4-22-2020 10:15 AM

A Visual Analytics System for Investigating Multimorbidity Using Supervised Machine Learning

Maede Sadat Nouri, *The University of Western Ontario*

Supervisor: Sedig, Kamran, *The University of Western Ontario*
: Lizotte, Daniel, *The University of Western Ontario*

A thesis submitted in partial fulfillment of the requirements for the Master of Science degree in Computer Science

Abstract

Patterns of multimorbidity are complex and difficult to summarise using static visualization techniques like tables and charts. We present a visual analytics system with the goal of facilitating the process of making sense of data collected from patients with multimorbidity. The system reveals underlying patterns in the data visually and interactively, which enables users to easily assess both prevalence and correlation estimates of different chronic diseases among multimorbid patients with varying characteristics. To do so, the system uses count-based conditional probability, binary logistic regression, softmax regression and decision tree models to dynamically compute and visualize prevalence and correlation estimates for subsets of the data characterized by a user-selected set of pre-existing chronic conditions. The system also allows the user to examine the impact of adjusting for characteristics like age and gender on both the prevalence estimates and on correlations among diseases. By dynamically changing patient characteristics of interest and examining the resulting visualizations, the user can explore how prevalence and correlation estimates change with disease diagnosis and with other patient characteristics. This thesis is therefore a significant effort in understanding high-dimensional joint distributions of random variables and the created system can be used in any domain, such as economics, politics or social sciences, in which investigating the relationships between several random variables is vital to drawing the right conclusion.

Keywords: multimorbidity, visual analytics system, conditional probability, binary logistic regression, softmax regression, decision tree

Summary for Lay Audience

Multimorbidity, which is defined as the presence of multiple chronic diseases, is a growing health care problem especially for older adults. The traditional single-disease-centric approaches are no longer efficient to address the challenge of multimorbidity and a holistic framework is required to create effective prevention and treatment strategies. Therefore, we designed a visual analytics system for investigating multimorbidity patterns. Visual analytics is defined as the science of analytical reasoning facilitated by interactive visual interfaces. Unlike many studies in multimorbidity whose patterns are represented using simple tables and graphs, our system employs interactive visualizations. Through these visualizations, users can interact with different subsets of data and select a set of chronic diseases as well as several categories of age, gender and socioeconomic scores for investigation. To do so, the system uses statistical and machine learning algorithms including count-based conditional probability, binary logistic regression, softmax regression and decision tree to compute and visualize prevalence and correlation estimates of the diseases. Machine learning models are trained on the data to perform learning tasks by relying on patterns and inference created from the observations. Every time by every selection, the visualizations update the prevalence and correlation of diseases. The visual analytics system can be used in different areas of healthcare or other disciplines where investigating the associations between random variables with joint probability distributions is interesting.

Acknowledgements

I would like to express my deep gratitude to my supervisors, Dr. Kamran Sedig and Dr. Dan Lizotte. They provided guidance, generosity, and valuable encouragement. They regularly contacted me and have met with me whenever I needed to do so. Dr. Sedig has believed in me during the process and provided me with valuable guidance regarding visualization strategies and interdisciplinary research opportunities. Dr. Lizotte has been inspiration and a great role model, actively involved with research and knowledgeable about the relevant literature.

I would like to thank my parents, my wonderful father who taught me everything that I know about the world and my beautiful mother who raised me with pure love and attention. I would also like to thank my family, my brother, his family and my amazing aunt who are thousands of kilometers away from me; however, their energy and love are always with me.

I dedicate this work to my husband, Mehdi, my best friend, who supported and encouraged me on this and everything else. He has always praised and reminded me that I can do anything I set my mind to.

Contents

List of Tables

List of Figures

Chapter 1

1 Introduction

1.1 Motivation

Multimorbidity, which is known as the presence of two or more chronic conditions (The Lancet, 2018), has been a persistent challenge for primary health care for many years (Nicholson, 2017). Almost one-third of adults in the world live with multimorbidity (Hajat & Stein, 2018). In 2012, 38 million (68%) deaths worldwide were due to chronic diseases, and according to The World Health Organization, this number will increase to 52 million by 2030 (World Health Organization, 2015).

Patients suffering from multiple chronic medical conditions are usually high-need, high-cost patients (Navickas et al. 2016). The higher the number of coexisting conditions and medications, the more challenges exist in managing people with multimorbidity in primary care (Wallace et al. 2015). Health care systems have mostly focused on single-disease-centric frameworks rather than practical solutions for the prevention of multiple medical conditions (Farmer et al. 2016, Wallace et al. 2015). Therefore, it is necessary to enhance the prevention efforts and develop more integrated models of care for multimorbid patients. For this purpose, a good knowledge of epidemiology and risk factors is needed. Patients with specific characteristics may develop a particular disease. These characteristics can be categorized according to gender, age, household income, household education, aboriginal status, immigration status, area of residence and risk factors like high blood pressure, obesity, high stress, etc. Besides, analysis of the association between chronic diseases plays an important role in the prevention and monitoring of these diseases, as the patients with multimorbidity face more health risks than those living with one chronic illness.

The volume of data generated by primary care practitioners, health care institutions, and patient self-reports is dramatically increasing as their tendency to update their services and use of Electronic Health Record (EHR) and Electronic Medical Record (EMR) systems is growing (Murdoch & Detsky, 2013). EHR and EMR databases are valuable, systematized platforms that can help researchers access more accurate and complete information about patients. The rapid growth of health data brings new challenges for physicians and policymakers who aim to manage and analyze extremely large and complex datasets. They need to communicate data effectively and extract patterns, associations, trends and gaps to improve and ensure the health of the public. Interactive visualization can be considered as an effective solution in the process of knowledge discovery and decision making (Shneiderman et al. 2013). The high prevalence and myriad combinations of chronic conditions present a good opportunity for visual analytics systems to examine the problem of multimorbidity and its underlying mechanisms. Visual analytics can deal with large, complex data extracted from EHR and EMR databases and aid stakeholders to make faster and more reliable decisions (Raghupathi & Raghupathi, 2018).

The main contribution of this thesis is to introduce a web application that is beyond simple charts and tabular presentation of the data and provides useful insights into multimorbidity. We use statistical and machine learning algorithms to identify the prevalence of different chronic diseases as well as the correlation estimate between each pair of diseases given varying patient characteristics and a set of pre-existing chronic conditions. This thesis attempts to provide a foundation for the design and use of visual analytics systems in examining the prevalence and patterns of multimorbidity. The visualizations in our system can be implemented for other purposes in the area of healthcare or other disciplines where high-dimensional joint distributions of random variables are of significance. For example, the system can be used to understand the relationship between economic growth and categorical variables containing economic freedom, political freedom and the level of income. As another example, the system can explore the effects of job involvement, job stress, job satisfaction, and organizational commitment on job burnout, adjusting for personal characteristics of gender, race, age, educational level, position, etc.

We present our analyses through an interactive bar chart and a dynamic correlation matrix in our visual analytics system. The bar graph displays disease prevalence estimated by count-based conditional probability, logistic regression, and decision tree models, while the correlation matrix employs softmax regression and decision tree to display disease association. By performing actions

such as selecting, filtering, ordering and comparing, the user can reach into the multimorbidity data to operate upon it. The actions contribute to the completion of user's tasks leading to a series of reactions that occur within the representation and computing spaces. The user may perform a series of tasks in order to make sense of data. During the process, the user engages in cognitive activities such as knowledge discovery, learning, decision making and problem solving which are made up of the tasks. Then the user may carry out another sequence of tasks until he/she achieves his/her goal.

1.2 Organization of the Thesis

In this research we present our visual analytics system which is an interactive platform for identifying and analyzing multimorbidity patterns. In this chapter, we explained the purpose of the thesis and the importance of uncovering the association between chronic conditions. The rest of this thesis is divided into four chapters:

Chapter 2 presents the general background concepts of multimorbidity, primary care EMRs, count-based conditional probability, supervised machine learning models and visual analytics. This chapter also briefly discusses the use of interactive visualizations in healthcare and the possible challenges for designing visual analytics systems using large and complex data. Chapter 3 provides a review of current research that is closely related to this thesis. We compare the contribution of our work with existing research and present some improvements. Chapter 4 outlines the research methodology, the description of the data, the prepossessing steps, the components of our visual analytics system and the ways in which the user can interact with different parts of it. This chapter also presents the results of the research obtained from performing several interactive tasks. Finally, we draw conclusions from the findings, describe the limitations and suggest corresponding possible future work in the fifth chapter.

Chapter 2

2 Background

2.1 Multimorbidity

Multimorbidity is defined as the co-existence of two or more chronic medical conditions within a single patient (Gallacher et al. 2019). There is an increase in the number of primary care patients with multimorbidity due to some factors like aging population or medical care improvement (Fortin et al. 2004). A chronic condition is a progressive, irreversible disease that lasts for a long time. Aetiology, duration, onset, recurrence/pattern, prognosis, sequelae, diagnosis, severity and prevalence are the factors by which the chronic conditions can be described (O'Halloran et al. 2004).

In the fiscal year 2011/12, the prevalence of multimorbidity was reported to be 26.5% among Canadian adult population at the age of 40 and over (Feely et al. 2017). Patients with multimorbidity usually come from the households with lower incomes and lower education levels (Roberts et al. 2015). In addition, the co-occurrence of multiple chronic diseases increases when patients get older (Feely et al. 2017, Roberts et al. 2015). Gender is another factor that influences the patterns of multimorbidity. According to a study conducted by Abad-Diez et al (2014), women present a higher prevalence in the different examined patterns of multimorbidity, and the reason may be in relation to their higher life expectancy and/or their worse health. Alimohammadian et al (2017) obtained similar results as the percentages of female patients and male patients with multiple disorders were 25% and 13.4%, respectively. They also highlighted that women suffer from multimorbidity at a younger age than men.

People with multi-morbid chronic diseases encounter significant challenges related to preventive care and self-management, since multimorbidity delays detection of early signs

of deterioration in patients and makes the management of taking prescription medications more difficult (Jowsey et al., 2009).

Traditional disease-focused guidelines are often not appropriate for patients with multimorbidity, based on the complex multiple conditions in which patients with multimorbidity are involved (Muth et al., 2019). These complex conditions affect the development of clinical decision-making skills as well (Muth et al., 2019).

Regarding the measurement of multimorbidity, there is a debate about whether routinely-collected data sources are more valuable and effective than self-reported datasets. It should be considered that both have their pros and cons; for instance, patients may be unaware that they have a condition and in turn they do not report it. On the other side, clinicians and health care providers who collect datasets may not collect accurate information about some conditions like depression or may not grade the severity of chronic pain precisely (Gallacher et al. 2019).

2.2 Primary Care

Primary care, known as a part within primary health care, provides accessible health care services. It plays an important part in improving health care and preventing and diagnosing diseases, disorders and injuries (Health Canada, 2012). Primary health care as a broader concept includes both services delivered to individuals (primary care services) and population level (public health care), such as income, housing, education, and environment (Health Canada 2012, Muldoon et al. 2006).

The integrated health care services and person-focused (not disease-oriented) care over time are provided by primary care clinicians who are responsible for tackling personal health care needs and developing sustained partnership with patients (Starfield, 1998).

2.3 Primary Care EMRs

An EMR is a system used to record and store patient's medical information electronically and leads to health improvement and less medical errors (Stewart et al. 2009). EMRs facilitate the process of diagnosis and treatment. Compared to EHR which is an inclusive version of EMR available throughout diverse health care settings, EMR is not designed to be shared outside an organization. Using EMR databases, health care providers can improve the quality of health

care delivery, reduce costs and fulfil timely preventative screening and examinations. EMRs bring a lot of advantages, especially in primary care research and surveillance with the goal of control and prevention of chronic diseases (Coleman et al., 2015).

The aggregation of EMR data sources from several primary care practices throughout Canada provides stakeholders and patients with a rich source of data at a regional, provincial, and national level (Birtwhistle & Williamson, 2015). To do this aggregation, CPCSSN (The Canadian Primary Care Sentinel Surveillance Network) was established as the first and largest primary care EMR-based database in the country. CPCSSN supports the management of eight chronic diseases including hypertension, diabetes, osteoarthritis, depression, chronic obstructive pulmonary disease, dementia, epilepsy and Parkinson's disease. The network is being expanded to investigate pelvic floor disorders in women, childhood asthma, speech disorders in the elderly, chronic kidney disease, chronic pain and heart failure (Garies et al. 2017).

2.4 Conditional Probability

Conditional probability is defined as the probability of an event occurring, given that one or more events have already occurred. If the event *A* is the interest event in the presence of event *B* in a sample space *S*, the formula for the conditional probability of *A* is defined as follows:

$$P(A|B) = \frac{P(A \cap B)}{P(B)} \quad \text{when } P(B) > 0 \tag{2.1}$$

$P(A, B)$ is the probability of the intersection of *A* and *B* and it means they both occur. Marginal probability is another term quantifying the likelihood that an event takes place regardless of other preceding events. As an example, a die is rolled and given the number is an even number, the probability that the die shows two is $\frac{1}{3}$, while the marginal probability of rolling a two is $\frac{1}{6}$.

Equation 2.1 can be generalized for the cases with more than one condition. For example, the probability of *C* under the multiple conditions *A* and *B* is an expanded form of the equation 2.1:

$$P(C|A \cap B) = \frac{P(A \cap B \cap C)}{P(B|A)P(A)} \tag{2.2}$$

Consider a binary random variable *X* that takes on values in $\{0,1\}$. Its Probability Mass Function (PMF) $P(x)$ gives the probability that *X* takes on the value *x*. The PMF satisfies constraints that $p(x) \geq 0$ for all *x*, and it satisfies $P(0) + P(1) = 1$.

A binary vector-valued random variable $[X_1, X_2, \ldots, X_p]$ takes on values in $\{0,1\}^P$. Its probability mass function $P([X_1, X_2, \ldots, X_p])$ gives the probability that the random vector takes on a particular (completely specified) sequence of values, i.e. it gives the probability that $X_1 = x_1$ and $X_2 = x_2$, ..., and $X_p = x_p$. The domain of the PMF is all binary strings of length *p*, it is nonnegative for every string of length *p*, and its sum over all possible binary strings of length *p* is 1.0. The *marginal* probability of one or more of the random variables in the random vector is computed by summing over all possible configurations of the unspecified variables. For example, for $p = 2$, the marginal probability that $X_1 = x_1$ is given by $P([X_1 = x_1, X_2 = 0]) + P([X_1 = x_1, X_2 = 1])$. We often write this simply by omitting the unspecified variables, so the marginal probability that $X_1 = x_1$ is written $P(X_1 = x_1)$.

A binary vector-valued random variable can be used to represent the chronic disease status of a patient. In this case, each X_i is an indicator variable that is set to 1 if the patient has been diagnosed with disease *i*, and 0 otherwise. There are 20 chronic diseases that are commonly considered in the study of multimorbidity (Nicholson et al., 2015). Given a population of patients, we can talk about the probability distribution of this binary vector-valued random variable, which reflects the distribution of co-occurring chronic diseases in the population.

Understanding this distribution is important from a health research perspective. Most basically, it is important to understand the relative prevalence for each disease, which is given by the marginal probabilities $P(X_i = 1)$. It is also important to understand when there are associations between diseases, meaning that the amount that they co-occur is either more or less likely than would be expected if they occurred independently. Statistically, this can be tested using known methods (e.g. Chi-squared test, Fisher's exact test.) These tests essentially compare estimates of the value of $P(X_i = 1, X_j = 1)$ to estimates of the value of $P(X_i = 1)P(X_j = 1)$. If in fact there is no association between diseases *i* and *j*, these quantities will be equal. (When the probabilities are estimated from data, they will be near-equal.) The more different the estimates are, the more

evidence there is that there is dependence between the diseases -- for example, they may have a common cause.

An equivalent way of assessing dependence is to compare $P(X_i = 1)$ to $P(X_i = 1 \mid X_j = 1)$, where the second quantity is the probability that $X_i = 1$ among patients for whom $X_j = 1$. This is known as a "conditional probability." For example, if the conditional probability is larger than the marginal probability, then there is a positive association between diseases *i* and *j*.

2.5 Supervised Machine Learning

It is useful to distinguish between three principal types of data: unstructured, semi-structured and structured data. Unstructured data is usually difficult to sort, store, manage and analyze through traditional databases and programs. Indeed, a value assignment (manually or automatically) to every data unit is needed for this type of data before analysis (Balducci & Marinova, 2018). Some examples of unstructured data are video and audio files, texts, social media activity and NoSQL databases. Semi-structured data represents a lower level of organization and predictability than structured data. However, semi-structured data types encompass semantic tags and markings which make them easier to group and analyze. XML and JSON data formats are two examples of semi-structured data. Structured data, by contrast, is searchable and organizable in tabular formats. Common examples of structured data contain characters, numbers, and strings whose patterns make them easily understandable. In this thesis we use structured data in order to identify the underlying patterns of multimorbidity in patients with one or more chronic conditions.

Machine learning models can perform a number of sophisticated learning tasks by relying on patterns and inference created from the observations (training set) (Japkowicz & Shah 2011, Mohri et al. 2018). Indeed, these techniques are trained on data from which they learn instead of being explicitly programmed. Due to this potential capability, the field of machine learning is growing rapidly in computer science (Alpaydin, 2014). Machine learning basically tackles a wide range of problems involving unstructured and structured data. Within this field, two main types of algorithms are applied: unsupervised learning algorithms, and supervised learning algorithms. Unsupervised learning algorithms aim to discover underlying structure and distribution in an unlabelled dataset. Most of unsupervised learning algorithms group inputs into clusters hidden in the data so that every input can belong to only one cluster. On the contrary, supervised machine

learning algorithms are utilized to learn a mapping function from input variables known as independent variables, covariates, predictors and features (X) to an output variable, also known by a variety of other names including dependent variable, response variable, and target (Y):

$$Y = f(X) \tag{2.3}$$

Supervised learning models recognize the patterns in a pre-existing labeled training data and predicts outputs for the corresponding input vectors. These models are also employed for class probability estimation. In other words, they predict the probability of an observation belonging to each known class.

The learning process continues until the mapping function is optimized and the machine learning model achieves the best possible level of performance (Schrider & Kern, 2018).

When training models for prediction and validating the predictive ability of those models, the data is sometimes split into three datasets: train, validation and test datasets. The model is initially trained on the training set and all parameters and weights are fit using this sample. The validation dataset evaluates the given models and is used to choose the best between them based on their performances and fine-tune the hyperparameters. Finally, the test set is used to assess the performance of the final model that is completely trained.

There are various, effective supervised machine learning algorithms including logistic regression, support vector machine, decision trees, k-nearest neighbor algorithm, neural networks and naïve bayes which are widely applied in classification problems. These algorithms are different in terms of the level of explainability/interpretability. For instance, linear regression, logistic regression and decision tree models are usually interpretable, while support vector machine, ensemble methods like random forests, and neural networks are mostly considered as less explainable models (Molnar 2019, Adadi & Berrada 2018). There is a strong demand of understanding the reasoning behind machine learning algorithms. Researchers attempt to satisfy this demand by creating a suite of machine learning techniques that generate more transparent and explainable models that continue to generate accurate predictions. Stakeholders in the domain of healthcare require researchers to justify and verify machine learning models and their results as this domain is faced with life and death decisions (Adadi & Berrada, 2018).

Another usage of machine learning algorithm is estimating the class probability of each observation. A value $P_k \in [0,1]$ is assigned to each class k, which indicates an estimated probability that the observation belongs to that class, based on the selected data samples.

In this thesis, three machine learning algorithms of decision tree, logistic regression and a generalization of the logistic regression model namely softmax regression will be used to provide three different ways to estimate conditional probabilities.

2.5.1 Decision Trees

Decision tree is a supervised learning algorithm widely used in decision analysis and has a flowchart-like or tree-like structure. This model comprises a number of nodes, leaves and branches; every node tests a feature, every leaf represents the class or label of the feature and every branch represents the connection of the feature coming to the outcome label (Shaikhina et al. 2017). Decision trees are constructed based on recursive partitioning. Recursive partitioning is a simple statistical method that creates decision trees with the goal of correctly classifying observations. To do so, it splits the observations into subsets given dichotomous features. Each of these subsets of observations may be split several times until the top-down construction stops after reaching a stopping criterion.

There are two criteria used to decide which feature to split on at each step in constructing a decision tree: entropy and information gain. Entropy indicates the degree of uncertainty or disorganization in a dataset. For example, in a sample with two classes, the entropy increases when the number of positive instances and the number of negative instances tend to be equal. The entropy decreases if the sample is homogeneous. The entropy is measured by the following equation:

$$Entropy(S) = \sum_{i=1}^{c} -P_i \log_2 P_i \qquad (2.4)$$

Where P_i is the fraction of instances belonging to class i^{th} and c is the number of possible classes in a sample *S*. Information gain is the reduction of entropy and determines the amount of information provided by a feature about the target. The features that maximize the information gain or minimize the entropy are tested first. The mathematical formula for information gain is as follows:

$$IG(Y|X) = Entropy(Y) - Entropy(Y|X) \quad (2.5)$$

Where $Entropy(Y|X) = \sum_x P(X = x) \times Entropy(Y|X = x)$. In other words, the information gain from *X* on *Y* is the reduction in entropy of target *Y* when the feature *X* is known and takes the value *X*=*x*. As mentioned in Section 2.5, decision tree is one of the three algorithms used in this thesis to predict the probability distribution of output classes. The decision tree model calculates the probability of a class k by returning the number of observations that belong to class k on a given leaf *l* over the total number of observations captured by that leaf $P(Y = k|X) = \frac{n_k^{(l)}}{n^{(l)}}$.

2.5.2 Binary Logistic Regression Models

Logistic regression is a classification algorithm in statistics and machine learning which is commonly employed to examine the relationship between a binary or dichotomous dependent variable and a set of independent variables (either continuous or categorical) (Manogaran & Lopez, 2018).

Logistic regression models can be applicable to any types of sampling: cross-sectional, prospective and retrospective; consequently, it is widely used in various disciplines including health, education, banking industry, and politics. (Wilson & Lorenz, 2015).

Logistic regression estimates the probability of an input belonging to the positive class. Let X denotes a *n* by *M*+1 matrix where *n* is the number of observations, *M* is the number of independent variables, and x_{ij} is the j^{th} independent variable in the i^{th} row of the matrix *X* allocated to i^{th} observation. In addition, let $\beta = (\beta_0, \beta_1, \ldots \beta_M)$ denotes a vector of *M*+1 coefficients. Using a sigmoid function, the probability that the i^{th} observation belongs to class 1 (positive class) can be computed by the following formula:

$$P(Y_i = 1|X_i, \beta) = \frac{1}{1 + e^{-(\beta_0 x_{i0} + \beta_1 x_{i1} + \cdots + \beta_M x_{iM})}} \quad (2.6)$$

Subsequently, the probability of *Y* becoming 0 is $P(Y = 0|X, \beta) = 1 - P(Y = 1|X, \beta)$.

2.5.3 Softmax Regression Models

Softmax regression, also known as multi class LR, multinomial logistic regression, and Maximum Entropy Classifier, is a supervised classification technique which predicts the probability of each particular value of the multi-class dependent variable through softmax function (Jiang et al. 2018). The softmax function is an extension of the sigmoid function to problems with more than two levels that takes a vector of K values and normalizes it into K outcomes which form a probability distribution and sum up to one. If we assume a two-class classification, the obtained probabilities from softmax regression equal to the probabilities estimated by the sigmoid function through the two-class logistic regression. There is no intrinsic ordering to the classes of the dependent variable. In addition, the independent variables can be either categorical (nominal, dichotomous, or ordinal) or continuous.

Given a *K* dimensional dependent variable *Y* and a vector of M covariates X_i collected from i^{th} observation, we estimate the probability of the output belonging to each *K* possible classes ($k = 1, \dots, K$) using a linear predictor function as follows:

$$f(k,i) = \beta_{0,k} + \beta_{1,k}x_{1,i} + \beta_{2,k}x_{2,i} + \cdots + \beta_{M,k}x_{M,i} \quad \text{for } k = 1, \dots, K \tag{2.7}$$

where $\beta_{m,k}$ is the coefficient of m^{th} independent variable and the k^{th} class of the outcome. If we group the coefficients and independent variables into vectors of size *M*+1, we can write the linear predictor function compactly:

$$f(k,i) = \beta_k . X_i \quad \text{for k=1,...,K} \tag{2.8}$$

where β_k denote the vector of *M*+1 coefficients associated with outcome *k*, and X_i denote the vector of independent variables corresponding to observation *i*. To clarify, based on the assumption that all *K* probabilities must sum to one, we can estimate the probability of each class from the following equations:

$$P(y_i = k) = \frac{e^{\beta_k . X_i}}{1 + \sum_{k=1}^{K-1} e^{\beta_k . X_i}} \qquad \text{for } k = 1, \dots, K-1 \tag{2.9}$$

$$P(y_i = k) = \frac{1}{1 + \sum_{k=1}^{K-1} e^{\beta_k . X_i}} \qquad \text{for } k = K \tag{2.10}$$

The softmax regression is broadly utilized in image analysis, text classification and generally deep neural networks (Yang et al. 2018, Jiang et al. 2018), since this model calculates the probability of each input belonging to a class and the obtained probabilities add up to 1.

2.6 Visual Analytics

As Sedig and Parsons (2016) mention in their book, until the second half of the 20th century, visualization was not considered as an effective method in the process of data analysis. This approach changed when some early researchers like Bertin, Tukey, Tufte and Cleveland indicated the importance of visualizations (Sedig & Parsons, 2016). Using visualizations, scientists can represent the data or information embedded in an object in many ways. Color and shape can show the difference between attributes and size can encode length, width, height or weight of components (Spence, 2014). A powerful visualization tool should be interactive and human-centered, should handle several tasks and have proper canvasses and number of variables. Innovative visualization techniques are beyond traditional graphs like bar charts, pie charts, or line graphs. They use complex data such as EHRs to discover information, patterns and variables without specific hypotheses (West et al. 2014).

Although existing computational systems bring a lot of benefits and tackle certain concerns, they cannot support various cognitive activities such as analytical reasoning, decision making, interpreting and problem solving. Visual analytics is a new approach of computational analysis that contains both data analytics and interactive visualizations and help users control their interactions with information (Ola & Sedig, 2014). Visual analytics systems play a mediator role between humans and information.

Thomas and Cook (2005) define Visual analytics as the science of analytical reasoning facilitated by interactive visual interfaces. Visual analytics systems allow users to gain reliable information from large and often complex data and discover patterns and outliers (Andrienko et al 2018) in order to improve the process of understanding, reasoning and decision making.

2.6.1 Components of Visual Analytics Systems

Analytics engine and interactive visualization engine are two main components of visual analytics systems with which stakeholders interact to perform various cognitive activities. Analytics engine encompasses two spaces of information and computation. In information space, data pre-processing and data transformation stages occur. During the stage of pre-processing, data is cleansed, integrated from diverse sources, fused and normalized. Following this, the data is transformed into a format or structure that is required for the process of analysis. In computing space, statistical and machine learning techniques are employed to recognize patterns in various types of data including integers, text, audio, images and video (Sedig & Parsons 2013, Sedig et al. 2017, Sedig & Parsons 2016).

Interactive visualization engine is the other component of visual analytics systems. It gets the results from analytics engine and creates representations to depict information in a mostly non-textual way (Ola & Sedig, 2014). These interactive visual representations allow users to interact with and reason about data. In other words, by visual analytics systems, users can examine several forms of display, change the subset of information and select and order analysis techniques. Representation space bridges the gap between stakeholders and processed information items by encoding them through interactive visualizations. Designers use navigational components, input controls, informational components and containers through representation space to let users easily navigate, select items, expand sections of content and filter options. Visual marks are atomic visual entities by which data items are encoded. Points, lines, shapes, colours, letters, digits and symbols are some examples of visual marks. These encoding units are classified based on the number of dimensions they display on the plane, for example, points (zero dimensions), lines (one dimension), surfaces or areas (two dimensions), and volumes (three dimensions). Visual structures are combinations of visual marks; hence they can communicate more dimensions of information. Visual structures can be either concrete or abstract. First, the designer chooses an abstract structure for representation and then implements it in the representation space with physical details and in a

concrete form. Visual marks and visual structures have certain properties called visual variables through which individual information items are encoded and represented in a meaningful manner (Sedig & Parsons, 2016). Visual variables are commonly listed as color intensity, color hue, size, motion, texture, orientation, shape, enclosure and curvature.

In interaction space, the user acts upon external representations exhibited at the interface of visual analytics system. This coupling between the user and the interface is accomplished through a series of epistemic actions, such as navigating, searching, filtering, comparing, measuring, selecting, linking, accelerating, etc (Sedig & Parsons, 2013). Following this, a subsequent reaction occurs within different components of the visual analytics system, particularly, computing, representation, interaction and mental spaces. The mental space is considered as the location through which a wide range of internal, complex cognitive activities emerges. Learning, planning, knowledge discovery, sense making, problem solving, analytical reasoning and decision making are some types of these cognitive activities. The process of action and reaction results in exploration and better discourse with the information in information space (Sedig et al. 2017). The whole process is shown in Figure 2-1 and is repeated until the user fulfills an overall activity and obtains satisfaction.

Visual analytics systems are used in various domains including ontology engineering, software evolution, security analysis (García-Peñalvo, 2015), healthcare (Simpao et al. 2015, Caban & Gotz, 2015) social media (Chen et al. 2017), economics (Evans & Basole, 2016) and management (Flood et al. 2016).

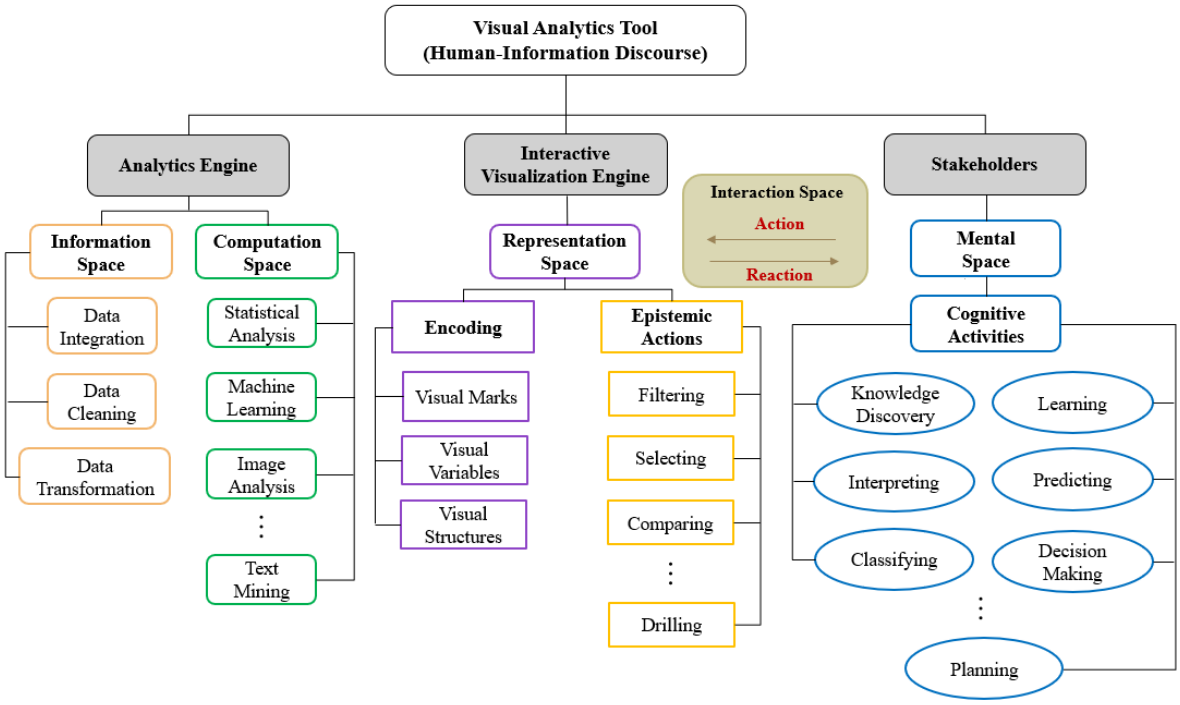

Figure 2-1: Human-information discourse through visual analytics systems

However, it should be considered that there is uncertainty inherent in real-world data which can originate from obsolete sources, missing values and noises, measurement limitation, inaccurate data entry and mixed data types. The data uncertainty which propagates in the visual analytics system through using machine learning, statistical analysis and visualization techniques, may lead to impaired reasoning and problem-solving abilities. In addition, the more complex the data is, the more complicated the visual analytics systems are designed, and in turn the more difficult analytical purposes are to achieve (Ceneda et al. 2017). Visualizing a large amount of data and summarizing the relationships between variables within the data in a single screenshot so that it can meet user cognitive needs and offer flexibility is a big challenge (West et al. 2014). Hence, Ceneta et al. (2017) introduce a general model that guides users to structure a goal and then find out a sequence of tasks to solve it. Their guidance also provides additional information about visualization techniques and algorithms with the aim of facilitating the process of insight generation and pattern exploration through visual analytics systems.

Chapter 3

3 Related Works

K. C. Roberts et al (2015) believe the traditional single-disease-focused approaches are no longer efficient to tackle the challenge of multimorbidity and a holistic approach is needed. They define multimorbidity as 2 or more or 3 or more of nine chronic diseases: asthma, arthritis, chronic obstructive pulmonary disease, diabetes, heart disease, mental disorder (mood disorder and/or anxiety), Alzheimer's disease and related dementias, cancer, and stroke. Their research is a valuable reference regarding the examination of prevalence and patterns of chronic diseases among Canadian adults. They present descriptive statistics of distribution and prevalence of multimorbidity given different groups of patient characteristics including gender, age, household income, aboriginal status, household education level, area of residence, immigration status and risk factors. They also provide analysis of the association between patient characteristics, behavioural risk factors and the chronic diseases. However, all of these analyses are presented in simple tables rather than using interactive visualizations.

Lian Leng Low et al. (2019) also emphasize on the importance of multimorbidity and its associations with epidemiologic characteristics and sociodemographic factors such as Socioeconomic Status (SES), age, race, and gender. They examine the patterns of multimorbidity among 1181024 Asian patients. The distribution of health care use and costs related to top 10 most common chronic diseases in 2016, and the physical and mental health diseases associated with sociodemographic factors are also investigated in their research. They do not rely only on tables to represent their results and employ static line charts a long with the tables to demonstrate the prevalence of multimorbidity varied by sociodemographic factors.

There is little research focusing on elaborate and interactive visualizations for enhancing the detection of multimorbidity patterns. Investigations in this area are mostly represented through static charts and tables without enabling users to filter, select, control and customize data points.

Nick Strayer et al (2019) used data derived from a combination of mega biobanks and EHRs to demonstrate the association between clinical multimorbidity patterns and a genetic variant. They designed a web application using Shiny package in R. The application lets users enhance their discourse with the data through its different views. The user selects a set of phenotypes and consequently the individuals who have one or more of the selected phenotypes are shown. Name, description, and statistical results of the phenotypes are displayed by hovering over them in the visualization. An information panel provides information about the application as well as the selected SNP (Single Nucleotide Polymorphism), the minor allele frequency, chromosome, and gene. A Manhattan plot displays the PheWAS (Phenome-Wide Association Study) analysis including phenotype diagnosis and statistical significance. Besides, an interactive upset plot demonstrates multimorbidity patterns accompanied by summary statistics, and a bipartite network plot represents the links between individuals and phenotypes.

Ingmar Schäfer et al (2014) have a different approach towards the analysis of chronic diseases and their associations. They use network analysis to explore the linking between disease combinations and clusters as well as the diseases which are responsible for overlapping these clusters. The study is conducted on insurance claims data set of the Gmünder ErsatzKasse that includes 43,632 women and 54,987 men aged 65 years and older in 2006. The chronic disease clusters are determined based on a previous study. The disease network created by Ingmar Schäfer et al is based on the combinations of three diseases (triads) with a prevalence greater than or equal to 1 percent. Thus, the associations between only two chronic diseases are not considered in the study. The static disease network represents the associations between diseases using lines with the same thickness for all the connections. Therefore, it is hard to discover the degree of associations between diseases using the network. However, the article offers some measures for the connectedness of a disease and its potential influence on the distribution of the other diseases in the format of a table.

W. Raghupathi and V. Raghupathi (2018) analyze the prevalence of chronic diseases and the relationship among them in the United States. They also investigate the associations between these diseases and behavioral habits, mental health, demographics, and overarching conditions. The study is conducted on a dataset from the Centers for Disease Control and Prevention which is the leading national public health institute of the United States. They use visualization techniques to apply descriptive analytics on the data and represents the patterns of multimorbidity. Both the title

and the content of the article give the reader an overview that a visual analytics system is created to explore multimorbidity and its patterns; while the visualizations offered in this article are single, static graphs that are potentially limited to let users interact with data, manipulate subsets of data and examine the distribution of multimorbidity.

3.1 Contributions

In this thesis, statistical and machine learning techniques are combined with interactive visualizations to help users make more sense of our multimorbidity data. Based on our knowledge, the visual analytics system offered in this thesis is the first attempt to interactively visualize the prevalence of and the associations between chronic diseases conditioned on a user-selected set of diseases and sociodemographic factors. The system allows users to interact with data and gain an extensive insight into the multimorbidity dataset. The user can perform a series of actions such as selecting, filtering, comparing and arranging on visual items to customize the data and increase cognitive load. In other words, this coupling between the user and the application facilitates analytical reasoning, knowledge discovery, problem solving and evidence-based decision making.

Chapter 4

4 Research Methods

Analyzing and evaluating data in the area of healthcare offer opportunities for developing visual analytics solutions. Besides, the increase in multimorbidity and the complexity of its underlying patterns and associations motivate researchers for further investigations. These motivations led us to design a visual analytics system in order to facilitate making sense of multimorbidity data.

In this chapter, we describe the data and its available variables used in this thesis as well as data pre-processing steps. Following this, we discuss our visual analytics system created to explore, analyze, compare and measure multimorbidity patterns. We utilize several statistical and machine learning models in our system to investigate the relationships between chronic diseases themselves and the associations between these diseases with patient attributes. The main goal of using different models in this thesis is to adjust for covariates including age, sex and SES when estimating disease diagnosis effects, rather than making predictions. This chapter ends with some screenshots of our visualization and the result section.

To design our visual analytics system, we have taken the following steps:

1. Those variables with more than two categories in the data are converted into dummy variables for further analysis.
2. The categories with small number of observations are merged together.
3. Five drop down lists are created to allow users select different categories of sociodemographic characteristics available in the data as well as the types of the models applied to the visualizations in the system.
4. A bar graph is designed interactively to represent the prevalence of chronic diseases. The three techniques, count-based conditional probability, logistic regression and decision tree

are being used by the bar graph. Based on the user selections, the model of interest would be chosen to represent the results on the bar chart.

5. A dynamic correlation matrix is created to show the pairwise correlations between chronic diseases. Two machine learning models, decision tree and softmax regression are employed to estimate these correlation values. The users can determine which of these two models represent the computed correlations by selecting the type of the model in the visual analytics system.
6. One more drop down menu is provided in the system to let users order the cells in the correlation matrix by disease name and correlation value.
7. When the data is filtered by the users, the sample size of the filtered data is shown in the system

4.1 Source of Data

Our dataset is drawn from DELPHI (Deliver Primary Healthcare Information) collection. DELPHI project is one of the eleven regional networks included in CPCSSN (Stewart, 2016). The database is the first Canadian primary care database derived from EMR data which coded symptoms and diagnoses for a subset of patient encounters using the International Classification of Primary Care. In 2005, the Centre for Studies in Family Medicine at Western University in London, Ontario started DELPHI project based on ten primary health care practices conducted across Southwestern Ontario. DELPHI collection provides surveillance data about chronic diseases to improve primary care research. It is now developed to contain information on 20 chronic diseases and 64,377 patients. This database includes more than 1.9 million patient-provider encounters and 60 family physicians from 18 practice sites (Centre for Studies in Family Medicine, 2020)

The data includes 20 chronic disease categories and a total of 13697 patients who have at least one chronic disease. Each patient is characterized by three features of age, gender, and socioeconomic score. Among 7565 females and 6132 males in the dataset, 6303 patients have only one disease, 3183 patients have developed two chronic diseases and 4211 patients face more than two chronic conditions. The individuals in DELPHI database are split into six age groups: 0 to 9 years (child), 10 to 19 years (adolescent), 20 to 29 years (young adult), 30 to 39 years (adult), 40 to 60 years (middle age), and over 60 (elder). SES is the other sociodemographic factor considered

in this thesis. It is a criterion for representing the level of education, income, occupation and wealth in a given population. SES is strongly involved in the prevalence of multimorbidity as lower SES results in an increase in multimorbidity prevalence (Salisbury et al. 2011, Violán et al. 2014) This factor is categorized into five equal-sized quintiles. First quintile represents to the lowest-income people and fifth quintile refers to the highest income group of individuals in the society. In our database there is no patient belonging to the first two quintiles and all individuals have been distributed among moderate, high and the highest income levels.

The distribution of sociodemographic factors among 13697 patients is represented in Table 4-1. 'Child' and 'Adolescent' age groups and 'Third (Moderate) Income Quintile' have the smallest population of patients among all age groups and socioeconomic categories according to our dataset.

Table 4-1: The distribution of patient characteristics (data is given as number of each category)

Female				Male				Total
Age Group	SES			Age Group	SES			
	Third quintile	Fourth quintile	Fifth quintile		Third quintile	Fourth quintile	Fifth quintile	
Child	0	58	20	Child	0	83	17	178
Adolescent	2	154	33	Adolescent	0	155	47	391
Young Adult	7	312	142	Young Adult	6	210	58	735
Adult	8	493	157	Adult	6	340	86	1090
Middle Age	20	2077	647	Middle Age	29	1678	491	4942
Elder	9	2423	1003	Elder	18	2106	802	6361
Total	46	5517	2002	Total	59	4572	1501	13697

Table 4-2 depicts the list of twenty chronic diseases ordered by patient counts according to the dataset, which was derived from the DELPHI database using the same methodology as Nicholson (2017). In the dataset, each row of the data table indicates an observation (a patient) and 20 columns of the data table are allocated to 20 chronic diseases. The data table has been created such that the patient i has the specific disease j when the corresponding cell C_{ij} in the data table has a

value 1 otherwise it is equal to 0. Therefore, by counting the number of cells with value 1 in the column corresponding to a disease we can conclude how common a disease is based on our dataset. As shown in table 4-2, 'Hypertension', 'Hyperlipidemia' and 'Bronchitis' are the most common diseases and 'Kidney Disease', 'Dementia', and 'Liver Disease' are the least common diseases among all patients in our database.

Table 4-2: The distribution of chronic disease among 13697 patients

	Chronic Disease	Patient Counts
1	Hypertension	4345
2	Hyperlipidemia	3442
3	Bronchitis	2617
4	Cardiovascular Disease	2332
5	Musculoskeletal Problem	2163
6	Diabetes	2161
7	Depression	1747
8	Arthritis	1718
9	Cancer	1589
10	Thyroid Disease	1510
11	Obesity	1266
12	Colon Problem	1216
13	Osteoporosis	926
14	Urinary Problem	861
15	Stomach Problem	804
16	Heart Failure	306
17	Stroke	231
18	Kidney Disease	212
19	Dementia	210
20	Liver Disease	45

We have chosen the ten most common chronic diseases based on our dataset to use for further analysis ('Hypertension', 'Hyperlipidemia', 'Bronchitis', 'Cardiovascular Disease', 'Musculoskeletal Problem', 'Diabetes', 'Depression', 'Arthritis', 'Cancer' and 'Thyroid Disease'). The main reason is that the dataset is not large enough to allow a good estimation of disease prevalence and correlations when multiple selections are made by the user. The number of data points becomes smaller with every selection until some variables possibly emerge with no observation or too small a number of observations.

4.2 Preprocessing

4.2.1 Creating Dummy Variables

Dummy variables are used in statistical analysis, particularly in regression models and they can take only two quantitative values, 1 or 0. 1 indicates the presence of the independent variable that means its coefficient has an effect on the dependent variable, while 0 represents the absence of the dependent variable leading to no impact on the prediction. A categorical variable with n categories is converted into *n* dummy variables when i^{th} dummy variable is equal to 1 if the observation belongs to i^{th} category, otherwise it is equal to 0. If the model has an intercept, one of the dummy variables should be dropped from the model (Garavaglia et al. 1998) as the n^{th} category can be represented when all other dummy variables get the value 0. Including the dummy variable corresponding to the last subgroup in the model adds redundant information that results in multicollinearity.

All chronic diseases as well as gender are already binary variables taking values either 0 or 1 in the dataset. Age is, however, a categorical variable with more than two categories as mentioned in Section 4.1 and needs to be converted into dummy variables so that it can be introduced into the regression equation.

4.2.2 Merging Categories with Few Observations

As mentioned in Section 4.1, the prevalence of chronic conditions decreases in children and adolescents as well as individuals belonging to the moderate-income level. These three categories have the lowest number of patients and it becomes lower when the user filters data and selects chronic diseases and other patient attributes in order to observe their associations. Indeed, we use different models to predict the prevalence and correlation of diseases. Since some dummy variables in the dataset have too few observations, the classification models in our visual analytics system are unable to fit models properly and return NaN values as coefficients which result in NaNs as prevalence and correlation estimates. One solution for tackling this problem could be to merge the categories of predictors with small number of patients together. To do so, the three groups of ‘Child’, ‘Adolescent’ and ‘Young Adult’ have been merged together and labeled as ‘Child and Young Adult’. We also merged ‘Adult’ and ‘Middle Aged’ to one category. Therefore,

the modified age variable in the dataset has three categories of 'Child and Young Adult', 'Adult and Middle-Aged' and 'Elder'.

'Third Income Quintile' and 'Fourth Income Quintile' have also been merged together. According to Statistics Canada (Statistics Canada, 2010) average adjusted after-tax income is divided into five quintiles in 2010 as it is shown in Table 4-3.

Table 4-3: Average Adjusted After-Tax Income by five quintiles for population in 2010

Quintile	Average adjusted after-tax income
Lowest income quintile	$16000
Second income quintile	$28000
Third income quintile	$38500
Fourth income quintile	$50600
Highest income quintile	$85500

Statistics Canada's income grouping is used in this thesis to label the new categories of SES after merging them. This attribute breaks down the patients into two groups of 'Less than or Equal to $50600' and 'Greater than $50600' average adjusted after-tax income.

Now, except the age variable, all other variables including chronic diseases, gender and SES are binary random variables in our dataset. Based on Sections 4.2.1 and 4.2.2, the age variable with three categories should be converted to dummy variables. So, we created two dummy variables *age1*, *age2* with the reference category being 'Child and Young Adult' (first category).

4.3 Covariate Adjustment

There are three different approaches for data assessment: explanatory modeling, descriptive modeling and predictive modeling (Shmueli, 2010). In explanatory modeling, we investigate the underlying causal relationships between independent variables and outcomes. Descriptive analysis refers to applying statistical models to the data in order to assess the association between one or more independent variables and a dependent variable. Predictive modeling is a process with the aim of predicting the target values for new observations given their input values (Shmueli, 2010). In this thesis, the goal is to examine the relationships between variables rather than predicting new

or future observations. We focus on including covariates for analyzing the association between the independent variable(s) of interest (disease diagnosis) and the outcome (disease prevalence and disease correlation). The techniques developed in this thesis are mostly descriptive, but they are moving towards explanatory modeling as we have included different covariates like age, gender and SES. We estimate the prevalence of chronic diseases as well as their correlations adjusting for these covariates. We could support explanatory analysis for some kinds of problems if we are given the right data and theoretical underpinning.

Covariate adjustment is a statistical strategy through which the covariates are held constant in order to capture the relative relationship of the dependent and independent variables. In other words, if the baseline covariate(s) is correlated with the dependent variable, their hidden effects on the dependent variable will be removed through this process. Controlling for covariates results in less bias and more precise estimates (Raab et al. 2000, Pocock et al. 2002).

4.4 Our Visual Analytics System

The visual analytics system in this thesis is designed using Flask and D3.js. Flask is a Python web application framework and D3.js is a library in JavaScript for creating interactive visualizations. We built our binary logistic regression and softmax regression models with Statsmodels and our decision tree model using Scikit-Learn library for Python. This visual analytics system encompasses two interactive graphs: a bar chart and a correlation matrix (see Figure 4-1). The bar graph depicts the prevalence of the chronic diseases and the correlation matrix represents the correlation between two diseases at the time. Both graphs assess the effects of the user-selected chronic conditions and patient characteristics on the target in their estimations.

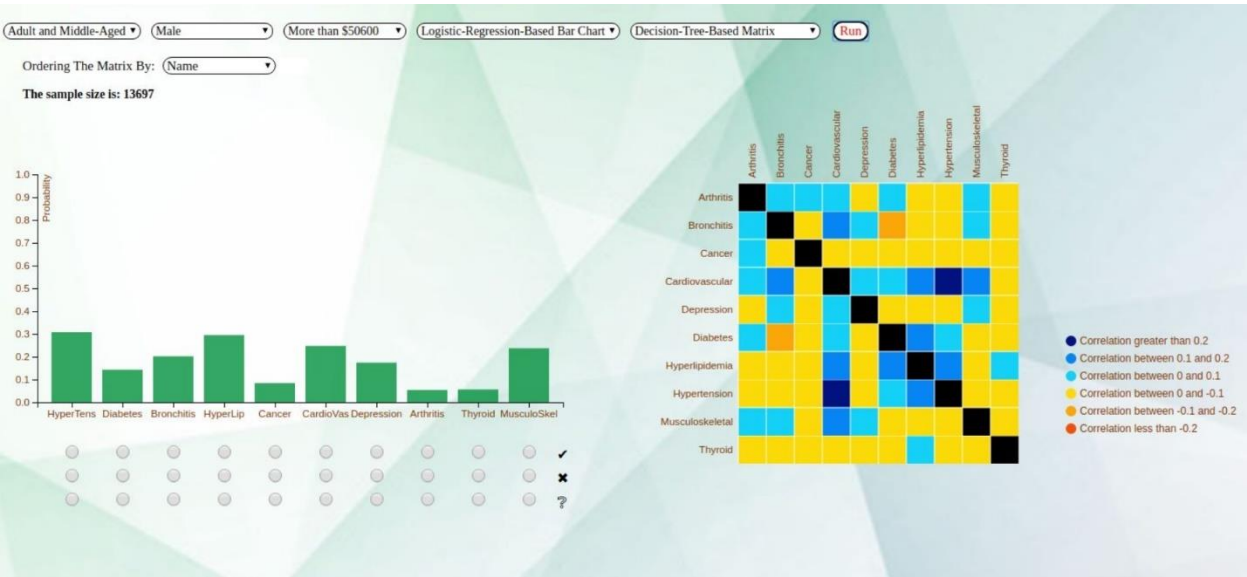

Figure 4-1: Screenshot of our visual analytics system and its components

As it is shown in Figure 4-2, three drop-down lists allow users to select different categories of patient characteristics, age, gender, and SES. There are two more drop-down lists on the top right corner of Figure 4-2 indicating the type of models the bar chart and the correlation matrix use to represent the results, respectively. The first or top left drop-down list represents three age groups including children and young adults, adults and middle aged, and elderly people. The drop-down list in relation to gender shows 'Female' and 'Male' categories and the third drop-down list enables users to select one of the two income groups. Right below the aforementioned dropdown lists, there is another drop-down list through which the user can order the pairwise correlations shown in the correlation matrix by their values and their names. The last line illustrates the sample size of filtered data changing based on user selection and filtering. As it is shown in Figure 4-2, the user has selected 'Adult and Middle-Aged', 'Male' and 'Less than $50600'. Therefore, the data would be filtered based on the selected attributes and by clicking on "Run" button, the prevalence of chronic diseases would be represented through count-based bar chart and the correlation estimates between each pair of diseases would be displayed on softmax-regression-based matrix plot. These correlation coefficients are estimated based on all patients in the data and in the presence of 'Adult and Middle-Aged' group and 'Male' category. More details about how the correlation matrix works are provided in Section 4.4.3 and 4.4.4. After filtering the data, the user can observe the change on the sample size that indicates 2053 patients in the dataset are adult or middle-aged and male with average income less than 50600 dollars.

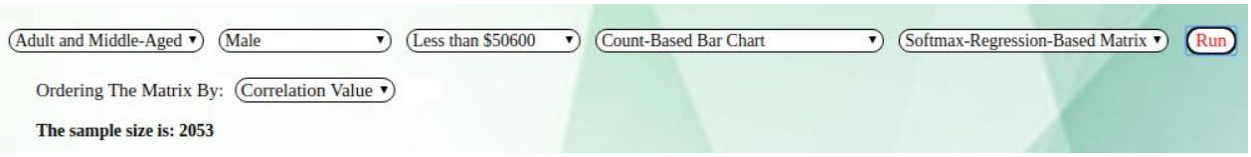

Figure 4-2: Screenshot of the system dropdown lists with 'Adult and Middle-Aged', 'Female' and 'Less than $50600' groups, Count-Based Bar Chart and Softmax-Regression-Based Correlation Matrix selected

Furthermore, three radio buttons are associated with each bar in the bar chart represented in our system allowing end-users to select a single item at a time. More clearly, every bar encodes the prevalence of a chronic disease and its corresponding radio buttons take one of the labels 1, 2, or null. if users select the first and nearest radio button to a bar with label 1, the presence of its corresponding disease is considered as the pre-existing condition for further analysis. In contrast, selecting the second radio button means the absence of the corresponding disease. Finally, a radio button with null label under each bar indicates the marginal probability of its corresponding disease in the count-based bar chart. It means the data is not filtered on a disease whose null radio button is selected for calculating the conditional probabilities. Using machine learning algorithms, the diseases with the third radio button selected, have no role in influencing the target. As an example, it is shown in Figure 4-3 that the user has selected the presence of bronchitis and 'Depression' and the absence of 'Diabetes' to identify their effects on the prevalence and correlation estimates.

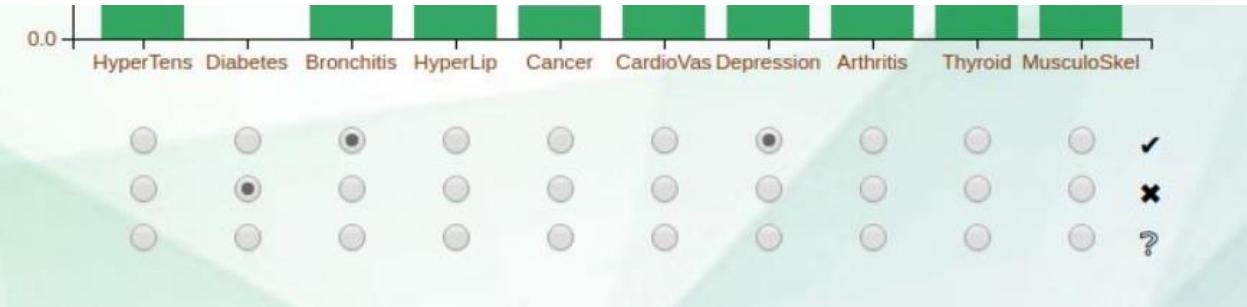

Figure 4-3: Screenshot of the system's radio button lists in the presence of bronchitis and depression and the absence of diabetes

The system analyzes and visualizes the underlying patterns in the multimorbidity data through its interactive graphs. The bar chart uses conditional probability as a measure for investigating the distribution of co-occurring chronic disease in the population. It also employs binary logistic

regression and decision tree model in order to examine the influence of sociodemographic factors and disease diagnosis on the prevalence of chronic diseases. Besides, the interactive correlation matrix in the system represents the pairwise correlation estimates of chronic conditions through two statistical and machine learning algorithms, softmax regression and decision tree. All these algorithms and methods are described in detail as follows:

4.4.1 Count-Based Bar Chart

Through selecting "Count-Based Bar Chart" from the drop-down list corresponding to the type of the interactive bar chart, the prevalence of chronic diseases is displayed on the bar chart in our visual analytics system. Each bar on the x-axis is allocated to one disease X_i and the prevalence of that disease $P(X_i = 1)$ is presented on the y-axis. When the user selects a disease X_i by clicking on its radio button with label 1, the bar graph would be animated to display the conditional probabilities of each disease (conditioned on X_i=1). Similar to it, by selecting the zero-labeled radio button associated to X_i, the probabilities of each disease conditioned on X_i=0 would be calculated and shown on the related bar to that disease. In former, the height of the bar allocated to X_i changes to 1, while in latter the bar shows the probability of X_i equal to 0.

If the user selects an additional disease, the system calculates the probability of each unselected disease conditioned on both selected diseases. Then, the system animates the change and updates the visualization. The selection process can be continued by the user to look for further associations within the subgroup who have the selected diseases, and so on. Therefore, the bar graph enables users to compare the original marginal probability of each disease with its conditional probabilities every time by every selection.

The user can also interact with the visualizations by selecting different age, gender and socioeconomic groups from the dropdown lists. As a result, the dataset of multimorbid patients would be filtered on the selected sociodemographic factors and the conditional probabilities would be updated. For example, if the user selects 'Child and Young Adult' as the age group, 'Male' from the gender groups, and the existence of diabetes, the dataset would be filtered and a subset of patients diagnosed with diabetes who are male and categorized as child and young adult would be chosen. Then the relative prevalence of each unselected disease X_j among child or young adult, male patients with diabetes would be computed and represented on its related bar and the

prevalence of diabetes would change to 1 in the bar graph. In this case, the conditional probability formula for the j^{th} unselected disease is as follows:

$$P\left(X_j = 1 \middle| diabetes = 1, age = \text{child and young adult}, gender = \text{male}\right)$$

Since all chronic diseases in the dataset are binary random variables taking value 0 or 1, the prevalence of a disease is equivalent with the mean of that binary random variable. Therefore, after user selections, the mean of every unselected disease is computed based on filtered data and shown in the related bar.

4.4.2 Logistic-Regression-Based Bar Chart

Selecting "Logistic-Regression-Based Bar Chart" from the corresponding drop-down list, the bar graph represents the prevalence of chronic diseases predicted by binary logistic regression. Similar to the count-based type, the user can reason with data through this predictive bar chart and discover the impacts of patient characteristics as well as one or a group of pre-existence diseases on the prevalence of other diseases. In other words, the binary logistic regression used in the system allows the user to estimate the association between diseases of interest as independent variables and an unselected disease as the outcome, while adjusting for (or controlling for) selected sociodemographic factors that are included in the model.

Each time the user makes a selection from the different dropdown lists and radio buttons with different labels available in the system, a logistic regression model would be built to determine the strength of the relationship between an unselected chronic disease as dependent variable and the selected attributes as independent variables. Therefore, in logistic-regression-based bar chart, the height of every bar corresponds to the prevalence estimated by relative logistic regression model, unless the bar is related to a selected disease whose radio button with label 0 or 1 is checked. The logistic regression model uses the entire data with all 13697 patients for estimating the prevalence and only the selected sociodemographic factor (factors) and pre-existing disease (diseases) are included in these models. The model would be changed and updated if the user changes the selection.

As an example, if the user clicks on the radio button with label 0 related to arthritis (the absence of arthritis), selects the radio button with label 1 corresponding to thyroid disease (the presence of

thyroid disease) and ‘Elder’ age group, the logistic regression model for finding a mathematical relationship between them and ‘Cancer’ as the target is as follows:

$$z = \log\left(\frac{P}{1-P}\right) = \beta_0 + \beta_1(arthritis) + \beta_2(age1) + \beta_3(age2) + \beta_4(thyroid\ disease)$$

Where P is the probability of developing cancer which can be defined as cancer prevalence in a given time period, and β_1 is the estimated coefficient that quantifies the association between arthritis, thyroid disease and cancer, adjusted for *age1*, *age2*. It is important to note that when age is in the list of selected attributes, one of its three dummy variables should be dropped from the model to avoid multicollinearity. More clearly, ‘Child and Young Adult’ would be excluded from the model and if the user selects this category, it can influence the target by assigning value 0 to both *age1* and *age2* dummy variables. In addition, both subgroups of each independent variable are used by the logistic regression to model an unselected disease. Then, the predictor variables in the standard model above are replaced by specific selected values to predict the probability of developing that disease. In this example, we change the independent variables in the model above such that arthritis is absent from the model, the age group is ‘Elder’ (*age1*=0 and *age2*=1), and thyroid disease is present in the model to estimate the conditional prevalence of cancer:

$$z = \log\left(\frac{P}{1-P}\right) = \beta_0 + \beta_1(arthritis = 0) + \beta_2(age1 = 0) + \beta_3(age2 = 1)$$

$$+ \beta_4(thyroid\ disease = 1)$$

The bar chart displays diseases on the x-axis and their prevalence recovered by exponentiating the log odds on the y-axis. Based on the following equation, the probability of developing cancer for the mentioned instance would be predicted and shown on the corresponding bar in the bar chart.

$$P(cancer = 1 | arthritis = \text{absent}, age = \text{elder}, thyroid\ disease = present) = \frac{1}{1+e^{-z}}$$

Following this, the prevalence of arthritis changes to 0, the prevalence of thyroid disease changes to 1, and the prevalence of other unselected diseases are also estimated through the aforementioned estimation process and represented in the logistic-regression-based bar chart.

4.4.3 Decision-Tree-Based Bar Chart

Decision tree is another model used in our system to assess the associations between multiple chronic diseases and patient attributes. If the user selects 'Decision-Tree-Based Bar Chart' from the dropdown list, a decision tree model would be created such that all diseases of interest as well as selected patient characteristics would be included in the model. It is important to note that if a categorical variable with more than two categories is selected (e.g. age), we do not use one-hot encoding to binarize each category, which converts the categorical variable into dummy variables. We avoid this process because dummy variables make a decision tree sparse and obscure the order of feature importance, which results in inefficiency and poor performance. We also build the model based on all patients included in the dataset.

Furthermore, we do not divide the dataset into train and test sets, since we aim to examine the relationships between binary random variables and sociodemographic factors rather than improving the prediction of the prevalence or the correlation of diseases. However, to avoid overfitting and reduce complexity, we utilize pruning methods by changing the parameters 'max_depth' (=3) and 'min_samples_leaf' (=200) in Python server, which refer to the maximum number of nodes in a branch and the minimum number of samples required at the leaf node (a node without further split), respectively. In this way we remove the sections of the tree that do not add significant value to the classification power of the tree and avoid unstable probability estimates.

The prevalence of disease X_i estimated using the decision tree model would be displayed on the corresponding bar in the bar chart. The height of the bar (bars) for the disease (diseases) whose radio button with label 1 or label 0 is selected changes to 1 and 0, respectively.

4.4.4 Softmax-Regression-Based Correlation Matrix

Correlation matrix is known as an appropriate statistical technique for describing the relationship between variables. It is a square matrix which each element demonstrates the association between a pair of variables. All values on the main diagonal of a correlation matrix are 1 since the correlation of a variable with itself is always 1. The strength and direction of a relationship between variables can be explained by correlation as a statistical measure. However, a correlation between variables does not imply causation. In other words, the correlation estimate does not assure that the change in the value of one variable is the cause of the change in the other variables. A

correlation value near 1 means the two variables have a strong positive correlation, while a value near -1 shows they are highly correlated but in the opposite direction. A correlation coefficient with value 0 indicates there is no linear relationship between the pair of variables.

The dynamic correlation matrix in our visual analytics system displays the pairwise correlations between ten chronic diseases. It calculates the correlation between two diseases at a time, conditioned on selected other diseases and covariates. A row and a column are allocated to each disease and each cell in the matrix shows the linear relationship between two diseases. Suppose we aim to measure the association between two chronic diseases D_1 and D_2. We create a new variable A having the following four levels:

$$A = 0 \text{ if } D_1 = 0 \text{ and } D_2 = 0$$
$$A = 1 \text{ if } D_1 = 0 \text{ and } D_2 = 1$$
$$A = 2 \text{ if } D_1 = 1 \text{ and } D_2 = 0$$
$$A = 3 \text{ if } D_1 = 1 \text{ and } D_2 = 1$$

Since our new target is variable A with four levels (K=4), we need to build a softmax regression in order to predict the pairwise correlation between these two diseases. As mentioned in Section 2.5.3, softmax regression utilizes a linear predictor function $f(k, i)$ to predict the probability that observation *i* belongs to class *k*:

$$f(k, i) = \beta_{0,k} + \beta_{1,k}x_{1,i} + \beta_{2,k}x_{2,i} + \cdots + \beta_{M,k}x_{M,i} \quad \text{for } k = 1, \ldots, K$$

Where *M* is the number of independent variables in the model and *i* is an observation from 13697 inputs in the data. We assign value 0 to 'Male' category and value 1 to 'Female' category, since in this dataset, gender is encoded as a binary variable.

If the user selects the presence depression and 'Male' group, the softmax regression model built for class zero is as follows:

$$f(0) = \beta_{0,0} + \beta_{1,0}(depression = 1) + \beta_{2,0}(gender = 0)$$

After computing the linear predictor function for all four classes of the dependent variable A, we can also compute the probability of each class as follows:

$$P(A = 0) = \frac{e^{\beta_{0,0}+\beta_{1,0}(depression=1)+\beta_{2,0}(gender=0)}}{1 + \sum_{k=1}^{3} e^{f(k)}}$$

$$P(A=1)=\frac{e^{\beta_{0,1}+\beta_{1,1}(depression=1)+\beta_{2,1}(gender=0)}}{1+\sum_{k=1}^{3}e^{f(k)}}$$

$$P(A=2)=\frac{e^{\beta_{0,2}+\beta_{1,2}(depression=1)+\beta_{2,2}(gender=0)}}{1+\sum_{k=1}^{3}e^{f(k)}}$$

$$P(A=3)=\frac{1}{1+\sum_{k=1}^{3}e^{f(k)}}$$

The correlation between the two random variables *X* and *Y* is calculated through the following formula:

$$\rho_{X,Y}=\frac{E(XY)-E(X)E(Y)}{\sigma_X\sigma_Y} \tag{4.1}$$

Where σ_X and $E(X)$ denote the standard deviation and the expected value of *X*, respectively, and $E(XY)$ is defined as follows when *X* and *Y* are discrete random variables and not independent:

$$E(XY)=\sum_{x\in X}\sum_{y\in Y}xy\,P(X=x,Y=y) \tag{4.2}$$

We name $P(A=0)=P_{00}$, $P(A=1)=P_{01}$, $P(A=2)=P_{10}$ and $P(A=3)=P_{11}$. We also define $P_{1.}=P_{10}+P_{11}$ and $P_{.1}=P_{01}+P_{11}$. Given that all chronic diseases in our thesis are random variables from Bernoulli distribution we have $E(D_1)=P_{1.}$, $\sigma_{D_1}^2=P_{1.}(1-P_{1.})$, $E(D_2)=P_{.1}$ and $\sigma_{D_2}^2=P_{.1}(1-P_{.1})$.

According to Equation 4.2 and given that D_1 and D_2 might influence each other, we calculate $E(D_1D_2)=(0\times0\times P_{00})+(0\times1\times P_{01})+\ (1\times0\times P_{10})+\ (1\times1\times P_{11})$=$P_{11}$. Then, the correlation between D_1 and D_2 is computed as follows:

$$\rho_{D1,D2}=\frac{P_{11}-P_{1.}P_{.1}}{\sqrt{P_{1.}(1-P_{1.})}\sqrt{P_{.1}(1-P_{.1})}} \tag{4.3}$$

This process would be repeated for each pair of chronic diseases, and their estimated correlation would be depicted by the corresponding cell in the interactive matrix. By hovering over each cell, the corresponding correlation value appears. The direction of the relationships between diseases are encoded by color. Blue and orange are used for positive and negative correlations, respectively. In addition, color intensity encodes the magnitude of the correlation coefficients such that a darker color represents a greater absolute value. The user can also re-arrange the correlation matrix by disease name and correlation value.

As mentioned in Sections 4.3.1, 4.3.2 and 4.3.3, the height of the bar corresponding to a selected disease X_i changes to 1 or 0, based on the selection. Similar to this process, if the user selects X_i, the color of all cells in the row i and the column i corresponding to X_i in the correlation matrix would change to black, which indicates the undefined correlations. The reason is that in calculating the correlation coefficient between two variables, if one variable does not vary, its standard deviation changes to zero that results in the denominator of the fraction to be zero and in turn the correlation coefficient is undefined.

4.4.5 Decision-Tree-Based Correlation Matrix

By selecting 'Decision-Tree-Based-Correlation Matrix' from the dropdown menu related to the type of matrix, a decision tree is made given the selected variables and with the parameters max_depth'=3 and 'min_samples_leaf'=200 to prevent overfitting. As mentioned in 4.3.4, the target in the correlation matrix is the variable A corresponding to a pair of chronic diseases and has four levels. For instance, suppose the user selects 'adult and middle aged' and the presence of hyperlipidemia, and aims to observe their influence on the association between cardiovascular disease and hypertension as the target. Therefore, the model would examine the relationship between hyperlipidemia and the target controlling for age. Then the probability of occurring each class of the target would be estimated using one instance (in this case *age*='Adult and Middle-Aged' and *hyperlipidemia*=1). According to equation 4.3, the four computed probabilities would be used in estimating the correlation coefficient between cardiovascular disease and hypertension. This analysis would be repeated for all other pairs of unselected diseases. The correlation of those pairs whose one or both diseases are selected, is undefined. In this example, all correlations between hyperlipidemia and the other nine chronic diseases would be undefined and their relative cells in the correlation matrix would change to black.

To design interaction, we considered the tasks that the user can perform with the data. These tasks encompass exploring the associations between the chronic diseases, analyzing the effects of different sociodemographic factors on the prevalence of diseases and comparing the correlation estimates in the presence of a set of diseases as well as patient attributes. Supporting these tasks, we have operationalized the following actions in the system: selecting, filtering, arranging, and comparing.

The visual marks such as shape, color, letter and digit along with their properties as visual variables including size, value and color saturation are utilized in our system to encode information items and represent quantity, association and order.

During performing the tasks, cognitive activities like analytical reasoning, sensemaking and decision-making can emerge. For example, a user decides to discover the effect of gender on the probability of developing thyroid disease. Therefore, he/she engages in testing the hypothesis if the prevalence of thyroid disease differs among woman and men and through selecting gender categories and filtering data, he/she would make sense of the relationship between these two variables.

4.5 Results

The goal of designing the visual analytics system in this thesis is to identify and analyze multimorbidity and its associations with patient attributes. Users can set their intentions, then select a group of items to achieve their purposes. There could be a large number of task sequences though which users can accomplish their overall goals. In our visual analytics system, through every user selection and filtering, several multimorbidity patterns would be explored and categorized. We aim to present one of these sequences, as an example, to clarify how multimorbidity patterns can appear in our visualization.

Analysis 1: Assume the user aims to observe the marginal probability of diseases for ‘Child and Young Adult’ category. The data would be filtered on the age group of interest and the results would be displayed on the Count-Based Bar Chart. As it is shown in Figure 4-4, bronchitis and depression are the most prevalent diseases among 1304 children and young adults in the data.

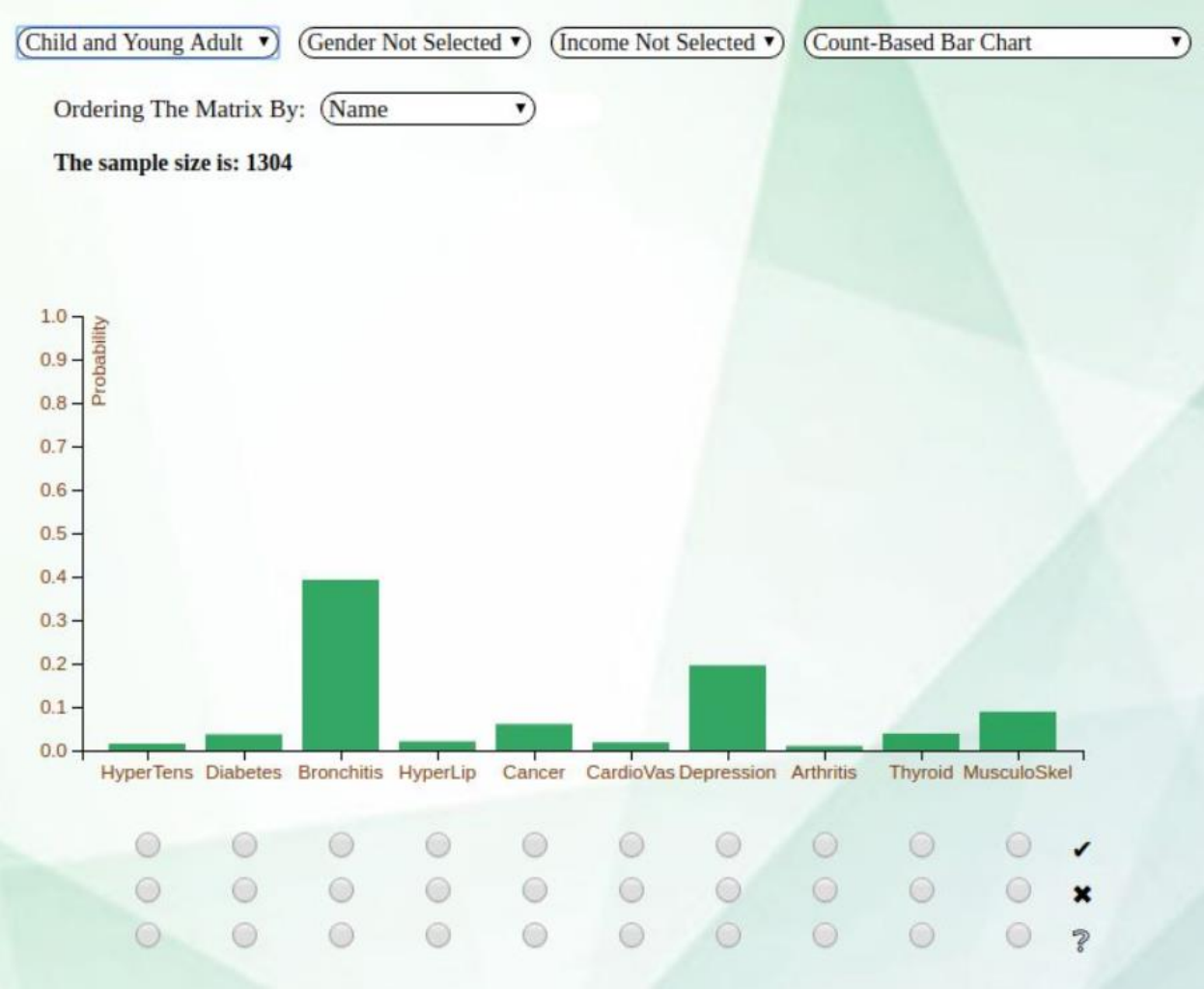

Figure 4-4: Screenshot of the Count-Based Bar Chart for Analysis 1 with 'Child and Young Adult' age category selected

Figure 4-5 shows the correlations between the ten diseases on Softmax-Regression-Based Correlation Matrix in Analysis 1. The user can observe the number of cells denoted by orange is more than the number of blue cells, though, all values are between -0.3 and 0.3 which indicate weak correlations between the diseases.

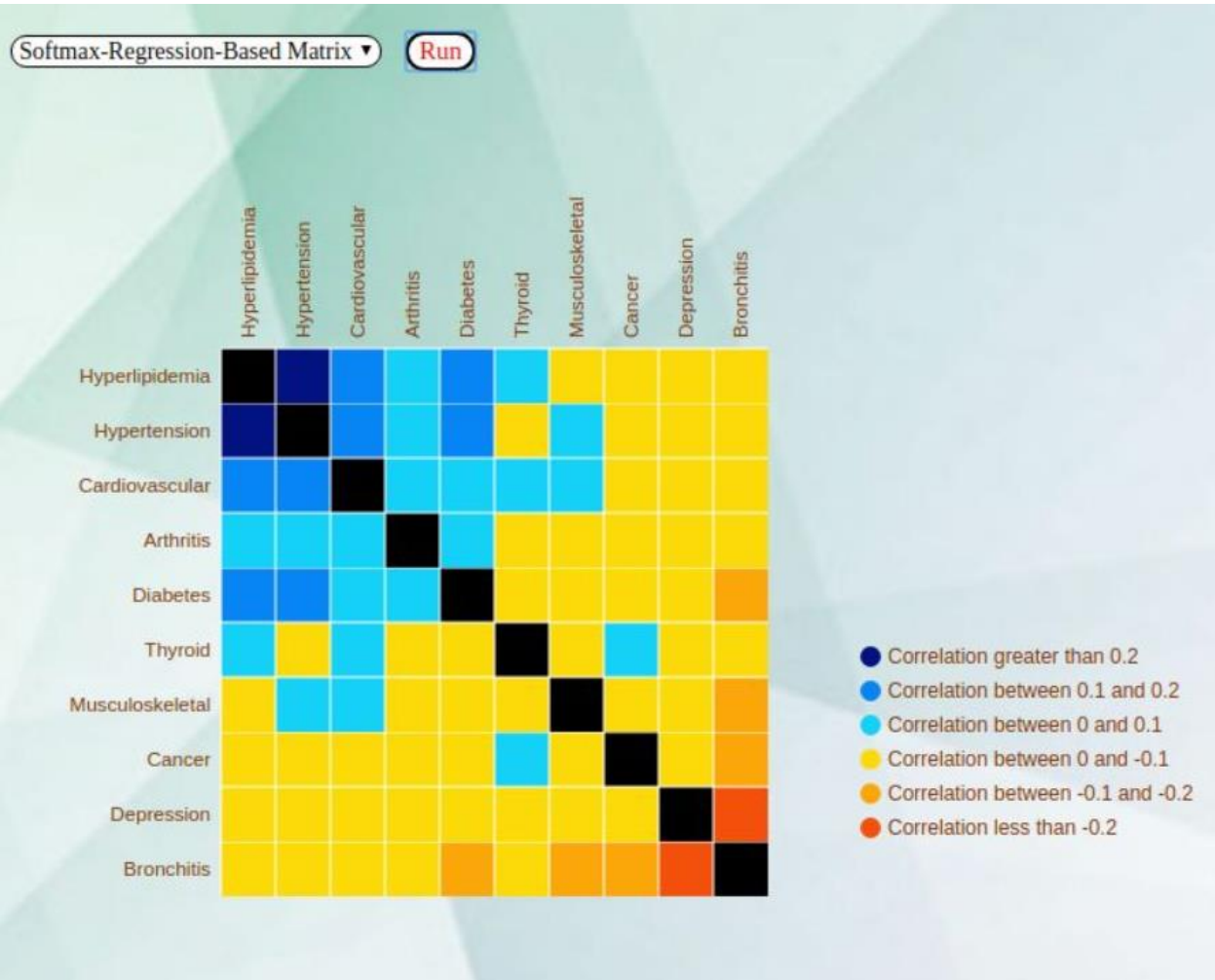

Figure 4-5: Screenshot of the Softmax-Regression-Based Correlation Matrix for Analysis 1 with 'Child and Young Adult' age category selected

Analysis 2: Looking at the original bar graph, if the user selects 'Elder' age group and Count-Based Bar chart and he/she does not select any of chronic diseases as a pre-existing condition, the data would be filtered on only the selected age group and the marginal probability for every disease is computed in the server and shown on the corresponding bar in the client side. Figure 4-6 depicts the updated bar chart when 'Elder' group is selected. The user can observe that the most common chronic diseases among 6361 older adults are hypertension, hyperlipidemia and cardiovascular disease. Besides, compared to Figure 4-4, the probabilities of developing almost all chronic diseases grow noticeably except depression and bronchitis which are less prevalent among older adults than among children and younger adults.

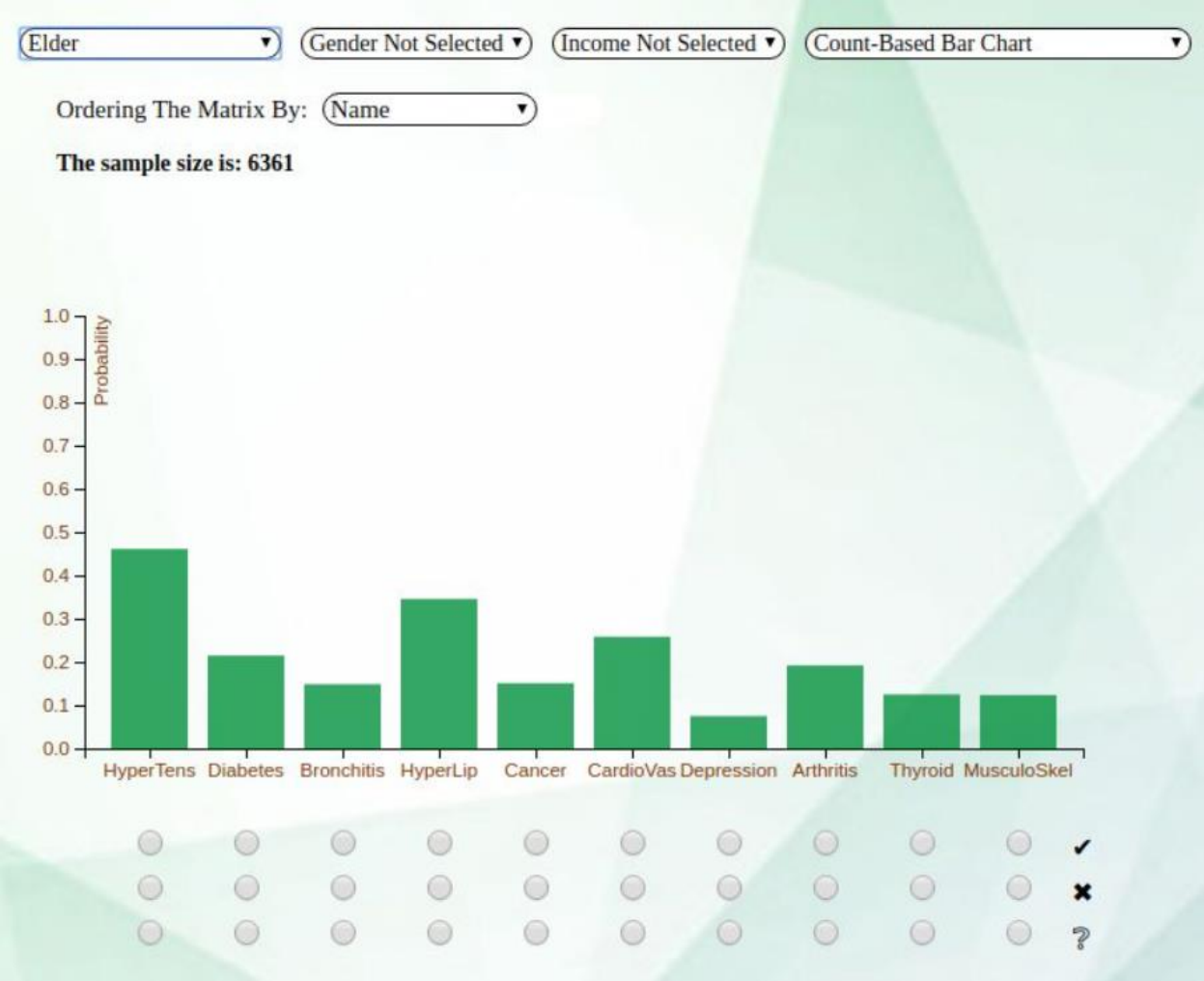

Figure 4-6: Screenshot of the Count-Based Bar Chart for Analysis 2 with 'Elder' age category selected

Analysis 3: Following this, suppose the user selects the presence of hypertension and chooses Decision-Tree Based Bar Chart from the fourth dropdown list to observe and interpret the results. Therefore, the prevalence of the other diseases conditioned on the presence of hypertension appears on this type of bar chart. The probabilities of diabetes, hyperlipidemia and 'Cardiovascular Disease' increase by 3.86%, 7.78% and 5.3%, respectively (see Figure 4-6 and Figure 4-7); while the probabilities for bronchitis and depression stay almost the same.

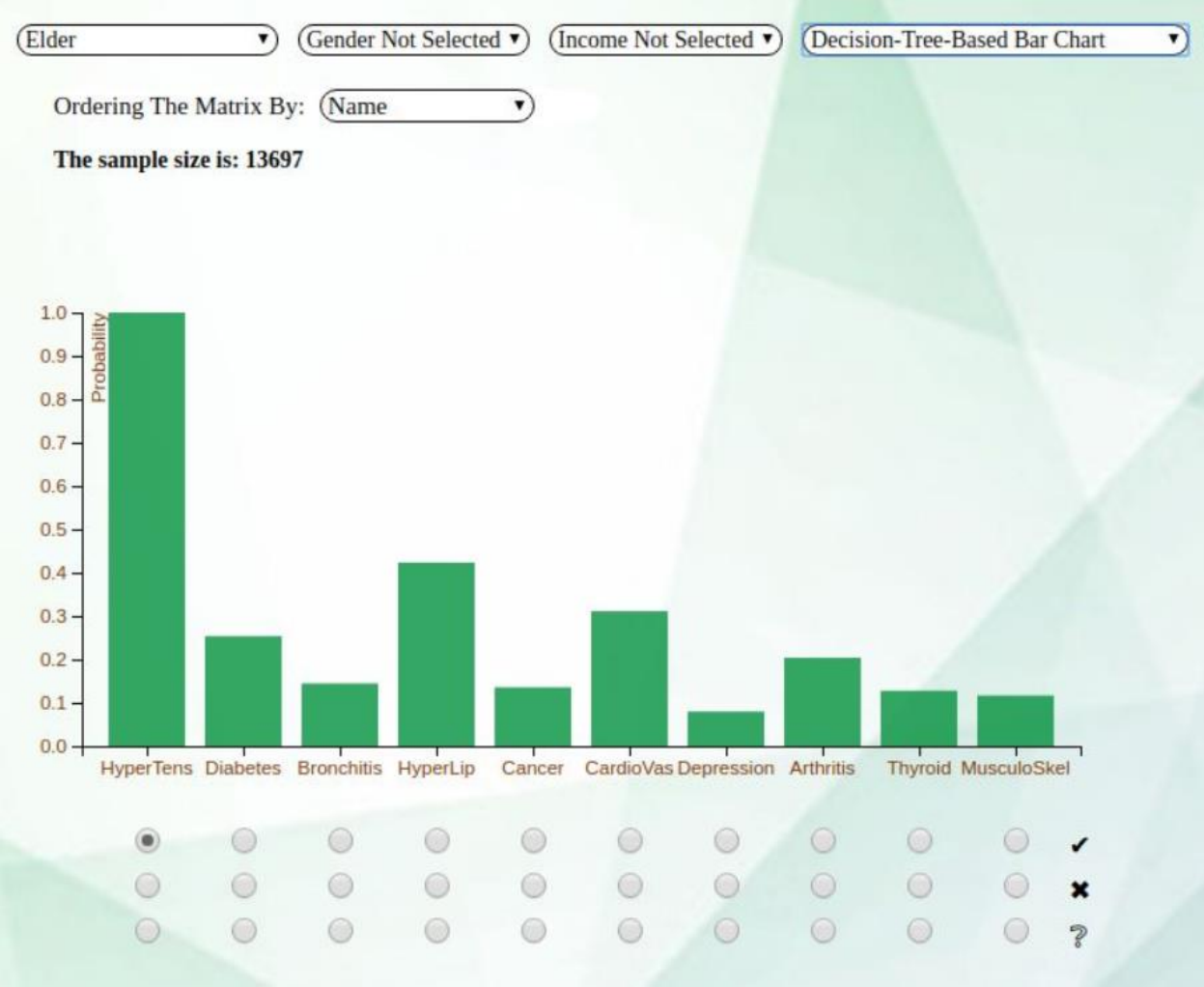

Figure 4-7: Screenshot of the Decision-Tree-Based Bar Chart for Analysis 3 with 'Elder' category and the presence of hypertension selected

In analysis 3, the user also selects Decision-Tree-Based Correlation Matrix to explore the associations between the chronic diseases. As shown in Figure 4-8, all cells corresponding to the correlation coefficients between hypertension and the other diseases change to black as these correlations are undefined. The user can hover over every cell to observe the exact correlation value of a disease pair. Compared to Figure 4-5, the number of positive correlations (blue cells) has been increased, which means for two diseases X_i and X_j, increased X_i results in an increase in X_j. Figure 4-8 also shows that for older people diagnosed with hypertension, the most correlated diseases are hyperlipidemia and cardiovascular disease with the value 0.227.

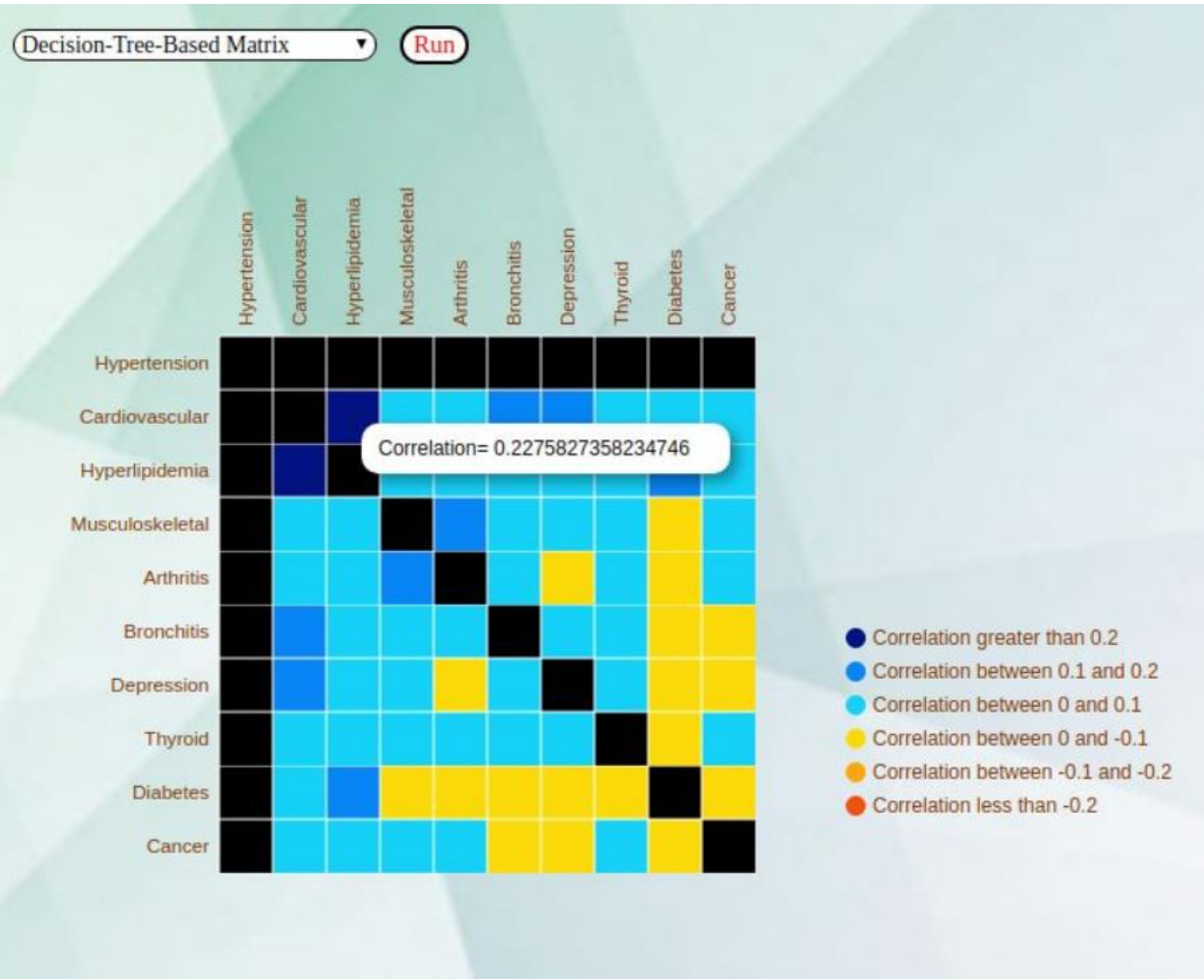

Figure 4-8: Screenshot of the Decision-Tree-Based Correlation Matrix for Analysis 3 with 'Elder' category and the presence of hypertension selected

Analysis 4: Then if the user selects 'Female' group, the prevalence of diabetes, hyperlipidemia and cardiovascular disease decreases, the prevalence of 'Arthritis', 'Thyroid Disease' and 'Musculoskeletal Problem' increases, and the prevalence of bronchitis, cancer and depression stays almost the same (see Figures 4-7 and 4-9, and Table 4-4). In other words, after the diagnosis of hypertension, gender does not affect the probability of developing bronchitis, cancer and depression. In contrast, the probability of living with arthritis, thyroid disease and musculoskeletal problem goes up among elderly adults who are female in the presence of hypertension. Of course, the actual increases and decreases depend on the estimated probabilities, which are derived from EHR data and DELPHI database.

Table 4-4: A comparison between the prevalence of the ten chronic diseases estimated based on the selections in Analysis 3 and 4. Abbreviations: HT=Hypertension, DB=Diabetes, BC=Bronchitis, HL=Hyperlipidemia, CC=Cancer, CD=Cardiovascular Disease, DP=Depression, AT=Arthritis, TD=Thyroid Disease, MP=Musculoskeletal Problem

	HT	DB	BC	HL	CC	CD	DP	AT	TD	MP
Analysis 3	1	0.253	0.144	0.423	0.136	0.311	0.080	0.204	0.128	0.117
Analysis 4	1	0.221	0.153	0.382	0.136	0.260	0.091	0.243	0.179	0.139

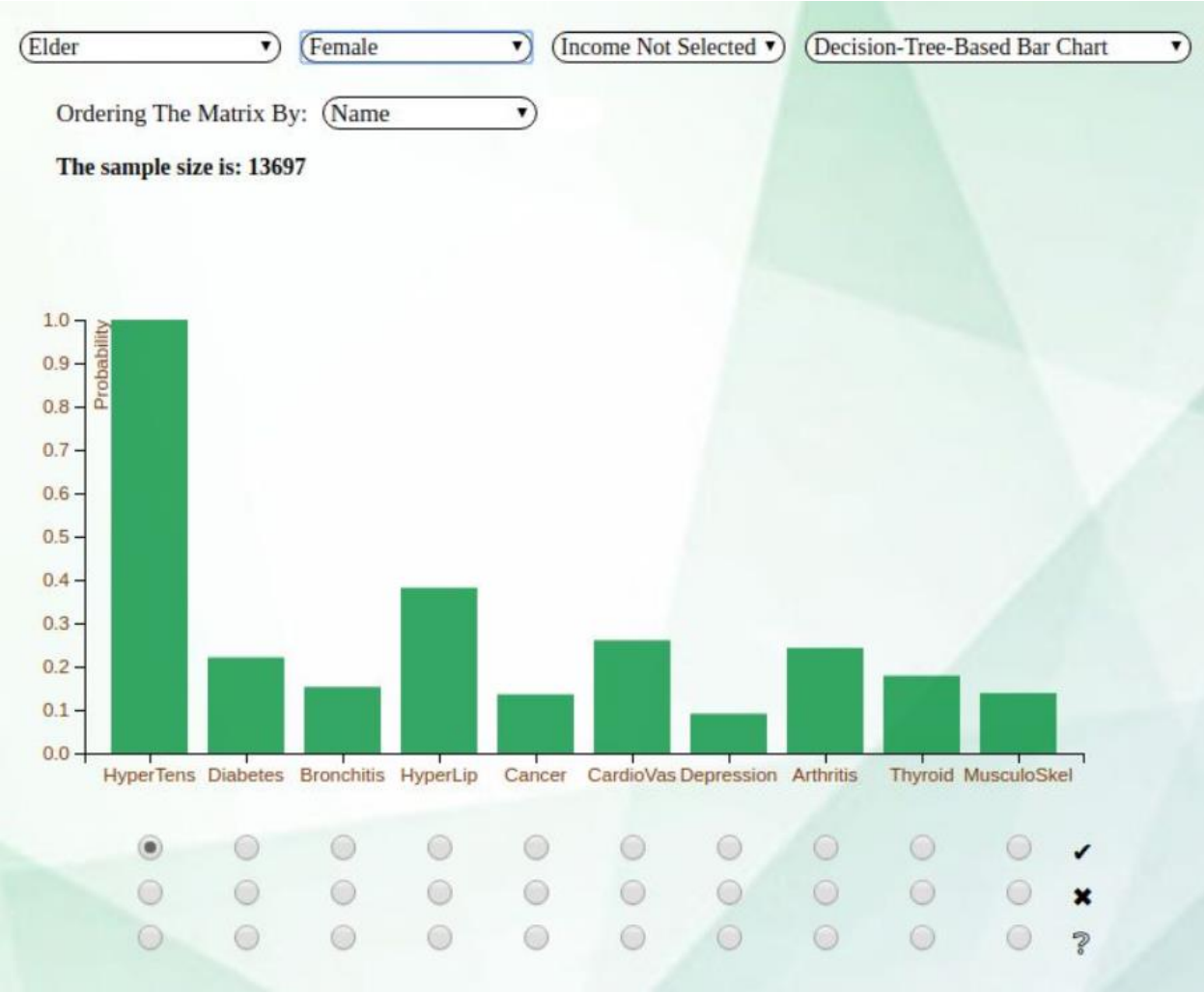

Figure 4-9: Screenshot of the Decision-Tree-Based Bar Chart for Analysis 4 with 'Elder' and 'Female' categories and the presence of hypertension selected

Analysis 5: As the next step, suppose the user selects the radio button with label 1 for arthritis, the probability of chronic diseases conditioned on the diagnosis of hypertension and arthritis would be represented on the corresponding bars. Depicted in Figure 4-10, the prevalence of musculoskeletal problem increases by six percent in the presence of hypertension and arthritis.

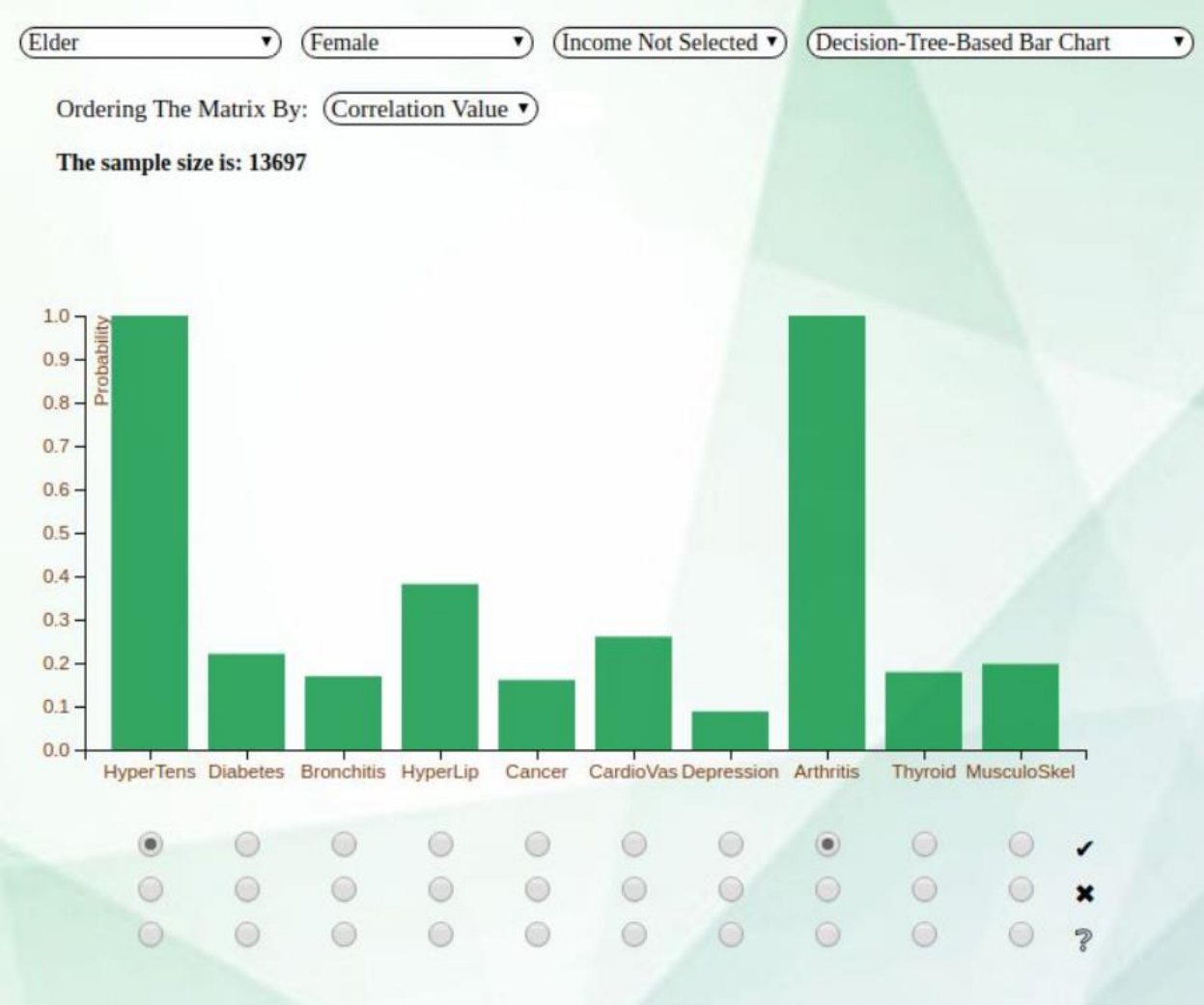

Figure 4-10: Screenshot of the Decision-Tree-Based Bar Chart for Analysis 5 with 'Elder' and 'Female' categories, the presence of hypertension and the presence of arthritis selected

Table 4-5 shows the prevalence estimates of the ten diseases obtained from count-based conditional probability, decision tree and logistic regression model based on the selections in Analysis 5. Comparing the two classifiers with count-based conditional probability, the estimated probabilities are close to each other, which means the models are generating the outputs accurately. Nevertheless, it is necessary to compare these models using statistical significance testing in order

to evaluate if there is no real difference. It is also important to mention that we use all of these models for stratifying for different kinds of covariates in the data. In other words, although the models improve the prediction of the prevalence of and the correlations between chronic diseases, they are being used for our statistical strategy which is investigating the relationships between the diseases and the sociodemographic characteristics.

Table 4-5: A comparison between three algorithms count-based Conditional Probability, Decision Tree and Binary Logistic Regression, by assessing the prevalence estimates based on the selections in Analysis 5. Abbreviations: HT=Hypertension, DB=Diabetes, BC=Bronchitis, HL=Hyperlipidemia, CC=Cancer, CD=Cardiovascular Disease, DP=Depression, AT=Arthritis, TD=Thyroid Disease, MP=Musculoskeletal Problem

	HT	DB	BC	HL	CC	CD	DP	AT	TD	MP
Conditional Probability (Count-Based)	1	0.235	0.171	0.468	0.165	0.331	0.085	1	0.233	0.220
Decision Tree	1	0.221	0.170	0.382	0.161	0.261	0.088	1	0.179	0.198
Logistic Regression	1	0.222	0.145	0.401	0.127	0.326	0.068	1	0.170	0.140

Furthermore, the two machine learning algorithms softmax regression and decision tree predict the same correlation coefficients in the cases the user only makes one selection. We examined the correlations between cardiovascular disease and the other chronic disease, as an example, to compare the performance of the two models through the five analyses. As it is shown in Table 4-6, for both Analysis 1 and Analysis 2 in which only one variable is selected, the correlations estimated by softmax regression and decision tree are the same. As the number of selections increases, the results obtained from the two models differ from each other for some pairs of the diseases. For instance, in Analysis 5, the correlation between cardiovascular disease and bronchitis estimated by the softmax regression differs from the correlation between these two diseases predicted by the decision tree. The reason could be that softmax regression is not a count-based model. It borrows information from the other examples that are not selected, especially in the cases

like Analysis 5 with multiple selections that the number of selected examples is low, and the model needs additional information. Since the goal of our visual analytics system is to explore the associations between variables rather than improving the predictions or determining the best classifier, we don't focus on the slight difference between the outputs of the decision tree and those of the softmax regression.

Table 4-6: A comparison between two machine learning models, Softmax Regression and Decision Tree, which are used for correlation estimation, for all five analyses. Cardiovascular disease is chosen as an example to compare the estimated correlations between this disease and the other nine diseases in the data. Abbreviations: HT=Hypertension, DB=Diabetes, BC=Bronchitis, HL=Hyperlipidemia, CC=Cancer, CD=Cardiovascular Disease, DP=Depression, AT=Arthritis, TD=Thyroid Disease, MP=Musculoskeletal Problem, SR=Softmax Regression, DT =Decision Tree.

	Type	HT	DB	BC	HL	CC	DP	AT	TD	MP
Analysis1	SR Model	0.122	0.004	0.087	0.140	0.035	0.010	0.044	0.002	0.017
	DT Model	0.122	0.004	0.087	0.140	0.035	0.010	0.044	0.002	0.017
Analysis2	SR Model	0.112	0.033	0.060	0.158	-0.016	0.107	0.041	0.027	0.047
	DT Model	0.112	0.033	0.060	0.158	-0.016	0.107	0.041	0.027	0.047
Analysis3	SR Model	-	0.044	0.107	0.228	0.0002	0.158	0.067	0.082	0.068
	DT Model	-	0.016	0.143	0.198	-0.003	0.193	0.051	0.091	0.128
Analysis4	SR Model	-	0.025	0.147	0.191	-.019	0.224	0.075	0.121	0.120
	DT Model	-	0.019	0.076	0.181	-.035	0.140	0.091	0.127	0.056
Analysis5	SR Model	-	0.061	0.206	0.248	0.029	0.251	-	0.239	0.196
	DT Model	-	0.018	0.076	0.181	-.035	0.140	-	0.127	0.056

As shown in Figure 4-11, the rows and columns corresponding to hypertension and arthritis in the matrix are represented by black color, which indicate these two diseases are selected and their correlations with the other diseases are not defined.

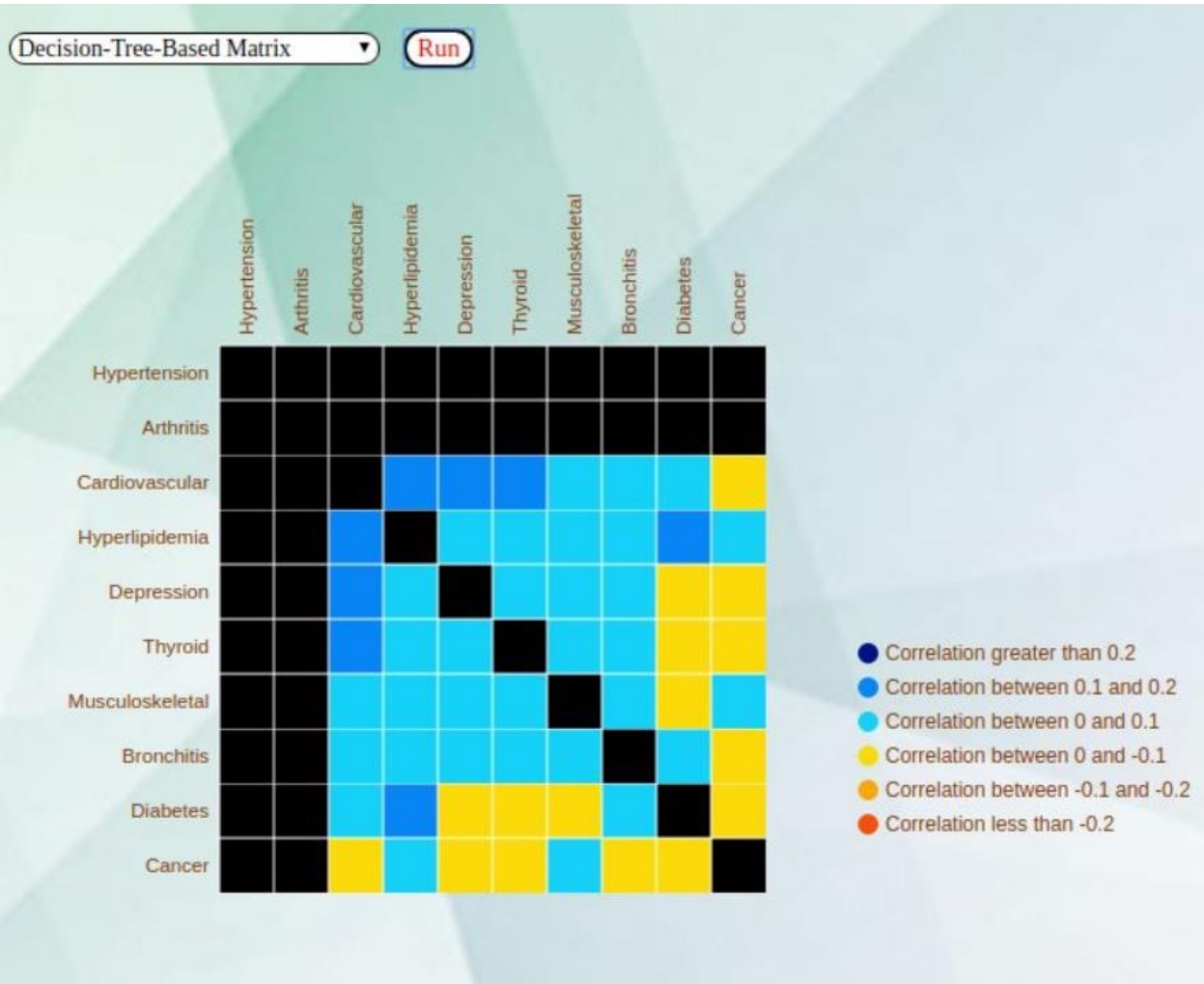

Figure 4-11: Screenshot of the Decision-Tree-Based Correlation Matrix for Analysis 5 with 'Elder' and 'Female' categories, the presence of hypertension and the presence of arthritis selected

Chapter 5

5 Conclusion

This chapter provides a summary of the purpose of the thesis, the visual analytics system designed in this thesis and its components. It also describes briefly how these visual and interactive components are developed and work, and how users can engage in a meaningful discourse with multimorbidity data through interaction. The chapter concludes by discussing the limitations of this research and the issues to be addressed in the future.

5.1 Thesis Summary

Multimorbidity is a growing healthcare challenge especially for older adults and results in greater vulnerability, higher risk of functional decline and disability and higher mortality (Low et al 2019, Schäfer et al 2014). Focusing on chronic diseases individually no longer meets the needs of healthcare providers in preventing and managing these chronic conditions. A holistic approach to chronic diseases and their associations with sociodemographic characteristics and risk factors is needed to design effectual prevention and control strategies. Therefore, we designed an application system for analyzing and exploring multimorbidity associations in a visual, interactive manner. Unlike many studies in the area of multimorbidity whose results are shown through simple charts, tables and flowcharts, our visual analytics system allows users to interact with several subsets of data and select a set of chronic diseases and specific categories of age, gender and socioeconomic scores for investigation. The system encompasses two dynamic graphs, a bar chart and a correlation matrix. The interactive bar chart uses count-based conditional probability, decision tree and binary logistic regression to show how a selected disease or a sociodemographic factor affects the prevalence of the other diseases. The correlation matrix builds a softmax regression and a decision tree for each pair of chronic diseases considering the disease diagnosis and selected

patient characteristics to estimate the correlations of disease pairs. This matrix indicates the magnitude of correlation coefficients by color intensity and their positivity or negativity using blue and orange, respectively. The user can also observe the estimated correlations of disease pairs by hovering over the related cell in the correlation matrix.

This thesis designs the foundation for the use of visual analytics systems in investigating multimorbidity patterns. The visualizations in our system can be implemented for other purposes in the area of healthcare or other disciplines where high-dimensional joint distributions of random variables are of significance. The system can also apply the other statistical and machine learning models and interpret more data with more available features.

5.2 Discussion

As mentioned in Chapter 3, most of the investigations on multimorbidity patterns are presented in static charts and tables. Besides, there are several visual analytics systems applied to healthcare data that investigate the relationships between chronic conditions. Some of these tools and applications enable users to select a set of sociodemographic characteristics, filter the data and consequently observe the associations of diseases. The difference between the visual analytics system designed in this thesis and other applications is that this system allows users to select not only different categories of patient characteristics but also a list of pre-existing chronic conditions to explore their effects on the prevalence and correlation of other unselected diseases in the data. For example, the user can select four chronic diseases as the diagnosed conditions. Therefore, these four diseases would be included in softmax regression, logistic regression and decision tree models for estimation. Then, these models estimate the prevalence and the pairwise correlation of the unselected diseases.

However, as mentioned in Section 4.4.4, the target variable A has 4 levels and is created based on the possible outcomes of two chronic diseases. In other words, the system provides the correlation between two diseases in the correlation matrix, while estimating the multiple correlation coefficients (the correlation between more than two unselected diseases) can be interesting as well.

Since some of the categories of patient characteristics have few observations in our data, we merged them together. Although merging small categories is a common way in preprocessing data

steps, it may hide some information. Thus, using count-based techniques like conditional probability (count-based bar chart) that show which categories are very small or missing can be useful. It is also important to note that our system shows disease correlation and prevalence estimates to help users discover more knowledge about the data provided by DELPHI collection. Therefore, changing the data that the system uses may result in completely different estimated values. The system may also represent different results in analyses with user-selected age and income groups if we shift these arbitrary categories slightly.

5.3 Limitations and Future Directions

Need for a larger dataset: The dataset used in this thesis includes 13697 patients with one or more chronic diseases. It is not a small number of observations, but through every selection by the user, the amount of data with certain attributes decreases quickly and the selected data can include categorical variables with too small a number of observations or even no observations. Due to our small dataset and to avoid more variables with too few observations through interaction, we presented the underlying patterns of the ten most common chronic diseases in our visual analytics system. The database provides information about a list of 20 chronic conditions among multimorbid patients across Ontario. Analyzing and visualizing all 20 diseases and their associations using a bigger dataset with more observations for every specific variable would be interesting and can be considered for future work.

Lack of risk factors and other sociodemographic attributes in the data: Apart from analysis of associations between chronic diseases and sociodemographic factors and associations among chronic diseases themselves, adding various risk factors like obesity, smoking, poor nutrition, lack of exercise and genetic and environmental factors to the system can be helpful for users in identifying the causal effect of risk factors on the likelihood of developing chronic diseases. Unfortunately, our dataset is not large enough and it does not provide any risk factors in the patients diagnosed with one or multiple chronic diseases. There is also a lack of other sociodemographic information such as aboriginal status, immigration status, area of residence and race/ethnicity.

Using other machine learning algorithms: As outlined in Chapter 4, decision tree, binary logistic regression and softmax regression models are employed to analyze multimorbidity and

explore its patterns. It should be mentioned that our visual analytics system is not limited to these three supervised learning algorithms and it can potentially comprise a wide variety of other statistical and machine learning models in order to enable users to compare the results obtained from different classifiers and their performances.

Quantifying the results by statistical significance: The correlation estimates represented by each cell in the correlation matrix plot are computed using softmax regression and decision tree models. It is necessary to determine if these association estimates are statistically significant. In this case, the null hypothesis states that there is no relationship between two chronic diseases, and they are independent. If the test result exceeds the significance level, which is often 0.05, the null hypothesis will be rejected and the alternative hypothesis is accepted, which indicates the two measured chronic conditions are strongly associated with each other.

References

Abad-Díez, J. M., Calderón-Larrañaga, A., Poncel-Falcó, A., Poblador-Plou, B., Calderón-Meza, J. M., Sicras-Mainar, A., … Prados-Torres, A. (2014). Age and gender differences in the prevalence and patterns of multimorbidity in the older population. *BMC Geriatrics*, *14*(1), 1–8.

Adadi, A., & Berrada, M. (2018). Peeking inside the black-box: A survey on explainable artificial intelligence (XAI). *IEEE Access, 6*, 52138–52160.

Alimohammadian, M., Majidi, A., Yaseri, M., Ahmadi, B., Islami, F., Derakhshan, M., … Malekzadeh, R. (2017). Multimorbidity as an important issue among women: Results of a gender difference investigation in a large population-based cross-sectional study in West Asia. *BMJ Open, 7*(5), 1–8.

Alpaydin, E. 2014. Introduction to machine learning. *MIT press*.

Andrienko, N., Lammarsch, T., Andrienko, G., Fuchs, G., Keim, D., Miksch, S., & Rind, A. (2018). Viewing visual analytics as model building. *Computer Graphics Forum*, *37*(6), 275–299.

Balducci, B., & Marinova, D. (2018). Unstructured data in marketing. *Journal of the Academy of Marketing Science*, 46(4), 557–590.

Birtwhistle, R., & Williamson, T. (2015). Primary care electronic medical records: A new data source for research in Canada. *Cmaj*, *187*(4), 239–240.

Ceneda, D., Gschwandtner, T., May, T., Miksch, S., Schulz, H. J., Streit, M., & Tominski, C. (2017). Characterizing guidance in visual analytics. *IEEE Transactions on Visualization and Computer Graphics*, *23*(1), 111–120.

Chen, S., Lin, L., & Yuan, X. (2017). Social media visual analytics. *Computer Graphics Forum*, *36*(3), 563–587.

Centre for Studies in Family Medicine (2020). DELPHI (Deliver Primary Healthcare Information) project. Available from: https://www.schulich.uwo.ca/familymedicine/research/csfm/research/current_projects/delphi.html

Coleman, N., Halas, G., Peeler, W., Casaclang, N., Williamson, T., & Katz, A. (2015). From patient care to research: A validation study examining the factors contributing to data quality in a primary care electronic medical record database. *BMC Family Practice*, *16*(11), 1–8.

Evans, P. C., & Basole, R. C. (2016). Economic and business dimensions: Revealing the API ecosystem and enterprise strategy via visual analytics. *Communications of the ACM*, *59*(2), 26–28.

Farmer, C., Fenu, E., O'Flynn, N., & Guthrie, B. (2016). Clinical assessment and management of multimorbidity: Summary of NICE guidance. *BMJ (Online)*, *354*, 1–5.

Feely, A., Lix, L. M., & Reimer, K. (2017). Estimating multimorbidity prevalence with the Canadian chronic disease surveillance system. *Health Promotion and Chronic Disease Prevention in Canada*, *37*(7), 215–222.

Flood, M. D., Lemieux, V. L., Varga, M., & William Wong, B. L. (2016). The application of visual analytics to financial stability monitoring. *Journal of Financial Stability*, *27*, 180–197.

Fortin, M., Lapointe, L., Hudon, C., Vanasse, A., Ntetu, A. & Maltais, D. (2004). Multimorbidity and quality of life in primary care: a systematic review. *Health and Quality of Life Outcomes*. 2 (51).

Gallacher, K. I., Jani, B. D., Hanlon, P., Nicholl, B. I., & Mair, F. S. (2019). Multimorbidity in Stroke. *Stroke*, *50*(7), 1919–1926.

Garavaglia, S., Dun, A. S., & Hill, B. M. (1998). A smart guide to dummy variables: Four applications and a macro.

García-Peñalvo, F. J. (2015). Issue on visual analytics. *Journal of Information Technology Research*, *8*(2), 1–2.

Garies, S., Birtwhistle, R., Drummond, N., Queenan, J., & Williamson, T. (2017). Data Resource Profile: National electronic medical record data from the Canadian Primary Care Sentinel Surveillance Network (CPCSSN). *International Journal of Epidemiology*, *46*(4), 1091-1092f.

Gilpin, L. H., Bau, D., Yuan, B. Z., Bajwa, A., Specter, M., & Kagal, L. (2018). Explaining explanations: An overview of interpretability of machine learning. *2018 IEEE 5th International Conference on Data Science and Advanced Analytics, DSAA 2018*, 80–89.

Hajat, C., & Stein, E. (2018). The global burden of multiple chronic conditions: A narrative review. *Preventive Medicine Reports*, *12*, 284–293.

Health Canada. (2012). Health care system. About primary health care. Available from: https://www.canada.ca/en/health-canada/services/primary-health-care/about-primary-health-care.html

Japkowicz, N., Shah, M, 2011. Evaluating learning algorithms: A classification perspective. *Cambridge University Press*.

Jiang, M., Liang, Y., Feng, X., Fan, X., Pei, Z., Xue, Y., & Guan, R. (2018). Text classification based on deep belief network and softmax regression. *Neural Computing and Applications*, *29*(1), 61–70.

Jowsey, T., Jeon, Y. H., Dugdale, P., Glasgow, N. J., Kljakovic, M., & Usherwood, T. (2009). Challenges for co-morbid chronic illness care and policy in Australia: A qualitative study. *Australia and New Zealand Health Policy*, *6*(1), 1–8.

Low, L. L., Kwan, Y. H., Ko, M. S. M., Yeam, C. T., Lee, V. S. Y., Tan, W. B., & Thumboo, J. (2019). Epidemiologic characteristics of multimorbidity and sociodemographic factors associated with multimorbidity in a rapidly aging Asian country. *JAMA Network Open*, *2*(11).

Manogaran, G., & Lopez, D. (2018). Health data analytics using scalable logistic regression with stochastic gradient descent. *International Journal of Advanced Intelligence Paradigms*, *10*(1–2), 118–132.

Mohri, M., Rostamizadeh, A. and Talwalkar, A. 2018. Foundations of machine learning. *MIT*

press.

Molnar, C. (2019). Interpretable Machine Learning: A guide for making black box models Explainable.

Muldoon, L. K., Hogg, W. E. & Levitt, M., (2006). Primary Care (PC) and Primary Health Care (PHC) what is the difference? *Canadian Journal of Public Health*. 97(5):409-11.

Murdoch, T. B., & Detsky, A. S. (2013). The inevitable application of big data to health care. *JAMA - Journal of the American Medical Association*, *309*(13), 1351–1352.

Muth, C., Blom, J. W., Smith, S. M., Johnell, K., Gonzalez-Gonzalez, A. I., Nguyen, T. S., … Valderas, J. M. (2019). Evidence supporting the best clinical management of patients with multimorbidity and polypharmacy: A systematic guideline review and expert consensus. *Journal of Internal Medicine*, *285*(3), 272–288.

Navickas, R., Petric, V.-K., Feigl, A. B., & Seychell, M. (2016). Multimorbidity: What do we know? What should we do? *Journal of Comorbidity*, *6*(1), 4–11.

Nicholson, K., Terry, A. L., Fortin, M., Williamson, T., Bauer, M., & Thind, A. (2015). Examining the prevalence and patterns of multimorbidity in Canadian primary healthcare: A methodologic protocol using a national electronic medical record database. *Journal of Comorbidity*, *5*, 150–161.

Nicholson, K. (2017). Multimorbidity among adult primary health care patients in Canada: Examining multiple chronic diseases using an electronic medical record database. (Doctoral thesis, The University of Western Ontario.)

O'Halloran, J., Miller, G. C. & Britt, H. (2004). Defining chronic conditions for primary care with ICPC-2. *Family Practice*, 21(4), 381–386.

Ola, O., & Sedig, K. (2014). The challenge of big data in public health: An opportunity for visual analytics. *Online Journal of Public Health Informatics*, *5*(3), 1–21.

Pocock, S. J., Assmann, S. E., Enos, L. E., & Kasten, L. E. (2002). Subgroup analysis, covariate adjustment and baseline comparisons in clinical trial reporting: Current practice and problems. *Statistics in Medicine*, *21*(19), 2917–2930.

Raab, G. M., Day, S., & Sales, J. (2000). How to select covariates to include in the analysis of a clinical trial. *Controlled Clinical Trials*, *21*(4), 330–342.

Raghupathi, W., & Raghupathi, V. (2018). An empirical study of chronic diseases in the united states: A visual analytics approach. *International Journal of Environmental Research and Public Health*, *15*(431), 1–24.

Roberts, K. C., Rao, D. P., Bennett, T. L., Loukine, L., & Jayaraman, G. C. (2015). Prevalence and patterns of chronic disease multimorbidity and associated determinants in Canada. *Health Promotion and Chronic Disease Prevention in Canada*, *35*(6), 87–94.

Salisbury, C., Johnson, L., Purdy, S., Valderas, J. M., & Montgomery, A. A. (2011). Epidemiology and impact of multimorbidity in primary care: A retrospective cohort study. *British Journal of General Practice*, *61*(582), 12–21.

Schäfer, I., Kaduszkiewicz, H., Wagner, H. O., Schön, G., Scherer, M., & Van Den Bussche, H. (2014). Reducing complexity: A visualisation of multimorbidity by combining disease clusters and triads. *BMC Public Health*, *14*(1285).

Schrider, D. R., & Kern, A. D. (2018). Supervised machine learning for population genetics: A new paradigm. *Trends in Genetics*, *34*(4), 301–312.

Sedig, K., Naimi, A., & Haggerty, N. (2017). Aligning information technologies with evidence-based health-care activities: A design and evaluation framework. *Human Technology*, *13*(2), 180–215.

Sedig, K., & Parsons, P. (2013). Interaction design for complex cognitive activities with visual representations: A pattern-based approach. *AIS Transactions on Human-Computer Interaction*, *5*(2), 84–133.

Sedig, K., & Parsons, P. (2016). Design of visualizations for human-information interaction: A pattern-based framework. In *Synthesis Lectures on Visualization*, *4*.

Shaikhina, T., Lowe, D., Daga, S., Briggs, D., Higgins, R., & Khovanova, N. (2017). Decision tree and random forest models for outcome prediction in antibody incompatible kidney transplantation. *Biomedical Signal Processing and Control*, *52*(2019), 456–462.

Shmueli, G. (2010). To explain or to predict? *Statistical Science*, *25*(3), 289–310.

Shneiderman, B., Plaisant, C., & Hesse, B. W. (2013). Improving healthcare with interactive visualization. *The IEEE Computer Society*, *46*(5), 58–66.

Simpao, A. F., Ahumada, L. M., & Rehman, M. A. (2015). Big data and visual analytics in anaesthesia and health care. *British Journal of Anaesthesia*, *115*(3), 350–356.

Spence, R. (2014). Information visualization. In *Springer International Publishing Switzerland*. Springer.

Starfield, B. (1998). Primary care: Balancing health needs, services and technology ,2nd Ed. *New York and Oxford: Oxford University Press*, 8(9).

Statistics Canada. (2010). Average adjusted after-tax income by after-tax income quintiles for all persons, 2006 to 2010. Available from: http://www.statcan.gc.ca/pub/75-202-x/2010000/ analysis-analyses-eng.htm

Stewart, M. (2016). CPCSSN(Canadian Primary Care Sentinel Surveillance Network) Project. Available from: http://cpcssn.ca/regional-networks/ delphi-deliver-primary-healthcare-information-project/

Stewart, M., Thind, A., Terry, A. L., Chevendra, V., & Marshall, J. N. (2009). Implementing and maintaining a researchable database from electronic medical records: A perspective from an academic family medicine department. *Healthcare Policy*, *5*(2), 26–39.

Strayer, N., Shirey-rice, J. K., Shyr, Y., Denny, J. C., Pulley, J. M., & Xu, Y. (2019). PheWAS-ME : A web-app for interactive exploration of multimorbidity patterns in PheWAS.

The Lancet. (2018). Making more of multimorbidity: an emerging priority. *The Lancet*, *391*(10131), 1637.

Thomas, J. J., & Cook, K. A. (2005). Illuminating the path: The research and development agenda for visual analytics. In *National Visualization and Analytics Ctr*.

Violán, C., Foguet-Boreu, Q., Roso-Llorach, A., Rodriguez-Blanco, T., Pons-Vigués, M., Pujol-Ribera, E., … Valderas, J. M. (2014). Burden of multimorbidity, SES and use of health services across stages of life in urban areas: A cross-sectional study. *BMC Public Health*,

14(530).

Wallace, E., Salisbury, C., Guthrie, B., Lewis, C., Fahey, T., & Smith, S. M. (2015). Managing patients with multimorbidity in primary care. *BMJ (Online)*, *350*, 1–6.

West, V. L., Borland, D., & Hammond, W. E. (2014). Innovative information visualization of electronic health record data: A systematic review. *Journal of the American Medical Informatics Association*, *22*(2), 330–339.

Wilson, J. R., & Lorenz, K. A. (2015). Short history of the logistic regression model. In: *Modeling Binary Correlated Responses using SAS, SPSS and R.* ICSA Book Series in Statistics.

World Health Organization. (2015). Health statistics and information systems: Projections of mortality and causes of death, 2015 and 2030. Available from: http://www.who.int/healthinfo/global_burden_disease/projections/en/

Yang, J., Bai, Y., Lin, F., Liu, M., Hou, Z., Liu, X. (2018). A novel electrocardiogram arrhythmia classification method based on stacked sparse auto-encoders and softmax regression. *International Journal of Machine Learning and Cybernetics*, *9*(10), 1733–1740.

Curriculum Vitae

Name: Maede Sadat Nouri

Post-secondary Education and Degrees:

Bachelor of Science in Statistics
Isfahan University of Technology, Isfahan, Iran
2006-2011

Master of Science in Social Economics Statistics
Allameh Tabataba'i University, Tehran, Iran
2012-2015

Master of Science in Computer Science (candidate)
The University of Western Ontario, London, Ontario
2018-Present

Related Work Experience

Teaching and Research Assistant (Computer Science)
The University of Western Ontario, London, Ontario
2018-Present

Data Scientist
Jooyeshgar Company, Isfahan, Iran
2014-2017

Teacher (Statistics)
Isfahan Mathematics House, Isfahan, Iran
2010-2016

Publications:

Nouri, M. S., Sedig, K. and Lizotte, D., 2019. "Interactive Visualization of Patterns of Multimorbidity", The International Multimorbidity Symposium 2019, London, Ontario.

Samadie, M., Gholami, M., Nouri, M. S. and Samaei, S., 2014. "Construction of QC LDPC Codes with Desired Girths", 1st National Industrial Mathematics Conference (NIMC 2014), Tabriz, Iran.