online]. Available: https://ojs.aaai.org/index.php/AIIDE/article/view/13003

[27] Tom Searle, "ICD9-CMS module python package."

[28] T. P. Peixoto, "Hierarchical Block Structures and High-resolution Model Selection in Large Networks," *Physical Review X*, vol. 4, no. 1, 10 2013. [Online]. Available: http://arxiv.org/abs/1310.4377 http://dx.doi.org/10.1103/PhysRevX.4.011047

[29] "martingerlach/hSBM_Topicmodel: Using stochastic block models for topic modeling." [Online]. Available: https://github.com/martingerlach/hSBM$_{T}opicmodel$

[30] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, A. Müller, J. Nothman, G. Louppe, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and Duchesnay, "Scikit-learn: Machine Learning in Python," 1 2012.

Appendix A

Bi partite network clusters

A.1 Boolean all icd9s

A.1.1 Top 15 clusters

(V498:Other specified conditions influencing health status:None, 1.0) 1.4926825869680564

(348:Other conditions of brain:None, 0.9213410702772404)
(378:Strabismus and other disorders of binocular eye movements:None, 0.032559638942617666)
(379:Other disorders of eye:None, 0.02869116698903933)
(377:Disorders of optic nerve and visual pathways:None, 0.017408123791102514)
1.3309392566396903

(570:Acute and subacute necrosis of liver:None, 1.0) 1.1793901877252353

(518:Other diseases of lung:None, 1.0) 0.9489059192337244

(286:Coagulation defects:None, 1.0) 0.9177023250763939

(785:Symptoms involving cardiovascular system:None, 1.0) 0.9124161734977879

(434:Occlusion of cerebral arteries:None, 1.0) 0.8840644900946553

(197:Secondary malignant neoplasm of respiratory and digestive systems:None, 0.5495689655172413)
(198:Secondary malignant neoplasm of other specified sites:None, 0.4504310344827586)
0.8251043693270308

(427:Cardiac dysrhythmias:None, 1.0) 0.7226021025427313

(331:Other cerebral degenerations:None, 0.6942875078468299)
(437:Other and ill-defined cerebrovascular disease:None, 0.21217827997489014)
(V452:Ventricular shunt status:Presence of cerebrospinal fluid drainage device, 0.04959196484620213)
(225:Benign neoplasm of brain and other parts of nervous system:None, 0.04394224733207784)
0.645270902093529

(584:Acute kidney failure:None, 1.0) 0.6397094898870879

(342:Hemiplegia and hemiparesis:None, 0.633175355450237)

(431:Intracerebral hemorrhage:None, 0.2672985781990521)
(430:Subarachnoid hemorrhage:None, 0.0995260663507109) 0.6365301605044799

(491:Chronic bronchitis:None, 0.6209453197405005)
(515:Postinflammatory pulmonary fibrosis:None, 0.31417979610750696)
(358:Myoneural disorders:None, 0.06487488415199258) 0.6265135442679464

(E888:Other and unspecified fall:None, 0.6250799744081894)
(E885:Accidental fall on same level from slipping tripping or stumbling:None, 0.37492002559181065)
0.5470116134245172

A.1.2 Bottom 15 clusters

(648:Other current conditions in the mother classifiable elsewhere but complicating pregnancy childbirth or the puerperium:None, 0.24898785425101214)
(V270:Deliver-single liveborn:Outcome of delivery, single liveborn, 0.0728744939271255)
(666:Postpartum hemorrhage:None, 0.06882591093117409)
(659:Other indications for care or intervention related to labor and delivery not elsewhere classified:None, 0.05263157894736842)
(642:Hypertension complicating pregnancy childbirth and the puerperium:None, 0.05263157894736842
(644:Early or threatened labor:None, 0.05060728744939271)
(646:Other complications of pregnancy not elsewhere classified:None, 0.048582995951417005)
(654:Abnormality of organs and soft tissues of pelvis:None, 0.04251012145748988)
(674:Other and unspecified complications of the puerperium not elsewhere classified:None, 0.04048582995951417)
(615:Inflammatory diseases of uterus except cervix:None, 0.038461538461538464)
(669:Other complications of labor and delivery not elsewhere classified:None, 0.032388663967611336)
(647:Infectious and parasitic conditions in the mother classifiable elsewhere but complicating pregnancy childbirth or the puerperium:None, 0.030364372469635626)
(649:Other conditions or status of the mother complicating pregnancy, childbirth, or the puerperium:None, 0.02631578947368421)
(670:Major puerperal infection:None, 0.020242914979757085)
(656:Other known or suspected fetal and placental problems affecting management of mother:None, 0.016194331983805668)
(665:Other obstetrical trauma:None, 0.016194331983805668)
(652:Malposition and malpresentation of fetus:None, 0.01417004048582996)
(639:Complications following abortion and ectopic and molar pregnancies:None, 0.012145748987854251
(641:Antepartum hemorrhage abruptio placentae and placenta previa:None, 0.012145748987854251)
(664:Trauma to perineum and vulva during delivery:None, 0.010121457489878543)
(V272:Deliver-twins, both live:Outcome of delivery, twins, both liveborn, 0.010121457489878543)
(655:Known or suspected fetal abnormality affecting management of mother:None, 0.008097165991902834)
(672:Pyrexia of unknown origin during the puerperium:None, 0.006072874493927126)
(V252:Sterilization:Sterilization, 0.006072874493927126)
(651:Multiple gestation:None, 0.006072874493927126)
(663:Umbilical cord complications during labor and delivery:None, 0.006072874493927126)
(635:Legally induced abortion:None, 0.004048582995951417)

(V271:Deliver-single stillborn:Outcome of delivery, single stillborn, 0.004048582995951417)
(658:Other problems associated with amniotic cavity and membranes:None, 0.004048582995951417)
(660:Obstructed labor:None, 0.004048582995951417)
(661:Abnormality of forces of labor:None, 0.004048582995951417)
(V181:Family history of other endocrine and metabolic diseases:None, 0.004048582995951417)
(668:Complications of the administration of anesthetic or other sedation in labor and delivery:None, 0.004048582995951417)
(121:Other trematode infections:None, 0.004048582995951417)
(V618:Family circumstances NEC:Other specified family circumstances, 0.004048582995951417)
(643:Excessive vomiting in pregnancy:None, 0.004048582995951417)
(V230:Preg w hx of infertility:Supervision of high-risk pregnancy with history of infertility, 0.0020242914979757085)
(V239:Suprv high-risk preg NOS:Supervision of unspecified high-risk pregnancy, 0.0020242914979757085)
(657:Polyhydramnios:None, 0.0020242914979757085)
(645:Late pregnancy:None, 0.0020242914979757085)
(V910:Twin gestation placenta status:None, 0.0020242914979757085) -1.5059617183528804

(411:Other acute and subacute forms of ischemic heart disease:None, 0.4784203102961918)
(413:Angina pectoris:None, 0.27193229901269395)
(V173:Fam hx-ischem heart dis:Family history of ischemic heart disease, 0.17376586741889985)
(746:Other congenital anomalies of heart:None, 0.07588152327221438) -1.442552266736624

(E950:Suicide and self-inflicted poisoning by solid or liquid substances:None, 0.386159169550173)
(E850:Accidental poisoning by analgesics antipyretics and antirheumatics:None, 0.10380622837370242
(969:Poisoning by psychotropic agents:None, 0.08858131487889273)
(965:Poisoning by analgesics antipyretics and antirheumatics:None, 0.0726643598615917)
(E980:Poisoning by solid or liquid substances undetermined whether accidentally or purposely inflicted:None, 0.056055363321799306)
(E854:Accidental poisoning by other psychotropic agents:None, 0.05259515570934256)
(E853:Accidental poisoning by tranquilizers:None, 0.03667820069204152)
(980:Toxic effect of alcohol:None, 0.02975778546712803)
(E855:Accidental poisoning by other drugs acting on central and autonomic nervous system:None, 0.019377162629757784)
(963:Poisoning by primarily systemic agents:None, 0.019377162629757784)
(970:Poisoning by central nervous system stimulants:None, 0.017993079584775088)
(966:Poisoning by anticonvulsants and anti-parkinsonism drugs:None, 0.017993079584775088)
(972:Poisoning by agents primarily affecting the cardiovascular system:None, 0.01660899653979239)
(E860:Accidental poisoning by alcohol not elsewhere classified:None, 0.01660899653979239)
(967:Poisoning by sedatives and hypnotics:None, 0.01522491349480969)
(E852:Accidental poisoning by other sedatives and hypnotics:None, 0.009688581314878892)
(977:Poisoning by other and unspecified drugs and medicinal substances:None, 0.006228373702422145
(968:Poisoning by other central nervous system depressants and anesthetics:None, 0.006228373702422145)
(971:Poisoning by drugs primarily affecting the autonomic nervous system:None, 0.006228373702422145)
(975:Poisoning by agents primarily acting on the smooth and skeletal muscles and

respiratory system:None, 0.005536332179930796)
(989:Toxic effect of other substances chiefly nonmedicinal as to source:None, 0.005536332179930796)
(962:Poisoning by hormones and synthetic substitutes:None, 0.004844290657439446)
(V614:Health problems within family:None, 0.004152249134948097)
(983:Toxic effect of corrosive aromatics acids and caustic alkalis:None, 0.0020761245674740486)
-1.0712689405295093

(787:Symptoms involving digestive system:None, 1.0) -0.9581544282375598

(455:Hemorrhoids:None, 0.5087209302325582)
(211:Benign neoplasm of other parts of digestive system:None, 0.49127906976744184)
-0.8451275170170894

(305:Nondependent abuse of drugs:None, 0.9297287113790504)
(V08:Asymptomatic human immunodeficiency virus [HIV] infection status:None, 0.04163526752072343)
(314:Hyperkinetic syndrome of childhood:None, 0.028636021100226075) -0.7978886654213461

(E870:Accidental cut puncture perforation or hemorrhage during medical care:None, 0.6439909297052154)
(552:Other hernia of abdominal cavity with obstruction but without mention of gangrene:None, 0.16780045351473924)
(614:Inflammatory disease of ovary fallopian tube pelvic cellular tissue and peritoneum:None, 0.11337868480725624)
(617:Endometriosis:None, 0.06575963718820861)
(551:Other hernia of abdominal cavity with gangrene:None, 0.009070294784580499)
-0.7924398132769419

(285:Other and unspecified anemias:None, 0.9909001206008113)
(726:Peripheral enthesopathies and allied syndromes:None, 0.009099879399188686)
-0.7414013124114299

(327:Organic sleep disorders:None, 1.0) -0.7280671859304481

(280:Iron deficiency anemias:None, 0.9166666666666666)
(282:Hereditary hemolytic anemias:None, 0.08333333333333333) -0.7136371206055536

(535:Gastritis and duodenitis:None, 0.7083333333333334)
(281:Other deficiency anemias:None, 0.2123015873015873)
(V113:Hx of alcoholism:Personal history of alcoholism, 0.07936507936507936) -0.6554083672941776

(296:Episodic mood disorders:None, 0.5287356321839081)
(292:Drug-induced mental disorders:None, 0.2220480668756531)
(309:Adjustment reaction:None, 0.15308254963427378)
(307:Special symptoms or syndromes not elsewhere classified:None, 0.0961337513061651)
-0.6134385871994231

(E878:Surgical operation and other surgical procedures as the cause of abnormal reaction of patient or of later complication without mention of misadventure at the time of operation:None, 1.0) -0.6103638155587235

(295:Schizophrenic disorders:None, 0.9325396825396826)
(E911:Inhalation and ingestion of food causing obstruction of respiratory tract or suffocation:None, 0.06746031746031746) -0.5451938653780912

(303:Alcohol dependence syndrome:None, 0.6923751095530236)
(291:Alcohol-induced mental disorders:None, 0.30762489044697633) -0.5286003769108019

Appendix B

Cooccurrence clusters

This appendix includes all the best clusters for the two significant assocaiton graphs.

B.1 All without sepsis and primary

B.1.1 Top 10 clusters

484:Pneumonia in infectious diseases classified elsewhere:None
902:Injury to blood vessels of abdomen and pelvis:None
1.2715061652038777

V716:Observ-inflicted inj NEC:Observation following other inflicted injury
V118:Hx-mental disorder NEC:Personal history of other mental disorders
V053:Need prphyl vc vrl hepat:Need for prophylactic vaccination and inoculation against viral hepatitis 988:Toxic effect of noxious substances eaten as food:None
E856:Accidental poisoning by antibiotics:None
760:Fetus or newborn affected by maternal conditions which may be unrelated to present pregnancy:None
V910:Twin gestation placenta status:None
992:Effects of heat and light:None
V016:Venereal dis contact:Contact with or exposure to venereal diseases
V902:Retain plastic fragments:Retained plastic fragments
125:Filarial infection and dracontiasis:None
V045:Vaccin for rabies:Need for prophylactic vaccination and inoculation against rabies
E908:Accident due to cataclysmic storms and floods resulting from storms:None
E800:Railway accident involving collision with rolling stock:None
V901:Retained metal fragments:None
V601:Inadequate housing:Inadequate housing 232:Carcinoma in situ of skin:None
915:Superficial injury of finger(s):None V608:Other specified housing or economic circumstances:None
848:Other and ill-defined sprains and strains:None V568:Dialysis encounter, NEC:Encounter for other dialysis 0.9144036788125248

049:Other non-arthropod-borne viral diseases of central nervous system:None

077:Other diseases of conjunctiva due to viruses and chlamydiae:None
V430:Eye replacement NEC:Eye globe replaced by other means 757:Congenital anomalies of the integument:None V694:Lack of adequate sleep:Lack of adequate sleep 551:Other hernia of abdominal cavity with gangrene:None
E988:Injury by other and unspecified means undetermined whether accidentally or purposely inflicted:None
046:Slow virus infection and prion diseases of central nervous system:None
639:Complications following abortion and ectopic and molar pregnancies:None
149:Malignant neoplasm of other and ill-defined sites within the lip oral cavity and pharynx:None 944:Burn of wrist(s) and hand(s):None
E831:Accident to watercraft causing other injury:None V530:Fitting and adjustment of devices related to nervous system and special senses:None
094:Neurosyphilis:None
306:Physiological malfunction arising from mental factors:None
039:Actinomycotic infections:None
0.7106680512918683

312:Disturbance of conduct not elsewhere classified:None
E955:Suicide and self-inflicted injury by firearms air guns and explosives:None
956:Injury to peripheral nerve(s) of pelvic girdle and lower limb:None
586:Renal failure, unspecified:None
V169:Family hx-malignancy NOS:Family history of unspecified malignant neoplasm
375:Disorders of lacrimal system:None
754:Certain congenital musculoskeletal deformities:None
0.6914754360797669

826:Fracture of one or more phalanges of foot:None
644:Early or threatened labor:None
854:Intracranial injury of other and unspecified nature:None
480:Viral pneumonia:None
0.6773615587140907

891:Open wound of knee leg (except thigh) and ankle:None
205:Myeloid leukemia:None
0.6760031211015498

900:Injury to blood vessels of head and neck:None
E956:Suicide and self-inflicted injury by cutting and piercing instrument:None
151:Malignant neoplasm of stomach:None
0.6189881597115917

490:Bronchitis, not specified as acute or chronic:None
485:Bronchopneumonia, organism unspecified:None
239:Neoplasms of unspecified nature:None
581:Nephrotic syndrome:None
0.5692642630944263

272:Disorders of lipoid metabolism:None
276:Disorders of fluid electrolyte and acid-base balance:None

518:Other diseases of lung:None
285:Other and unspecified anemias:None
401:Essential hypertension:None
427:Cardiac dysrhythmias:None
0.564939885636589

V850:BMI less than 19,adult:Body Mass Index less than 19, adult
V113:Hx of alcoholism:Personal history of alcoholism
E882:Accidental fall from or out of building or other structure:None
0.5078259356032

B.1.2 Bottom 10 clusters

907:Late effects of injuries to the nervous system:None
815:Fracture of metacarpal bone(s):None
-1.6114244934662316

595:Cystitis:None
394:Diseases of mitral valve:None
-1.33422866780642

831:Dislocation of shoulder:None
720:Ankylosing spondylitis and other inflammatory spondylopathies:None
-1.1557033199875202

534:Gastrojejunal ulcer:None
540:Acute appendicitis:None
E855:Accidental poisoning by other drugs acting on central and autonomic nervous system:None
970:Poisoning by central nervous system stimulants:None
245:Thyroiditis:None
966:Poisoning by anticonvulsants and anti-parkinsonism drugs:None
-1.1227551288751891

916:Superficial injury of hip thigh leg and ankle:None -1.1185007164593597

233:Carcinoma in situ of breast and genitourinary system:None
647:Infectious and parasitic conditions in the mother classifiable elsewhere but complicating pregnancy childbirth or the puerperium:None
526:Diseases of the jaws:None
706:Diseases of sebaceous glands:None
308:Acute reaction to stress:None
669:Other complications of labor and delivery not elsewhere classified:None
-1.1102378895324112

488:Influenza due to certain identified influenza viruses:None
448:Disease of capillaries:None
832:Dislocation of elbow:None

445:Atheroembolism:None
834:Dislocation of finger:None
E824:Other motor vehicle nontraffic accident while boarding and alighting:None
520:Disorders of tooth development and eruption:None
971:Poisoning by drugs primarily affecting the autonomic nervous system:None
-1.0851339370475563

B.2 Non nested hsbm elixhauser features - association graph

B.2.1 Top 10 clusters

276:Disorders of fluid electrolyte and acid-base balance:None 0.714087270897209

427:Cardiac dysrhythmias:None 0.6314383074843525

496:Chronic airway obstruction, not elsewhere classified:None
780:General symptoms:None
197:Secondary malignant neoplasm of respiratory and digestive systems:None
196:Secondary and unspecified malignant neoplasm of lymph nodes:None
286:Coagulation defects:None
331:Other cerebral degenerations:None
996:Complications peculiar to certain specified procedures:None 0.6256294439475076

311:Depressive disorder, not elsewhere classified:None
190:Malignant neoplasm of eye:None
263:Other and unspecified protein-calorie malnutrition:None
None
287:Purpura and other hemorrhagic conditions:None
493:Asthma:None
181:Malignant neoplasm of placenta:None
425:Cardiomyopathy:None
278:Overweight, obesity and other hyperalimentation:None
None
None
244:Acquired hypothyroidism:None
426:Conduction disorders:None
571:Chronic liver disease and cirrhosis:None
None
V56:Encounter for dialysis and dialysis catheter care:None
None
424:Other diseases of endocardium:None
505:Pneumoconiosis, unspecified:None
785:Symptoms involving cardiovascular system:None
V450:Cardiac device in situ:None
143:Malignant neoplasm of gum:None
416:Chronic pulmonary heart disease:None

443:Other peripheral vascular disease:None
348:Other conditions of brain:None
165:Malignant neoplasm of other and ill-defined sites within the respiratory system and intrathoracic organs:None
None
093:Cardiovascular syphilis:None
503:Pneumoconiosis due to other inorganic dust:None
070:Viral hepatitis:None
None
504:Pneumonopathy due to inhalation of other dust:None
202:Other malignant neoplasms of lymphoid and histiocytic tissue:None
300:Anxiety, dissociative and somatoform disorders:None
345:Epilepsy and recurrent seizures:None
None
570:Acute and subacute necrosis of liver:None
291:Alcohol-induced mental disorders:None
403:Hypertensive chronic kidney disease:None
293:Transient mental disorders due to conditions classified elsewhere:None
415:Acute pulmonary heart disease:None
243:Congenital hypothyroidism:None
456:Varicose veins of other sites:None
303:Alcohol dependence syndrome:None
280:Iron deficiency anemias:None
572:Liver abscess and sequelae of chronic liver disease:None
440:Atherosclerosis:None 0.4241625626420227

531:Gastric ulcer:None
501:Asbestosis:None
446:Polyarteritis nodosa and allied conditions:None
340:Multiple sclerosis:None
261:Nutritional marasmus:None
447:Other disorders of arteries and arterioles:None
281:Other deficiency anemias:None
198:Secondary malignant neoplasm of other specified sites:None
588:Disorders resulting from impaired renal function:None
344:Other paralytic syndromes:None
333:Other extrapyramidal disease and abnormal movement disorders:None
533:Peptic ulcer site unspecified:None
710:Diffuse diseases of connective tissue:None
342:Hemiplegia and hemiparesis:None
728:Disorders of muscle ligament and fascia:None 0.252363533014831

V420:Kidney transplant status:Kidney replaced by transplant
240:Simple and unspecified goiter:None
532:Duodenal ulcer:None
189:Malignant neoplasm of kidney and other and unspecified urinary organs:None
200:Lymphosarcoma and reticulosarcoma and other specified malignant tumors of

lymphatic tissue:None
V422:Heart valve transplant:Heart valve replaced by transplant
188:Malignant neoplasm of bladder:None
V533:Fitting and adjustment of cardiac device:None
150:Malignant neoplasm of esophagus:None
199:Malignant neoplasm without specification of site:None
402:Hypertensive heart disease:None
334:Spinocerebellar disease:None
720:Ankylosing spondylitis and other inflammatory spondylopathies:None
V113:Hx of alcoholism:Personal history of alcoholism
157:Malignant neoplasm of pancreas:None
335:Anterior horn cell disease:None
404:Hypertensive heart and chronic kidney disease:None
494:Bronchiectasis:None
336:Other diseases of spinal cord:None
174:Malignant neoplasm of female breast:None
262:Other severe protein-calorie malnutrition:None
725:Polymyalgia rheumatica:None
711:Arthropathy associated with infections:None
154:Malignant neoplasm of rectum rectosigmoid junction and anus:None 0.2145587923644663

201:Hodgkin's disease:None 0.19454173218746038

164:Malignant neoplasm of thymus heart and mediastinum:None
701:Other hypertrophic and atrophic conditions of skin:None
298:Other nonorganic psychoses:None
297:Delusional disorders:None
508:Respiratory conditions due to other and unspecified external agents:None
153:Malignant neoplasm of colon:None
195:Malignant neoplasm of other and ill-defined sites:None
191:Malignant neoplasm of brain:None
490:Bronchitis, not specified as acute or chronic:None
171:Malignant neoplasm of connective and other soft tissue:None
161:Malignant neoplasm of larynx:None
534:Gastrojejunal ulcer:None
184:Malignant neoplasm of other and unspecified female genital organs:None
417:Other diseases of pulmonary circulation:None
V654:Other counseling not elsewhere classified:None
341:Other demyelinating diseases of central nervous system:None
343:Infantile cerebral palsy:None
147:Malignant neoplasm of nasopharynx:None
151:Malignant neoplasm of stomach:None
193:Malignant neoplasm of thyroid gland:None
182:Malignant neoplasm of body of uterus:None
500:Coal workers' pneumoconiosis:None
395:Diseases of aortic valve:None
246:Other disorders of thyroid:None

405:Secondary hypertension:None
V427:Liver transplant status:Liver replaced by transplant
394:Diseases of mitral valve:None
203:Multiple myeloma and immunoproliferative neoplasms:None
172:Malignant melanoma of skin:None
156:Malignant neoplasm of gallbladder and extrahepatic bile ducts:None
158:Malignant neoplasm of retroperitoneum and peritoneum:None
183:Malignant neoplasm of ovary and other uterine adnexa:None 0.18598540633956123

194:Malignant neoplasm of other endocrine glands and related structures:None
398:Other rheumatic heart disease:None
179:Malignant neoplasm of uterus, part unspecified:None 0.12329376619282272

502:Pneumoconiosis due to other silica or silicates:None
142:Malignant neoplasm of major salivary glands:None
146:Malignant neoplasm of oropharynx:None
170:Malignant neoplasm of bone and articular cartilage:None
180:Malignant neoplasm of cervix uteri:None
V434:Blood vessel replac NEC:Blood vessel replaced by other means
506:Respiratory conditions due to chemical fumes and vapors:None
141:Malignant neoplasm of tongue:None
163:Malignant neoplasm of pleura:None
140:Malignant neoplasm of lip:None
260:Kwashiorkor:None
160:Malignant neoplasm of nasal cavities middle ear and accessory sinuses:None
149:Malignant neoplasm of other and ill-defined sites within the lip oral cavity and pharynx:None
148:Malignant neoplasm of hypopharynx:None
192:Malignant neoplasm of other and unspecified parts of nervous system:None
495:Extrinsic allergic alveolitis:None
145:Malignant neoplasm of other and unspecified parts of mouth:None
186:Malignant neoplasm of testis:None
586:Renal failure, unspecified:None
265:Thiamine and niacin deficiency states:None
152:Malignant neoplasm of small intestine including duodenum:None
176:Kaposi's sarcoma:None 0.11945972278887447

B.2.2 Bottom 10 clusters

305:Nondependent abuse of drugs:None
250:Diabetes mellitus:None
401:Essential hypertension:None
428:Heart failure:None -0.20425498553312474

573:Other disorders of liver:None
535:Gastritis and duodenitis:None
292:Drug-induced mental disorders:None
296:Episodic mood disorders:None

799:Other ill-defined and unknown causes of morbidity and mortality:None
783:Symptoms concerning nutrition metabolism and development:None
719:Other and unspecified disorders of joint:None
585:Chronic kidney disease (ckd):None
042:Human immunodeficiency virus [HIV] disease:None
V433:Heart valve replac NEC:Heart valve replaced by other means
491:Chronic bronchitis:None
253:Disorders of the pituitary gland and its hypothalamic control:None
437:Other and ill-defined cerebrovascular disease:None
397:Diseases of other endocardial structures:None
304:Drug dependence:None
295:Schizophrenic disorders:None
238:Neoplasm of uncertain behavior of other and unspecified sites and tissues:None
714:Rheumatoid arthritis and other inflammatory polyarthropathies:None -0.12418584091556974

980:Toxic effect of alcohol:None
187:Malignant neoplasm of penis and other male genital organs:None
144:Malignant neoplasm of floor of mouth:None
159:Malignant neoplasm of other and ill-defined sites within the digestive organs and peritoneum:None
175:Malignant neoplasm of male breast:None -0.1234571320235372

557:Vascular insufficiency of intestine:None
357:Inflammatory and toxic neuropathy:None
784:Symptoms involving head and neck:None
309:Adjustment reaction:None
185:Malignant neoplasm of prostate:None
162:Malignant neoplasm of trachea bronchus and lung:None
155:Malignant neoplasm of liver and intrahepatic bile ducts:None
332:Parkinson's disease:None
492:Emphysema:None
396:Diseases of mitral and aortic valves:None
V451:Postsurgical renal dialysis status:None
746:Other congenital anomalies of heart:None
441:Aortic aneurysm and dissection:None
729:Other disorders of soft tissues:None -0.0342502608381811

502:Pneumoconiosis due to other silica or silicates:None
142:Malignant neoplasm of major salivary glands:None
146:Malignant neoplasm of oropharynx:None
170:Malignant neoplasm of bone and articular cartilage:None
180:Malignant neoplasm of cervix uteri:None
V434:Blood vessel replac NEC:Blood vessel replaced by other means
506:Respiratory conditions due to chemical fumes and vapors:None
141:Malignant neoplasm of tongue:None
163:Malignant neoplasm of pleura:None
140:Malignant neoplasm of lip:None
260:Kwashiorkor:None

160:Malignant neoplasm of nasal cavities middle ear and accessory sinuses:None
149:Malignant neoplasm of other and ill-defined sites within the lip oral cavity and pharynx:None
148:Malignant neoplasm of hypopharynx:None
192:Malignant neoplasm of other and unspecified parts of nervous system:None
495:Extrinsic allergic alveolitis:None
145:Malignant neoplasm of other and unspecified parts of mouth:None
186:Malignant neoplasm of testis:None
586:Renal failure, unspecified:None
265:Thiamine and niacin deficiency states:None
152:Malignant neoplasm of small intestine including duodenum:None
176:Kaposi's sarcoma:None 0.11945972278887447

194:Malignant neoplasm of other endocrine glands and related structures:None
398:Other rheumatic heart disease:None
179:Malignant neoplasm of uterus, part unspecified:None 0.12329376619282272

164:Malignant neoplasm of thymus heart and mediastinum:None
701:Other hypertrophic and atrophic conditions of skin:None
298:Other nonorganic psychoses:None
297:Delusional disorders:None
508:Respiratory conditions due to other and unspecified external agents:None
153:Malignant neoplasm of colon:None
195:Malignant neoplasm of other and ill-defined sites:None
191:Malignant neoplasm of brain:None
490:Bronchitis, not specified as acute or chronic:None
171:Malignant neoplasm of connective and other soft tissue:None
161:Malignant neoplasm of larynx:None
534:Gastrojejunal ulcer:None
184:Malignant neoplasm of other and unspecified female genital organs:None
417:Other diseases of pulmonary circulation:None
V654:Other counseling not elsewhere classified:None
341:Other demyelinating diseases of central nervous system:None
343:Infantile cerebral palsy:None
147:Malignant neoplasm of nasopharynx:None
151:Malignant neoplasm of stomach:None
193:Malignant neoplasm of thyroid gland:None
182:Malignant neoplasm of body of uterus:None
500:Coal workers' pneumoconiosis:None
395:Diseases of aortic valve:None
246:Other disorders of thyroid:None
405:Secondary hypertension:None
V427:Liver transplant status:Liver replaced by transplant
394:Diseases of mitral valve:None
203:Multiple myeloma and immunoproliferative neoplasms:None
172:Malignant melanoma of skin:None
156:Malignant neoplasm of gallbladder and extrahepatic bile ducts:None

158:Malignant neoplasm of retroperitoneum and peritoneum:None
183:Malignant neoplasm of ovary and other uterine adnexa:None 0.18598540633956123

201:Hodgkin's disease:None 0.19454173218746038

V420:Kidney transplant status:Kidney replaced by transplant
240:Simple and unspecified goiter:None
532:Duodenal ulcer:None
189:Malignant neoplasm of kidney and other and unspecified urinary organs:None
200:Lymphosarcoma and reticulosarcoma and other specified malignant tumors of lymphatic tissue:None
V422:Heart valve transplant:Heart valve replaced by transplant
188:Malignant neoplasm of bladder:None
V533:Fitting and adjustment of cardiac device:None
150:Malignant neoplasm of esophagus:None
199:Malignant neoplasm without specification of site:None
402:Hypertensive heart disease:None
334:Spinocerebellar disease:None
720:Ankylosing spondylitis and other inflammatory spondylopathies:None
V113:Hx of alcoholism:Personal history of alcoholism
157:Malignant neoplasm of pancreas:None
335:Anterior horn cell disease:None
404:Hypertensive heart and chronic kidney disease:None
494:Bronchiectasis:None
336:Other diseases of spinal cord:None
174:Malignant neoplasm of female breast:None
262:Other severe protein-calorie malnutrition:None
725:Polymyalgia rheumatica:None
711:Arthropathy associated with infections:None
154:Malignant neoplasm of rectum rectosigmoid junction and anus:None 0.2145587923644663

531:Gastric ulcer:None
501:Asbestosis:None
446:Polyarteritis nodosa and allied conditions:None
340:Multiple sclerosis:None
261:Nutritional marasmus:None
447:Other disorders of arteries and arterioles:None
281:Other deficiency anemias:None
198:Secondary malignant neoplasm of other specified sites:None
588:Disorders resulting from impaired renal function:None
344:Other paralytic syndromes:None
333:Other extrapyramidal disease and abnormal movement disorders:None
533:Peptic ulcer site unspecified:None
710:Diffuse diseases of connective tissue:None
342:Hemiplegia and hemiparesis:None
728:Disorders of muscle ligament and fascia:None 0.252363533014831