ISSN 2278-3091

Volume 9, No.4, July – August 2020

International Journal of Advanced Trends in Computer Science and Engineering

Available Online at http://www.warse.org/IJATCSE/static/pdf/file/ijatcse353942020.pdf

https://doi.org/10.30534/ijatcse/2020/353942020

Predicting the Number of Multiple Chronic Conditions in Arizona State Using Data Mining Algorithms

Najah Al-Shanableh[1], Mohammed Salem Atoum[2]

[1]Department of Computer Science, Al Albayt University, Jordan, Najah2746@aabu.edu.jo

[2]Department of Computer Science, Irbid National University, Jordan, cs@inu.edu.jo

ABSTRACT

Multiple chronic diseases are prevalent within the older adult population. The varying numbers and patterns of co-morbidity create a challenge for healthcare providers; the coordination of treatment and care is an essential part of an effective plan. Knowing more about the factors that affect multiple chronic conditions (MCC) can help the authorities with enforcing such policies. Nonetheless, more information about border areas and the distribution of MCC in its population can help too. The US-Mexico border has its natural and ethnic distribution, which makes using the general plan a failure. This research comes to investigate both MCC inside the US-Mexico border. We used the data mining process to predict the number of chronic conditions in Arizona State as part of the US. Mexico border use the inpatient Data set. Our preliminary analysis showed that the median number of chronic diseases in Arizona is 5 for older adults; while the amount reported for the whole USA is 2. Besides, the first diagnosis identified upon hospital admission was a chronic mental health condition. For this study, 5 models have been tested on a data sample, each represented by an individual process in the Rapidminer data mining tool. The top algorithms that gave the best results decision tree algorithm with an accuracy of (100%).

Key words: Data Mining, Informatics, Multiple Chronic Conditions, , U.S.-Mexico Border Area.

1. INTRODUCTION

Chronic diseases are the biggest problem of healthcare in America, as 7 out of 10 people in America die due to chronic diseases, according to the Center for Disease Control (CDC) statistic [8]. When the patient has two or more chronic conditions, this disease is called multiple chronic conditions (MCC) [9]. It is imperative to intervene in order to improve the lives of those with more than two chronic conditions among the elderly because most of them suffer from more than two chronic conditions according to the CDC statistic [7]. The complexity of having a single form for MCC medicament is also making this progression more difficult. The process of finding the most common patterns of multimorbidity and predicting disease risk of MCC will help to develop centers for healthcare providers to create plans and laws that enhance patients' quality of life and reduce the risk of developing more chronic conditions.

The United States - Mexico border has a unique and private nature that differs from the whole U.S. The different population distribution has affected the health disparities in this area [13]. Several states from the U.S. (California, Arizona, New Mexico, and Texas) are part of the border, which consists of a large population. Figure 1 shows how large is the border area. This large population has its gender, age, income, and economic variations [13]. These circumstances make applying and enforcing general health policies hard to maintain. Patterns of comorbidity are also part of the health disparities in this area since it is affected by many factors such as gender, age group, and race [10]. According to the United States-Mexico Border Health Commission [2], this area is also recognized for less income and lower education level [14-16].

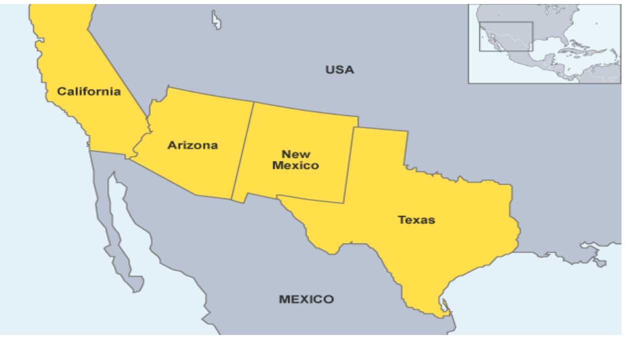

Figure 1: US-Mexico Borders [2]

The most recent research done on multimorbidity focuses on prevalence, risk ratio, prediction model for individual chronic disease with the presence of another disease, most common diseases, and the most frequent pairwise patterns of disease co-occurrence [6,7]. Less is done about the multimorbidity prediction model as a whole entity and patterns that have more than 3 diseases. One research has reported multimorbidity patterns in its broader aspect but it depends mainly on patient self-reported disease.

The predominant part of the conducted research has targeted a specific disease or the co-occurrence of a small fraction of illnesses, such as cardiovascular diseases, diabetes, and cancer, instead of the whole extent of chronic morbidity present in older adults [3]. A small number of studies has investigated multimorbidity distribution or co-occurrences in the same individual. Most of them used different approaches to address this issue. To sum up, The multimorbidity studies have focused on five major perspectives:

1. Patterns of multimorbidity: Most of the research focused only on either pairwise disease associations or three disease co-occurrence.
2. Prevalence of multimorbidity in different populations
3. The effect of multimorbidity in elderly life span and functionality.
4. Recommendations to conduct more research on multiple chronic conditions.
5. Risk factor of multimorbidity.

Methods used to find multimorbidity in the literature varied from using simple statistical analysis to data mining algorithms [9]. A combination of these methods has also been used. Most of these studies have targeted only one illness or two occurrences of diseases related to each other [4]. Most of the MCC's previous research targeted risk prediction for known chronic diseases like cardiovascular and breast cancer [11]. The previous studies have a huge gap about multimorbidity prediction models, focusing on diseases among the U.S. Mexico border area.

2. METHODS

This study is an observational retrospective cohort study.

2.1 Data Source

The data for this study were retrieved from the Healthcare Cost and Utilization Project (HCUP). HCUP data consist of hospital discharge data files. These discharge files include all patient care, ambulatory care, and emergency visits information. The State Inpatient Databases (SID) for Arizona State (AZ) was mainly used in this research. It has all the needed information about the patients' health status, income, and county they live in. This information helped in identifying MCC as well as which patients live in border counties.

2.2 Data mining

Data mining is mainly used as part of the Knowledge and Discovery Process (KDD) to analyze the data. KDD process consists of the following stages as figure 2 shows [5]:

1. Data were retrieved from patients' discharge records according to age from AZ SID files.
2. Data-preprocessing consists of data cleaning, integration, selection, and transformation steps [4]. Data from different sources come in different formats and attributes may be named differently. As part of this step, the researchers transformed the data into a proper format, select the data that apply to our research (people 65 years or older, etc.), and handle inconsistent and missing data.
3. The data mining stage consists of selecting a specific task, algorithms, and applying them to the data. The prediction task was used here to predict the number of chronic diseases based on available variables. Prediction algorithms in data mining also give weight to each used variable in the model. For predicting the number of chronic conditions, there are several models to choose from in data mining. Mainly there are two main categories of prediction in data mining.
4. The evaluation step includes a comparison of obtained cluster and prediction models, selecting the best models, and visualization of the results. Accuracy and Root Mean Squared Error measures were used to evaluate the selected model.

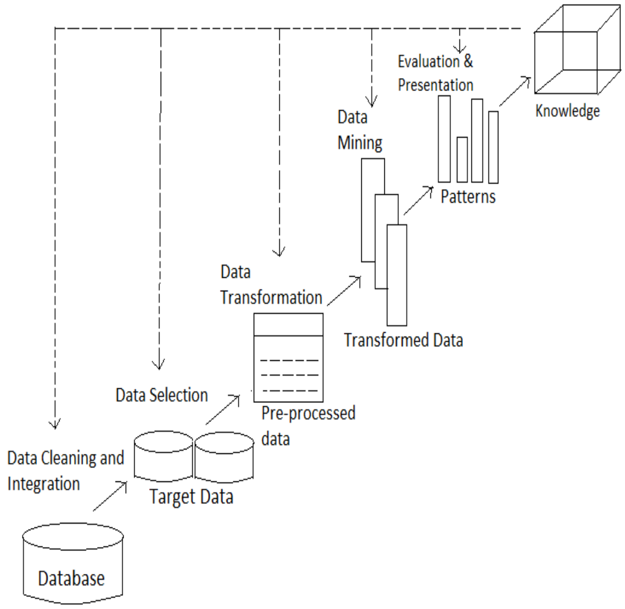

Figure 2: Knowledge and Discovery Process Steps [5].

3. RESULTS AND DISCUSSION

Records for people aged 65 or more and lived in AZ are enclosed in this research. Also, MCC was identified as the existence of two or more chronic conditions in patients' records. The total number of records in SID who were included in this study was 246,450. Table 1 encapsulates the main attributes of the included population with regard to being a resident in a border county, income, gender, and race, as well as the length of stay.

Race distribution among the included records is presented in figure 3. One noticeable difference between the border and non-border counties was a higher percentage of Hispanics in border counties. This is a well-known property for the border area [2]. This, in fact, affects disease distribution since most chronic illnesses are affected by race [12].

Disease count is used in this research as an indicator of existing MCC in a patient file. Disease count was defined as a simple count of chronic diseases in the patients' records and it is a standard measure of multimorbidity [1]. Figure 4 shows the distribution of the total number of chronic conditions (disease counts) in border counties vs. non-border counties. Also, figure 5 shows the distribution of MCC per age group in all patients.

Table 1: Population Characteristics

Characteristics	**Non-Border counties N=191797 (0.78%)**	**Border counties n= 54653 (0.22%)**	***p*-value**
Gender			-.064**
Male	45.97%	45.05%	
Female	54.03%	54.95%	
Race			-.005**
Hispanic	8.09%	15.03%	
Age Group			-.150**
65-74	42.72%	40.71%	
75-84	37.50%	38.28%	
85-94	17.63%	18.67%	
95-104	2.14%	2.33%	
>105	0.00%	0.01%	
Median household income			-.007**
1	21.57%	33.72%	
2	32.79%	35.28%	
3	25.32%	12.06%	
4	20.32%	18.95%	
Length of Stay (Average)	5	5	.233**

**. Correlation is significant at the 0.01 level (2-tailed).

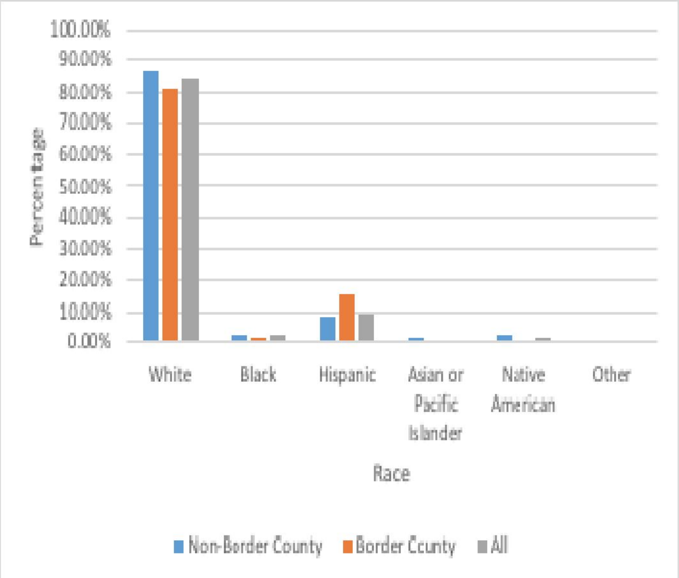

Figure 3: Race distribution

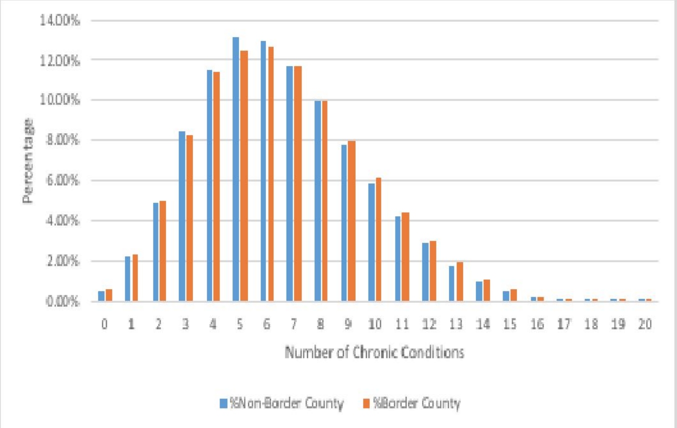

Figure 4: Number of chronic conditions Distribution.

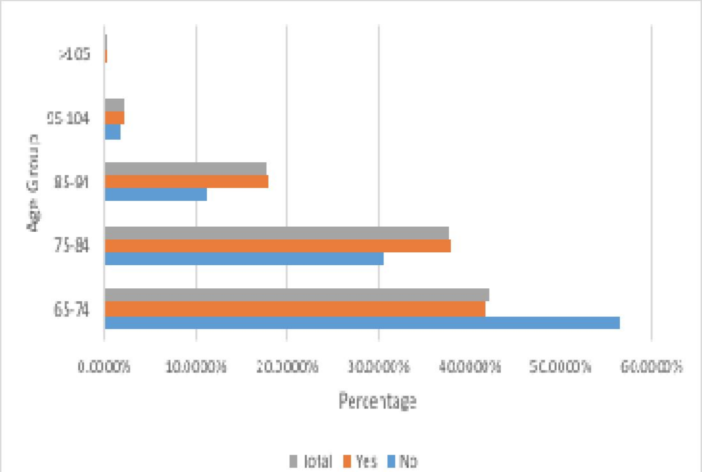

Figure 5: MCC per Age Group (yes=have MCC, No= No MCC)

In this study, six data mining prediction algorithms have been tested on the data. These algorithms were Decision Tree, Gradient Boosted Trees, Random Forest, Deep Learning, Support Vector Machine, and Generalized Linear Model. The selection of these algorithms was based on the most used algorithms in the literature. Every algorithm testing run was represented as a single process in the RapidMiner tool.

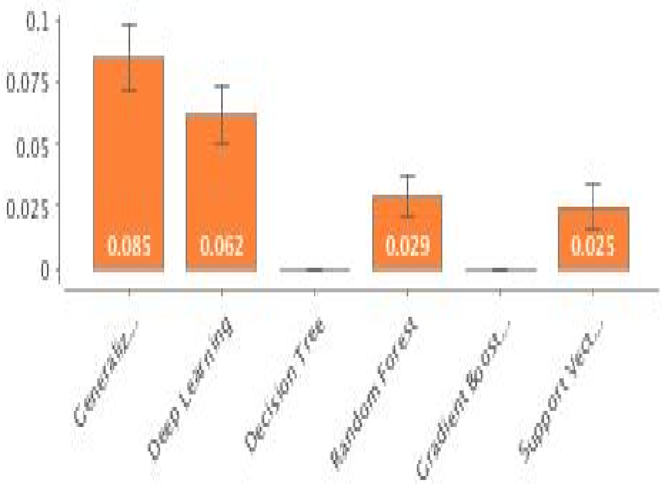

Figure 6: Absolute Error for Tested Algorithms.

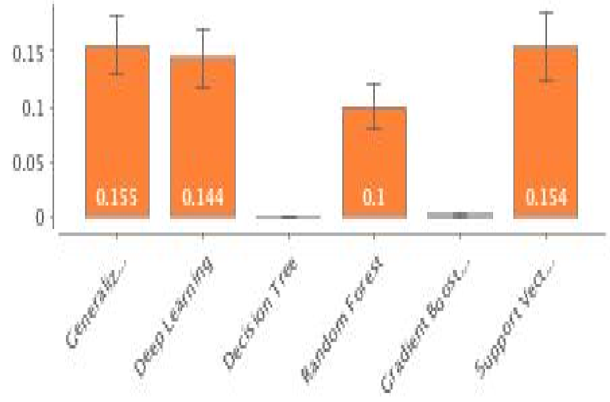

Figure 7: Root Mean Squared Error for Tested Algorithms.

Table 2 illustrates the resulting accuracy of the tested models along with the Root Mean Squared Error and Standard Deviation. Table 2 shows that the top two algorithms were decision tree and Gradient boosted tree. The minimum achieved. Root mean squared error and Absolute Error among the tested algorithms were recorded by the Decision Tree, and the Gradient Boosted Trees, as shown in figures 6 and 7.

Table 2: Data Mining Algorithms Accuracy

Model	Accuracy	Root Mean Squared Error	Standard Deviation
Decision Tree	100.00%	0.000	0.000
Gradient Boosted Trees	99.97%	0.000	0.000
Random Forest	80.55%	0.100	0.021
Deep Learning	71.96%	0.144	0.027
Support Vector Machine	69.89%	0.154	0.030
Generalized Linear Model	69.72%	0.155	0.027

Based on the top accurate algorithms, the data mining model also identified related attributes based and their weight. According to this result, shown in table 3 below, the significant findings were that gender, race, and median household income differs among the hospitalized population in relation to the number of chronic diseases, especially in border counties. The correlation between sex and the number of chronic conditions was significant in addition to the relationship between the length of stay and the number of chronic diseases. Age and alcoholism were expected to affect the MCC along with income. But the surprising result that marital status also affects patients' number of chronic conditions besides being a border county's resident.

These findings, especially the unexpected ones, stress on carrying out more studies targeting border counties. Health care investigators should consider several conditions related to being in the U.S.-Mexico border area in their future studies besides regular researched factors.

Table 3: Identified Related Attributes Based on the Data Mining Algorithms and their Weight

Attribute	**Weight**
Number of Diseases in Record	0.247
Length of Stay	0.074
Obesity	0.050
Total charge	0.044
Race	0.041
Gender	0.041
Age Group	0.039
Marital Status	0.038
Alcoholic	0.030
Border County	0.027
In Hospital Death	0.026
Median Household Income	0.026

4. CONCLUSION

Multimorbidity or MCC has a significant impact on the length of hospitalization as well as different healthcare outcomes. It also influences the use of health services as well as the cost of health insurance. In this research, the researchers predicted the count of chronic diseases taking into account the border residency. According to the study findings, gender, race, and living in a border county affected patients and the risks of developing MCC. Further researches are highly recommended for more investigations of multimorbidity among people who live in border areas. Data mining revealed that there are different variables besides gender and race, like marital status and living in specific areas to consider when we discuss about MCC. Data mining models can be tailored to model health disparities in such areas.

REFERENCES

1. E. a. Bayliss, D. E. Bonds, C. M. Boyd, M. M. Davis, B. Finke, M. H. Fox, R. E. Glasgow, R. a. Goodman, S. Heurtin-Roberts, S. Lachenmayr, C. Lind, E. a. Madigan, D. S. Meyers, S. Mintz, W. J. Nilsen, S. Okun, S. Ruiz, M. E. Salive, and K. C. Stange, **Understanding the context of health for persons with multiple chronic conditions: Moving from what is the matter to what matters**, *Ann. Fam. Med.*, vol. 12, no. 3, pp. 260–269, 2014.
https://doi.org/10.1370/afm.1643
2. K. E. Bliss, **The Challenge of Chronic Diseases on the U.S. -Mexico Border**, no. June, 2010.
3. K. Christensen, G. Doblhammer, R. Rau, and J. W. Vaupel, **Ageing populations: the challenges ahead**, *Lancet*, vol. 374, no. 9696, pp. 1196–1208, 2009.

4. H.Koh and G.Tan. **Data Mining Applications in Healthcare**, *Journal of Healthcare Information Management*, Vol. 19, No. 2.
5. 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
6. S. M. Dy, E. R. Pfoh, M. E. Salive, and C. M. Boyd, **Health-related quality of life and functional status quality indicators for older persons with multiple chronic conditions**, *Journal of the American Geriatrics Society*, vol. 61, no. 12. pp. 2120–2127, 2013. https://doi.org/10.1111/jgs.12555
7. D. a. Hanauer and N. Ramakrishnan, **Modeling temporal relationships in large scale clinical associations**, *J. Am. Med. Informatics Assoc.*, pp. 332–341, 2012.
8. P. Healthdesign, **Tracking and Sharing Observations from Daily Life Could Transform Chronic Care Management Project HealthDesign selects five teams to test use of personal health applications to capture and integrate patient-recorded data into clinical**, pp. 30–32, 2010.
9. D. J. Kim, A. O. Westfall, E. Chamot, A. L. Willig, M. J. Mugavero, C. Ritchie, G. a Burkholder, H. M. Crane, J. L. Raper, M. S. Saag, and J. H. Willig, **Multimorbidity patterns in HIV-infected patients: the role of obesity in chronic disease clustering**., *J. Acquir. Immune Defic. Syndr.*, vol. 61, no. 5, pp. 600–5, 2012. https://doi.org/10.1097/QAI.0b013e31827303d5
10. 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.
11. Najah Al-shanableh, Mofleh Al Diabat, **Multimorbidity Prediction Using Data Mining Model**, *The World of Computer Science and Information Technology Journal* (WSCIT). 2019 Volume 9, Issue 2, pp.7.11.
12. M.Mehdi .**The Significance of the Race Factors in Breast Cancer Prognosis**,*Int'l Conf. Data Mining DMIN'14,* 2014.
13. National Rural Health Association. (2010). **Addressing the Health Care Needs in the U .S . -Mexico Border Region: Policy Brief**. pp. 2–5, 2000.
14. Munya, A. A., &Sangita, B. (2019). **Survey of Machine Learning Techniques in Medical Imaging**. *International Journal of Advanced Trends in Computer Science and Engineering*, 8(5), 2107-2116. https://doi.org/10.30534/ijatcse/2019/39852019
15. Mauritsius, T, Braza, A, &Fransisca (2019). **Bank Marketing Data Mining using CRISP-DM Approach**. *International Journal Of Advanced Trends In Computer Science And Engineering, 8(5), 2322-2329.* https://doi.org/10.30534/ijatcse/2019/71852019
16. Christ, Mogi &Rahmanto, Nikolaus. (2019). **Lending Club Default Prediction using Naïve Bayes and Decision Tree.** *International Journal of Advanced Trends in Computer Science and Engineering.* 8, 2528 – 2534. https://doi.org/10.30534/ijatcse/2019/99852019