World of Computer Science and Information Technology Journal (WCSIT)
ISSN: 2221-0741
Vol. 9, No. 2, 7-11, 2019

Multimorbidity Prediction Using Data Mining Model

Najah Al-shanableh, Mofleh Al Diabat
Department of Computer Science
Al Albayt University
Al Mafraq- Jordan

Abstract— This research aims to use data mining to predict health care outcomes. We will investigate patterns of multiple chronic conditions (MCCs), or multimorbidity, among the US elderly population. The multimorbidity prediction model, as a general aspect, was not found in the literature, although some researchers have been exploring the risk of developing further chronic conditions after reporting an index disease. Data mining can provide richer results compared to those produced using a statistical approach and greater depth and breadth. It can also help professionals to identify the best time to intervene. In this research, the primary focus was on building disease knowledge using data mining algorithms for MCCs in the elderly. We identified potential morbidity groups using clustering and tested several prediction models on HCUP real data with high accuracy, where the highest accuracy of 99.05% was achieved by Logistic Regression.

Keywords- Multimorbidity; Data mining; Classification; Clustering; Prediction; Chronic Diseases.

I. INTRODUCTION

Little is known about the pattern of MCCs and their occurrence in older adult patients, especially in the United States. We do know that such conditions are more widespread among older adults, those aged 65 and over. Furthermore, the varying patterns of comorbidity create a challenge for health care providers in that they affect the delivery of effective treatment and care coordination plans [5].

MCC, or multimorbidity, is defined as the "coexistence of two or more chronic conditions in one patient" [9]. Determining the patterns of multimorbidity and the differences between population characteristics can help in identifying better treatment options.

Most older adults have more than two such conditions, a situation that increases the urgency of intervention to enhance their lives [7]. This intervention is rendered more complicated by having only one model for MCC treatment. Identifying the most common patterns of MCCs and being able to predict patients' risk of developing MCCs will help health care providers to enact plans that enhance the quality of patients' lives and minimize their risk of developing further chronic conditions.

The importance of this research comes from several facts that affect US populations. According to the Centers for Disease Control (CDC), seven in ten Americans die from chronic conditions, which makes chronic illness one of the nation's most significant health care problems [8]. The top three causes of death in the United States are heart disease, cancer, and chronic lower respiratory diseases, all of which are chronic conditions [8].

"Older Adults" is one of the newly added topics in Healthy People 2020, a national benchmark that sets out US health objectives for the next decade [11]. Objective QA-3 of this topic aims to enhance the ability of older adults with one or more chronic diseases to manage their conditions and increase the proportion of the population who report greater control over their health. By finding patterns of multimorbidity in older adults, health care providers can provide better care for these patients because the providers will know what they are dealing with in advance; it can also assist them in helping patients gain control over their health.

Due to an aging population and the evolution of health care technology, we see an increase in life expectancy, which also brings further health issues. The prevalence of multimorbidity is growing, and its effect on individuals' lives and community performance is worsening [20].

Also, based on several reports from the Centers for Medicare & Medicaid Services (2012), chronic diseases are the main reason for disabilities and deaths throughout the United States; moreover, 66% of Medicare recipients in 2012 had two or more chronic conditions. Patients with multimorbidity were at higher risk of hospitalization and had longer stays in the hospital over a year. Furthermore, health care expenses increased thanks to the greater number of chronic conditions from which patients suffered [24].

Recently, in addition to the rise in the co-occurrence of chronic conditions, the burden of prevalence and management of multimorbidity is increasing [20]. The outcomes of many studies illustrate that multimorbidity occurs in more than 65% of adults aged 65 years or older. Thus, the management of MCCs is a crucial part of today's health care systems. Nonetheless, there has been a shortage of studies that focus on

multimorbidity compared to those looking at chronic diseases in general. As such, it is necessary to research MCCs in general and multimorbidity patterns individually.

The methods used to find multimorbidity patterns have varied from Chi-square tests to data mining [20] and generalized association rule mining [25]. Combinations of methods (prevalence figures, conditional count, logistic regression, and cluster analysis) have also been used to find disease co-occurrences. Table 1 summarizes the research methods that scholars have used to find disease patterns in either pairwise or triadic relationships. Although one paper featured an investigation of MCC patterns in general, that research depended mainly on patients reporting their disease [23].

The most commonly applied prediction model for chronic diseases has targeted only one disease or two combinations of diseases related to each other; only in rare cases have several disease risks been calculated in regard to an index disease [13]. Most research on the risk prediction model has targeted the most common chronic conditions, such as diabetes, cardiovascular diseases, and breast cancer [22] [1].

The multimorbidity prediction model was not found in the literature, although some research has offered an investigation of the risk of developing more chronic conditions after reporting an index disease [20][24].

In general, the methods used to develop risk prediction models fall into one of the following categories:

- Association rules
- Decision trees
- Multiple linear regression
- Regression model

The identification of useful information on the co-occurrence of chronic diseases will inform a health care plan that serves older adults.

TABLE 1. RESEARCH METHODS USED TO FIND DISEASE ASSOCIATIONS AND PATTERNS

Analysis methods	**Number of diseases in an association**	**Disease**
Exploratory factor analysis [6]	2	Physical and mental conditions
Exploratory factor analysis [21]	2	Cardio-metabolic, mechanical, and psychiatric-substance abuse
Patient self-reported [15]	≥2	Not specific
Prevalence and risk ratio [22]	3	The most common disease in the population
Person X^2 [13]	2	Not specific
Chi-square test [17]	2	Not specific
Exploratory factor analysis [23]	2	Chronic conditions and geriatric syndrome
Chi-square test [19]	3	Not specific
Data mining model [20]	2	Charcot foot

The massive data generated by health care organizations are too complicated, extensive, and numerous to be processed and analyzed by traditional methods [2], but they could be a rich source of information were they investigated in more depth. Data mining provides the necessary models to transform these comprehensive data into useful information for health care decision-making [2].

II. METHODOLOGY

The methodology used in this research was based entirely on the Knowledge Discovery in Databases (KDD) process [3]. KDD is defined as the process of information extraction from raw data. KDD uses data mining techniques as the primary method of such extraction and is the most widely used data mining process [3]. As Figure 1 shows, KDD consists of five stages, with the ability to go back to a previous step if needed. In this research, we followed all five stages:

1. Problem understanding
2. Target data selection and extraction
3. Data preparation
4. Data mining tasks
5. Evaluation

Data Set

The data set used was developed by the Healthcare Cost and Utilization Project (HCUP) [4]. This database is available for public use and does not include any patient identifiers.

The exploratory data analysis was carried out after several data preparation steps were performed to clean the data. These steps were implemented in RapidMiner software. We used only diagnosis variables from this data set.

Data Mining Step

Phase 1: Clustering was used in this research to create a cluster label to improve classification models. This step was used to identify possible MCC groups in the data set to enhance MCC prediction accuracy. We used the k-means algorithm to find groups within the data set and to label the new clustering for later use in classification. The k-means clustering algorithm works by dividing observations into k clusters. Furthermore, we used cluster density measures to check the quality of clusters found by k-means.

Phase 2: To overcome the issue of unbalanced classes, we trained and tested several classifiers on a balanced sample. The classifiers were trained on a subset of the data that contained the whole minority class and a random subset of the majority class so that both classes were balanced.

Six classifier models were trained and validated on a sample of HCUP data; then, tenfold cross-validation was used to check the model's performance. The tested classifiers were

1. Generalized Linear Model
2. Logistic Regression
3. Fast Large Margin
4. Decision Tree
5. Random Forest
6. Support Vector Machine

In classification, the experiments were set to monitor how a classifier performed on HCUP data. The sample data were divided into training data and a test data set to verify the accuracy of the algorithms and to evaluate the performance of a classifier. Of the sample data, 70% were used for the

algorithms' training, and the rest were employed for testing; this is one of data mining's best practices.

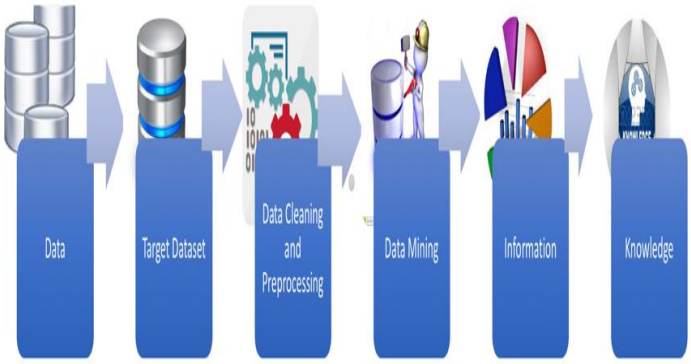

Figure 1. KDD process.

III. Results and Discussion

For this study, MCC, or multimorbidity, is defined as the coexistence of two or more chronic conditions in one patient [9]. Only patients aged 65 or older with MCCs were included in this research. There were 115,944 records in the data set included in this research, of which 1,075 were for patients with MCCs.

The mean age of patients was 76 ± 10 years and ranged from 65 to 99 years. The majority of MCC patients were in the 80–90 age group, accounting for 98.14% of the total records, which is similar to a national sample with a corresponding figure of 99% [9]. Approximately half of the MCC records were male patients (55%), and the majority were white (63.21%), both of which are consistent with previous literature [6].

To define the number of clusters into which the data could be divided and the optimal number of clusters to use (k), the Davies-Bouldin index (DBI) was used for each generated number of clusters. The lowest DBI values were found on k = 8, and k-means decided these values.

Figure 2 shows one of the many processes used to prepare data in RapidMiner, and Figure 3 shows the resultant clusters. Also, Table 2 shows the resulted clusters' description in term of some key variables.

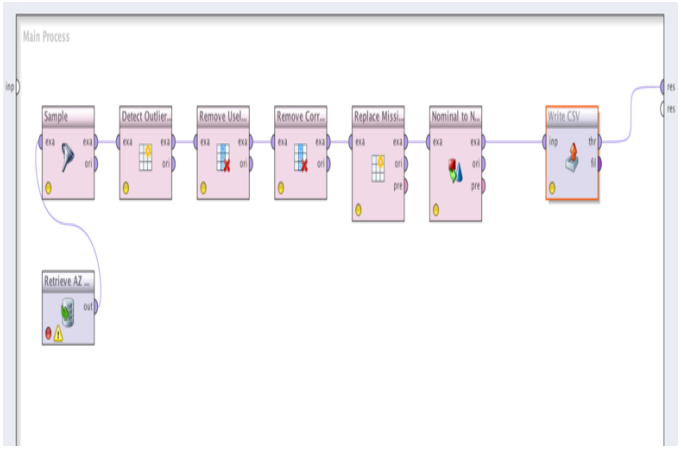

Figure 2. Data preparation process in RapidMiner.

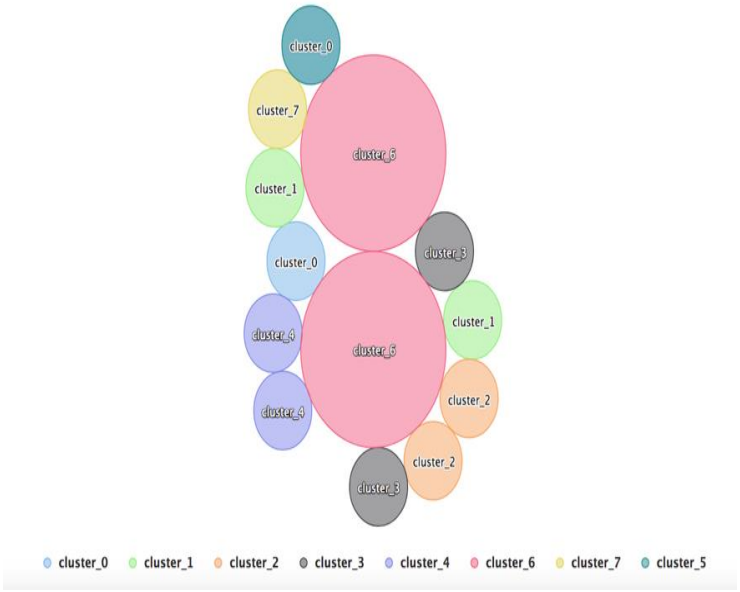

Figure 3 Resulted clusters

TABLE 2 RESULTED CLUSTERS' DESCRIPTION.

Label	Age average	Age SD	LOS average	LOS SD	NCHRONIC average	NCHRONIC SD
cluster_0	80.74	+/- 8.52	8.08	+/- 15.24	7.86	+/- 3.26
cluster_1	78.63	+/- 8.24	6.85	+/- 6.99	9.66	+/- 2.58
cluster_2	78.24	+/- 8.45	5.08	+/- 7.23	7.06	+/- 3.00
cluster_3	77.67	+/- 8.42	4.56	+/- 4.78	6.30	+/- 2.71
cluster_4	78.33	+/- 8.56	5.29	+/- 5.57	7.49	+/- 2.62
cluster_5	78.41	+/- 8.37	4.87	+/- 6.42	6.18	+/- 3.18
cluster_6	78.21	+/- 7.96	7.89	+/- 8.86	10.93	+/- 2.66
cluster_7	76.87	+/- 7.86	4.97	+/- 7.16	6.49	+/- 3.26

Accuracy, which is used as an evaluation measure in prediction, is defined as the degree to which the result of an estimation or calculation comply with the correct value of

prediction [3]. The top three algorithms were random forest, fast large margin, and neural network respectively as shown in Table 3; which shows the accuracy of the tested models. Along with table 3, figure 4 shows the tested models' accuracy, and figure 5 shows the classification errors.

TABLE 3 ACCURACY OF ALGORITHMS

Model	Accuracy	SD
Generalized Linear Model	95.59%	+/- 2.22%
Logistic Regression	99.05%	+/- 0.08%
Fast Large Margin	53.44%	+/- 0.35%
Decision Tree	98.49%	+/- 0.12%
Random Forest	89.00%	+/- 0.33%
Support Vector Machine	38.23%	+/- 1.20%

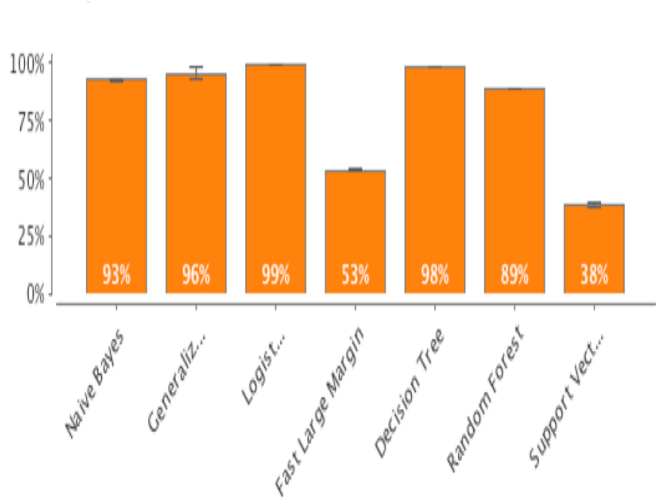

Figure 4 Models' accuracy

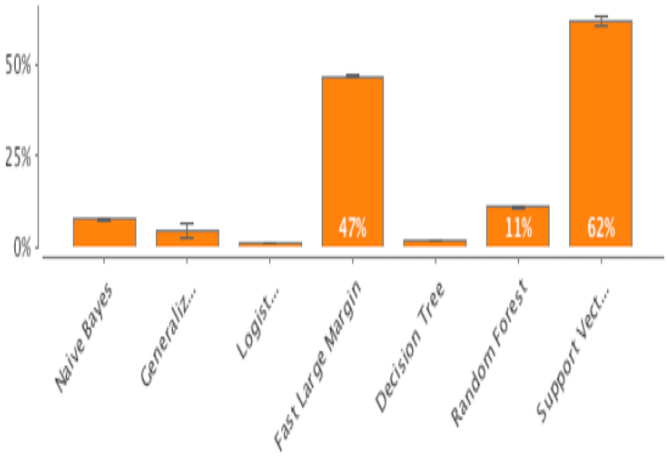

Figure 5 Classification error for tested models

In this paper, we used the KDD process to find MCCs patterns in the US elderly patients and predicted the type of MCCs in this population. The top listed diseases in the clusters were neoplasms in cluster_0, cluster_1, Cluster_2, Cluster_4, Cluster_6, and Cluster_7 while Echinococcosis was the first listed disease in cluster_3 followed by malignant tumors.

Moreover, a significant relationship between the length of stay (LOS) and the number of chronic diseases (NCHRONIC) were found along with the case that the presence of MCCs was affected by age. Figure 6 shows that the NCHRONIC and LOS relation. We also saw a significant association between LOS, gender, and in-hospital death to the MCCs clusters. Table 4 shows that the top MCCs related variables resulted from prediction models. Additionally, The results suggest that age and LOS are the top critical variables for predicting the total cost per hospitalization stay, while in-hospital death, and the NCHRONIC, are also related to the LOS.

Furthermore, the first listed diagnosis in patients records (DXCCS1) was a critical factor in distinguishing MCCs clusters. Figure 7 shows that the resulted clusters were affected by DXCCS1 which also determines the length of stay. This finding contributes to the current research about a relationship between multimorbidity pattern and the existence of specific diseases that increased the risk of developing further chronic illness.

TABLE 4 LIST OF TOP VARIABLES RELATED WITH READMISSION PREDICTION

Variables	P value
Gender	1.00
DIED	0.81
NCHRONIC	0.47
LOS	0.15
AGE	0.13

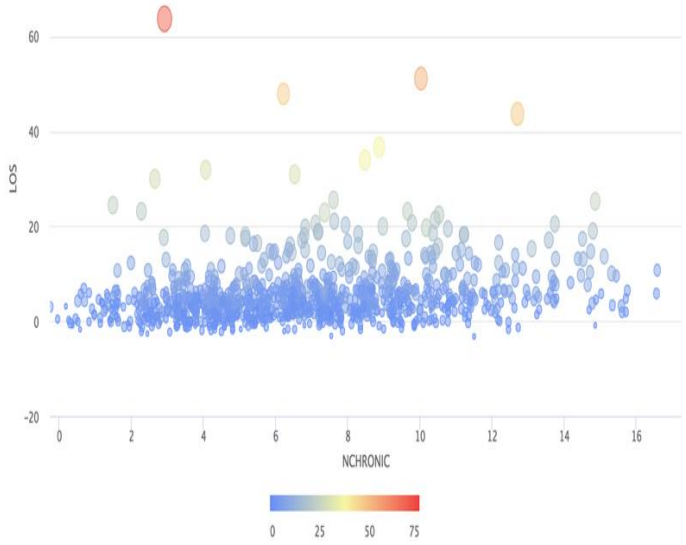

Figure 6 Number of chronic conditions (NCHRONIC) vs Length of stay (LOS)

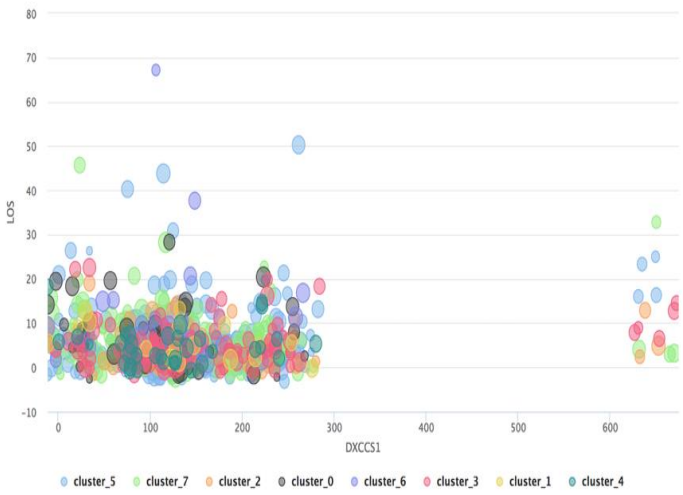

Figure 7 Resulted clusters were affected by the first listed disease (DXCCS1) and length of stay.

IV. CONCLUSION

Identifying which factors are significant to disease development and the associated risks is essential to serve older patients better. This research could be the start of a more substantial focus on multimorbidity modeling and a better way of using data mining for such tasks. Our results can be of significance to the comprehension of the distinct nature of MCCs by presenting the background of describing it in different ways. In future research, we plan to study the impact of varying multimorbidity clusters on various related health variables and outcomes. Moreover, future inquiries about MCCs might guide more personalized treatment plans. Particularly, due to a massive gap in the knowledge about the healthcare needs of MCCs patients and their disease patterns, there are several research questions can be targeted to help to enhance the health of elderly persons.

REFERENCES

[1] a. Abbasi, L. M. Peelen, E. Corpeleijn, Y. T. van der Schouw, R. P. Stolk, a. M. W. Spijkerman, D. L. van der A, K. G. M. Moons, G. Navis, S. J. L. Bakker, and J. W. J. Beulens, "Prediction models for risk of developing type 2 diabetes: systematic literature search and independent external validation study," Bmj, vol. 345, no. sep18 2, pp. e5900–e5900, 2012.

[2] H.Koh and G.Tan. "Data Mining Applications in Healthcare." Journal of Healthcare Information Management, Vol. 19, No. 2. pp. 64–72, 2005.

[3] Fayyad, Piatetsky-Shapiro, Smyth, "From Data Mining to Knowledge Discovery: An Overview", in Fayyad, Piatetsky-Shapiro, Smyth, Uthurusamy, Advances in Knowledge Discovery and Data Mining, AAAI Press / The MIT Press, Menlo Park, CA, 1996, pp.1-34

[4] Healthcare Cost and Utilization Project (HCUP). August 2017. Agency for Healthcare Research and Quality, Rockville, MD. www.hcup-us.ahrq.gov/ Accessed 2017. [5] E. E. Nolte and M. McKee, "Caring for people with chronic conditions : a health system perspective," Eur. Obs. Heal. Care Syst. Ser., p. XXI, 259 p., 2008.

[6] N. Garin, B. Olaya, J. Perales, M. V. Moneta, M. Miret, J. L. Ayuso-Mateos, and J. M. Haro, "Multimorbidity patterns in a national representative sample of the Spanish adult population.," PLoS One, vol. 9, no. 1, p. e84794, 2014.

[7] C. for M. and M. Services, "Chronic conditions among Medicare beneficiaries, chart book," Balt. MD, 2012.

[8] Centers for Disease Control and Prevention, National Center for Health Statistics. http:/www.cdc.gov/ . Accessed in 2017.

[9] M. E. Salive, "Multimorbidity in older adults," Epidemiol. Rev., vol. 35, no. 1, pp. 75–83, 2013.

[10] P. D. S. John, S. L. Tyas, V. Menec, and R. Tate, "Multimorbidity , disability , and mortality in community-dwelling older adults Recherche Multi-morbidité , incapacité et mortalité chez les personnes âgées vivant dans la communauté," vol. 60, pp. 272–280, 2014.

[11] Healthy People 2020. http://www.healthypeople.gov/2020/ . Accessed in 2018.

[12] K. R. Ananthapadmanaban and G. Parthiban, "Prediction of Chances - Diabetic Retinopathy using Data Mining Classification Techniques," Indian J. Sci. Technol., vol. 7, no. October, pp. 1498–1503, 2014.

[13] C. Y. Chin, M. Y. Weng, T. C. Lin, S. Y. Cheng, Y. H. K. Yang, and V. S. Tseng, "Mining Disease Risk Patterns from Nationwide Clinical Databases for the Assessment of Early Rheumatoid Arthritis Risk," PLoS One, vol. 10, no. 4, p. e0122508, 2015.

[14] P. R. Hachesu, M. Ahmadi, S. Alizadeh, and F. Sadoughi, "Use of data mining techniques to determine and predict length of stay of cardiac patients," Healthc. Inform. Res., vol. 19, no. 2, pp. 121–129, 2013.

[15] A. B. Holmes, A. Hawson, F. Liu, C. Friedman, H. Khiabanian, and R. Rabadan, "Discovering disease associations by integrating electronic clinical data and medical literature," PLoS One, vol. 6, no. 6, 2011.

[16] M. Jakovljević and L. Ostojić, "Comorbidity and multimorbidity in medicine today: challenges and opportunities for bringing separated branches of medicine closer to each other.," Psychiatr. Danub., vol. 25 Suppl 1, no. 1, pp. 18–28, 2013.

[17] M. Khalilia, S. Chakraborty, and M. Popescu, "Predicting disease risks from highly imbalanced data using random forest.," BMC Med. Inform. Decis. Mak., vol. 11, no. 1, p. 51, 2011.

[18] I. Kirchberger, C. Meisinger, M. Heier, A. K. Zimmermann, B. Thorand, C. S. Autenrieth, A. Peters, K. H. Ladwig, and A. Döring, "Patterns of multimorbidity in the aged population. results from the KORA-Age study," PLoS One, vol. 7, no. 1, pp. 1–8, 2012.

[19] A. Marengoni, S. Angleman, R. Melis, F. Mangialasche, A. Karp, A. Garmen, B. Meinow, and L. Fratiglioni, "Aging with multimorbidity: A systematic review of the literature," Ageing Res. Rev., vol. 10, no. 4, pp. 430–439, 2011.

[20] M. E. Munson, J. S. Wrobel, C. M. Holmes, and D. a. Hanauer, "Data mining for identifying novel associations and temporal relationships with charcot foot," J. Diabetes Res., vol. 2014, 2014.

[21] A. Prados-Torres, B. Poblador-Plou, A. Calderón-Larrañaga, L. A. Gimeno-Feliu, F. González-Rubio, A. Poncel-Falcó, A. Sicras-Mainar, and J. T. Alcalá-Nalvaiz, "Multimorbidity patterns in primary care: Interactions among chronic diseases using factor analysis," PLoS One, vol. 7, no. 2, 2012.

[22] G. Purusothaman and P. Krishnakumari, "A Survey of Data Mining Techniques on Risk Prediction : Heart Disease," ndian J. Sci. Technol., vol. 8, no. June, 2015.

[23] I. Schäfer, E. C. von Leitner, G. Schön, D. Koller, H. Hansen, T. Kolonko, H. Kaduszkiewicz, K. Wegscheider, G. Glaeske, and H. van den Bussche, "Multimorbidity patterns in the elderly: A new approach of disease clustering identifies complex interrelations between chronic conditions," PLoS One, vol. 5, no. 12, 2010.

[24] C. for M. and M. Services, "Chronic conditions among Medicare beneficiaries, chart book," Balt. MD, 2012.

[25] R. Kost, B. Littenberg, ES. Chen, "Exploring generalized association rule mining for disease co-occurrences" AMIA Annu Symp Proc. 2012; 2012:1284-93. Epub 2012.