Discovering Type 1 Diabetes Patient Subgroups through Integrative Analysis of Heterogeneous Data

by

© *S. Sadra Mirhendi*

A thesis submitted to the

School of Graduate Studies

in partial fulfilment of the requirements for the degree of

Master of *Computer Science*

Department of *Computer Science*

Memorial University of Newfoundland

May 2018

St. John's Newfoundland

Abstract

Type 1 diabetes (T1D) is a disease in which the body immune system attacks the β-cells. As a result, very little, or no insulin is released to control the level of glucose in the blood. Our research investigates whether groups of patients at higher risk for developing T1D complications can be identified by integrating demographic, clinical and genetic data. Regarding this purpose, we explore two methods including Generalized Low Rank Models (GLRM) and Similarity Network Fusion (SNF) to investigate our T1D dataset and to determine groups of patients at higher risk of developing complications related to T1D.

By applying the stated methods, we have identified groups of patients suffering from nerve damage, high blood pressure, dyslipidemia, and thyroid diseases. This result could be used as the basis to achieve a predictive model that could allow patients and health-care providers to take preemptive steps to reduce the risk of developing T1D related complications.

Acknowledgements

Grateful to the God, completing the master has been an incredibly rewarding experience, and I am very grateful to have had the opportunity to finish this program.

I would like to thank my amazing supervisor Dr. Lourdes Peña-Castillo for her tremendous support and guidance. She created an open and supportive working environment where she was always available and willing to help. Her enthusiasm and incredible work ethic were both inspiring and motivating and I consider myself very fortunate to have had the opportunity to study under her supervision. I would also like to thank my co-supervisor Dr. Ting Hu for her support and guidance.

Additionally, I would like to express my sincere gratitude to my parents for their unconditional support and my wonderful wife for her endless love. I am very fortunate to have such wonderful people in my life and I truly value the encouragement and faith that they have always shown in me.

— S. Sadra Mirhendi

Contents

List of Tables

List of Figures

Chapter 1

Introduction

Diabetes mellitus type 1 also known as type 1 diabetes (T1D) is a disease in which the body immune system attacks the β-cells. As a result, very little, or no insulin is released to control the level of glucose in the blood. Thus the amount of glucose obtained from foods will be built up in the body instead of being used for energy.

This research investigates whether groups of patients at a higher risk for developing complications or secondary disease related to T1D can be identified by integrating demographic, clinical and genetic data. We have a T1D dataset which contains 239 features concerning demographic, clinical and genetic factors from 196 patients (details are available in Section 1.4). We will explore two methods including Generalized Low Rank Modelling (GLRM) [1] and Similarity Network Fusion (SNF) [2] to analyze this dataset.

As a result of our research, we have taken first steps to identify groups of patients at higher risk of developing T1D complications. This results could be taken as the basis to develop a predictive model that could allow patients and health-care providers

to take preemptive steps to reduce the risk of developing T1D related complications based on each patient characteristics

To sum up, we have a heterogeneous dataset that contains demographic, clinical and genetic data from T1D patients. Our research goal is to determine groups of patients at higher risk of developing complications or secondary disease related to T1D by analyzing this dataset.

1.1 Biological Background

T1D is not a preventable disease and is not related to eating an excessive amount of sugar. Scientists could not determine any particular agent for the cause of T1D yet [3]. Many factors may contribute to T1D, including genetic susceptibility and exposure to specific antigens. Hence, T1D is considered a "complex disease" which a combination of numerous risk factors may lead to it. T1D occurs when the body's immune system destroys the β-cells in the pancreas [4, 5]. T1D is a polygenic disorder, with about 50 loci so far known to influence this disease susceptibility [6]. In this research we investigate following complications related to T1D.

Thyroid Disease: Thyroid disease affects the thyroid gland which controls various metabolic processes in the body. Thyroid dysfunction in patients with T1D is two - to three fold higher than in the general population [7, 8].

Dyslipidemia: Dyslipidemia is an abnormal amount of lipids in the blood. People with T1D have increased rates of vascular disease in which dyslipidemia is a major risk factor [9].

High Blood Pressure: High blood pressure is common in people with diabetes and about 25% of people with T1D develop high blood pressure at some stage. Having diabetes and high blood pressure together, increases risk of other health problems [10].

Nerve Damage: Nerve damage can occur in people with T1D which is called Diabetic neuropathy. Depending on the types of nerve damage it causes different symptoms. More than 50% of all diabetics patients suffer from some types of nerve damage [11].

Retinopathy: Retinopathy is the impair to the retina of eyes, which may leads to vision problems. During the first two decades of T1D disease, nearly all patients suffers from diabetic retinopathy [12].

Diabetic ketoacidosis: Diabetic ketoacidosis occurs when the body produces high levels of blood acids called ketones. Diabetic ketoacidosis (DKA) are common serious complications of T1D [13].

Hyperglycemia: Hyperglycemia is a condition in which an excessive amount of glucose is in the blood plasma. Low insulin levels in T1D patients cause hyperglycemia [14].

Hypoglycemia: Hypoglycemia is when blood sugar decreases to below normal levels in the blood. It is a common and dangerous occurrence with T1D patients [15].

Anxiety and Depression: Mental health problems are frequent in youth with T1D, and they are at an increased risk of mental health conditions, such as anxiety, eating and behavioral disorders, as well as depressive symptoms [16].

1.2 Related Works

We review related research to this study from three different approaches: 1- Type 1 Diabetes genetics, 2- Heterogeneous data analyzing challenges, 3- Patients subgroup discovery strategies.

1.2.1 Type 1 Diabetes Genetics

There are many papers published in the literature about T1D. T1D is one of the most common chronic diseases of childhood [17]. Genetic studies of T1D have identified 50 loci (susceptibility regions) that affect risk of T1D [6, 18, 19]. Atkinson et al. [20] released a survey that reviews current flow in epidemiology, pathology, diagnosis, and treatment of T1D, and its prospects for an improved future for individuals dealing with this disorder. Davies et al. [21] searched the human genome for genes that influence T1D. Additionally, Barrot et al. [22] reported findings of a genome-wide association study of T1D, combined in a meta-analysis. Roizen et al. [23] compared variants associated with increased risk for T1D with those variants identified in other autoimmune diseases and revealed genetic overlap between T1D and other autoimmune diseases.

1.2.2 Heterogeneous Data Challenges

Integrating data from different sources such as clinical, environmental, and demographic data with genomic data is an ongoing part of current research in genomics. In our research, we have chosen to use two mentioned approaches (GLRM and SNF) which are able to handle heterogeneous data. However, we acknowledge that there

are several efforts underway to deal with heterogeneous data. For example, Hamid et al. [24] proposed a conceptual framework for integrating data as well as a review of current approaches for combining genomic data. As another example, Ren et al. [25] evaluated the possible challenges in the integrative analysis of the heterogeneous disease data types. They proposed a computational method (named iBFE) based on a feature extraction perspective. They showed that iBFE could recognize disease subtypes in genomic data.

1.2.3 Patients Subgroup Discovery

Diagnosing and defining subtypes is a difficult challenge for complex diseases. Higdon et al. [26] described how different disease subtypes could be identified through the combination of clinical and multi-omics data. In the article, they clustered various types of omics data and then, the results were integrated with clinical data to identify disease subtypes. They applied this method to Autism Spectrum Disorder (ASD) to facilitate subtype identification.

1.2.3.1 Network Approaches

A computational framework is presented by Zhang et al. [27] to stratify a biological network into function-specific network layers, which transform the network analysis from gene level to the functional level by integrating expression data, the gene/protein network and gene ontology information.

A large scale of studies in complex disease is classifying patients based on their genomic mutations, but these mutations are rarely shared across patients for some diseases. Zhong et al. [28] used network-based stratification approaches on thirteen

major cancer types to classify tumours based on exome-level mutations.

Beforehand, the most common approach to integrative data analyzing was a separate clustering followed by a manual integration. Shen et al. [29] developed a joint latent variable model for integrative clustering (called iCluster). They could identify subtypes in breast cancer and lung cancer, characterized by concordant DNA copy number changes and gene expression using the iCluster algorithm. Kim et al. [30] proposed a method to improve feature selection on iCluster factor model using prior knowledge of inter-omics regulatory flows.

Hillenmeyer et al. [31] used a combination of an algorithm for weighted-edge module searching and a probabilistic interaction network in order to explain a method for designating genes with strong associations to the phenotype.

Cho et al. [32] showed how networks could be used to represent clinical data such as genotype and gene expression to distinguish dysregulated pathways and to understand the connections between genotype and phenotype, and to explain disease heterogeneity. Their article showed how to analyze complex disease using similarity network fusions since genetic variations in affected individuals might be different. Wang et al. [2] used similarity network fusions for disease data obtained from a group of patients. Yang et al. [33] proposed an integrative method based on Similarity Network Fusion (SNF), named ndmaSNF (network diffusion model assisted SNF). This method can be used for cancer subtype discovery with making use of somatic mutation data and other discrete data.

Wang et al. [34] proposed a network-based approach for the integrative analysis of heterogeneous omics data. They represented a network-based solution in which each type of data is treated independently and tested the method on the subtypes

identification of a brain tumor.

1.2.3.2 Machine Learning Approaches

Speicher et al. [35] extended multiple kernel learning for dimensionality reduction. They could identify biologically meaningful subgroups for five different cancer types.

Schuler et al. [36] applied Generalized Low Rank Modelling presented by Udell et al. [1] to discover phenotypes in two datasets of patient information related to two different diseases. The method is used to overcome barriers such as missing data, data sparsity, and data heterogeneity in input data. As shown in this paper, the result of GLRM method is remarkably different in comparison to other machine learning methods in applications of discovering patient phenotypes.

Young et al. [37] used unsupervised deep learning to learn the hierarchical structure of cancer gene expression data. They showed that a deep learning model can be trained to represent biologically and clinically important concepts of cancer genes. Lasko et al. [38] introduced new deep learning methods used for phenotype discovery in clinical data.

Wei et al. [39] tested Support Vector Machine (SVM) on three large-scale GWAS dataset generated on the Affymetrix genotyping platform for type 1 diabetes (T1D) and demonstrated a risk assessment for this disorder.

1.3 Research Question

This research is an interdisciplinary study across computer science, molecular biology, and medicine. The aim of conducting this research is to improve knowledge regarding

complications associated with T1D and its risk factors and, to eventually achieve an efficient, trusted preemptive strategy for T1D.

Our research is distinct from previous research by three different aspects: 1- We have a unique dataset from T1D patients that comprise demographic, clinical and genetic data, from both diagnosis stage and patients current stage. 2- We are using two state-of-the-art methods to identify patient subgroups. 3- We include several T1D complications instead of focusing on a single complication. To investigate our T1D dataset, we apply two methods namely Generalized Low Ranks Modelling (GLRM), and Similarity Network Fusion (SNF). GLRM advantages are handling missing values and compressing data. SNF profits from capturing both shared and complementary information in the fused network. Our results can be used to identify patients at higher risk of developing T1D complications and could be taken as the basis to create a predictive model of developing T1D complications.

1.4 Dataset Overview

Our dataset is collected from a cohort study by Newhook et al. [40] regarding the incidence of childhood T1D in children aged 0-14 years who were diagnosed with T1D on the Avalon Peninsula of Newfoundland, Canada. Subjects for this study were a cohort of individuals with T1D who participated in a genetics study between 2001 and 2006. At the time of that study, demographic and clinical information from each individual had been entered in the Newfoundland and Labrador Diabetes Genetics Database and was used as a basis for patient contact. Later, given the passage of time, the most up to date demographic information about the cohort was collected.

According to Newhook et al. [40] paper "the Avalon Peninsula of Newfoundland has one of the highest incidences of T1D reported worldwide". The obtained raw T1D dataset comprises 239 features concerning demographical, clinical and genotype factors from 196 patients. This dataset is heterogeneous and have missing values especially in its genotype information. These imperfections drive us to perform the following pre-processing steps before applying the two selected methods.

Irrelevant features elimination: Not all of 239 features in the raw dataset is related to our research. We eliminate 25 irrelevant features such as "number of patient visits to the hospital" or "patient insurance status".

Sparse raws elimination: In the raw dataset, we have overall 12724 (27.2%) missing entries. We eliminate 43 patients (rows) which have more than 50% missing entries to reduce data sparsity.

Complications matrix extraction: 20 features are representing patients complication data. We extracted these columns and named them as complications matrix. Complications (columns) with more than 75% missing entries are eliminated from the obtained matrix. We consider that minimum sample size to analyze a complication is at least ten patients; thus complications with less than eleven patients are excluded. Eventually, we obtained a 153×10 complications matrix. Missing values in complications are replaced with zero (healthy status).

Substituting values and merging categories: Some of the raw dataset entries are text. For each unique string, a number is assigned and strings were replaced

by their corresponding number for better computational handling. Additionally, in each column, low population categories are merged to achieve a larger category. Additionally, patients date of birth is converted to age and its calculated until 2016.

Genotype features: We have 122 genotype features in total which 98 of them are categorical with three categories (homozygous, heterozygous type 1, heterozygous type 2) in each column. However, the rest (24 columns) are categorical while each data entry holds the position of an allele in the chromosome. We rearranged these features into binary features which whether they have an allele in the given position or not. This results in removing 24 features and adding 330 binary genotype features.

Sparse raws elimination: For the final step, we eliminate 64 features (columns) which have more than 30% missing entries to reduce data sparsity. This number is found empirically as a compromise between the number of rows eliminated and amount of missing data.

Following the mentioned steps we have a 153×436 T1D pre-processed dataset and a 153×10 complications data matrix. We use the T1D pre-processed dataset as our input data. The complications matrix is used for evaluating the obtained clusters to identify clusters with higher incidence of having complications. Table 1.1 represents a summary of pre-processed T1D dataset features and Appendix A includes all dataset features details and their specification.

Table 1.1: T1D Dataset Features Summary

Features Category	Features Type	No. of Features
Patient Clinical Data	Ordinal, Numeric, Binary	24
Relative Clinical Data	Binary	24
Patient Demographic Data	Numeric	4
Patient Genotype Data	Categorical, Binary	384

1.5 Study Overview

In this chapter, we introduced T1D and provided an overview of our input dataset. In the second chapter, we describe the basis of GLRM and the methods used to identify over enriched clusters. Then we present GLRM result and discuss its outcomes. In the third chapter, we present the principles of SNF, then we present the achieved result and discuss its features. Finally, in the last chapter, we give a comprehensive summary of our research. Figure 1.1 illustrates our work-flow and thesis organization.

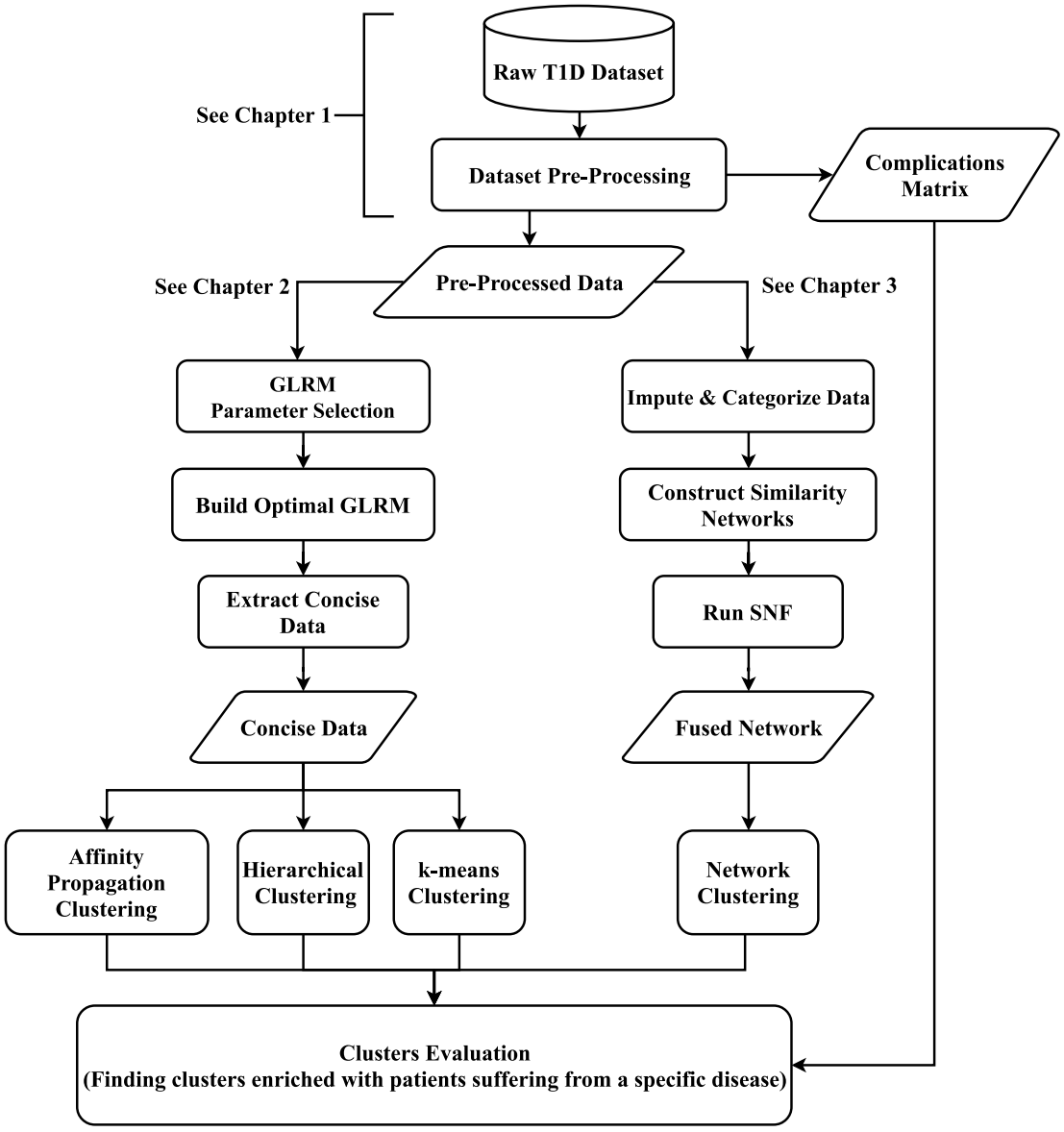

Figure 1.1: Research work flow diagram

Chapter 2

Generalized Low Rank Modeling

2.1 Introduction

Artificial Intelligence (AI) and Machine Learning (ML) methods have been used extensively for analyzing health records data and patient stratification. Researchers dealing with these methods are thwarted by imperfect data characteristics such as missing data records, mixed type of features, heterogeneity, and sparsity. We use Generalized Low Rank Modeling (GLRM) as a framework for analyzing our pre-processed Type 1 Diabetes (T1D) dataset described in Section 1.4. Unlike typical machine learning algorithms, this framework offers flexible solutions to overcome data barriers such as missing data and heterogeneity. The GLRM framework was first introduced by Udell et al. [1]. It extends Principle Component Analysis (PCA) technique [41] to design a framework which can handle heterogeneous data with mixed feature types (numerical and categorical). This framework transforms high-dimensional data into lower dimension space by solving an optimization problem.

Prior to applying the GLRM to the T1D raw dataset, we perform data preprocessing steps discussed in Section 1.4 including merging, replacing and segregating features values as well as eliminating sparse samples. Following the data preparation step, we apply the GLRM method to the pre-processed T1D dataset and optimize its parameters using cross-validation. Consequently, the low dimensional data with minimum error obtained after cross-validation is used for further analysis. Three clustering algorithms (K-means [42], Hierarchical [43], and Affinity Propagation [44] clustering) are applied to the optimal low-dimensional data, and then, the results are evaluated with various statistical tests.

In this chapter, we describe the principles of GLRM framework and clustering algorithms in the methods section. Next, we discuss the outcome of each procedure in the results section and finally, we illustrate how GLRM helped us to achieve a patient stratification strategy for T1D patients.

2.2 Methods

In this section, we describe the procedure that we followed to group T1D patients from pre-processed dataset. We will discuss each of the following topics in a subsection:

- Summary of the GLRM framework principles, the philosophy behind it and the software package used for it.
- Description about optimal low-dimensional concise data extraction and how we find the proper parameter set for building the model.
- Reviewing the basis of three clustering methods, including K-means, hierarchical, and affinity propagation clustering, which we use for analyzing low-dimensional data.
- Finally, investigating clustering results to find the relation between clusters and complications.

2.2.1 GLRM Framework

Unavoidable imperfections in data such as noise, missing entries, sparsity, and heterogeneity have challenged common machine learning methods in previous studies for finding patterns in clinical health-related data [45]. Udell et al. [1] extended the idea behind Principal Components Analysis (PCA) into a generalized method that can handle different types of data sets including numerical, boolean, categorical, ordinal, and other data types. Generalized Low Rank Modeling (GLRM) handles heterogeneous datasets, compresses and denoises data, and imputes missing records. We apply

this method to deal with a large number of input data features and to utilize most of the samples even if they have missing values in some features.

GLRM represents high dimensional bulk data in a lower-dimensional space. Suppose we have a matrix A that has m rows representing samples and n columns that represent n features while these features have different types of data; for instance, one column may take float values while the others have categorical values. By solving an optimization problem, we can approximate A by X as a "tall and skinny" matrix and Y as a "short and wide" matrix (Figure 2.1). X represents k new latent features for m samples, and Y encodes the transformation of n original features into the k new latent features.

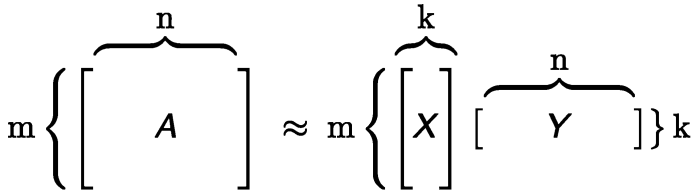

Figure 2.1: Matrix transformation with GLRM framework (obtained from Udell et al. [1])

To find X and Y the following optimization problem should be solved:

$$minimize \quad \sum_{(i,j)\in\Omega} L_{ij}(x_i y_j, A_{ij}) + \sum_{i=1}^{m} r_i(x_i) + \sum_{j=1}^{n} \tilde{r}_j(y_j) \tag{2.1}$$

Where $L_{ij}(x_i y_j, A_{ij})$ is the loss function, $x_i y_j$ is the predicted entry which is obtained by matrix production of the i^{th} row of estimated X matrix and the j^{th} column of estimated Y matrix, A_{ij} is the observed entry in the input data; r_x and r_y are regularizers used to limit output matrices. The loss function (first sigma operand)

measures the accuracy of data approximation, and the problem-solving algorithm will try to minimize this part. Different loss functions are appropriate for various types of data inputs. Thus, GLRM gives the flexibility to define different functions for each column of data (features) based on their type. The loss should be calculated only over the set Ω which represents non-missing entries. The regularizers r_x and r_y limits latent feature values. Choosing appropriate regularization can improve the model and prevent it from over-fitting. Furthermore, appropriate k value which represents the number of latent features can be estimated using cross validation over the observed data and investigating test and train errors.

We use GLRM to estimate and fill out the missing values in our dataset and transform our big heterogeneous dataset into a smaller homogeneous one. For this purpose, we use H_2O.ai package (version 3.14.0.2)[46] in the R programming language [47]. This package has built-in implementations of GLRM framework as well as popular machine learning algorithms.

Following building an appropriate model for the input dataset, we extract the tall and skinny matrix X and use it to cluster samples (patients) based on the k latent features obtained by GLRM.

2.2.2 GLRM Parameter Setting

To make a low-rank model converge efficiently, we need to choose proper input parameters. Udell et al. [1] thoroughly described the impact and purpose of each parameter in GLRM implementation. To achieve an optimal performance, parameters must be fitted based on the dataset. Here we briefly describe input parameters including loss

functions (L_j), regularizers($r, \tilde{r}$) gamma (γ), and output matrix rank (k).

Loss function (L_j): The loss functions is defined for each column (feature) based on its data nature. According to Udell et al. [1] and GLRM implementation in H_2O [46], we use "quadratic", "logistic", "categorical" and "ordinal" loss functions for numerical, boolean, categorical, and sequential features, respectively.

Regularizers($r, \tilde{r}$) and Gamma (γ) The regularization functions r and $\tilde{r}$ are used to prevent overfitting or to enforce constraints on the values of low-rank matrices X and Y. These regularizations can be scaled by γ. Thus, the GLRM optimization problem would be adjusted to:

$$minimize \quad \sum_{(i,j)\in\Omega} L_{ij}(x_i y_j, A_{ij}) + \gamma_x \sum_{i=1}^{m} r_i(x_i) + \gamma_y \sum_{j=1}^{n} \tilde{r}_j(y_j) \qquad (2.2)$$

Where all the terms are as defined in Equation 2.1, γ_x and γ_y are scaling values for the regularizers. Our input dataset has many missing values, and this can prevent the model from overfitting itself. Therefore, we use no regularizer for building the low rank model. Additionally, by adding the regularizers to the model, we observed that test and train errors in cross validation were significantly increased.

Rank (k): Rank of a model is the number of columns in the concise low-rank matrix (X). We use cross validation over the input data and find a proper rank based on the train and test error.

In addition to these parameters, we need to set the low-rank matrices initialization method, Udell et all [1] showed that a suitable approach for matrices initialization is

"Singular Value Decomposition (SVD)" [48] which performs much better than other random initialization methods.

Lastly, following the determination of well-fitted parameters for building the model, we can extract low-rank matrices X and Y. X represents sample features in a different domain while it has homogeneous data with no missing entries. It has m rows which is equal to the number of samples and k columns which represent k latent features. Y with the size of $k \times n$, represents the relation between k latent features and n original features. We call X matrix **"concise data"** and will use it to cluster patients.

2.2.3 Data Clustering

Patients grouping helps clinicians to investigate the diseases cause in a group [36]. Our input dataset, contains patients complications data as well as clinical, demographical and genetic data. As indicated in Section 1.4 we separate features set into two categories: 1- complications features, 2- Other features. The second features category is used as the input data for building the GLRM model, extracting the concise data and clustering samples based on this concise matrix. The first features category is used to evaluate the clustering results. In this section, we first describe the algorithms used for clustering the concise data (each cluster represents a group of patients), and then we explain how we evaluate clustering results to discover the relationship between clusters and complications.

K-means Clustering: k-means clustering aims to partition m samples into k clusters while each sample fits the cluster with the nearest mean [42]. We use

`kmeans` function in Matlab® Statistics and Machine Learning Toolbox for clustering the concise data [49]. This function needs the number of clusters to be determined as the input; we empirically set this number to 10. Other clustering methods which do not require the number of clusters as an input indicated roughly the same number of clusters as well. Euclidean distance is used for distance measurement between data points.

Hierarchical Clustering: Hierarchical clustering groups data by forming a cluster tree or dendrogram [43]. The tree is a multilevel hierarchy where clusters at one level are joined as clusters at the next level [50]. There are two strategies for hierarchical clustering: Agglomerative and Divisive. The agglomerative strategy is a "bottom up" approach in which each sample has its own cluster, and two clusters are merged as one moves up the hierarchy. On the other hand, the divisive strategy is a "top down" approach which all samples are gathered in one cluster, and then splits are performed as one moves down the hierarchy [51]. We use `clusterdata` function in Matlab® Statistics and Machine Learning Toolbox for clustering concise data [49]. This function supports agglomerative clustering. We use Euclidean distance as the distance metric, and inner squared distance (minimum variance algorithm) as the algorithm for computing distance between clusters, more information about the function inputs are available at Matlab user guide in hierarchical clustering [50]. Finally, after applying hierarchical clustering to the concise data, we choose a proper value for the maximum number of clusters based on the obtained dendrogram. We select a cutting level on the dendrogram where clusters are neither too small (at

least three members in each cluster) nor too big (not more than 40 members in each cluster).

Affinity Propagation Clustering: Affinity propagation clustering is an algorithm based on the concept of "message passing" between samples, and it does not require the number of clusters to be determined before running the algorithm [44, 52]. We use `apcluster` function in Matlab® for clustering our data, more information about the algorithm and function parameters are available at Frey et al. article [44]. This function requires two inputs: 1- A square matrix representing pairwise similarities between two samples. We use the negative of Euclidean distance as the pairwise similarity measure. 2- An input preference p which is a real-valued vector. p_i indicates the preference that data point i be chosen as an exemplar. We set all preferences to a same value since we have no preferences among the samples.

2.2.4 Clusters Evaluation

As indicated in Section 1.4, we extract patients complications information from raw input dataset and name this extracted data as the complications matrix. The complications matrix is a binary data which contains 153 rows corresponding to the number of patients and ten columns corresponding to the number of accessible complications information.

The results of each clustering algorithm is a vector indicating the cluster assignment for each patient. For each clustering result, we apply three statistical measures per complication to investigate if there is any relation between a group of patients

and a disease. These measures are hypergeometric test, odds ratio, and risk ratio.

Hypergeometric Test: The hypergeometric test is based on the hypergeometric distribution to calculate the significance of having drawn a specified number of successes from a specified population. This test can be used to distinguish which subsets of the population are over- or under-represented in a clustering scheme [53].

We use Matlab® `hygecdf` function to "compute the complement hypergeometric cdf at the value of x using the corresponding size of the population (M), total number of items with the desired characteristic in the population (K), and number of samples drawn (N). The result, p, is the complement probability of drawing up to x of a possible K items in N drawings without replacement from a group of M objects" [50].

$$p = 1 - \sum_{i=0}^{x} \frac{\binom{K}{i}\binom{M-K}{N-i}}{\binom{M}{N}} \tag{2.3}$$

We define x as the number of patients with specified complication in the cluster, M as the total number of patients (population size), K as the total number of patients with specified complication, and N as the cluster size. The obtained p value represents the probability in the null distribution of observing up to x patients with the specified complication in the given cluster. Thus, the lower probability means a more reliable cluster that could adequately capture patients with a specified disease. However, this probability is very low in the small clusters (such as clusters with less than four patients), which is not desired; therefore we add other methods to evaluate clustering results.

Risk Ratio: Risk ratio or relative risk (RR) is the probability of an event occurring in an exposed group divided by the probability of the event occurring in a comparison, non-exposed group [54].

Table 2.1: Exposed and diseased population ratio definition

	Diseased	Healthy	Total
Exposed	D_E	H_E	$N_E = D_E + H_E$
Not exposed	D_N	H_N	$N_N = D_N + H_N$

Considering Table 2.1 the risk ratio can be calculated using Equation 2.4.

$$RR = \frac{D_E/N_E}{D_N/N_N} \tag{2.4}$$

A $RR = 1$ represents no difference in risk between the exposed and non-exposed group. However, $RR < 1$ means the event is less likely to occur in the exposed group than in the not exposed group, and a $RR > 1$ means the event is more likely to occur in the exposed group. We count "Diseased" population as the patients with a specific complication and "Exposed" population as the patients inside a cluster. Consequently, clusters with higher risk ratio are more reliable clusters which they could represent patients more likely to develop a given complication.

Odds Ratio: The odds ratio (OR) is used commonly to quantify how strongly the presence or absence of an exposure is associated with an outcome in a given population. If $OR = 1$ it means that the exposure does not affect odds of the

outcome. $OR > 1$ indicating that the exposure is associated with higher odds of outcome and $OR < 1$ is the opposite case [55, 56].

Considering Table 2.1 the odds ratio is calculated using Equation 2.5.

$$OR = \frac{D_E/H_E}{D_N/H_N} \tag{2.5}$$

When the OR is one, the RR will be equal to one as well and OR approximates RR if the probability of the disease occurrence is low. However, the OR is always bigger compared to RR; thus, OR can better represent slight differences. We use OR for plotting our results, but it should be kept in mind that a small cluster (such as clusters with less than four patients) may have very high OR.

To find the significant level of odds ratio in each cluster given a complication, we bootstrap obtained clustering result for each clustering method and determine the p-value for the odds ratio per cluster given a complication. For this purpose, 10000 random set of cluster labels are produced while the total number of clusters and cluster's size are identical to the original clustering result. Then, for each complication given a cluster, the p-value is measured by dividing the number of results from the random sampling with an OR greater or equal to the real OR, divided by total number of results from the random sampling.

By evaluating the clustering results from k-means, hierarchical and affinity propagation clustering algorithms using the three outlined measurements, we can find out whether risk factors for developing a secondary disease related to T1D can be identified by integrating demographic, clinical and genetic data using the GLRM method.

2.3 Results

As described in Section 1.4 our pre-processed T1D dataset is segmented into two separate datasets. First one has 153 samples and 436 features; we call this dataset as the **"T1D pre-processed input data"** since we use it for machine learning algorithms input. The second part of the dataset is a binary matrix with 153 samples and ten columns representing the presence of ten complications in each patient. This matrix is used for evaluating the clustering results. We call this matrix the **"complications data"**. The input data has 14.75% missing entries, however, the GLRM algorithm can tolerate this amount of missing data. Table 1.1 represents a summary of input data characteristics and a detailed table is available in Appendix A.

In this section, we present the cross-validation results to set the GLRM optimal parameters. Next, we illustrate the results of the clustering algorithms and, finally, we discuss whether or not discovered clusters are over-enriched with patients having a given complication.

2.3.1 GLRM Parameter Selection

In section 2.2.2 we discussed methods for finding proper GLRM parameters including loss functions, regularizers, and rank. Appendix A includes list of all features and their types, we use "quadratic", "logistic", "categorical" and "ordinal" loss functions for numerical, boolean, categorical and sequential features, respectively. We use no regularizer for building the low rank model since by adding the regularizers to the model, we observed that cross validation test and train errors were significantly increased.

Cross validation is used for finding a proper rank (k). We use ω portion of observed samples which is randomly selected as the validation set and the remaining observations $(1-\omega)$ as the training set. Additionally, cross validation is repeated five times with distinct sets of ω for each set of parameters in GLRM model. Since our input data is heterogeneous, we use two types of error for evaluating model performance. Mean Square Error (MSE) used for calculating model errors in numerical features (only 11 features) and Misclassification Ratio (MCR) used for the rest. Considering the few number of numerical features, MSE is not a reliable measure for selecting the parameters, and we decide based on MCR.

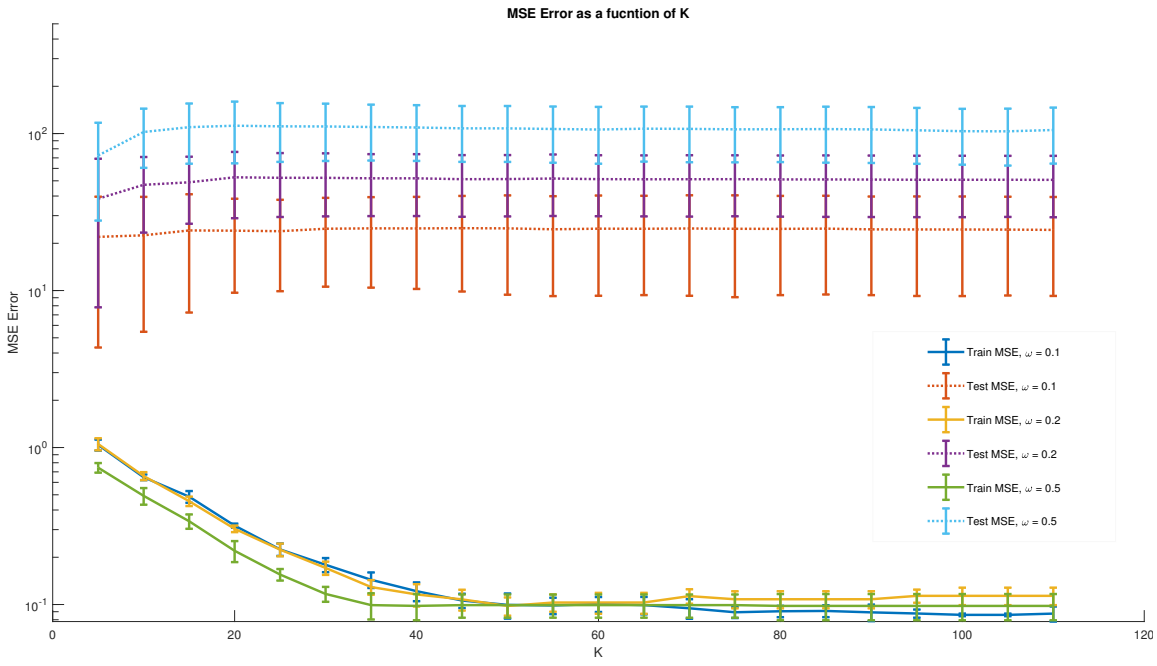

Figure 2.2: GLRM model average Mean Square Errors (MSE) on five fold cross validation results. The horizontal solid lines indicate the MSE on the training data and the horizontal dotted lines indicate the MSE on the test data. The vertical lines indicate the standard error.

Figure 2.2 illustrates the average of MSE, over five cross validations, on train and test set, for three different ω. Figure 2.3 illustrates MCR with same properties as the previous figure. As we can observe in the Figure 2.3, train and test errors are both in their minimum values after $k = 70$. Thus, we choose $k = 70$ as the proper rank and extract the X matrix which has 153 samples and 70 latent features. As it mentioned, we call this matrix "concise data".

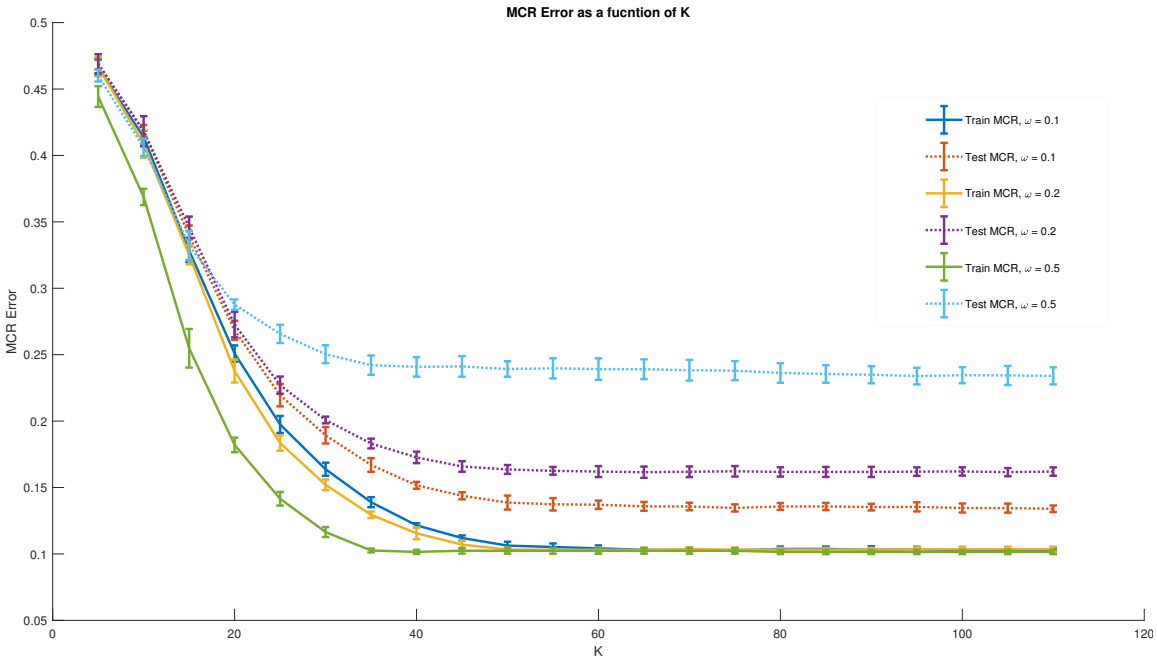

Figure 2.3: GLRM model average Misclassification Ratio (MCR) on five fold cross validation results. The horizontal solid lines indicate the MCR on the training data and the horizontal dotted lines indicate the MCR on the test data. The vertical lines indicate the standard error.

2.3.2 Clustering Concise Data

Three algorithms described in Section 2.2.3 are used to cluster the concise data. In this section, we first represent and analyze each clustering results, then, we address clusters over-enriched with patients having T1D secondary diseases. Table 2.2 illustrates an overview of three clustering methods' results.

Table 2.2: Overview of GLRM three clustering methods' results

	k-means	Hierarchical	Affinity Propagation
No. of clusters	10	9	11
Clusters' Size Average ± SD	15.3±18.2	17±11.5	13.9±10.3
Median Clusters' Size	9.5	15	17
Maximum Clusters' Size	63	38	36
Minimum Clusters' Size	3	3	1

2.3.2.1 K-means Clustering

As it explained in Section 2.2.3, we use `kmeans` function in Matlab® for clustering the concise data. Number of desired clusters (k) is empirically set to ten; other clustering methods which do not require to specify the number of clusters a priori indicated roughly the same amount for the number of clusters as well. Table 2.3 illustrates demographic statistics regarding k-means output clusters.

We evaluate clustering outcome using three statistical measures including Hypergeometric test, odds ratio, and risk ratio which are described in Section 2.2.4.

Figure 2.4 represents odds ratio for each complication and cluster. Numbers in each cell represents total number of patients with the specified complication in the

Table 2.3: Statistics of the patients clusters obtained using k-means clustering

	Cluster Size	Male(%)	Female(%)	Weight (kg)	Age
Cluster #1	14	7.1	92.9	81.0 ± 18.1	33.1 ± 7.2
Cluster #2	3	100.0	0.0	91.7 ± 20.6	25.7 ± 2.9
Cluster #3	11	54.5	45.5	80.1 ± 14.9	30.7 ± 6.0
Cluster #4	12	0.0	100.0	62.8 ± 7.2	27.1 ± 4.9
Cluster #5	7	14.3	85.7	68.7 ± 12.2	51.6 ± 10.7
Cluster #6	27	51.9	48.1	82.5 ± 16.0	47.9 ± 11.0
Cluster #7	4	50.0	50.0	111.1 ± 35.0	31.5 ± 7.6
Cluster #8	8	75.0	25.0	92.2 ± 15.7	37.6 ± 15.1
Cluster #9	63	46.0	54.0	77.5 ± 15.3	29.5 ± 6.4
Cluster #10	4	50.0	50.0	76.3 ± 15.1	29.8 ± 8.4

Weight and age columns values indicate corresponding average ± standard deviation.

specified cluster.

Figure 2.5 illustrates the three clusters with the highest odds ratio for each complication while clusters with less than four members are filtered out. Each bubble represents the odds ratio for a cluster concerning the specified complication, and bubbles sizes are proportional to the number of patients with a given complication.

Table 2.4 shows statistical measures for the cluster with highest odds ratio concerning each complication while clusters with less than four members are filtered out. Entire clustering evaluation measures are available in Section B.1 in Appendix B. Table B.41 shows odds ratio p-value for each cluster given a complication.

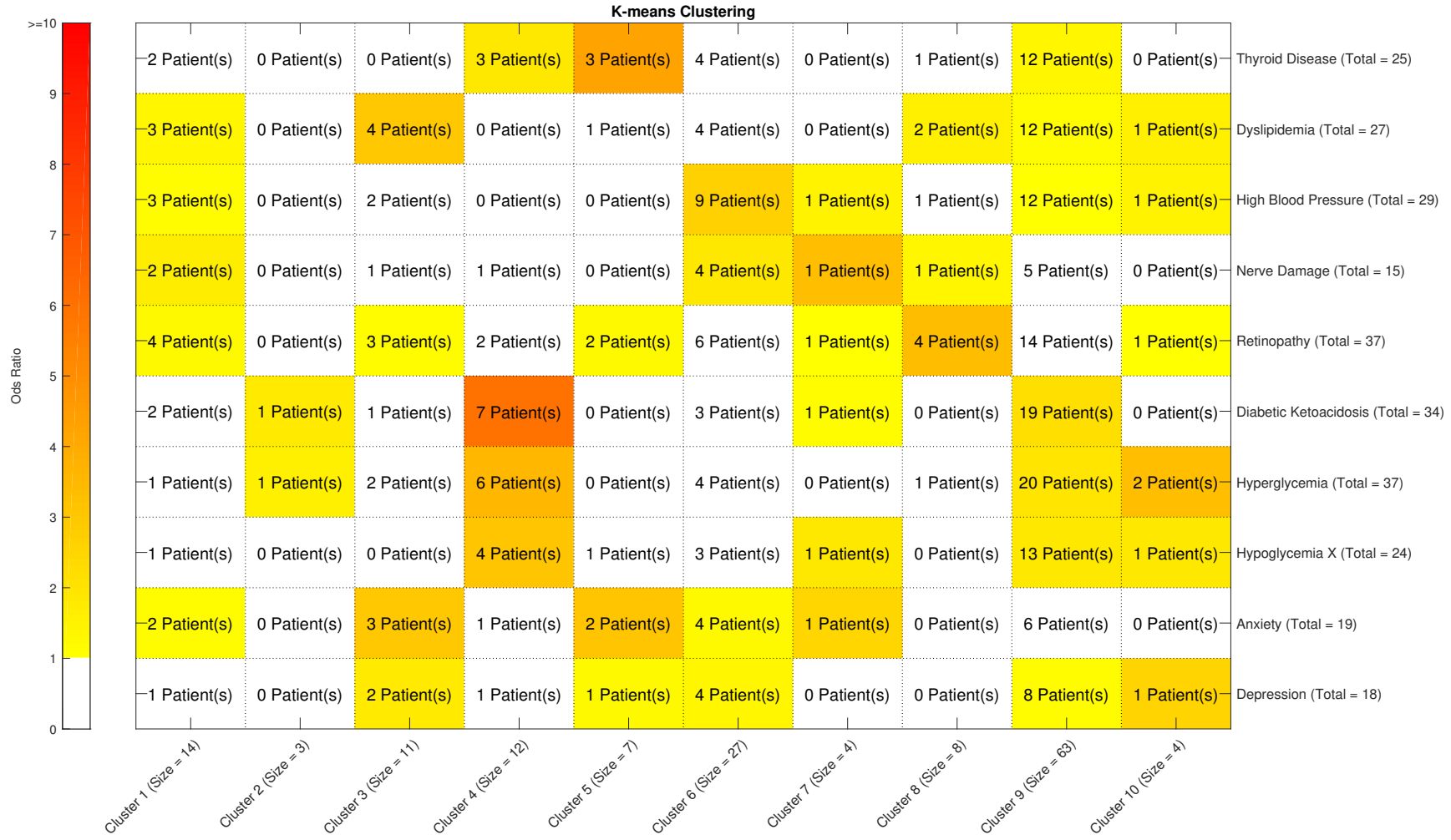

Figure 2.4: Heatmap for k-means clustering indicating the calculated odds ratio per complication for obtained clusters

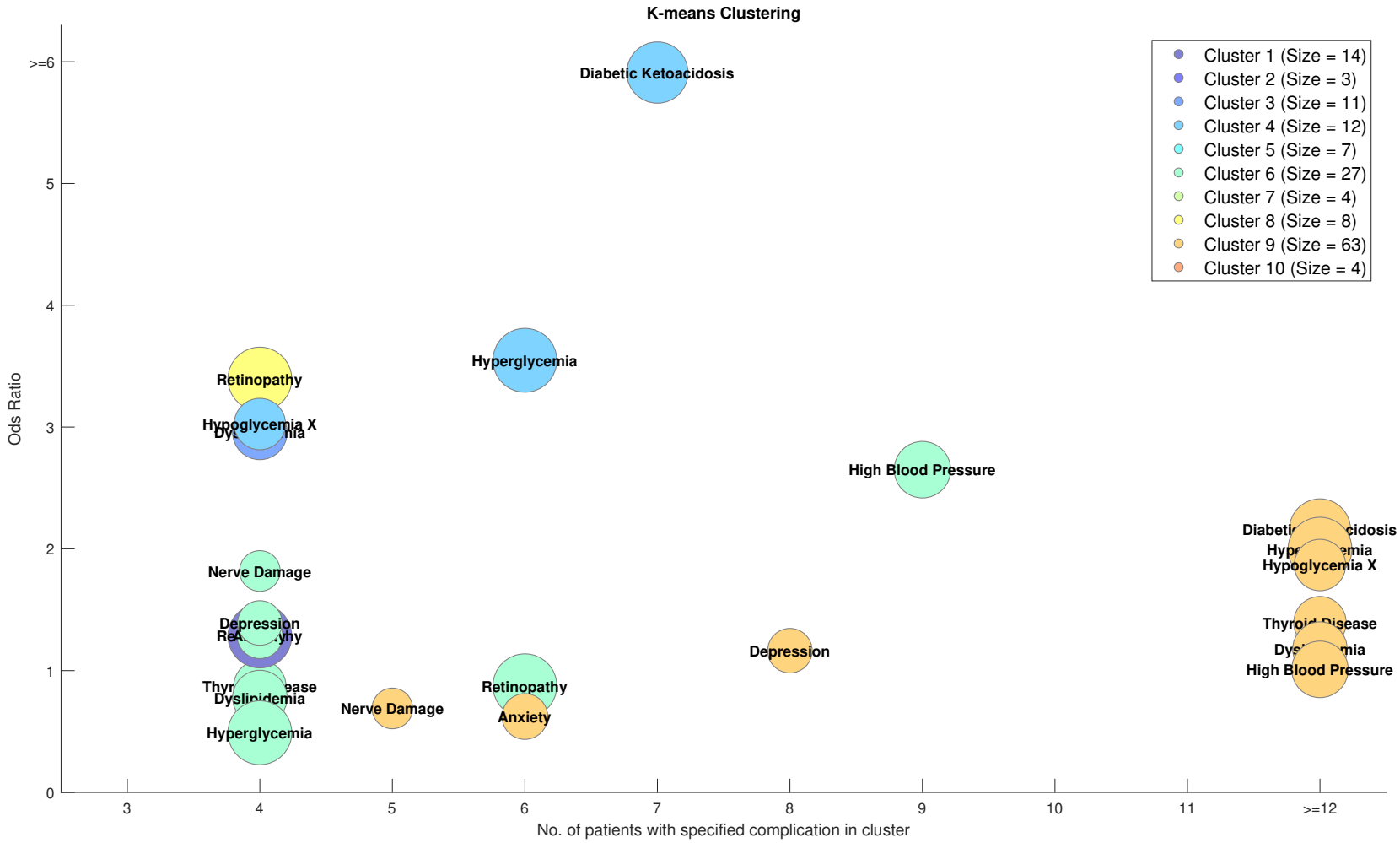

Figure 2.5: Bubble graph for k-means clustering result

Table 2.4: Most significant results per complication obtained using k-means clustering

	OR	p-value	Cluster #	#PCC	#CLS	#CMP
Diabetic Ketoacidosis	5.91	0.006	4	7	12	34
Hyperglycemia	3.55	0.041	4	6	12	37
Retinopathy	3.39	0.097	8	4	8	37
Hypoglycemia X	3.02	0.095	4	4	12	24
Dyslipidemia	2.96	0.107	3	4	11	27
High Blood Pressure	2.65	0.036	6	9	27	29
Nerve Damage	1.82	0.260	6	4	27	15
Thyroid Disease	1.39	0.294	9	12	63	25
Depression	1.39	0.389	6	4	27	18
Anxiety	1.29	0.437	6	4	27	19

#PCC means number of patients with specified complication in the cluster. #CLS represents total number of patients in the cluster and #CMP shows total number of patients with specified complication.

2.3.2.2 Hierarchical Clustering

Agglomerative hierarchical clustering is the second method used for clustering the concise data. Figure 2.6 illustrates dendrogram of the hierarchical clustering. Each leaf in the tree corresponds to one data sample. The number of clusters is obtained by partitioning the dendrogram in a level in which cluster are neither too small (less than three members) nor too large (more than 40 members). The red line in the figure shows the chosen level which allows nine clusters as the algorithm output.

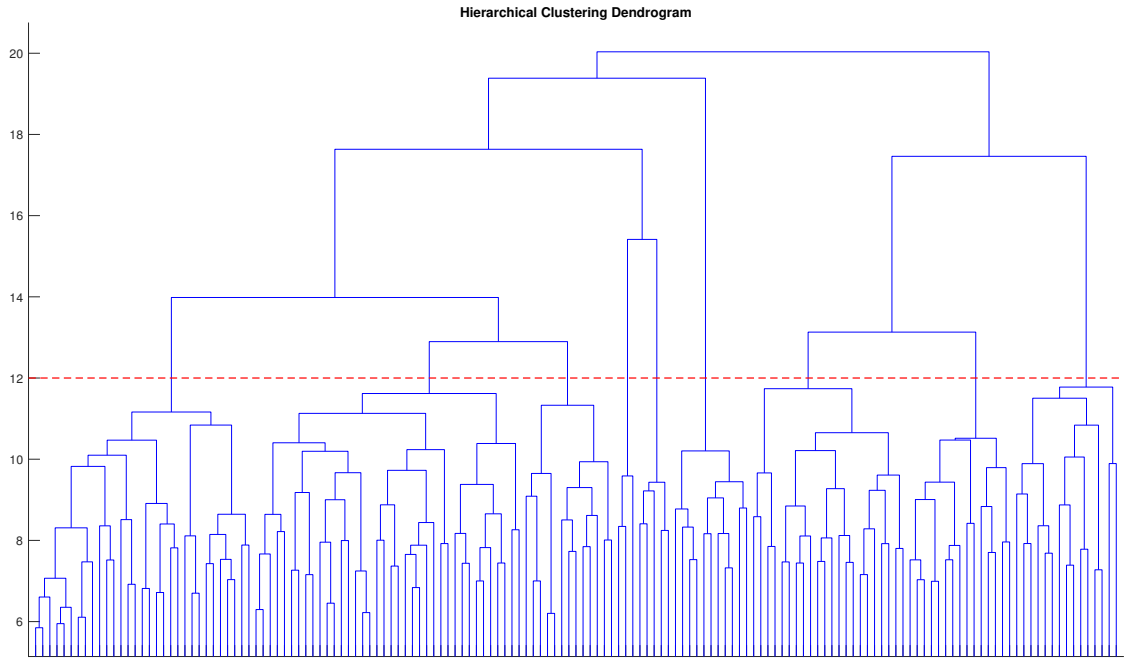

Figure 2.6: Hierarchical Clustering Dendrogram Leaf nodes at the bottom of the dendrogram present patients. The height of each U represents the distance between the two connected patient clusters.

Table 2.5 illustrates demographic statistics regarding hierarchical clustering extracted clusters.

Based on the hierarchical clustering result, we illustrate the algorithms outcome

Table 2.5: Statistics of the patients clusters obtained using hierarchical clustering

	Cluster Size	Male(%)	Female(%)	Weight (kg)	Age
Cluster #1	13	30.8	69.2	88.8 ± 13.9	32.2 ± 7.1
Cluster #2	38	39.5	60.5	73.5 ± 12.9	29.0 ± 4.6
Cluster #3	15	66.7	33.3	78.2 ± 10.8	35.6 ± 9.6
Cluster #4	22	45.5	54.5	78.4 ± 11.9	54.3 ± 7.3
Cluster #5	31	16.1	83.9	68.2 ± 10.5	28.7 ± 5.8
Cluster #6	5	40.0	60.0	80.5 ± 16.1	34.4 ± 7.3
Cluster #7	3	33.3	66.7	62.0 ± 4.7	29.7 ± 6.7
Cluster #8	15	60.0	40.0	111.2 ± 18.5	35.6 ± 13.6
Cluster #9	11	72.7	27.3	82.8 ± 13.3	29.0 ± 5.6

Weight and age columns values indicate corresponding average ± standard deviation.

using a heatmap (Figure 2.7), a bubble plot (Figure 2.8) and a table (Table 2.6) with same characteristics of the corresponding elements in Section 2.3.2.1. Entire clustering evaluation measures are available in Section B.1 in Appendix B. Table B.42 shows odds ratio p-value for each cluster given a complication.

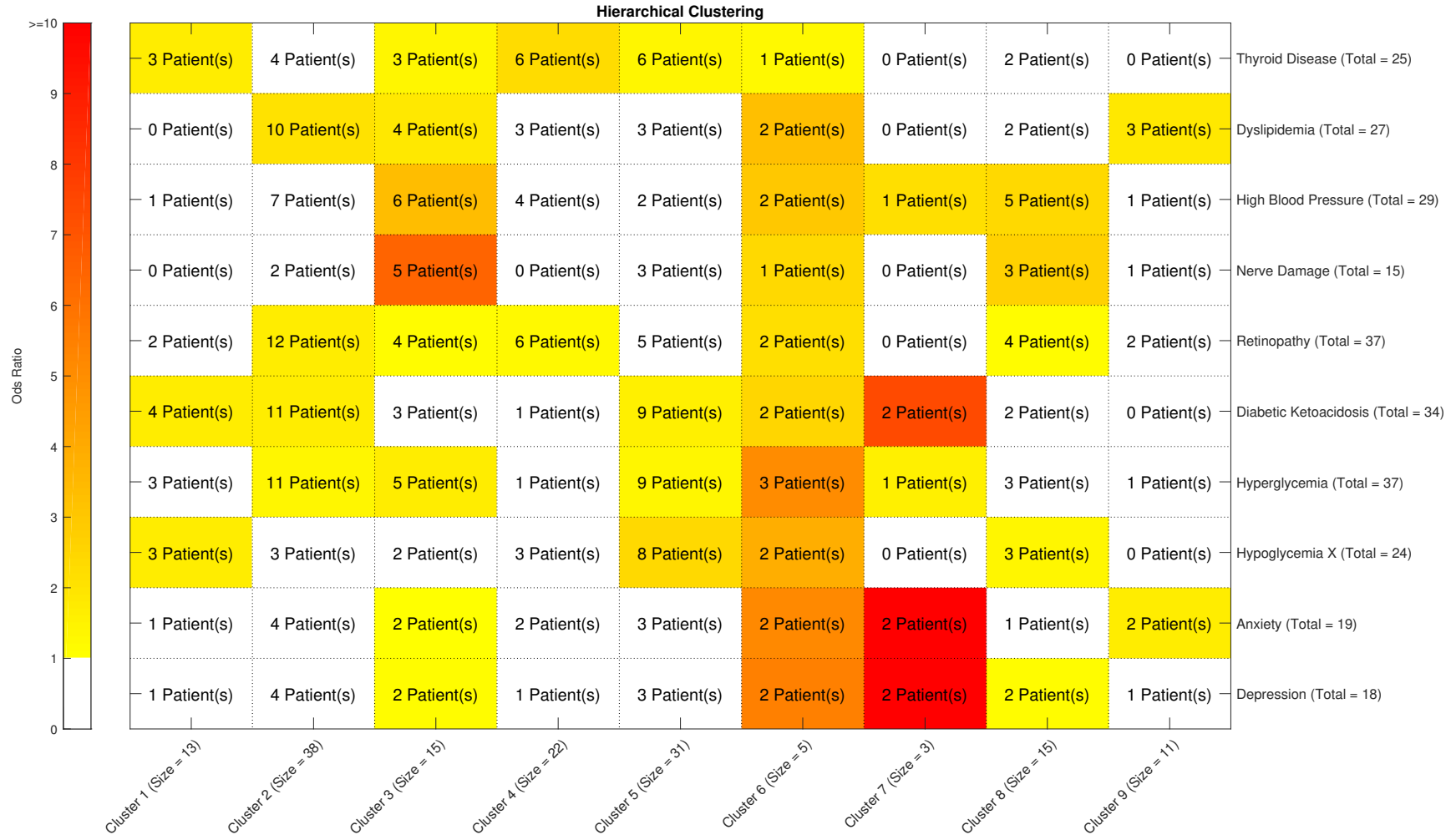

Figure 2.7: Heatmap for hierarchical clustering indicating the calculated odds ratio per complication for obtained clusters

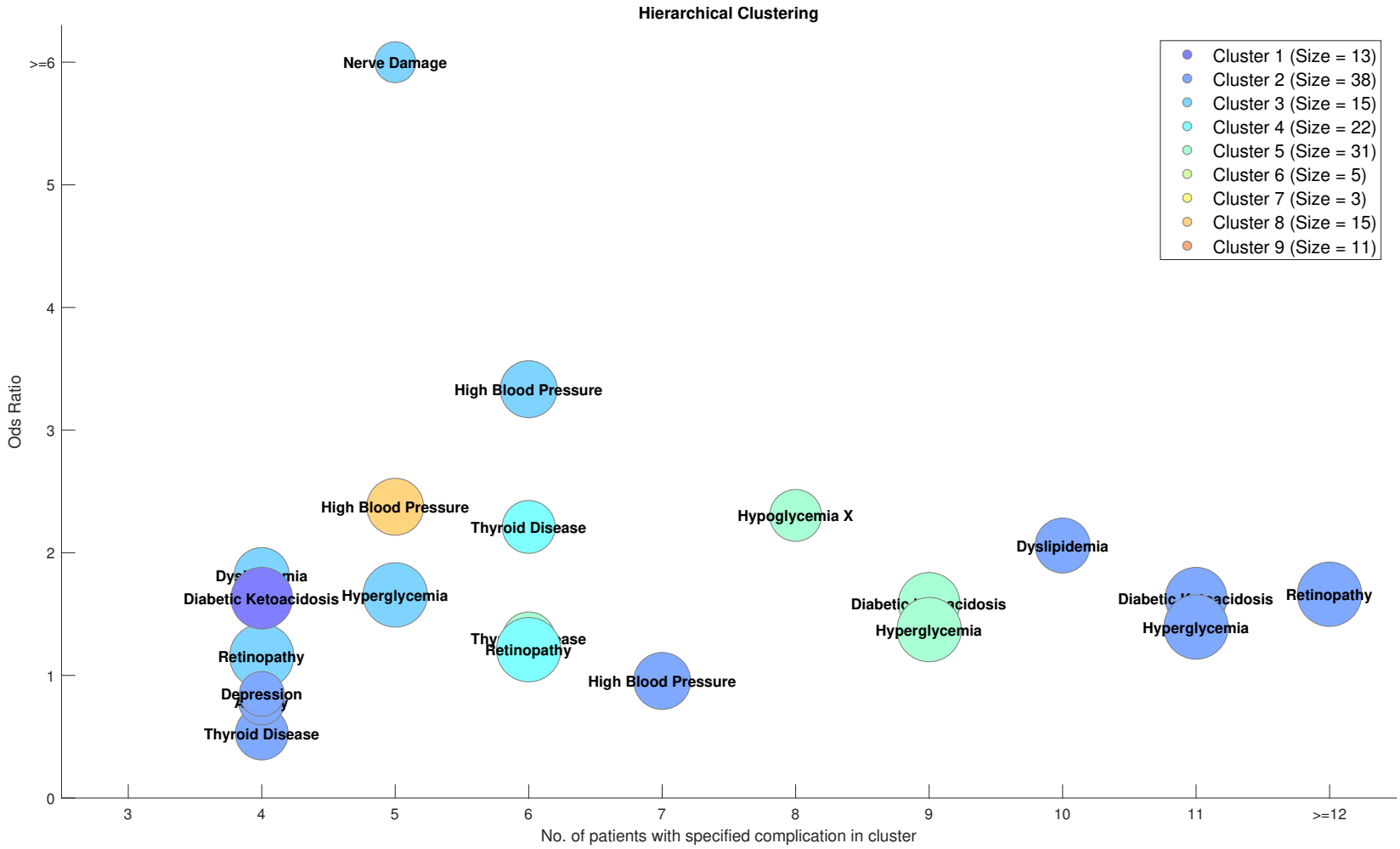

Figure 2.8: Bubble graph for hierarchical clustering result

Table 2.6: Most significant results per complication obtained using hierarchical clustering

	OR	p-value	Cluster #	#PCC	#CLS	#CMP
Nerve Damage	6.40	0.007	3	5	15	15
High Blood Pressure	3.33	0.041	3	6	15	29
Hypoglycemia X	2.30	0.075	5	8	31	24
Thyroid Disease	2.21	0.124	4	6	22	25
Dyslipidemia	2.06	0.090	2	10	38	27
Retinopathy	1.66	0.160	2	12	38	37
Hyperglycemia	1.66	0.280	3	5	15	37
Diabetic Ketoacidosis	1.63	0.323	1	4	13	34
Depression	0.85	0.706	2	4	38	18
Anxiety	0.78	0.751	2	4	38	19

#PCC means number of patients with specified complication in the cluster. #CLS represents total number of patients in the cluster and #CMP shows total number of patients with specified complication.

2.3.2.3 Affinity Propagation Clustering

Affinity Propagation clustering is the third methods used for clustering the concise data. This method does not need the number of clusters to be determined. Table 2.7 illustrates demographic statistics regarding affinity propagation clustering extracted clusters.

Table 2.7: Statistics of the patients clusters obtained using affinity propagation clustering

	Cluster Size	Male(%)	Female(%)	Weight (kg)	Age
Cluster #1	3	33.3	66.7	106.7 ± 27.6	57.7 ± 5.1
Cluster #2	1	0.0	100.0	159.1 ± 0.0	42.0 ± 0.0
Cluster #3	12	33.3	66.7	72.7 ± 15.7	30.1 ± 5.5
Cluster #4	17	52.9	47.1	75.2 ± 10.5	31.4 ± 13.7
Cluster #5	17	23.5	76.5	67.1 ± 12.8	30.2 ± 9.3
Cluster #6	20	50.0	50.0	77.0 ± 15.4	29.7 ± 8.4
Cluster #7	36	50.0	50.0	81.8 ± 15.3	42.0 ± 11.2
Cluster #8	19	15.8	84.2	85.0 ± 19.9	34.5 ± 10.8
Cluster #9	1	0.0	100.0	56.8 ± 0.0	34.0 ± 0.0
Cluster #10	18	66.7	33.3	85.6 ± 12.8	31.2 ± 9.2
Cluster #11	9	33.3	66.7	72.8 ± 10.3	31.1 ± 4.6

Weight and age columns values indicate corresponding average ± standard deviation.

We illustrate the algorithms outcome using a heatmap (Figure 2.9), a bubble plot (Figure 2.10) and a table (Table 2.8) with same characteristics of the corresponding elements in Section 2.3.2.1. Entire clustering evaluation measures are available in Section B.1 in Appendix B. Table B.43 shows odds ratio p-value for each cluster given a complication.

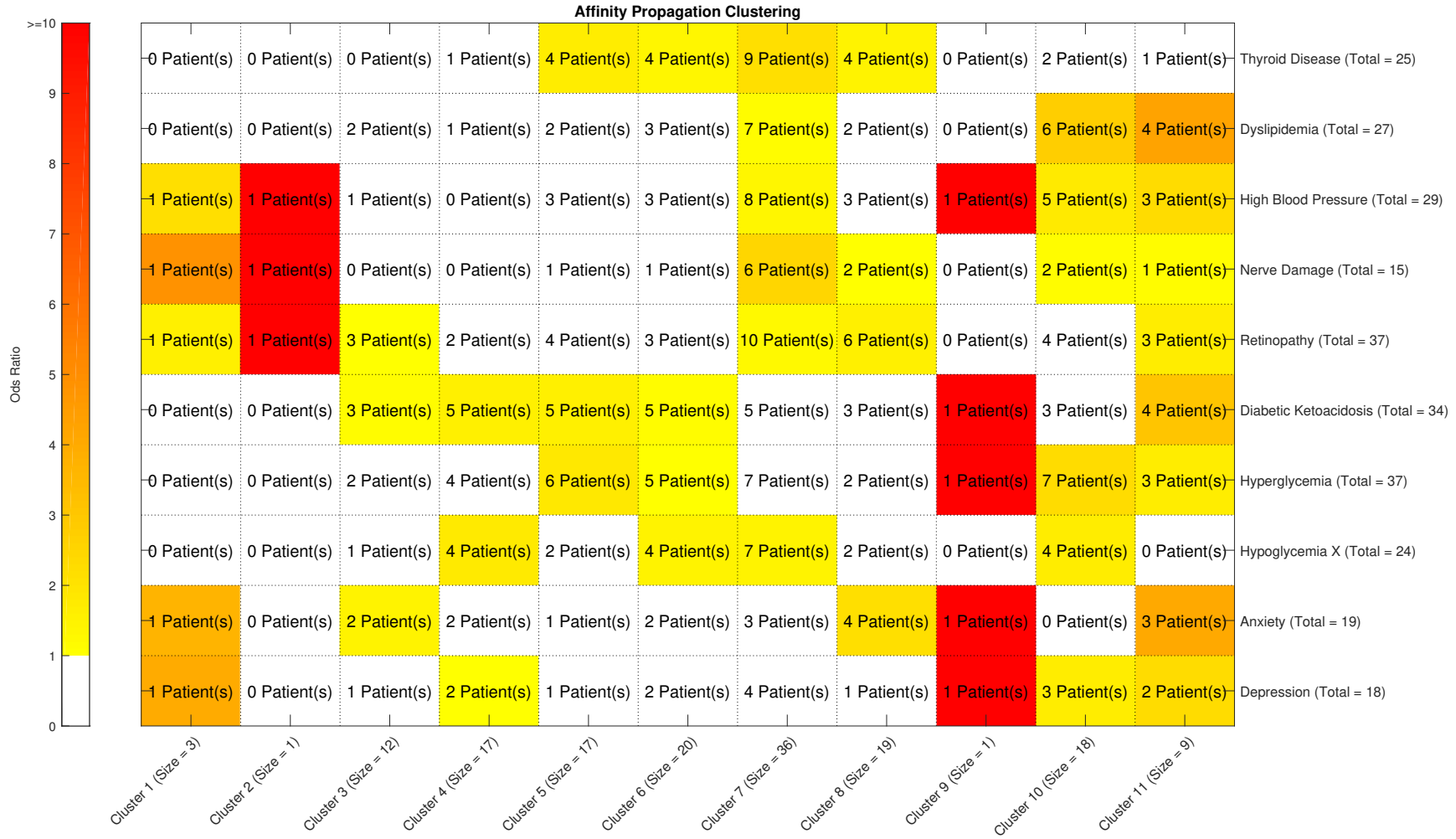

Figure 2.9: Heatmap for affinity propagation clustering indicating the calculated odds ratio per complication for obtained clusters

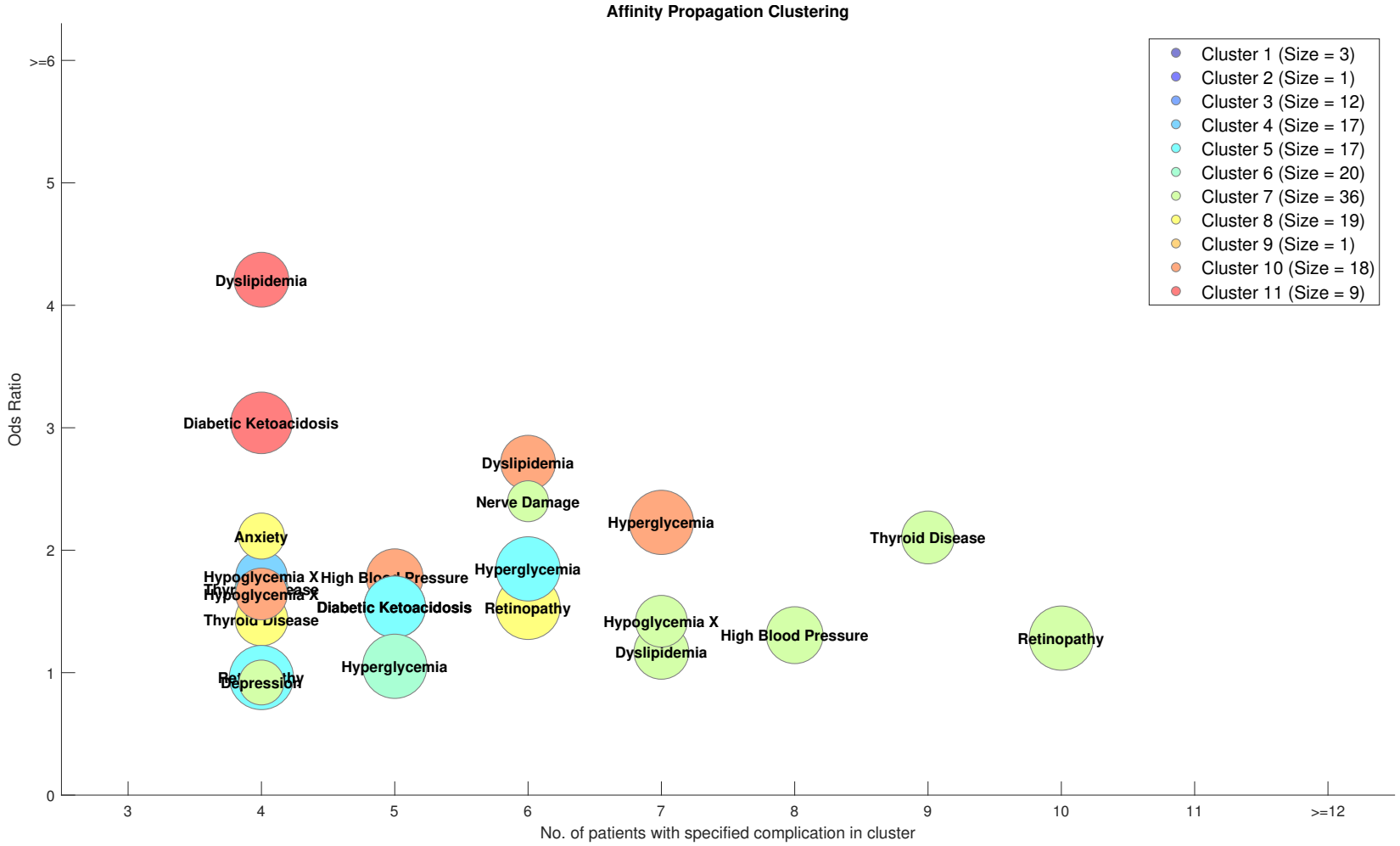

Figure 2.10: Bubble graph for affinity propagation clustering result

Table 2.8: Most significant results per complication obtained using affinity propagation clustering

	OR	p-value	Cluster #	#PCC	#CLS	#CMP
Dyslipidemia	4.21	0.052	11	4	9	27
Diabetic Ketoacidosis	3.04	0.111	11	4	9	34
Nerve Damage	2.40	0.099	7	6	36	15
Hyperglycemia	2.23	0.108	10	7	18	37
Anxiety	2.12	0.199	8	4	19	19
Thyroid Disease	2.10	0.090	7	9	36	25
Hypoglycemia X	1.78	0.254	4	4	17	24
High Blood Pressure	1.78	0.233	10	5	18	29
Retinopathy	1.53	0.291	8	6	19	37
Depression	0.92	0.657	7	4	36	18

#PCC means number of patients with specified complication in the cluster. #CLS represents total number of patients in the cluster and #CMP shows total number of patients with specified complication.

2.4 Discussion

We have an input dataset from T1D patients with no control data (no healthy samples), we compressed this dataset using GLRM and applied three clustering algorithms to identify groups of patients at higher risk of developing T1D complications. We found that, k-means clustering result is not stable; it produces a different clustering result on each run since the initialization is random and the algorithm is not able to converge. Therefore, it seems that k-means is not a proper clustering algorithm for our data and we avoid further analysis on the k-means clustering result.

Based on the hierarchical clustering result, by considering the odds ratio value for each cluster given a compilation and its corresponding p-value (Table B.42), we can state that Cluster #3 with 15 members is enriched with patients who suffer from

nerve damage and high blood pressure. These two complications have the odds ratio of 6.4 (with p-value = 00.7) and 3.3 (with p-value = 0.041), respectively in Cluster #3.

On the other hand, according to affinity propagation clustering result, by considering the odds ratio value for each cluster given a compilation and its corresponding p-value (Table B.43), Cluster #11 with 9 members, is enriched with Dyslipidemia patients and this complication's odds ratio is 4.2 (with p-value = 0.052) in this cluster.

Table 2.9: Concordance between affinity propagation clustering and hierarchical clustering based on NMI percentage

	Cluster #1	*Cluster #2*	*Cluster #3*	*Cluster #4*	*Cluster #5*	*Cluster #6*	*Cluster #7*	*Cluster #8*	*Cluster #9*
Cluster #1	1.1	2.4	1.2	1.5	2.0	0.6	0.4	27.0	0.9
Cluster #2	0.5	1.3	0.6	0.8	1.1	0.3	0.2	13.7	0.5
Cluster #3	0.0	12.3	2.8	3.8	0.1	1.4	1.0	2.8	2.3
Cluster #4	1.7	0.4	0.1	0.1	0.7	1.7	1.9	3.6	2.6
Cluster #5	0.2	0.4	3.6	1.2	11.0	0.5	1.9	3.6	2.9
Cluster #6	0.4	0.2	0.0	0.3	0.0	0.3	1.4	0.6	0.2
Cluster #7	0.0	4.8	5.5	10.7	0.8	0.0	2.3	0.8	1.4
Cluster #8	0.1	0.1	0.0	0.2	0.0	0.3	1.4	0.7	0.1
Cluster #9	0.5	1.3	0.6	0.8	1.1	0.3	43.6	0.6	0.5
Cluster #10	0.3	2.7	0.5	1.3	0.1	0.4	1.3	0.9	3.1
Cluster #11	0.1	0.0	2.3	3.1	0.6	1.1	0.8	2.3	22.9

Rows and columns indicate affinity propagation clustering and hierarchical clustering, respectively. A higher NMI percentage implies greater concordance between clusters.

Table 2.9 shows the concordance between the affinity propagation and the hierarchical clustering results based on NMI percentage. We use percentage of Normalized Mutual Information (NMI) [57, 58] for measuring the concordance of two clustering result. NMI is a value between 0 and 1 and NMI percentage is between 0 to 100. NMI

can be calculated using NMI $= \frac{I(U,V)}{\sqrt{H(U)H(V)}}$, where U and V are two sets of clustering results from the same input data. $I(U,V)$ is the mutual information between U and V; and H is the entropy of the given clustering result; calculation details are available at Vinh et al. [57]. As we can observe in Table 2.9, the two clustering methods does not have any significant concordance, this is supported by the fact that the obtained result from the two methods are completely different.

As a future work, one would need to investigate what factors cause patients to develop T1D related complications such as nerve damage and Dyslipidemia, since we found clusters enriched with these complications.

2.5 Conclusion

Machine Learning algorithms have been used broadly for analyzing data, however, researchers dealing with those methods are challenged by imperfect data characteristics such as missing data and heterogeneity. We used Generalized Low Rank Modeling (GLRM) for analyzing our input dataset to discover patients subgroups more likely to have a given complication. This framework can overcome data barriers like missing data and heterogeneity. In a nutshell, we took following steps to identify groups of patients at higher risk of developing T1D complications using demographic, clinical and genetic data.

1. Cleansed raw dataset (we call the clean data as the "input data")

2. Applied GLRM framework with a proper parameter set to the input data.

3. Extracted the "concise data" from GLRM output.

4. Applied three clustering algorithm: k-means, hierarchical and affinity propagation clustering to the concise data.

5. Evaluated the clustering with patients complications information and found clusters that are over-enriched with patients having secondary diseases related to T1D.

According to the achieved result, we identified clusters enriched with patients having nerve damage, high blood pressure and Dyslipidemia. Consequently, we have taken first steps to identify groups of patients at higher risk of developing T1D complications. This results could be taken as the basis to develop a predictive model that could allow patients and health-care providers to take preemptive steps to reduce the risk of developing T1D related complications based on each patient characteristics.

Chapter 3

Similarity Network Fusion

3.1 Introduction

Current technology delivers the opportunity to efficiently collect diverse clinical, demographic and genetic data to address biological questions. However, powerful computational methods are needed to investigate these data and create a comprehensive view of a biological progresses such as developing a diseases. Systems biology approaches and more specifically, network-based techniques have emerged as powerful tools for studying complex diseases. We use Similarity Network Fusion (SNF) [2] to reveal the links among clinical, demographic and genetic data to facilitate patients stratification and to distinguish those patients at a higher risk of developing T1D complications.

SNF uses networks of samples as a basis for integration. The fused output network captures both shared and complementary information from all input datasets and offers the main view about how informative is each dataset by comparing the ob-

served similarity between samples. SNF can present valuable information from even a small number of samples. This algorithm is robust to noise and can be used with heterogeneous data. The fused network is used to classify subtypes among patients using network clustering.

Prior to applying the SNF to the raw dataset, we perform data preprocessing steps discussed in Section 1.4 including merging, replacing and segregating features values as well as reducing sparse samples. Following the data preparation step, we implement data imputation to eliminate missing values. Based on the features types, six sub-datasets are derived and each one is used to generate a patients similarity network. We apply the SNF method to merge all patients networks together and produce a single network. Finally, spectral clustering is applied to the fused network and then the result is evaluated with various statistical tests.

In this chapter, we describe the data imputation and sub-dataset extracting step, then we present the principles of the SNF and network clustering. Next, we see the outcome of each procedure in the results section and finally, we discuss how SNF helps us to stratify T1D patients.

3.2 Methods

Section 1.4 explains how we cleansed the original raw data and segmented it into two datasets. One entitled as "pre-processed data" which contains 436 features from 153 patients and the other entitled as "complications data" which presents ten complications information for each patient. Here, we cluster the input data by applying Similarity Network Fusion (SNF). Following topics are discussed in this section:

1. Data pre-processing for Similarity Network Fusion
2. Principles of Similarity Network Fusion
3. Network Clustering

Following the clustering step, we evaluate the results using methods outlined in Section 2.2.4 to investigate if there is any relation between a cluster patients and a disease.

3.2.1 Data Pre-Processing

SNF requires each data type as an individual similarity network. Thus, we segment the input data features into six categories while each category comprises identical data types - i.e., we partition the input data into six matrices while each one holds all samples from chosen features set. Appendix A illustrates all data features and their types and Table 3.1 presents a summary of six derived features' categories and their attributes. We name each category's data a "classified data".

SNF does not tolerate missing data. Thus, we impute missing entries in the input data using two approaches based on the features' types: 1- Mode of each column is

Table 3.1: Summary of six derived features' categories

	No. Features	Data Type	Description
Category #1	5	Ordinal	Overall health status data
Category #2	7	Numeric	Clinical data
Category #3	12	Binary	Clinical data
Category #4	4	Numeric	Demographic data
Category #5	24	Binary	Relatives clinical data
Category #6	384	Categorical	Genetic variables data

used to fill missing values in ordinal, categorical and binary features. 2- Observed entries average in a column is used to fill missing values in the numerical features.

Additionally, ordinal and numerical features are normalized using Equation 3.1 since we need to calculate samples' distances based on the futures values.

$$\tilde{f} = \frac{f - E(f)}{\sqrt{Var(f)}} \tag{3.1}$$

Where f implies any feature and $\tilde{f}$ is the corresponding feature after normalization. $E(f)$ and $Var(f)$ represent mean and variance of f, respectively [2].

3.2.2 Principles of SNF

Similarity network fusion (SNF) uses networks of samples as a basis for data integration. Although networks of samples have been used in other contexts previously, Wang et al. [2] intended patients-similarity networks for biological data integration. SNF algorithm can be wrapped-up into two steps: 1- Construction of patients' similarity network for each data type, 2- Fusing these networks toward a single similarity network by applying a nonlinear integration method. The driven fused network com-

prises both shared and complementary data among all feed sources.

Figure 3.1 illustrates a schematic example of SNF steps. Two types of data including mRNA expression and DNA methylation from the same cohort of patients are used to demonstrate this schematic. In this figure, we can observe: "(a) Example representation of mRNA expression and DNA methylation data sets for the same cohort of patients. (b) Patient-by-patient similarity matrices for each data type. (c) Patient-by-patient similarity networks, equivalent to the patient-by-patient data. Patients are represented by nodes and patients' pairwise similarities are represented by edges. (d) Network fusion by SNF iteratively updates each of the networks with information from the other networks, making them more similar with each step. (e) The iterative network fusion results in convergence to the final fused network. Edge color indicates which data type has contributed to the given similarity." This figure and its descrption obtained from Wang et al. [2].

We use Matlab® SNF software package developed by Wang et al. [2] for applying SNF algorithm to our T1D input data.

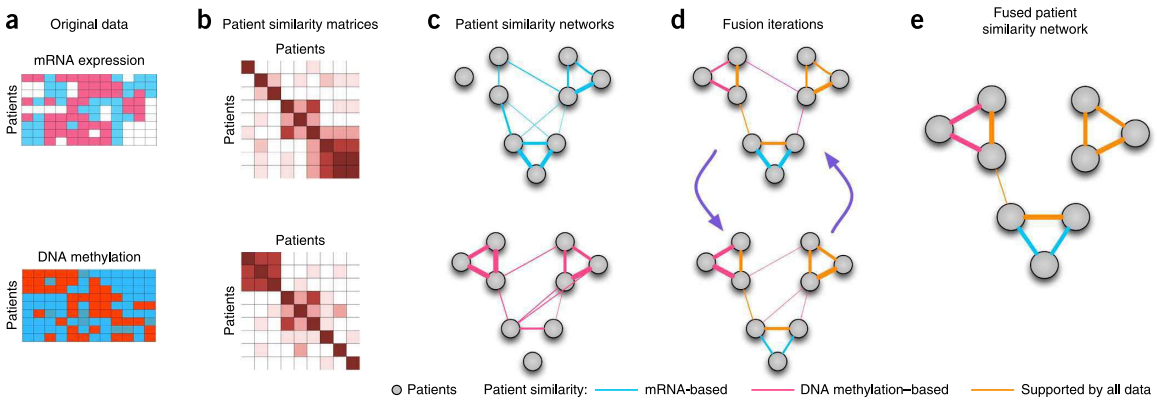

Figure 3.1: Schematic representation of SNF steps (obtained from Wang et al. [2])

The first step in SNF is to create a network for each data type. Suppose we have n samples and m features from each input data types. A patient similarity network is represented as a graph $G = (V, E)$ where vertices V present the patients $x_1, x_2, ..., x_n$ and edges E represent the similarity weight between patients. Edge weights are described by an $n \times n$ matrix W where $W(i, j)$ indicates the similarity between two corresponding patients. $W(i, j)$ is determined based on patients pairwise distance using `affinityMatrix` function in Matlab® SNF package. This function requires three inputs: a square matrix P representing patients pairwise distance, K as the number of neighbourhoods and, μ as a hyper parameter in the SNF method. A comprehensive explanation of these parameters and this function is available at Wang et al. paper [2].

As Wang et al. [2] proposed, SNF method is not sensitive to these two free parameters (μ and K). The suggested range for μ is between 0.3 to 0.8 and the rule of thumb for choosing parameter K is $K = N/C$, where N is the number of patients, and C is the number of clusters that is assumed to be in the input data. We take $\mu = 0.8$ and $K = 15$ since we have 153 patients and roughly ten clusters. We will describe a method for choosing an accurate number of the clusters in the next section. Matrix P is obtained using Matlab® `pdist` function which requires data samples and distance measurement type as inputs. We choose Squared Euclidean distance for ordinal and numerical data types and Hamming distance for categorical and binary data types.

By following the stated steps for each sub-dataset, we obtain six edge weights matrices W, each one representing a similarity network. We use `SNF` function in Matlab® SNF package to fuse all networks into a single network. This function requires three inputs: an array of obtained W matrices, K as the number of neighbourhoods and

T as the number of iterations. K is already decided earlier, and T is set to 25 as it suggested by Wang et al. paper [2]. The output is an $n \times n$ matrix which illustrates the fused network. Complete information about methods used to achieve the fused network is available at Wang et al. paper [2]

Using the `Concordance_Network_NMI` function in Matlab® SNF package, we determine the concordance among input networks and fused network based on normalized mutual information (NMI) which is described in Section 2.4.

3.2.3 Network Clustering

We obtained $n \times n$ matrix representing the fused network, and now we want to identify C clusters out of it. According to the Wang et al. paper [2] we use spectral clustering [59] which is beneficial for capturing the global structure of the given graph. This method aims to minimize RatioCut [60] by solving an optimization problem. It provides two main approaches to decide the best number of clusters: 1- Analyzing the Eigengap based on the connectivity of the network [59] 2- Analyzing the Eigenvectors of the Laplacian L [61]. We use `SpectralClustering` in Matlab® SNF package to cluster the fused network. This function requires the fused network and number of clusters C as inputs while C is determined based on the two mentioned methods.

Following the network clustering step, we have a vector which represents a cluster label for each patient. We evaluate this result by applying three statistical measures per complication to investigate if there is any relation between a cluster patients and a disease. These measures include hypergeometric test, odds ratio, and risk ratio. Detailed description of these methods are available in Section 2.2.4.

3.3 Results

The T1D input data is imputed and normalized using the methods outlined in Section 3.2.1. We segment the input data features into six categories while each category comprises identical data types. Following that, a similarity network is constructed for each sub-dataset. Obtained networks are merged by applying SNF, and finally, the fused network is clustered using network clustering. Table 3.2 represents networks concordance based on NMI percentage.

Table 3.2: Concordance among similarity networks based on NMI percentage

	Fused Network	*Network #1*	*Network #2*	*Network #3*	*Network #4*	*Network #5*	*Network #6*
Fused Network	100.0	-	-	-	-	-	-
Network #1	49.3	100.0	-	-	-	-	-
Network #2	17.9	13.5	100.0	-	-	-	-
Network #3	18.9	15.2	13.2	100.0	-	-	-
Network #4	24.9	12.5	16.0	24.2	100.0	-	-
Network #5	33.3	11.2	13.4	13.2	13.8	100.0	-
Network #6	17.6	12.9	10.6	13.1	16.6	15.9	100.0

In this section, we present the obtained clustering result; later we evaluate the result, and finally, we discuss clusters over-enriched with patients having a given complication.

3.3.1 Network Clustering

Network clustering (described in Section 3.2.3) is used to cluster the fused network. This method requires the number of clusters to be determined as an input. Two estimation methods including eigengap and rotation cost (eigenvectors of the Laplacian) suggest two and nine as the number of believed clusters, respectively. We accept nine because the other clustering approaches discovered a number of clusters closer to nine. Table 3.3 illustrates an overview of the clustering result and Table 3.4 illustrates statistics of the extracted clusters.

Table 3.3: Overview of SNF network clustering result

	Network Clustering
No. of clusters	9
Clusters' Size Average ± SD	17±9.7
Median Clusters' Size	15
Maximum Clusters' Size	37
Minimum Clusters' Size	6

Table 3.4: Statistics of the patients clusters obtained using network clustering

	Cluster Size	Male(%)	Female(%)	Weight (kg)	Age
Cluster #1	29	69.0	31.0	89.5 ± 20.8	40.0 ± 13.5
Cluster #2	16	25.0	75.0	70.3 ± 15.9	34.9 ± 5.5
Cluster #3	12	16.7	83.3	71.7 ± 10.9	53.5 ± 8.9
Cluster #4	15	46.7	53.3	84.7 ± 19.6	32.3 ± 5.5
Cluster #5	12	16.7	83.3	72.3 ± 8.9	30.5 ± 8.8
Cluster #6	37	32.4	67.6	80.5 ± 14.7	30.4 ± 7.5
Cluster #7	6	83.3	16.7	91.9 ± 18.0	30.7 ± 1.6
Cluster #8	15	53.3	46.7	70.9 ± 10.8	26.9 ± 8.9
Cluster #9	11	36.4	63.6	74.0 ± 19.2	30.5 ± 11.5

Weight and age columns values indicate corresponding average ± standard deviation.

3.3.2 Clusters Evaluation

We evaluate network clustering outcome using three statistical measures including Hypergeometric test, odds ratio, and risk ratio which are described in Section 2.2.4.

Figure 3.2 represents odds ratio per complication for each cluster. Value in each cell represents number of patients with the specified complication in the given cluster. Figure 3.3 illustrates the three most significant clusters (with the highest odds ratio) for each complication. Clusters with less than four members are filtered out. Each bubble represents the odds ratio for a cluster concerning the specified complication, and bubbles sizes are proportional to the number of patients with a given complication.

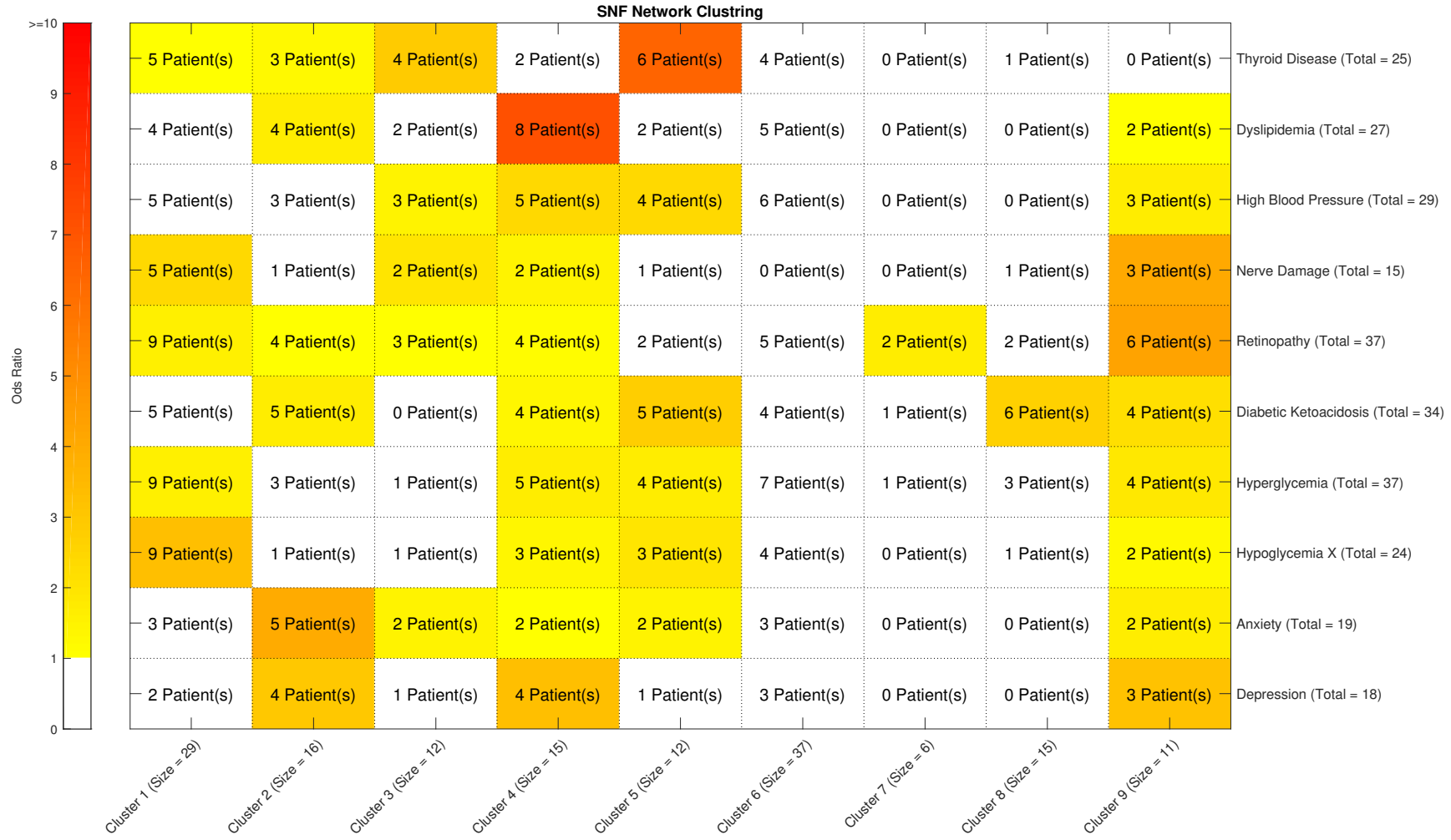

Figure 3.2: Heatmap for network clustering indicating the calculated odds ratio per complication for obtained clusters

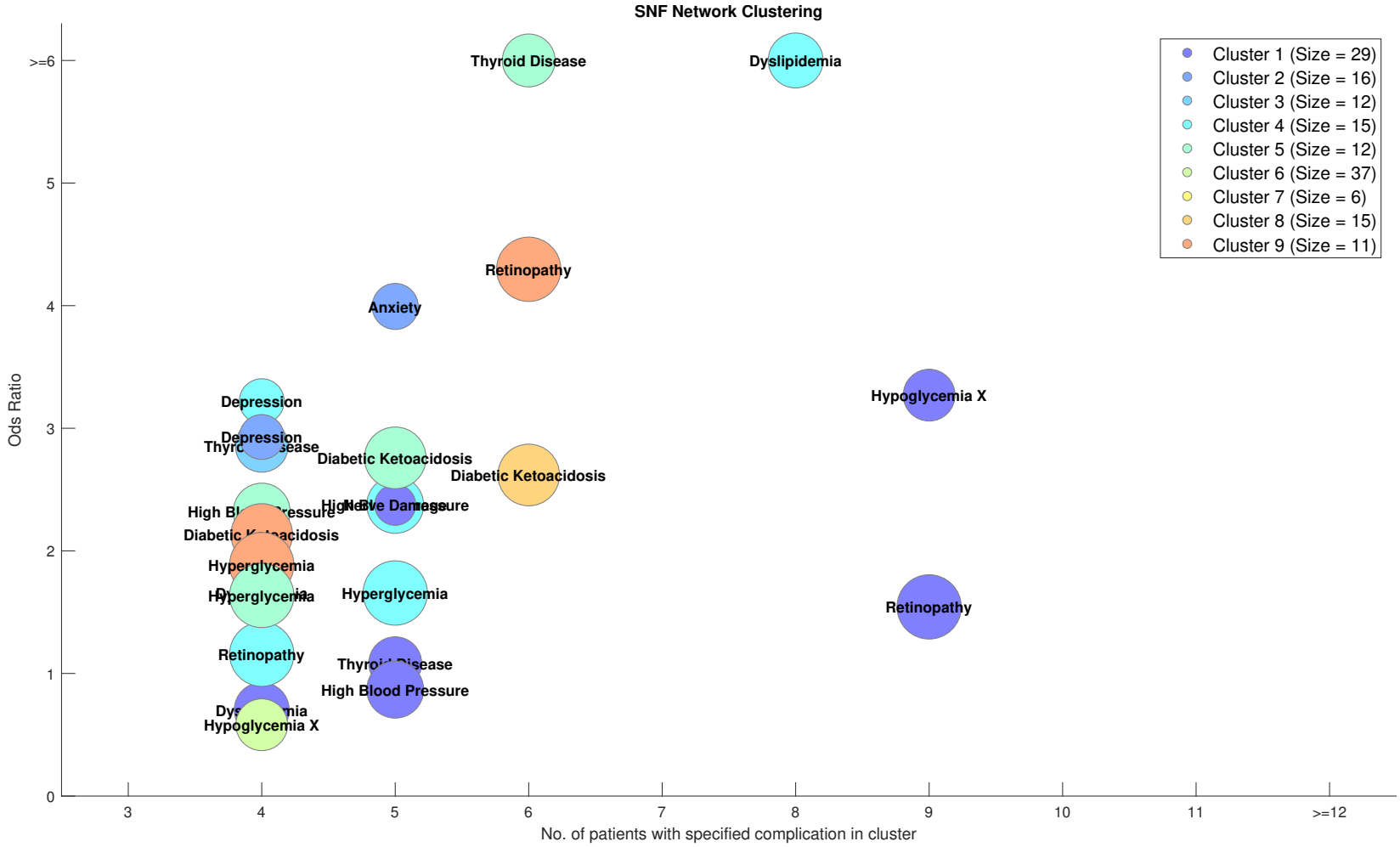

Figure 3.3: Bubble graph for network clustering result

Table 3.5 shows statistical measures for the cluster with highest odds ratio concerning each complication while clusters with less than four members are filtered out. Entire clustering evaluation measures are available Section B.1 in Appendix B. Table B.44 shows odds ratio p-value for each cluster given a complication.

Table 3.5: Most significant results per complication obtained using network clustering

	OR	p-value	Cluster #	#PCC	#CLS	#CMP
Dyslipidemia	7.16	0.001	4	8	15	27
Thyroid Disease	6.42	0.004	5	6	12	25
Retinopathy	4.30	0.023	9	6	11	37
Anxiety	3.99	0.031	2	5	16	19
Hypoglycemia X	3.27	0.016	1	9	29	24
Depression	3.22	0.080	4	4	15	18
Diabetic Ketoacidosis	2.76	0.094	5	5	12	34
Nerve Damage	2.38	0.122	1	5	29	15
High Blood Pressure	2.38	0.127	4	5	15	29
Hyperglycemia	1.89	0.257	9	4	11	37

#PCC means number of patients with specified complication in the cluster. #CLS represents total number of patients in the cluster and #CMP shows total number of patients with specified complication.

3.4 Discussion

We have an input dataset from T1D patients with no control data (no healthy samples). We segment this dataset into six categories and construct a similarity network for each data category. These networks are fused into a single network using SNF. Following that, network clustering is applied to identify groups of patients at higher risk of developing T1D complications.

By investigating Table 3.2, we can observe that Network #1 has the highest

concordance (49.3%) with the final fused network. This indicates that the fused network is significantly influenced by Network #1 which contains ordinal values from the features that represents patients overall health status (features list is available at Appendix A).

Based on the achieved network clustering result, Cluster #1, Cluster #2, Cluster #4, Cluster #5 and Cluster #9 are promising candidates for further research. Cluster #1 is enriched with Hypoglycemia patients with the odds ratio of 3.3 (with p-value = 0.016). This cluster holds 29 patients where eight patients are suffering from both Hypoglycemia and Hyperglycemia. Cluster #2 enriched with Anxiety patients with an odds ratio of 4 (with p-value = 0.031). Cluster #4 is enriched with Dyslipidemia patients with an odds ratio of 7.2 (with p-value = 0.001) and Cluster #5 is enriched with patients suffering from thyroid diseases with and an odds ratio of 6.4 (with p-value = 0.004). Cluster #9 is enriched with Retinopathy patients with an odds ratio of 4.3 (with p-value = 0.023).

As described in the second chapter, we have two significant clustering results from GLRM method (described in Section 2.3), Table 3.6 shows concordance between hierarchical clustering (GLRM) and network clustering (SNF) based on Normalized Mutual Information (NMI) percentage (described in Section 2.4). Similarly, Table 3.7 shows the concordance between affinity propagation clustering (GLRM) and network clustering (SNF). As we can observe in the mentioned tables, there is no significant concordance between network clustering and the other two methods. This low concordance may indicate that each method is interpreting data uniquely, by detecting distinct signals from the data. This is supported by the fact that distinct diseases are over-enriched in the clusters found by each approach.

Table 3.6: Concordance between hierarchical clustering and network clustering based on NMI percentage

	Cluster #1	*Cluster #2*	*Cluster #3*	*Cluster #4*	*Cluster #5*	*Cluster #6*	*Cluster #7*	*Cluster #8*	*Cluster #9*
Cluster #1	1.2	0.3	2.6	3.0	0.0	2.7	4.7	0.1	2.4
Cluster #2	4.7	0.0	6.0	0.0	3.2	0.4	0.2	0.5	0.7
Cluster #3	0.0	3.5	0.7	1.6	2.8	0.1	0.4	0.2	0.0
Cluster #4	5.0	0.0	18.9	4.4	0.4	4.9	2.4	0.9	3.6
Cluster #5	1.6	2.3	1.2	0.3	1.2	0.0	3.1	0.3	1.5
Cluster #6	1.5	0.6	1.4	1.6	1.3	0.1	0.9	1.6	1.3
Cluster #7	1.9	2.1	1.0	1.2	3.3	2.4	0.6	2.4	0.9
Cluster #8	3.3	3.5	0.0	0.2	2.8	0.1	0.4	3.3	0.9
Cluster #9	0.5	2.8	2.3	0.9	0.0	0.8	1.4	0.0	2.2

Rows and columns indicate hierarchical clustering result and network clustering result, respectively. A higher NMI percentage implies greater concordance between clusters.

Table 3.7: Concordance between affinity propagation clustering and network clustering based on NMI percentage

	Cluster #1	*Cluster #2*	*Cluster #3*	*Cluster #4*	*Cluster #5*	*Cluster #6*	*Cluster #7*	*Cluster #8*	*Cluster #9*
Cluster #1	0.5	1.2	3.3	1.2	1.0	2.4	0.6	1.2	3.7
Cluster #2	7.9	0.6	0.5	0.6	0.5	1.2	0.3	0.6	0.5
Cluster #3	1.0	2.5	2.4	2.8	2.4	0.5	5.3	2.8	1.7
Cluster #4	0.0	3.8	0.1	3.6	0.4	0.2	2.0	6.0	0.1
Cluster #5	2.1	0.5	0.1	0.1	0.4	0.9	2.0	0.1	0.6
Cluster #6	0.4	0.0	3.5	4.1	1.4	0.3	0.1	0.0	0.2
Cluster #7	2.4	0.0	6.2	0.6	5.7	0.9	0.2	2.6	1.4
Cluster #8	2.6	0.0	0.2	4.8	0.2	1.8	2.4	2.4	3.2
Cluster #9	1.0	13.0	0.5	0.6	0.5	1.2	0.3	0.6	0.5
Cluster #10	0.1	0.6	3.2	0.5	0.3	0.6	2.0	0.5	2.3
Cluster #11	0.4	1.3	2.0	1.6	2.6	0.0	1.3	2.3	1.9

Rows and columns indicate affinity propagation clustering result and network clustering result, respectively. A higher NMI percentage implies greater concordance between clusters.

As a future work, one would need to investigate what factors increase the risk of developing T1D related complications such as Dyslipidemia and thyroid diseases since we found clusters enriched with these complications.

3.5 Conclusion

Computational methods are needed to investigate large heterogeneous data and create a comprehensive view of a biological process such as a diseases developing. We use Similarity Network Fusion (SNF) to integrate clinical, demographic and genetic data to aid patients stratification and to distinguish those at a higher risk of developing T1D complications. We take following steps to achieve this:

1. We impute T1D input dataset to eliminate missing entries and then segment this dataset into six parts while each segment comprises identical feature types.

2. A sample similarity network is constructed for each of the six sub-datasets.

3. These similarity networks are fused to a single network by applying a nonlinear integration method.

4. The fused network samples are clustered using the network clustering method.

5. Finally, the obtained clustering result is evaluated with the patients' complications information and clusters that are over-enriched with patients having secondary diseases related to T1D are discussed.

SNF can extract valuable information from even a small number of samples, this algorithm is robust to noise and can be used with heterogeneous data. According

to the achieved result, we identified clusters enriched with patients suffering from Dyslipidemia and thyroid diseases. Consequently, we have taken first steps to identify groups of patients at higher risk of developing T1D complications. This results can be taken as the basis to develop a predictive model that can allow patients and health-care providers to take preemptive steps to reduce the risk of developing T1D related complications based on each patient characteristics.

Chapter 4

Summary

Diabetes mellitus type 1 also known as type 1 diabetes (T1D) is a disease in which the body immune system attacks the β-cells. As a result, very little, or no insulin is released to control the level of glucose in the blood. This research investigates whether groups of patients at higher risk for developing of complications or secondary disease related to T1D can be identified by integrating demographic, clinical and genetic data.

Our dataset is collected from a cohort study regarding the incidence of childhood T1D on the Avalon Peninsula of Newfoundland, Canada This region has one of the highest incidences of T1D reported worldwide [62, 63]. The obtained raw T1D dataset is a heterogeneous dataset, with 27.2% missing values, which has 239 features from 196 patients. We first perform some pre-processing steps on the raw T1D dataset to extract a pre-processed dataset and a complication matrix. The pre-processes dataset comprises 153 rows (patients) with 436 features regarding patients' demographic, clinical and genetic data. The complications matrix is a binary data with 153 rows

(patients) and ten columns (complications). It indicates whether a patient is suffering from the given T1D secondary diseases or not.

We investigate T1D preprocessed dataset using two approaches namely Generalized Low Rank Modeling (GLRM) and Similarity Network Fusion (SNF), then we evaluate the methods outcomes to find clusters enriched with patients suffering from a specific complication. Figure 4.1 illustrates our work-flow.

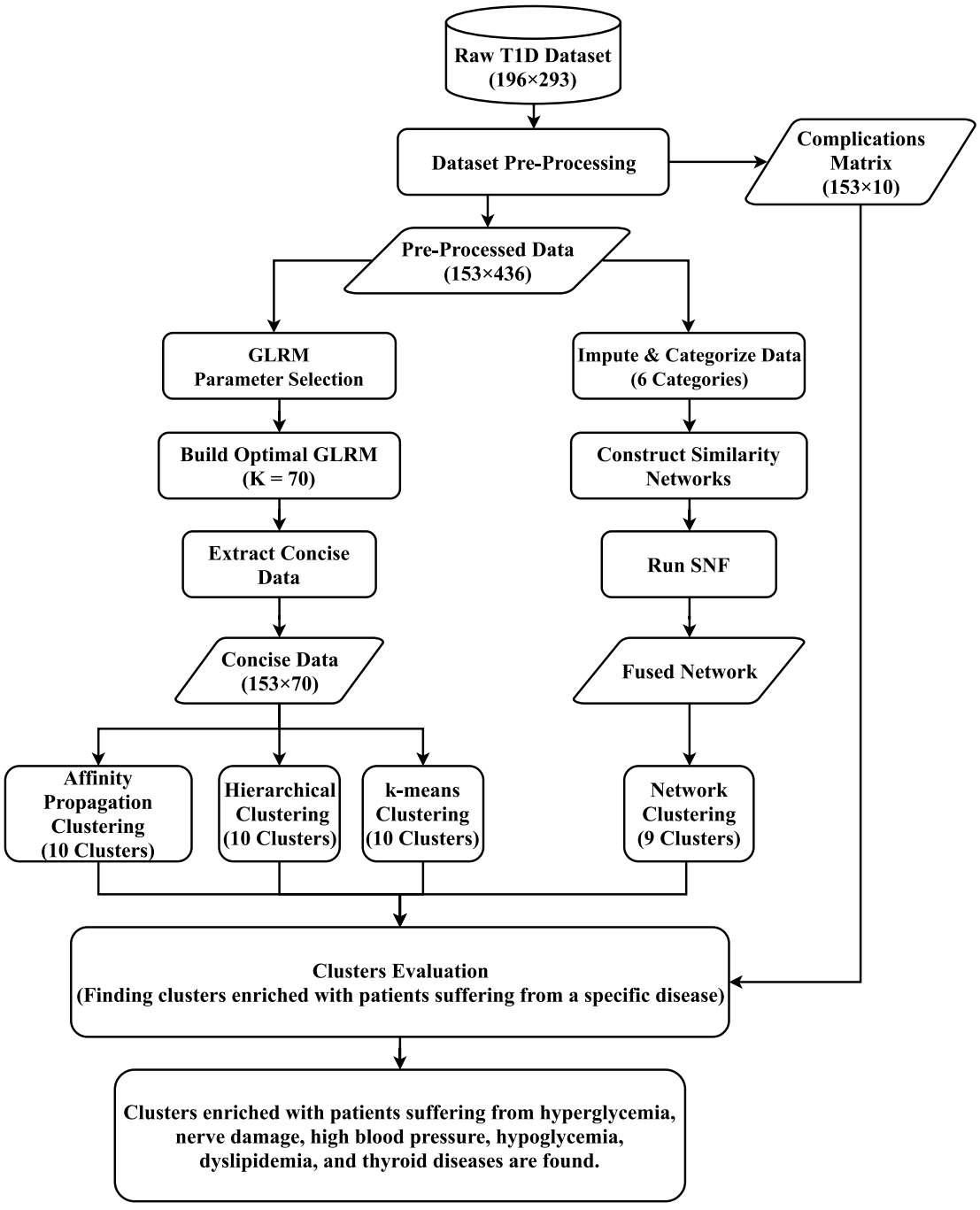

Figure 4.1: Research summary diagram

We use GLRM approach for analyzing our input dataset and to discover patients subgroups more likely to have a given complication. This framework handles heterogeneous datasets, compresses, and de-noises data, and imputes missing records. It represents high dimensional bulk data into a lower-dimensional space. The framework output is two matrices: a tall $m \times k$ matrix and a wide $k \times n$ matrix; where m is the number of patients, n is the number of features in the input dataset, and k is the number of latent features produced by GLRM. We name the tall matrix as "concise data" which represents the original data feature into k new latent features. Next, we cluster concise data using three clustering methods including k-means clustering, hierarchical clustering, and affinity propagation clustering. Each clustering result is then evaluated using the complications matrix to find clusters enriched with a specified complication. With the GLRM approach, we could identify clusters enriched with patients suffering from Hyperglycemia, nerve damage, and high blood pressure.

The other used approach is SNF. SNF uses networks of samples as a basis for data integration. Its algorithm can be wrapped-up into two steps: 1- Construction of sample similarity network for each data type, 2- Fusing these networks toward a single similarity network by applying a nonlinear integration method. The driven fused network comprises both shared and complementary data among all feed sources. We categorize our T1D pre-processed dataset into six parts (each part consists of similar data types). Then, we construct a similarity network for each sub-dataset and use SNF to fuse all these similarity networks into a single network. The fused network is clustered using network clustering, and then the result is evaluated using the complications matrix to find clusters enriched with a specified complication. With the SNF approach, we could identify clusters enriched with patients suffering from

Hypoglycemia, Dyslipidemia, and thyroid diseases.

As a result of our research, we have taken first steps to identify groups of patients at higher risk of developing T1D complications. This results could be taken as the basis to create a predictive model that could allow patients and health-care providers to take preemptive steps to reduce the risk of developing T1D related complications based on each patient characteristics. It also could be used by clinical experts to guide further investigation on T1D. As an extension of this research, other approaches could also be applied to the same dataset and compare the obtained results.

Bibliography

[1] Madeleine Udell, Corinne Horn, Reza Zadeh, and Stephen Boyd. Generalized low rank models. *Foundations and Trends in Machine Learning*, 9(1):1–118, 2016.

[2] Bo Wang, Aziz M Mezlini, Feyyaz Demir, Marc Fiume, Zhuowen Tu, Michael Brudno, Benjamin Haibe-Kains, and Anna Goldenberg. Similarity network fusion for aggregating data types on a genomic scale. *Nature methods*, 11(3):333–337, 2014.

[3] World Health Organization (WHO). Diabetes. `http://www.who.int/mediacentre/factsheets/fs312/en/`, June 2016.

[4] The National Institute of Diabetes and Digestive and Kidney Diseases. Causes of diabetes. In *National Diabetes Information Clearinghouse*, number 14-5164. NIH Publication, June 2014.

[5] Denis Daneman. Type 1 diabetes. *The Lancet*, 367(9513):847–858, 2006.

[6] Jonathan P Bradfield, Hui-Qi Qu, Kai Wang, Haitao Zhang, Patrick M Sleiman, Cecilia E Kim, Frank D Mentch, Haijun Qiu, Joseph T Glessner, Kelly A Thomas, et al. A genome-wide meta-analysis of six type 1 diabetes cohorts identifies multiple associated loci. *PLoS genetics*, 7(9):e1002293, 2011.

[7] Guillermo E. Umpierrez, Kashif A. Latif, Mary Beth Murphy, Helen C. Lambeth, Frankie Stentz, Andrew Bush, and Abbas E. Kitabchi. Thyroid dysfunction in patients with type 1 diabetes. *Diabetes Care*, 26(4):1181–1185, 2003.

[8] John E Hall. *Guyton and Hall textbook of medical physiology e-Book*. Elsevier Health Sciences, 2015.

[9] David M. Maahs and Robert H. Eckel. *Type 1 Diabetes Mellitus and Dyslipidemia*, pages 115–135. Humana Press, Totowa, NJ, 2015.

[10] Dr Colin Tidy. Diabetes and high blood pressure. `https://patient.info/health/diabetes-mellitus-leaflet/diabetes-and-high-blood-pressure`, Oct. 2017.

[11] National Institute of Diabetes and Digestive and Kidney Diseases (NIDDK). Diabetic neuropathy. `https://www.niddk.nih.gov/health-information/diabetes/overview/preventing-problems/nerve-damage-diabetic-neuropathies`, 2018.

[12] Donald S. Fong, Lloyd Aiello, Thomas W. Gardner, George L. King, George Blankenship, Jerry D. Cavallerano, Fredrick L. Ferris, and Ronald Klein. Retinopathy in diabetes. *Diabetes Care*, 27(suppl 1):s84–s87, 2004.

[13] Cengiz Eda, Xing Dongyuan, Wong Jenise C., Wolfsdorf Joseph I., Haymond Morey W., Rewers Arleta, Shanmugham Satya, Tamborlane William V., Willi Steven M., Seiple Diane L., Miller Kellee M., DuBose Stephanie N., and Beck Roy W. and. Severe hypoglycemia and diabetic ketoacidosis among youth

with type 1 diabetes in the t1d exchange clinic registry. *Pediatric Diabetes*, 14(6):447–454, 2013.

[14] Diagnosis and classification of diabetes mellitus. *Diabetes Care*, 37(Supplement 1):S81–S90, 2014.

[15] Rory J. McCrimmon and Robert S. Sherwin. Hypoglycemia in type 1 diabetes. *Diabetes*, 59(10):2333–2339, 2010.

[16] Michele Herzer and Korey K Hood. Anxiety symptoms in adolescents with type 1 diabetes: association with blood glucose monitoring and glycemic control. *Journal of pediatric psychology*, 35(4):415–425, 2009.

[17] E. A. M. Gale. Type 1 diabetes in the young: the harvest of sorrow goes on. *Diabetologia*, 48(8):1435–1438, Aug 2005.

[18] Jason D Cooper, Deborah J Smyth, Adam M Smiles, Vincent Plagnol, Neil M Walker, James E Allen, Kate Downes, Jeffrey C Barrett, Barry C Healy, Josyf C Mychaleckyj, et al. Meta-analysis of genome-wide association study data identifies additional type 1 diabetes risk loci. *Nature genetics*, 40(12):1399, 2008.

[19] Lorenza Nisticò, Raffaella Buzzetti, Lynn E Pritchard, Bart Van der Auwera, Claudio Giovannini, Emanuele Bosi, Maria Teresa Martinez Larrad, Manuel Serrano Rios, CC Chow, Clive S Cockram, et al. The ctla-4 gene region of chromosome 2q33 is linked to, and associated with, type 1 diabetes. *Human molecular genetics*, 5(7):1075–1080, 1996.

[20] Mark A Atkinson, George S Eisenbarth, and Aaron W Michels. Type 1 diabetes. *The Lancet*, 383(9911):69 – 82, 2014.

[21] June L Davies, Yoshihiko Kawaguchi, Simon T Bennett, James B Copeman, Heather J Cordell, Lynn E Pritchard, Peter W Reed, Stephen CL Gough, Suzanne C Jenkins, Sheila M Palmer, et al. A genome-wide search for human type 1 diabetes susceptibility genes. *Nature*, 371(6493):130–136, 1994.

[22] Jeffrey C Barrett, David G Clayton, Patrick Concannon, Beena Akolkar, Jason D Cooper, Henry A Erlich, Cécile Julier, Grant Morahan, Jørn Nerup, Concepcion Nierras, et al. Genome-wide association study and meta-analysis find that over 40 loci affect risk of type 1 diabetes. *Nature genetics*, 41(6):703–707, 2009.

[23] Jeffrey D. Roizen, Jonathan P. Bradfield, and Hakon Hakonarson. Progress in understanding type 1 diabetes through its genetic overlap with other autoimmune diseases. *Current Diabetes Reports*, 15(11):1–7, 2015.

[24] Jemila S Hamid, Pingzhao Hu, Nicole M Roslin, Vicki Ling, Celia MT Greenwood, and Joseph Beyene. Data integration in genetics and genomics: methods and challenges. *Human genomics and proteomics: HGP*, 2009, 2009.

[25] Xianwen Ren, Hua Fu, and Qi Jin. Integrating heterogeneous genomic data to accurately identify disease subtypes. *BMC medical genomics*, 8(1):78, 2015.

[26] Roger Higdon, Rachel K Earl, Larissa Stanberry, Caitlin M Hudac, Elizabeth Montague, Elizabeth Stewart, Imre Janko, John Choiniere, William Broomall, Natali Kolker, et al. The promise of multi-omics and clinical data integration

to identify and target personalized healthcare approaches in autism spectrum disorders. *OMICS a Journal of Integrative Biology*, 19(4):197–208, 2015.

[27] Chuanchao Zhang, Jiguang Wang, Chao Zhang, Juan Liu, Dong Xu, and Luonan Chen. Network stratification analysis for identifying function-specific network layers. *Molecular BioSystems*, 12(4):1232–1240, 2016.

[28] Xue Zhong, Hushan Yang, Shuyang Zhao, Yu Shyr, and Bingshan Li. Network-based stratification analysis of 13 major cancer types using mutations in panels of cancer genes. *BMC genomics*, 16(7):1–8, 2015.

[29] Ronglai Shen, Adam B Olshen, and Marc Ladanyi. Integrative clustering of multiple genomic data types using a joint latent variable model with application to breast and lung cancer subtype analysis. *Bioinformatics*, 25(22):2906–2912, 2009.

[30] Sunghwan Kim, Steffi Oesterreich, Seyoung Kim, Yongseok Park, and George C Tseng. Integrative clustering of multi-level omics data for disease subtype discovery using sequential double regularization. *Biostatistics*, 18(1):165–179, 2017.

[31] Sara Hillenmeyer, Lea K Davis, Eric R Gamazon, Edwin H Cook, Nancy J Cox, and Russ B Altman. Stams: String-assisted module search for genome wide association studies and application to autism. *Bioinformatics*, pages 1–8, 2016.

[32] DongYeon Cho, Yoo-Ah Kim, and Teresa M Przytycka. Network biology approach to complex diseases. *PLoS Computational Biology*, 8(12), 2012.

[33] Chao Yang, Shu-Guang Ge, and Chun-Hou Zheng. ndmasnf: cancer subtype discovery based on integrative framework assisted by network diffusion model. *Oncotarget*, 8(51):89021, 2017.

[34] H. Wang, H. Zheng, J. Wang, C. Wang, and F. X. Wu. Integrating omics data with a multiplex network-based approach for the identification of cancer subtypes. *IEEE Transactions on NanoBioscience*, 15(4):335–342, June 2016.

[35] Nora K Speicher and Nico Pfeifer. Integrating different data types by regularized unsupervised multiple kernel learning with application to cancer subtype discovery. *Bioinformatics*, 31(12):i268–i275, 2015.

[36] Alejandro Schuler, Vincent Liu, Joe Wan, Alison Callahan, Madeleine Udell, David E Stark, and Nigam H Shah. Discovering patient phenotypes using generalized low rank models. In *Pacific Symposium on Biocomputing*, volume 21, pages 144–155, 2016.

[37] Jonathan D. Young, Chunhui Cai, and Xinghua Lu. Unsupervised deep learning reveals prognostically relevant subtypes of glioblastoma. *BMC Bioinformatics*, 18(11):381, Oct 2017.

[38] Thomas A Lasko, Joshua C Denny, and Mia A Levy. Computational phenotype discovery using unsupervised feature learning over noisy, sparse, and irregular clinical data. *PloS one*, 8(6), June 2013.

[39] Zhi Wei, Kai Wang, Hui-Qi Qu, Haitao Zhang, Jonathan Bradfield, Cecilia Kim, Edward Frackleton, Cuiping Hou, Joseph T. Glessner, Rosetta Chiavacci, Charles Stanley, Dimitri Monos, Struan F. A. Grant, Constantin Polychronakos, and

Hakon Hakonarson. From disease association to risk assessment: An optimistic view from genome-wide association studies on type 1 diabetes. *PLOS Genetics*, 5(10):1–11, 10 2009.

[40] Leigh A. Newhook, Joseph Curtis, Donna Hagerty, Marie Grant, Andrew D. Paterson, Cheryl Crummel, Tracey Bridger, and Patrick Parfrey. High incidence of childhood type 1 diabetes in the avalon peninsula, newfoundland, canada. *Diabetes Care*, 27(4):885–888, 2004.

[41] Ian T Jolliffe. Principal component analysis and factor analysis. *Principal component analysis*, pages 150–166, 2002.

[42] James MacQueen et al. Some methods for classification and analysis of multivariate observations. In *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*, volume 1, pages 281–297. Oakland, CA, USA., 1967.

[43] Jerome Friedman, Trevor Hastie, and Robert Tibshirani. *The elements of statistical learning*, volume 1. Springer series in statistics New York, 2001.

[44] Brendan J Frey and Delbert Dueck. Clustering by passing messages between data points. *science*, 315(5814):972–976, 2007.

[45] Jan Van den Broeck, Solveig Argeseanu Cunningham, Roger Eeckels, and Kobus Herbst. Data cleaning: detecting, diagnosing, and editing data abnormalities. *PLoS medicine*, 2(10):e267, 2005.

[46] The H2O.ai team. *h2o: R Interface for H2O*, 2017. R package version 3.14.0.2.

[47] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2015.

[48] Raghunandan H Keshavan, Andrea Montanari, and Sewoong Oh. Matrix completion from a few entries. *IEEE Transactions on Information Theory*, 56(6):2980–2998, 2010.

[49] MATLAB. *version 9.1.0.441655 (R2016b)*. The MathWorks Inc., Natick, Massachusetts, 2016.

[50] MATLAB. *MathWorks, (2016) Statistics and Machine Learning Toolbox User's Guide (R2016b)*. The MathWorks Inc., 2016.

[51] Pradeep Rai and Shubha Singh. A survey of clustering techniques. *International Journal of Computer Applications*, 7(12):1–5, 2010.

[52] Ulrich Bodenhofer, Andreas Kothmeier, and Sepp Hochreiter. Apcluster: an r package for affinity propagation clustering. *Bioinformatics*, 27(17):2463–2464, 2011.

[53] AleÅ¡ Berkopec. Hyperquick algorithm for discrete hypergeometric distribution. *Journal of Discrete Algorithms*, 5(2):341 – 347, 2007. 2004 Symposium on String Processing and Information Retrieval.

[54] Christopher L. Sistrom and Cynthia W. Garvan. Proportions, odds, and risk. *Radiology*, 230(1):12–19, 2004. PMID: 14695382.

[55] Magdalena Szumilas. Explaining odds ratios. *Journal of the Canadian Academy of Child and Adolescent Psychiatry*, 19(3):227, 2010.

[56] Margaret Sullivan Pepe, Holly Janes, Gary Longton, Wendy Leisenring, and Polly Newcomb. Limitations of the odds ratio in gauging the performance of a diagnostic, prognostic, or screening marker. *American journal of epidemiology*, 159(9):882–890, 2004.

[57] Nguyen Xuan Vinh, Julien Epps, and James Bailey. Information theoretic measures for clusterings comparison: Variants, properties, normalization and correction for chance. *Journal of Machine Learning Research*, 11(Oct):2837–2854, 2010.

[58] Alexander Strehl and Joydeep Ghosh. Cluster ensembles—a knowledge reuse framework for combining multiple partitions. *Journal of machine learning research*, 3:583–617, 2002.

[59] Andrew Y Ng, Michael I Jordan, and Yair Weiss. On spectral clustering: Analysis and an algorithm. In *Advances in neural information processing systems*, pages 849–856, 2002.

[60] Yen-Chuen Wei and Chung-Kuan Cheng. Towards efficient hierarchical designs by ratio cut partitioning. In *Computer-Aided Design, 1989. ICCAD-89. Digest of Technical Papers., 1989 IEEE International Conference on*, pages 298–301. IEEE, 1989.

[61] Lihi Zelnik-Manor and Pietro Perona. Self-tuning spectral clustering. In *Advances in neural information processing systems*, pages 1601–1608, 2005.

[62] LA Newhook, M Grant, S Sloka, M Hoque, AD Paterson, D Hagerty, and J Curtis. Very high and increasing incidence of type 1 diabetes mellitus in newfoundland and labrador, canada. *Pediatric diabetes*, 9(3pt2):62–68, 2008.

[63] Leigh A Newhook, Sharon Penney, Jackie Fiander, and Jeff Dowden. Recent incidence of type 1 diabetes mellitus in children 0–14 years in newfoundland and labrador, canada climbs to over 45/100,000: a retrospective time trend study. *BMC research notes*, 5(1):628, 2012.

Appendix A

Dataset Features Information

This appendix represents a pre-processed T1D dataset features details and their specification.

Table A.1: Dataset Features Details

Feature Description	Feature Type	Feature Category	GLRM Loss Type	SNF Network No.
Amount of insulin taken	Ordinal	Patient Clinical Data	Ordinal	1
Most recent urinalysis result	Ordinal	Patient Clinical Data	Ordinal	1
Rate of glucose monitoring	Ordinal	Patient Clinical Data	Ordinal	1
Current health rating	Ordinal	Patient Clinical Data	Ordinal	1
Current rating of diabetes management	Ordinal	Patient Clinical Data	Ordinal	1
Most recent cholesterol/HDL ratio	Numeric	Patient Clinical Data	Quadratic	2
Most recent creatinine	Numeric	Patient Clinical Data	Quadratic	2
Most recent fasting blood glucose	Numeric	Patient Clinical Data	Quadratic	2
Most recent hemoglobin A1C	Numeric	Patient Clinical Data	Quadratic	2
Most recent Microalbumin/Creatinine ratio	Numeric	Patient Clinical Data	Quadratic	2
Most recent thyroid stimulating hormone level	Numeric	Patient Clinical Data	Quadratic	2
Blood glucose if taken at diagnosis	Numeric	Patient Clinical Data	Quadratic	2
ACE inhibitor used	Binary	Patient Clinical Data	Logistic	3
BP meds used?	Binary	Patient Clinical Data	Logistic	3
Statins used?	Binary	Patient Clinical Data	Logistic	3
Gender	Binary	Patient Clinical Data	Logistic	3
Flu vaccinated?	Binary	Patient Clinical Data	Logistic	3
Pneummococcal vaccinated	Binary	Patient Clinical Data	Logistic	3
Exercise regularly	Binary	Patient Clinical Data	Logistic	3
Self reported difficulty with hypoglycemia	Binary	Patient Clinical Data	Logistic	3
On dialysis?	Binary	Patient Clinical Data	Logistic	3
DKA present on diagnosis	Binary	Patient Clinical Data	Logistic	3
Are Ketones in urine	Binary	Patient Clinical Data	Logistic	3
Whether he/she smokes	Binary	Patient Clinical Data	Logistic	3
Weight (lbs)	Numeric	Patient Demographic Data	Quadratic	4

Table A.1: Dataset Features Details

Feature Description	Feature Type	Feature Category	GLRM Loss Type	SNF Network No.
Height (kg)	Numeric	Patient Demographic Data	Quadratic	4
Date of birth (Age)	Numeric	Patient Demographic Data	Quadratic	4
Date of diagnosis (Age)	Numeric	Patient Demographic Data	Quadratic	4
Relative with T1D (Brother)	Binary	Relative Clinical Data	Logistic	5
Relative with T1D (Sister)	Binary	Relative Clinical Data	Logistic	5
Relative with T1D (Father)	Binary	Relative Clinical Data	Logistic	5
Relative with T1D (Maternal Aunt)	Binary	Relative Clinical Data	Logistic	5
Relative with T1D (Paternal Aunt)	Binary	Relative Clinical Data	Logistic	5
Relative with T1D (Maternal Uncle)	Binary	Relative Clinical Data	Logistic	5
Relative with T1D (Paternal Uncle)	Binary	Relative Clinical Data	Logistic	5
Relative with T1D (Maternal Cousin)	Binary	Relative Clinical Data	Logistic	5
Relative with T1D (Paternal Cousin)	Binary	Relative Clinical Data	Logistic	5
First degree relative has had Lupus	Binary	Relative Clinical Data	Logistic	5
First degree relative has had Thyroid Disease	Binary	Relative Clinical Data	Logistic	5
Crohns disease Complication	Binary	Relative Clinical Data	Logistic	5
Family History of Graves Disease	Binary	Relative Clinical Data	Logistic	5
Family History of Organ Transplantation	Binary	Relative Clinical Data	Logistic	5
Autoimmune disease (Maternal Grandmother)	Binary	Relative Clinical Data	Logistic	5
Autoimmune disease (Maternal Grandfather)	Binary	Relative Clinical Data	Logistic	5
Autoimmune disease (Paternal Grandmother)	Binary	Relative Clinical Data	Logistic	5
Autoimmune disease (Paternal Grandfather)	Binary	Relative Clinical Data	Logistic	5
Parent with MI	Binary	Relative Clinical Data	Logistic	5
Parent with angina	Binary	Relative Clinical Data	Logistic	5
Parent with CVA	Binary	Relative Clinical Data	Logistic	5
Parent with hypertension	Binary	Relative Clinical Data	Logistic	5

Table A.1: Dataset Features Details

Feature Description	Feature Type	Feature Category	GLRM Loss Type	SNF Network No.
Parent with hyperlipidemia	Binary	Relative Clinical Data	Logistic	5
Parent with PVD	Binary	Relative Clinical Data	Logistic	5
Genetic Marker (95 features)	Categorical	Patient Genotype Data	Categorical	6
Genetic Marker (289 features)	Binary	Patient Genotype Data	Logistic	6

Appendix B

Supplementary Clustering Data

B.1 Clustering Results

This section includes result tables from four clustering approaches used in this report. P_HG indicates hypergeometric test p-value, RR indicates Risk Ratio and OR presents Odds Ratios. #PCC means number of patients with specified complication in the cluster. #CLS represents total number of patients in the cluster and #CMP stands for total number of patients with specified complication.

B.1.1 K-means Clustering

Table B.1: k-means clustering results for Thyroid Disease

P_{HG}	RR	OR	Cluster #	#PCC	#CLS	#CMP
1.41E-02	2.84	4.23	5	3	7	25
1.10E-01	1.60	1.80	4	3	12	25
1.63E-01	1.32	1.39	9	12	63	25
4.63E-01	0.89	0.87	6	4	27	25
4.09E-01	0.86	0.84	1	2	14	25
3.87E-01	0.76	0.72	8	1	8	25
4.17E-01	0.00	0.00	2	0	3	25
8.70E-01	0.00	0.00	3	0	11	25
5.14E-01	0.00	0.00	7	0	4	25
5.14E-01	0.00	0.00	10	0	4	25

Table B.2: k-means clustering results for Dyslipidemia

P_{HG}	RR	OR	Cluster #	#PCC	#CLS	#CMP
2.57E-02	2.25	2.96	3	4	11	27
1.49E-01	1.45	1.60	8	2	8	27
1.44E-01	1.43	1.58	10	1	4	27
2.16E-01	1.24	1.31	1	3	14	27
2.74E-01	1.14	1.18	9	12	63	27
5.43E-01	0.81	0.78	6	4	27	27
3.58E-01	0.80	0.77	5	1	7	27
4.44E-01	0.00	0.00	2	0	3	27
9.12E-01	0.00	0.00	4	0	12	27
5.44E-01	0.00	0.00	7	0	4	27

Table B.3: k-means clustering results for High Blood Pressure

P_{HG}	RR	OR	Cluster #	#PCC	#CLS	#CMP
1.20E-02	2.10	2.65	6	9	27	29
1.63E-01	1.33	1.44	7	1	4	29
1.63E-01	1.33	1.44	10	1	4	29
2.60E-01	1.15	1.19	1	3	14	29
4.05E-01	1.01	1.01	9	12	63	29
3.47E-01	0.96	0.95	3	2	11	29
4.69E-01	0.65	0.60	8	1	8	29
4.70E-01	0.00	0.00	2	0	3	29
9.28E-01	0.00	0.00	4	0	12	29
7.78E-01	0.00	0.00	5	0	7	29

Table B.4: k-means clustering results for Nerve Damage

P_{HG}	RR	OR	Cluster #	#PCC	#CLS	#CMP
4.81E-02	2.66	3.21	7	1	4	15
9.81E-02	1.70	1.82	6	4	27	15
1.43E-01	1.53	1.62	1	2	14	15
1.78E-01	1.29	1.34	8	1	8	15
2.94E-01	0.92	0.91	3	1	11	15
3.33E-01	0.84	0.82	4	1	12	15
6.41E-01	0.71	0.69	9	5	63	15
2.68E-01	0.00	0.00	2	0	3	15
5.22E-01	0.00	0.00	5	0	7	15
3.41E-01	0.00	0.00	10	0	4	15

Table B.5: k-means clustering results for Retinopathy

P_{HG}	RR	OR	Cluster #	#PCC	#CLS	#CMP
2.06E-02	2.20	3.39	8	4	8	37
2.26E-01	1.20	1.28	1	4	14	37
2.23E-01	1.19	1.27	5	2	7	37
2.59E-01	1.14	1.19	3	3	11	37
2.46E-01	1.03	1.05	7	1	4	37
2.46E-01	1.03	1.05	10	1	4	37
4.95E-01	0.90	0.88	6	6	27	37
6.09E-01	0.87	0.83	9	14	63	37
5.91E-01	0.67	0.61	4	2	12	37
5.67E-01	0.00	0.00	2	0	3	37

Table B.6: k-means clustering results for Diabetic Ketoacidosis

P_{HG}	RR	OR	Cluster #	#PCC	#CLS	#CMP
7.23E-04	3.05	5.91	4	7	12	34
1.53E-02	1.81	2.16	9	19	63	34
1.24E-01	1.52	1.77	2	1	3	34
2.14E-01	1.13	1.17	7	1	4	34
6.42E-01	0.62	0.56	1	2	14	34
9.03E-01	0.45	0.38	6	3	27	34
7.49E-01	0.39	0.33	3	1	11	34
8.35E-01	0.00	0.00	5	0	7	34
8.73E-01	0.00	0.00	8	0	8	34
6.38E-01	0.00	0.00	10	0	4	34

Table B.7: k-means clustering results for Hyperglycemia

P_{HG}	RR	OR	Cluster #	#PCC	#CLS	#CMP
8.89E-03	2.27	3.55	4	6	12	37
4.41E-02	2.13	3.26	10	2	4	37
2.22E-02	1.68	2.00	9	20	63	37
1.45E-01	1.39	1.58	2	1	3	37
5.24E-01	0.74	0.68	3	2	11	37
8.43E-01	0.57	0.49	6	4	27	37
6.18E-01	0.50	0.43	8	1	8	37
8.98E-01	0.28	0.22	1	1	14	37
8.62E-01	0.00	0.00	5	0	7	37
6.74E-01	0.00	0.00	7	0	4	37

Table B.8: k-means clustering results for Hypoglycemia X

P_{HG}	RR	OR	Cluster #	#PCC	#CLS	#CMP
2.32E-02	2.35	3.02	4	4	12	24
5.20E-02	1.69	1.87	9	13	63	24
1.16E-01	1.62	1.83	7	1	4	24
1.16E-01	1.62	1.83	10	1	4	24
3.02E-01	0.91	0.89	5	1	7	24
6.52E-01	0.67	0.63	6	3	27	24
6.82E-01	0.43	0.39	1	1	14	24
4.03E-01	0.00	0.00	2	0	3	24
8.57E-01	0.00	0.00	3	0	11	24
7.53E-01	0.00	0.00	8	0	8	24

Table B.9: k-means clustering results for Anxiety

P_{HG}	RR	OR	Cluster #	#PCC	#CLS	#CMP
4.15E-02	2.45	3.04	5	2	7	19
3.23E-02	2.42	2.95	3	3	11	19
7.55E-02	2.07	2.43	7	1	4	19
2.23E-01	1.24	1.29	6	4	27	19
2.42E-01	1.17	1.20	1	2	14	19
7.43E-01	0.66	0.62	9	6	63	19
4.55E-01	0.65	0.62	4	1	12	19
3.30E-01	0.00	0.00	2	0	3	19
6.63E-01	0.00	0.00	8	0	8	19
4.15E-01	0.00	0.00	10	0	4	19

Table B.10: k-means clustering results for Depression

P_{HG}	RR	OR	Cluster #	#PCC	#CLS	#CMP
6.82E-02	2.19	2.59	10	1	4	18
1.23E-01	1.61	1.75	3	2	11	18
1.88E-01	1.33	1.39	6	4	27	18
1.92E-01	1.23	1.26	5	1	7	18
2.87E-01	1.14	1.16	9	8	63	18
4.25E-01	0.69	0.66	4	1	12	18
5.12E-01	0.58	0.55	1	1	14	18
3.15E-01	0.00	0.00	2	0	3	18
3.97E-01	0.00	0.00	7	0	4	18
6.42E-01	0.00	0.00	8	0	8	18

B.1.2 Hierarchical Clustering

Table B.11: Hierarchical clustering results for Thyroid Disease

P_{HG}	RR	OR	Cluster #	#PCC	#CLS	#CMP
4.16E-02	1.88	2.21	4	6	22	25
1.41E-01	1.47	1.61	1	3	13	25
2.12E-01	1.25	1.32	3	3	15	25
2.13E-01	1.24	1.30	5	6	31	25
1.88E-01	1.23	1.29	6	1	5	25
4.58E-01	0.80	0.77	8	2	15	25
8.04E-01	0.58	0.53	2	4	38	25
4.17E-01	0.00	0.00	7	0	3	25
8.70E-01	0.00	0.00	9	0	11	25

Table B.12: Hierarchical clustering results for Dyslipidemia

P_{HG}	RR	OR	Cluster #	#PCC	#CLS	#CMP
3.87E-02	2.37	3.28	6	2	5	27
3.48E-02	1.78	2.06	2	10	38	27
1.05E-01	1.61	1.84	9	3	11	27
9.81E-02	1.60	1.82	3	4	15	27
5.73E-01	0.74	0.70	4	3	22	27
5.16E-01	0.74	0.70	8	2	15	27
8.52E-01	0.49	0.44	5	3	31	27
9.29E-01	0.00	0.00	1	0	13	27
4.44E-01	0.00	0.00	7	0	3	27

Table B.13: Hierarchical clustering results for High Blood Pressure

P_{HG}	RR	OR	Cluster #	#PCC	#CLS	#CMP
9.43E-03	2.40	3.33	3	6	15	29
4.73E-02	2.19	2.99	6	2	5	29
3.98E-02	1.92	2.38	8	5	15	29
9.23E-02	1.79	2.18	7	1	3	29
4.34E-01	0.96	0.95	2	7	38	29
4.07E-01	0.95	0.94	4	4	22	29
6.56E-01	0.46	0.41	9	1	11	29
7.50E-01	0.38	0.33	1	1	13	29
9.66E-01	0.29	0.24	5	2	31	29

Table B.14: Hierarchical clustering results for Nerve Damage

P_{HG}	RR	OR	Cluster #	#PCC	#CLS	#CMP
9.47E-04	4.60	6.40	3	5	15	15
4.30E-02	2.30	2.63	8	3	15	15
7.57E-02	2.11	2.39	6	1	5	15
3.59E-01	0.98	0.98	5	3	31	15
2.94E-01	0.92	0.91	9	1	11	15
7.73E-01	0.47	0.44	2	2	38	15
7.53E-01	0.00	0.00	1	0	13	15
9.14E-01	0.00	0.00	4	0	22	15
2.68E-01	0.00	0.00	7	0	3	15

Table B.15: Hierarchical clustering results for Retinopathy

P_{HG}	RR	OR	Cluster #	#PCC	#CLS	#CMP
9.16E-02	1.69	2.15	6	2	5	37
7.63E-02	1.45	1.66	2	12	38	37
2.57E-01	1.15	1.21	4	6	22	37
2.80E-01	1.12	1.16	3	4	15	37
2.80E-01	1.12	1.16	8	4	15	37
5.24E-01	0.74	0.68	9	2	11	37
6.52E-01	0.62	0.55	1	2	13	37
8.25E-01	0.61	0.54	5	5	31	37
5.67E-01	0.00	0.00	7	0	3	37

Table B.16: Hierarchical clustering results for Diabetic Ketoacidosis

P_{HG}	RR	OR	Cluster #	#PCC	#CLS	#CMP
1.02E-02	3.12	7.38	7	2	3	34
7.31E-02	1.85	2.42	6	2	5	34
1.32E-01	1.44	1.63	1	4	13	34
8.68E-02	1.45	1.63	2	11	38	34
1.05E-01	1.42	1.59	5	9	31	34
4.37E-01	0.89	0.86	3	3	15	34
6.93E-01	0.57	0.51	8	2	15	34
9.79E-01	0.18	0.14	4	1	22	34
9.43E-01	0.00	0.00	9	0	11	34

Table B.17: Hierarchical clustering results for Hyperglycemia

P_{HG}	RR	OR	Cluster #	#PCC	#CLS	#CMP
1.24E-02	2.61	5.03	6	3	5	37
1.19E-01	1.44	1.66	3	5	15	37
1.45E-01	1.39	1.58	7	1	3	37
1.56E-01	1.28	1.39	2	11	38	37
1.73E-01	1.26	1.37	5	9	31	37
3.87E-01	0.95	0.94	1	3	13	37
5.14E-01	0.81	0.76	8	3	15	37
7.96E-01	0.36	0.29	9	1	11	37
9.87E-01	0.17	0.13	4	1	22	37

Table B.18: Hierarchical clustering results for Hypoglycemia X

P_{HG}	RR	OR	Cluster #	#PCC	#CLS	#CMP
2.77E-02	2.69	3.82	6	2	5	24
2.67E-02	1.97	2.30	5	8	31	24
1.25E-01	1.54	1.70	1	3	13	24
1.90E-01	1.31	1.39	8	3	15	24
4.66E-01	0.85	0.83	4	3	22	24
4.28E-01	0.84	0.81	3	2	15	24
9.01E-01	0.43	0.38	2	3	38	24
4.03E-01	0.00	0.00	7	0	3	24
8.57E-01	0.00	0.00	9	0	11	24

Table B.19: Hierarchical clustering results for Anxiety

P_{HG}	RR	OR	Cluster #	#PCC	#CLS	#CMP
1.66E-03	5.88	15.65	7	2	3	19
1.40E-02	3.48	5.14	6	2	5	19
1.41E-01	1.52	1.63	9	2	11	19
2.79E-01	1.08	1.10	3	2	15	19
5.35E-01	0.81	0.78	2	4	38	19
5.66E-01	0.74	0.71	5	3	31	19
5.38E-01	0.70	0.67	4	2	22	19
5.00E-01	0.60	0.56	1	1	13	19
5.84E-01	0.51	0.48	8	1	15	19

Table B.20: Hierarchical clustering results for Depression

P_{HG}	RR	OR	Cluster #	#PCC	#CLS	#CMP
1.39E-03	6.25	16.75	7	2	3	18
1.19E-02	3.70	5.50	6	2	5	18
2.50E-01	1.15	1.17	3	2	15	18
2.50E-01	1.15	1.17	8	2	15	18
4.78E-01	0.86	0.85	2	4	38	18
5.17E-01	0.79	0.76	5	3	31	18
3.80E-01	0.76	0.74	9	1	11	18
4.69E-01	0.63	0.60	1	1	13	18
7.72E-01	0.35	0.32	4	1	22	18

B.1.3 Affinity Propagation Clustering

Table B.21: Affinity Propagation clustering results for Thyroid Disease

P_{HG}	RR	OR	Cluster #	#PCC	#CLS	#CMP
3.51E-02	1.83	2.10	7	9	36	25
1.18E-01	1.52	1.68	5	4	17	25
1.75E-01	1.34	1.43	8	4	19	25
2.06E-01	1.27	1.33	6	4	20	25
4.50E-01	0.67	0.63	11	1	9	25
5.95E-01	0.65	0.61	10	2	18	25
8.09E-01	0.33	0.29	4	1	17	25
4.17E-01	0.00	0.00	1	0	3	25
1.63E-01	0.00	0.00	2	0	1	25
8.92E-01	0.00	0.00	3	0	12	25
1.63E-01	0.00	0.00	9	0	1	25

Table B.22: Affinity Propagation clustering results for Dyslipidemia

P_{HG}	RR	OR	Cluster #	#PCC	#CLS	#CMP
9.15E-03	2.78	4.21	11	4	9	27
1.99E-02	2.14	2.71	10	6	18	27
2.77E-01	1.14	1.17	7	7	36	27
3.57E-01	0.94	0.93	3	2	12	27
4.87E-01	0.83	0.80	6	3	20	27
6.11E-01	0.64	0.59	5	2	17	27
6.94E-01	0.56	0.51	8	2	19	27
8.45E-01	0.31	0.26	4	1	17	27
4.44E-01	0.00	0.00	1	0	3	27
1.76E-01	0.00	0.00	2	0	1	27
1.76E-01	0.00	0.00	9	0	1	27

Table B.23: Affinity Propagation clustering results for High Blood Pressure

P_{HG}	RR	OR	Cluster #	#PCC	#CLS	#CMP
0.00E+00	5.43	Inf	2	1	1	29
0.00E+00	5.43	Inf	9	1	1	29
6.65E-02	1.85	2.27	11	3	9	29
9.23E-02	1.79	2.18	1	1	3	29
9.51E-02	1.56	1.78	10	5	18	29
2.05E-01	1.24	1.31	7	8	36	29
4.07E-01	0.92	0.91	5	3	17	29
5.05E-01	0.81	0.78	8	3	19	29
5.51E-01	0.77	0.73	6	3	20	29
7.06E-01	0.42	0.37	3	1	12	29
9.78E-01	0.00	0.00	4	0	17	29

Table B.24: Affinity Propagation clustering results for Nerve Damage

P_{HG}	RR	OR	Cluster #	#PCC	#CLS	#CMP
0.00E+00	10.86	Inf	2	1	1	15
2.55E-02	3.57	4.86	1	1	3	15
3.40E-02	2.17	2.40	7	6	36	15
2.50E-01	1.15	1.17	10	2	18	15
2.16E-01	1.14	1.16	11	1	9	15
2.79E-01	1.09	1.10	8	2	19	15
5.18E-01	0.57	0.54	5	1	17	15
6.15E-01	0.48	0.45	6	1	20	15
7.24E-01	0.00	0.00	3	0	12	15
8.44E-01	0.00	0.00	4	0	17	15
9.80E-02	0.00	0.00	9	0	1	15

Table B.25: Affinity Propagation clustering results for Retinopathy

P_{HG}	RR	OR	Cluster #	#PCC	#CLS	#CMP
0.00E+00	4.22	Inf	2	1	1	37
1.44E-01	1.41	1.62	11	3	9	37
1.45E-01	1.39	1.58	1	1	3	37
1.38E-01	1.37	1.53	8	6	19	37
2.10E-01	1.20	1.28	7	10	36	37
3.22E-01	1.04	1.05	3	3	12	37
3.93E-01	0.97	0.96	5	4	17	37
4.50E-01	0.91	0.88	10	4	18	37
7.67E-01	0.59	0.51	6	3	20	37
8.33E-01	0.46	0.38	4	2	17	37
2.42E-01	0.00	0.00	9	0	1	37

Table B.26: Affinity Propagation clustering results for Diabetic Ketoacidosis

P_{HG}	RR	OR	Cluster #	#PCC	#CLS	#CMP
0.00E+00	4.61	Inf	9	1	1	34
2.63E-02	2.13	3.04	11	4	9	34
1.44E-01	1.38	1.54	4	5	17	34
1.44E-01	1.38	1.54	5	5	17	34
2.63E-01	1.15	1.20	6	5	20	34
2.62E-01	1.14	1.18	3	3	12	34
6.02E-01	0.73	0.67	10	3	18	34
6.51E-01	0.68	0.62	8	3	19	34
8.76E-01	0.56	0.49	7	5	36	34
5.32E-01	0.00	0.00	1	0	3	34
2.22E-01	0.00	0.00	2	0	1	34

Table B.27: Affinity Propagation clustering results for Hyperglycemia

P_{HG}	RR	OR	Cluster #	#PCC	#CLS	#CMP
0.00E+00	4.22	Inf	9	1	1	37
3.75E-02	1.75	2.23	10	7	18	37
7.99E-02	1.55	1.85	5	6	17	37
1.44E-01	1.41	1.62	11	3	9	37
3.44E-01	1.04	1.05	6	5	20	37
3.93E-01	0.97	0.96	4	4	17	37
6.99E-01	0.76	0.70	7	7	36	37
5.91E-01	0.67	0.61	3	2	12	37
8.89E-01	0.40	0.33	8	2	19	37
5.67E-01	0.00	0.00	1	0	3	37
2.42E-01	0.00	0.00	2	0	1	37

Table B.28: Affinity Propagation clustering results for Hypoglycemia X

P_{HG}	RR	OR	Cluster #	#PCC	#CLS	#CMP
1.02E-01	1.60	1.78	4	4	17	24
1.26E-01	1.50	1.64	10	4	18	24
1.65E-01	1.34	1.42	7	7	36	24
1.81E-01	1.33	1.41	6	4	20	24
5.20E-01	0.73	0.69	5	2	17	24
6.05E-01	0.64	0.60	8	2	19	24
5.92E-01	0.51	0.47	3	1	12	24
4.03E-01	0.00	0.00	1	0	3	24
1.57E-01	0.00	0.00	2	0	1	24
1.57E-01	0.00	0.00	9	0	1	24
7.94E-01	0.00	0.00	11	0	9	24

Table B.29: Affinity Propagation clustering results for Anxiety

P_{HG}	RR	OR	Cluster #	#PCC	#CLS	#CMP
0.00E+00	8.44	Inf	9	1	1	19
1.46E-02	3.00	4.00	11	3	9	19
4.08E-02	2.78	3.67	1	1	3	19
6.40E-02	1.88	2.12	8	4	19	19
1.73E-01	1.38	1.46	3	2	12	19
3.55E-01	0.94	0.93	4	2	17	19
4.67E-01	0.78	0.76	6	2	20	19
7.02E-01	0.61	0.57	7	3	36	19
6.58E-01	0.44	0.41	5	1	17	19
1.24E-01	0.00	0.00	2	0	1	19
9.21E-01	0.00	0.00	10	0	18	19

Table B.30: Affinity Propagation clustering results for Depression

P_{HG}	RR	OR	Cluster #	#PCC	#CLS	#CMP
0.00E+00	8.94	Inf	9	1	1	18
3.67E-02	2.94	3.91	1	1	3	18
7.34E-02	2.00	2.29	11	2	9	18
1.41E-01	1.50	1.60	10	3	18	18
3.21E-01	1.00	1.00	4	2	17	18
4.22E-01	0.93	0.92	7	4	36	18
4.28E-01	0.83	0.81	6	2	20	18
4.25E-01	0.69	0.66	3	1	12	18
6.26E-01	0.47	0.44	5	1	17	18
6.91E-01	0.41	0.38	8	1	19	18
1.18E-01	0.00	0.00	2	0	1	18

B.1.4 Network Clustering

Table B.31: Network clustering results for Thyroid Disease

P_{HG}	RR	OR	Cluster #	#PCC	#CLS	#CMP
6.32E-04	3.71	6.42	5	6	12	25
2.77E-02	2.24	2.86	3	4	12	25
2.51E-01	1.17	1.21	2	3	16	25
3.24E-01	1.07	1.08	1	5	29	25
4.58E-01	0.80	0.77	4	2	15	25
7.81E-01	0.60	0.55	6	4	37	25
7.44E-01	0.38	0.34	8	1	15	25
6.64E-01	0.00	0.00	7	0	6	25
8.70E-01	0.00	0.00	9	0	11	25

Table B.32: Network clustering results for Dyslipidemia

P_{HG}	RR	OR	Cluster #	#PCC	#CLS	#CMP
1.14E-04	3.87	7.16	4	8	15	27
1.25E-01	1.49	1.65	2	4	16	27
3.02E-01	1.03	1.04	9	2	11	27
3.57E-01	0.94	0.93	3	2	12	27
3.57E-01	0.94	0.93	5	2	12	27
6.18E-01	0.74	0.70	1	4	29	27
6.87E-01	0.71	0.67	6	5	37	27
6.95E-01	0.00	0.00	7	0	6	27
9.54E-01	0.00	0.00	8	0	15	27

Table B.33: Network clustering results for High Blood Pressure

P_{HG}	RR	OR	Cluster #	#PCC	#CLS	#CMP
3.98E-02	1.92	2.38	4	5	15	29
5.19E-02	1.88	2.32	5	4	12	29
1.31E-01	1.49	1.67	9	3	11	29
1.70E-01	1.36	1.47	3	3	12	29
3.58E-01	0.99	0.99	2	3	16	29
4.85E-01	0.89	0.87	1	5	29	29
5.88E-01	0.82	0.78	6	6	37	29
7.23E-01	0.00	0.00	7	0	6	29
9.64E-01	0.00	0.00	8	0	15	29

Table B.34: Network clustering results for Nerve Damage

P_{HG}	RR	OR	Cluster #	#PCC	#CLS	#CMP
1.34E-02	3.23	4.06	9	3	11	15
3.98E-02	2.14	2.38	1	5	29	15
9.78E-02	1.81	1.97	3	2	12	15
1.68E-01	1.42	1.48	4	2	15	15
3.33E-01	0.84	0.82	5	1	12	15
4.47E-01	0.66	0.63	8	1	15	15
4.83E-01	0.61	0.59	2	1	16	15
9.88E-01	0.00	0.00	6	0	37	15
4.67E-01	0.00	0.00	7	0	6	15

Table B.35: Network clustering results for Retinopathy

P_{HG}	RR	OR	Cluster #	#PCC	#CLS	#CMP
4.52E-03	2.50	4.30	9	6	11	37
1.53E-01	1.40	1.60	7	2	6	37
1.17E-01	1.37	1.54	1	9	29	37
2.80E-01	1.12	1.16	4	4	15	37
3.36E-01	1.04	1.05	2	4	16	37
3.22E-01	1.04	1.05	3	3	12	37
5.91E-01	0.67	0.61	5	2	12	37
7.55E-01	0.53	0.45	8	2	15	37
9.40E-01	0.49	0.41	6	5	37	37

Table B.36: Network clustering results for Diabetic Ketoacidosis

P_{HG}	RR	OR	Cluster #	#PCC	#CLS	#CMP
2.63E-02	2.03	2.76	5	5	12	34
2.45E-02	1.97	2.62	8	6	15	34
6.75E-02	1.72	2.13	9	4	11	34
1.11E-01	1.48	1.69	2	5	16	34
2.17E-01	1.23	1.31	4	4	15	34
4.00E-01	0.74	0.69	7	1	6	34
6.72E-01	0.74	0.68	1	5	29	34
9.60E-01	0.42	0.35	6	4	37	34
9.57E-01	0.00	0.00	3	0	12	34

Table B.37: Network clustering results for Hyperglycemia

P_{HG}	RR	OR	Cluster #	#PCC	#CLS	#CMP
9.39E-02	1.56	1.89	9	4	11	37
1.19E-01	1.44	1.66	4	5	15	37
1.32E-01	1.42	1.64	5	4	12	37
1.17E-01	1.37	1.54	1	9	29	37
5.14E-01	0.81	0.76	8	3	15	37
5.73E-01	0.76	0.70	2	3	16	37
7.34E-01	0.73	0.67	6	7	37	37
4.48E-01	0.68	0.62	7	1	6	37
8.37E-01	0.33	0.27	3	1	12	37

Table B.38: Network clustering results for Hypoglycemia X

P_{HG}	RR	OR	Cluster #	#PCC	#CLS	#CMP
4.23E-03	2.57	3.27	1	9	29	24
9.66E-02	1.68	1.90	5	3	12	24
1.90E-01	1.31	1.39	4	3	15	24
2.38E-01	1.17	1.21	9	2	11	24
7.45E-01	0.63	0.58	6	4	37	24
5.92E-01	0.51	0.47	3	1	12	24
7.21E-01	0.40	0.36	8	1	15	24
7.57E-01	0.37	0.33	2	1	16	24
6.47E-01	0.00	0.00	7	0	6	24

Table B.39: Network clustering results for Anxiety

P_{HG}	RR	OR	Cluster #	#PCC	#CLS	#CMP
6.03E-03	3.06	3.99	2	5	16	19
1.41E-01	1.52	1.63	9	2	11	19
1.73E-01	1.38	1.46	3	2	12	19
1.73E-01	1.38	1.46	5	2	12	19
2.79E-01	1.08	1.10	4	2	15	19
5.05E-01	0.80	0.78	1	3	29	19
7.26E-01	0.59	0.55	6	3	37	19
5.55E-01	0.00	0.00	7	0	6	19
8.77E-01	0.00	0.00	8	0	15	19

Table B.40: Network clustering results for Depression

P_{HG}	RR	OR	Cluster #	#PCC	#CLS	#CMP
1.82E-02	2.63	3.22	4	4	15	18
2.65E-02	2.58	3.17	9	3	11	18
2.44E-02	2.45	2.93	2	4	16	18
4.25E-01	0.69	0.66	3	1	12	18
4.25E-01	0.69	0.66	5	1	12	18
6.80E-01	0.63	0.59	6	3	37	18
7.07E-01	0.53	0.50	1	2	29	18
5.34E-01	0.00	0.00	7	0	6	18
8.61E-01	0.00	0.00	8	0	15	18

B.2 Clustering Evaluation p-value

This section includes obtained p-value per clusters for given complications obtained from four clustering approaches used in this report. p-values less than 0.1 are highlighted.

Table B.41: k-means Clustering p-value

	Cluster #1	*Cluster #2*	*Cluster #3*	*Cluster #4*	*Cluster #5*	*Cluster #6*	*Cluster #7*	*Cluster #8*	*Cluster #9*	*Cluster #10*
Thyroid Disease	0.705	1.000	1.000	0.316	0.084	0.680	1.000	0.766	0.294	1.000
Dyslipidemia	0.472	1.000	0.107	1.000	0.742	0.751	1.000	0.424	0.430	0.543
High Blood Pressure	0.510	1.000	0.669	1.000	1.000	0.036	0.579	0.823	0.568	0.573
Nerve Damage	0.415	1.000	0.683	0.729	1.000	0.260	0.343	0.569	0.818	1.000
Retinopathy	0.449	1.000	0.520	0.841	0.532	0.689	0.669	0.097	0.740	0.682
Diabetic Ketoacidosis	0.867	0.534	0.942	0.006	1.000	0.968	0.641	1.000	0.042	1.000
Hyperglycemia	0.984	0.567	0.796	0.041	1.000	0.936	1.000	0.902	0.053	0.245
Hypoglycemia X	0.918	1.000	1.000	0.095	0.699	0.840	0.498	1.000	0.125	0.498
Anxiety	0.547	1.000	0.139	0.814	0.204	0.437	0.415	1.000	0.881	1.000
Depression	0.844	1.000	0.377	0.794	0.591	0.389	1.000	1.000	0.480	0.396

Table B.42: Hierarchical Clustering p-value

	Cluster #1	*Cluster #2*	*Cluster #3*	*Cluster #4*	*Cluster #5*	*Cluster #6*	*Cluster #7*	*Cluster #8*	*Cluster #9*
Thyroid Disease	0.360	0.921	0.459	0.124	0.389	0.590	1.000	0.745	1.000
Dyslipidemia	1.000	0.090	0.260	0.785	0.946	0.209	1.000	0.790	0.304
High Blood Pressure	0.946	0.623	0.041	0.632	0.994	0.237	0.466	0.129	0.910
Nerve Damage	1.000	0.932	0.007	1.000	0.627	0.398	1.000	0.171	0.701
Retinopathy	0.873	0.160	0.514	0.446	0.926	0.344	1.000	0.509	0.794
Diabetic Ketoacidosis	0.323	0.167	0.691	0.998	0.223	0.309	0.120	0.889	1.000
Hyperglycemia	0.650	0.280	0.280	0.999	0.313	0.093	0.570	0.751	0.957
Hypoglycemia X	0.334	0.970	0.717	0.716	0.075	0.173	1.000	0.425	1.000
Anxiety	0.838	0.751	0.587	0.799	0.790	0.115	0.038	0.877	0.401
Depression	0.820	0.706	0.555	0.948	0.747	0.104	0.034	0.568	0.765

Table B.43: Affinity Propagation Clustering p-value

	Cluster #1	*Cluster #2*	*Cluster #3*	*Cluster #4*	*Cluster #5*	*Cluster #6*	*Cluster #7*	*Cluster #8*	*Cluster #9*	*Cluster #10*	*Cluster #11*
Thyroid Disease	1.000	1.000	1.000	0.958	0.287	0.421	0.090	0.386	1.000	0.838	0.813
Dyslipidemia	1.000	1.000	0.669	0.970	0.851	0.734	0.458	0.886	1.000	0.069	0.052
High Blood Pressure	0.473	0.195	0.926	1.000	0.671	0.783	0.358	0.742	0.189	0.233	0.234
Nerve Damage	0.266	0.102	1.000	1.000	0.843	0.890	0.099	0.594	1.000	0.554	0.623
Retinopathy	0.571	0.248	0.583	0.953	0.630	0.914	0.355	0.291	1.000	0.682	0.372
Diabetic Ketoacidosis	1.000	1.000	0.518	0.312	0.315	0.469	0.950	0.845	0.221	0.814	0.111
Hyperglycemia	1.000	1.000	0.841	0.620	0.195	0.557	0.834	0.974	0.238	0.108	0.382
Hypoglycemia X	1.000	1.000	0.880	0.254	0.782	0.383	0.323	0.836	1.000	0.309	1.000
Anxiety	0.325	1.000	0.454	0.662	0.907	0.752	0.871	0.199	0.129	1.000	0.085
Depression	0.313	1.000	0.789	0.620	0.898	0.717	0.657	0.922	0.120	0.353	0.287

Table B.44: Network Clustering p-value

	Cluster #1	*Cluster #2*	*Cluster #3*	*Cluster #4*	*Cluster #5*	*Cluster #6*	*Cluster #7*	*Cluster #8*	*Cluster #9*
Thyroid Disease	0.539	0.503	0.110	0.748	0.004	0.910	1.000	0.942	1.000
Dyslipidemia	0.806	0.304	0.657	0.001	0.664	0.839	1.000	1.000	0.616
High Blood Pressure	0.695	0.621	0.399	0.127	0.168	0.764	1.000	1.000	0.343
Nerve Damage	0.122	0.822	0.328	0.454	0.730	1.000	1.000	0.808	0.074
Retinopathy	0.224	0.577	0.593	0.516	0.828	0.980	0.446	0.920	0.023
Diabetic Ketoacidosis	0.830	0.267	1.000	0.439	0.094	0.989	0.787	0.082	0.208
Hyperglycemia	0.238	0.795	0.969	0.276	0.324	0.860	0.813	0.756	0.257
Hypoglycemia X	0.016	0.942	0.880	0.427	0.282	0.895	1.000	0.929	0.527
Anxiety	0.748	0.031	0.452	0.580	0.462	0.891	1.000	1.000	0.396
Depression	0.896	0.105	0.795	0.080	0.785	0.858	1.000	1.000	0.130