International Journal of Statistics and Probability; Vol. 7, No. 6; November 2018
ISSN 1927-7032 E-ISSN 1927-7040
Published by Canadian Center of Science and Education

Unsupervised Machine Learning for Co/Multimorbidity Analysis

Shatrunjai P. Singh[1], Swagata Karkare[2], Sudhir M. Baswan[3] & Vijendra P. Singh[4]

[1] Lindner College of Business, University of Cincinnati, Ohio, USA

[2] School of Public Health, Boston University, Boston, MA, USA

[3] Independent Researcher, Grand Rapids, MI, USA

[4] Department of Internal Medicine, Baptist Easley Hospital, Easley, SC, USA

Correspondence: Shatrunjai P. Singh, Lindner College of Business, University of Cincinnati, Ohio, USA

Received: July 12, 2018 Accepted: August 21, 2018 Online Published: September 12, 2018

doi:10.5539/ijsp.v7n6p23 URL: https://doi.org/10.5539/ijsp.v7n6p23

Abstract

Although co/multimorbidities are associated with a significant increase in mortality, lack of quantitative exploratory techniques often impedes an in-depth analysis of their association. In the current study, we explore the clustering of co/multimorbid patients in the Texas patient population. We employ unsupervised agglomerative hierarchical clustering to find clusters of co/multimorbid patients within this population. Our analysis revealed the presence of nine distinct, clinically relevant clusters of co/multimorbidities within the study population of interest. This technique provides a quantitative exploratory analysis of the co/multimorbidities present in a specific population.

1. Introduction

One in four Americans have two (comorbid) or more (multimorbid) chronic conditions (Hoffman, Rice, & Sung, 1996). Projections estimate that more than 81 million Americans will suffer from these co/multimorbidities by 2020 (Anderson, 2003). Further, new data suggests that when routine clinical procedures are applied to patients with a co/multimorbidity, it can lead to unintended adverse events if healthcare professionals are unaware of the patient's history (Fried et al., 2014). Thus, there is a critical need to identify these patients with complex co/multimorbidities to enable proper support and intervention. Previous research to classify subgroups of composite patients with co/multimorbidities have conventionally depended on complex multivariate regression techniques (Cheng, Dy, Fang, Chen, & Chiu, 2013; Ilesanmi & Fatiregun, 2014). Although, newer supervised and unsupervised machine learning algorithms have been successfully adopted in many other spheres of biomedical data analysis, they have not been applied for the characterization of co/multimorbidities in patient data (Clifton, Niehaus, Charlton, & Colopy, 2015; Deo, 2015).

Clustering is an unsupervised machine learning technique that aims to group analogous entities into one cluster and partitions dissimilar objects into another cluster (Becker, 2005; Hofstetter, Dusseldorp, van Empelen, & Paulussen, 2014; Whiteman & Whiteman, 1949). A cluster is defined as a subset of similar objects, defined by certain parameters, within a larger set. The threshold definition of similarity cutoff is often subjective and is usually determined by the study design. With respect to medical conditions, a co/multimorbid clustering can be defined as an unsupervised technique to find patients with similar medical conditions. In the current study, we use correlational clustering analysis to find the key groups of diseases present in the population. We further employ hierarchical clustering analysis on patients from the Texas health care patient data to describe inherent patterns in clusters of multimorbid patients. The clustering approach identifies cohorts of co/multimorbidities and presents opportunities for better management of these patients.

2. Methods

2.1 Study Population

We used open access, de-identified aggregate data provided by the Texas Department of State Health Services (http://healthdata.dshs.texas.gov/Home) to conduct this analysis. Inpatient and Outpatient datasets were combined to generate a composite dataset consisting of more than 15,000 data points and the inpatient procedure code was used to identify different clinical conditions. The training cohort consisted of patients who were 21 years or older as of January 1, 2015, with two (comorbidities) or more (multimorbidities) identified by inpatient procedure code on first examination. Members with admits to hospice, a long-term care facility, or with a pregnancy reported in the last 3 months were excluded from the study. After exclusions, our final study population was 13,920 patients. We isolated the list of 75 most common conditions reported by the Center for Disease Control, USA and used it to further filter out input dataset ("CDC

- NCHS - National Center for Health Statistics," 2018). A literature search was also performed to further identify conditions which could be included in the study based on the general Texas population, our specific study cohort and disease with relatively high prevalence. Identification of conditions within cohort members were based on an outpatient data cross-referenced to the *International Classification of Diseases, Tenth Revision (ICD-10)* diagnosis and procedure codes in 2015 ("WHO | International Classification of Diseases, 11th Revision (ICD-11))

2.2 Exploratory Data Analysis and Feature Generation

Microsoft SQL Server (version 2012) was used to extract, transform, load and query the dataset. Binary outcome variables were created for the selected conditions. Age, gender, income and other demographic variables were also included in the input dataset. Input variables were scanned for outliers. Imputations to median/mode were performed for the non-binary continuous/categorical variables. We observed less than 2% imputations overall in the dataset. A non-zero variance analysis was performed on the binary disease variables and variables with less than 2% variance were further excluded from the analysis.

2.3 Statistical Analysis

All statistical analysis was performed using R-statistical software (Version 0.98.109). The R packages 'Cluster' (v 2.0.7.1) and '`Dendextend' (1.8.0)` was employed for this analysis. Significance testing for normal, non-normal and binary data was performed as described previously by the authors in other studies (Fleming et al., 2018; Hester et al., 2016; Karkare et al., 2014; Singh, 2015, 2016; Singh et al., 2016; Singh, He, McNamara, & Danzer, 2013; Singh, Karkare, Baswan, & Singh, 2018; Singh, LaSarge, An, McAuliffe, & Danzer, 2015; Singh, Singh, Fatima, Kubo, & Singh, 2008; Singh & Karkare, 2017; Singh & Singh, 2017). De-identified data and the statistical analysis R code used was uploaded to an online repository.

2.4 Correlation Algorithm

Correlation clustering involves the creation of a weighted matrix X= (P, E), such that the edge weight specifies the similarity ($+^{ve}$ edge weight) or dissimilarity ($-^{ve}$ edge weight). The goal is to find an optimal cluster which maximizes similarity or minimizes dissimilarity (Becker, 2005). The method of minimizing disagreement was chosen for the current study based on the characteristics of the input data. A spearman's correlation matrix was generated for binary variables created from different input conditions. A k-means clustering algorithm based on Jaccard's distance was created and analyzed. The optimal cluster number was chosen to minimize the goodness of fit criterion. The cluster of conditions was analyzed for similarity of condition based on origin, organ system and patient demographic.

2.5 Clustering Algorithm

An agglomerative hierarchical clustering (AHC) algorithm with a bottom up approach was used to separate clinically appropriate clusters within the study population. The bottom up reproach to AHC initiates with each member starting at an isolated cluster, followed by serial merging of similar members to form similarity clusters until only once cluster remains. After the clustering procedure terminates, subject matter expertise, clinical relevance and study design criterion are used to select a cutoff/threshold which produces the final clusters. The process can be visualized using dendrogram. We used Ward's method along with Gower's distance matrix for similarity calculations as it has shown to be more reliable for mixed data with a preponderance of weighted binary data (like condition related binary variables) (Gower, 1971).

2.6 Figure Preparation

The results from R-software were exported into csv files, which were imported into Tableau (version 8.0) or Microsoft Excel (version 2013), which were then used to create graphs and visualizations. Tables were created in Microsoft Word (version 2013).

3. Results

3.1 Correlation Analysis Reveals Expected Cluster of Major Conditions in the Patient Population

A spearman's correlation analysis was performed on the dataset of more than 70 different conditions to check for the correlation between different conditions in the population. The resulting correlation matrix was further clustered to produce grouping of similar conditions with a correlation coefficient cutoff greater than 0.60 (Figure 1).

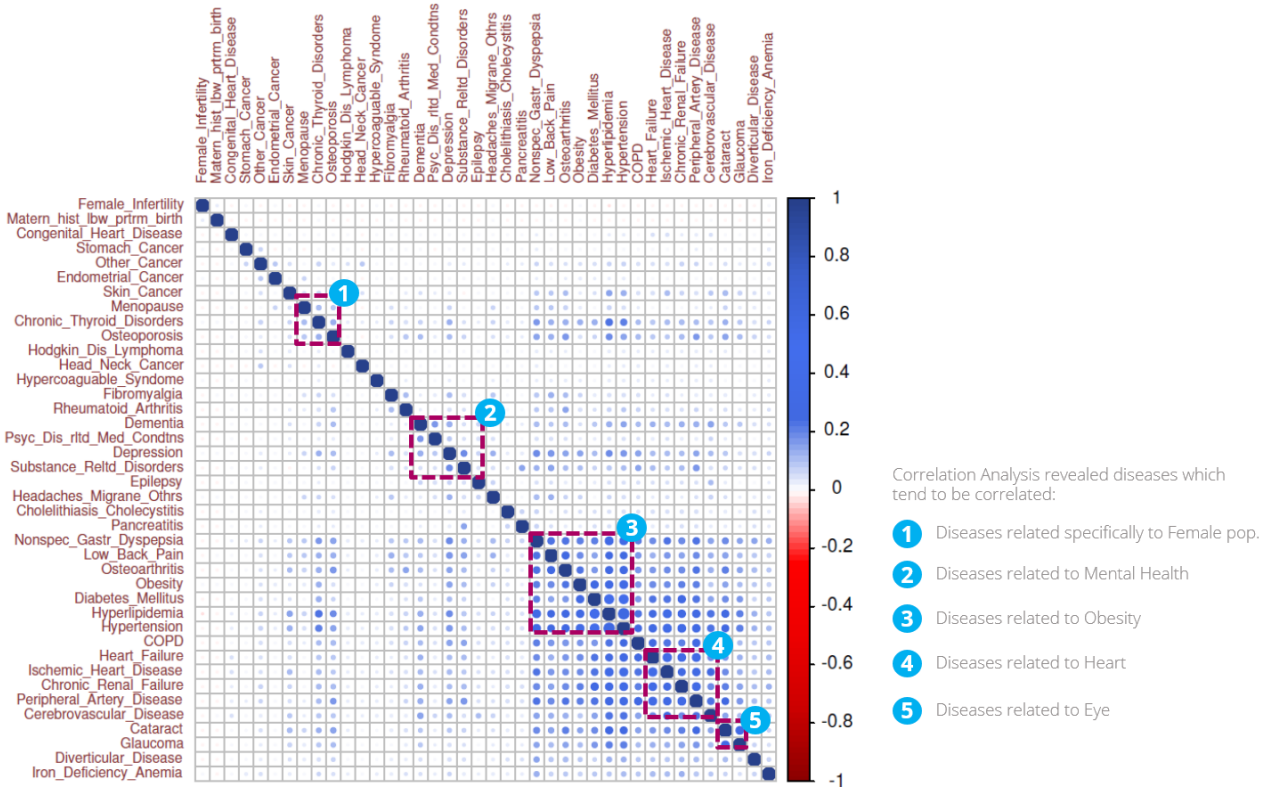

Figure 1. A matrix showing the results of spearman's correlation analysis followed by k-means clustering on the results. Conditions are represented on the top and left panel. The size of the circles depict the strength of correlation between diseases. An additional color coding of the spearman's correlation coefficient (red to blue signifies correlation from -1 to +1) was added to increase interpretability. Clustering analysis using the wards method reveals five clusters of clinically related conditions shown in the dotted red box.

We identified 5 broad cluster of multi-morbid conditions: (1) diseases overrepresented in the female population including menopause, Chronic Thyroid disorder and Osteoporosis; (2) Neurological and Psychiatric conditions including substance related psychiatric conditions, epilepsy, dementia and depression; (3) Disease related metabolic syndrome including hypertension, lower back pain, hyperlipidemia, obesity and diabetes mellitus; (4) conditions related to the cardiovascular system including heart failure, ischemic heart disease, peripheral artery disease and cerebrovascular disease; (5) conditions of the eye including cataract and glaucoma. The clusters identified were homogenous and overall had low demographic variance. A radial dendrogram was created to further visualize the similarity of conditions within the population (Figure 2).

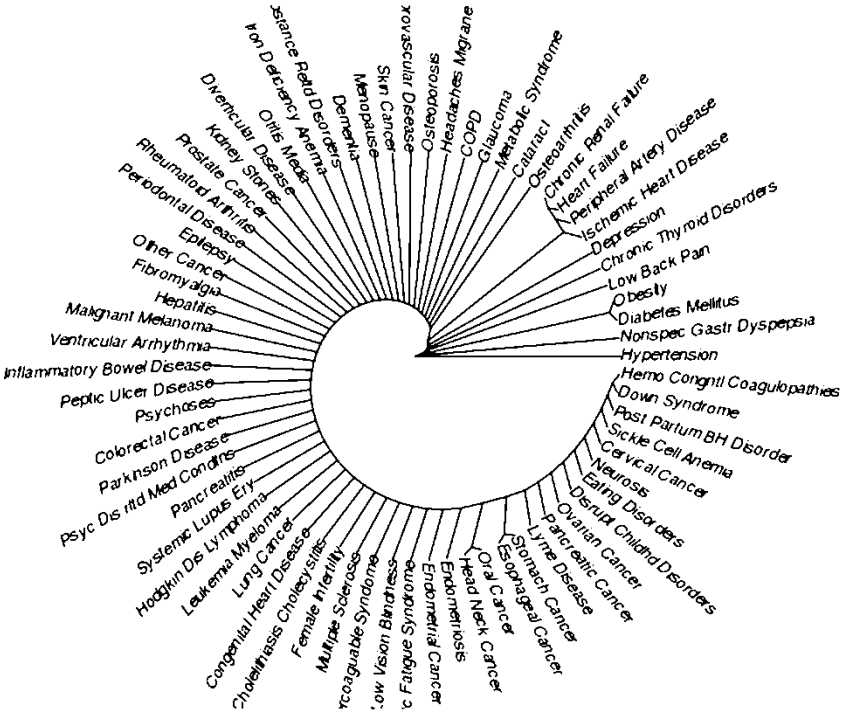

Figure 2. A closed radial dendrogram shows the structure of different conditions as inferred from the data. Conditions which frequently co-occur in patients share a common node (branch) in the dendrogram.

3.2 Clustering Analysis Reveals 9 Broad Cluster of Multi-morbidity Patients in the Population

Agglomerative Hierarchical Clustering revealed 9 broad clusters in the population data of 13,920 patients (Figure 3).

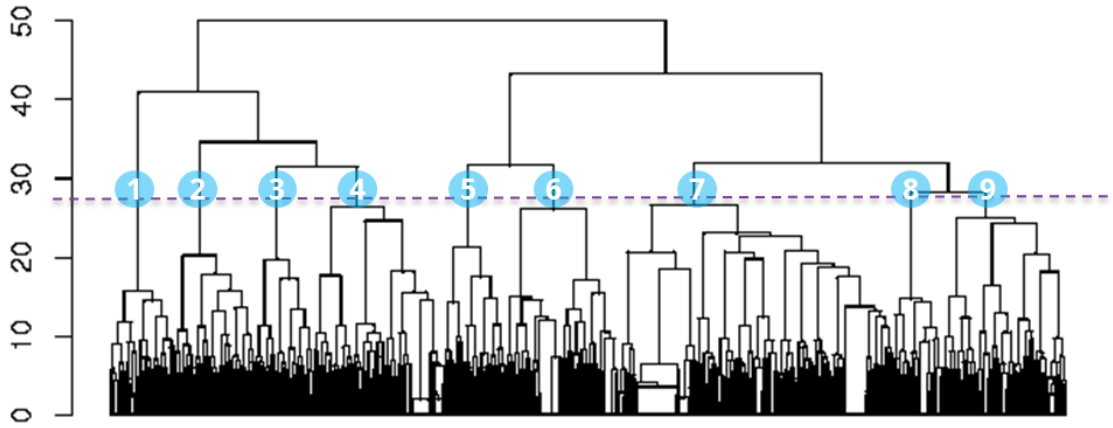

Figure 3. Results from hierarchical agglomerative clustering reveal nine distinct clusters. Clustering was performed using Ward's method with Gower's distance and a threshold (h=27; shown as a dotted line) was used to isolate 9 clusters.

The average age of patients was 54.9 years and contained 50.1 % males and 49.9 % females. Descriptive statistics revealed clinical homogeneity within the clusters (Table 1). The clusters (numbered randomly) were divergent based on the mix of co/multimorbidities observed and age/gender demographics.

Table 1. Summaries of disease distributions in different clusters. This table shows cluster summaries for age (median; years), male and female composition (%), and the proportion of people identified with different medical conditions (%; number of members in cluster with the disease/ total number of members with the disease), for the nine clusters.

CLUSTER NUMBER	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**
SIZE	4532	976	733	1564	895	765	1655	1101	1699
AGE	33.9	66.5	58.3	51	46.8	58	74.6	69	54.5
MALES	52.6	59.8	48.1	49.2	34.9	23.9	42.2	44.2	44.3
FEMALES	47.4	40.2	51.9	50.8	65.1	76.1	57.8	55.8	55.7
DIABETES MELLITUS	0.1	55.8	8.6	11.5	9.6	12.7	60.9	17.7	16.1
HEART FAILURE	0	17.5	0.7	0.3	1.9	1.8	41.8	5.5	3.7
COPD	0	6.8	1.8	1.2	3.2	1.1	35.9	23.4	2.8
ISCHEMIC HEART DISEASE	0	35.7	1.2	3.1	1.9	3.6	51.6	6.7	2.6
OTITIS MEDIA	2	0.6	0.7	0.9	0.3	1.6	1.8	0.7	0
PEPTIC ULCER DISEASE	0	0.5	0.3	0	0.5	0.2	3.1	0.6	2
HYPERTENSION	0.2	80.1	58.4	35	34.9	33.3	96.9	64	38.4
EPILEPSY	0	0.5	2	0	2.3	1.1	2.6	1	0.2
CHRONIC THYROID DISORDERS	0.2	13.9	7.6	7.4	8.4	99.1	29	15.6	23.8
DEPRESSION	0.1	6.8	4.5	8.4	64.1	15.2	30	11.3	10.5
CHOLELITHIASIS	0.2	0.6	0.3	0.6	0.3	0.4	1.6	0.6	0.2
CHOLECYSTITIS									
IRON DEFICIENCY ANEMIA	0.1	1.3	3.3	1.2	0.8	0.4	12.2	0.4	2.4
OSTEOARTHRITIS	0.1	5.4	3.2	0.6	3.1	9.8	37.6	49.6	5.5
RHEUMATOID ARTHRITIS	0.2	1	0.1	0.3	0.6	1.8	4	2.9	3.1
COLORECTAL CANCER	0.1	0.6	0.6	0.6	0.2	0.4	2.8	0.7	0.7
LUNG CANCER	0	0.1	0.3	0	0.2	0.4	1.2	1.1	0
HYPERLIPIDEMIA	3.7	74	43.1	29.1	34.3	44	89.1	61.6	42.4
CEREBROVASCULAR DISEASE	0	9.4	5.1	0.3	1.9	1.8	26.6	5.5	3.7
HEADACHES & MIGRAINES	0.3	1.2	3.6	4	39	1.3	8.2	1.9	5
CHRONIC RENAL FAILURE	0	19.7	2.5	0.3	1.3	2	47.5	6.6	1.7
KIDNEY STONES	0	2	3.9	1.5	0.5	0.4	3.1	2.2	1.3
DIVERTICULAR DISEASE	0	1.2	3.9	0	1.3	0.7	9.7	2.7	1.7
LOW BACK PAIN	0.1	9.8	6.5	99.7	14	6.7	49.4	25.4	20.5
NONSPECIFIC GASTRIC DYSPEPSIA	0.4	15.9	13.2	12.1	22.2	5.6	69.1	28.8	97.4
SICKLE CELL ANEMIA	0	0	0	0.9	0	0	0.1	0.1	0.4
MULTIPLE SCLEROSIS	0.2	0	0.3	0	0.3	0	0.1	0.1	0.2
INFLAMMATORY BOWEL DISEASE	0.3	0.7	0.6	0.3	0.2	0	1	0.4	0.7
HEMO-CONGENITAL COAGULOPATHIES	0	0	0.2	0	0	0	0.1	0.1	0.2
SYSTEMIC LUPUS	0	0.1	0.6	0.3	0.3	0	0.9	0.4	0.4
PROSTATE CANCER	0	3	3.7	0.9	0.5	0.9	7.2	2.6	0.4
OVARIAN CANCER	0	0.1	0	0	0.2	0.2	0.4	0.4	0.2
ENDOMETRIAL CANCER	0	0.1	0	0.3	0.2	0.2	0.7	0.6	0
CERVICAL CANCER	0.1	0	0.1	0	0	0.2	0	0.1	0
HODGKIN DIS LYMPHOMA	0.1	0.5	0.4	0.3	0	0	0.7	0.5	0.4
LEUKEMIA MYELOMA	0.1	0.4	0.3	0.3	0.2	0	1	0.2	0.4
MALIGNANT MELANOMA	0	0.5	1.9	0	0.2	0.2	1.5	0.4	1.3
HEAD NECK CANCER	0	0.1	0.3	0	0.2	0.2	0.3	0	0.4
ESOPHAGEAL CANCER	0	0.1	0.1	0	0	0	0	0.2	0.2
STOMACH CANCER	0	0	0	0	0	0	0.3	0.1	0.2
PANCREATIC CANCER	0	0.4	0	0	0.2	0	0.4	0	0
PANCREATITIS	0	0.6	0.1	0	1.3	0	2.4	0.5	0.4
HEPATITIS	0	0.1	1.8	0	1	0	1.5	0.9	1.1
PERIPHERAL ARTERY DISEASE	0	22.2	2.1	0	1.8	0.4	50	11.5	2.4
ENDOMETRIOSIS	0.1	0	0.1	0	0.2	0.2	0	0	0
VENTRICULAR ARRHYTHMIA	0	2.8	0.3	0.9	0.3	0	6.5	0.4	0.2
LYME DISEASE	0	0	0.1	0	0	0	0	0.1	0
FEMALE INFERTILITY	0.3	0	0	0.6	0.5	0.2	0	0	0.2
MENOPAUSE	0	0.8	10.4	2.8	1.9	3.3	4.9	4.5	2.2
GLAUCOMA	0.1	7.3	2.5	0.9	1.1	2	28.2	29.8	2.8

LOW VISION BLINDNESS	0	0.5	0.2	0.3	0.2	0	1.2	1.2	0.6
CATARACT	0.1	10.2	4	0.6	2.1	3.8	40.3	36.2	9.4
OTHER CANCER	0	1.3	1.1	0.3	0.6	3.1	4.6	1	2
DEMENTIA	0	2.4	3	0.3	0.8	0	12.8	2.1	0.4
OSTEOPOROSIS	0.1	4.3	7.7	4.6	0.6	0.9	17.2	8.2	6.1
OBESITY	0.2	23.4	20.3	8.4	15.7	16.5	52.1	13.6	5
ORAL CANCER	0	0	0	0	0	0	0.1	0	0
CYSTIC FIBROSIS	0	0	0.1	0	0	0	0	0	0
NEUROSIS	0	0	0.1	0.3	0.5	0	0.3	0	0
PSYCHOSES	0	0.2	0.1	0.3	3.4	0.7	2.1	0.1	0
EATING DISORDERS	0	0	0.1	0	0.3	0	0	0	0
DISRUPT CHILDHOOD DISORDERS	0.2	0	0.1	0	0	0	0	0	0
SUBSTANCE RELATED DISORDERS	0	1.4	0.5	0.3	9.1	0.7	12.5	3	0.6
SKIN CANCER	0.1	5.3	10	0.6	0.5	0.9	11.2	5	1.3
CONGENITAL HEART DISEASE	0.2	0.5	0	0	0	0	0.7	0.5	0
PERIODONTAL DISEASE	0	1	5	0.3	0.2	0	0.3	0.9	0.2
CHRONIC FATIGUE SYNDROME	0.1	0.2	0.2	0	0.3	0	0.1	0.4	0.6
FIBROMYALGIA	0	0.6	0.2	1.2	4.7	0.9	3.4	2.1	2.2
PARKINSON DISEASE	0.2	1	0.6	0.9	0.2	0.2	1.3	0.4	0.2
HYPERCOAGUABLE SYNDOME	0.1	0.2	0.2	0	0.3	0	0.6	0.2	0.2
POST-PARTUM BH DISORDER	0	0	0.1	0	0	0	0	0	0
MATERNAL LOW BIRTH WEIGHT	0	0	0	0	0	0	0	0	0
METABOLIC SYNDROME	0	3.5	12.8	8.4	8.3	7.4	6.6	5.9	3.1
PSYCHIATRIC DISORDER	0	0.4	1.6	0.3	0.5	0.2	2.1	0.6	0.2

Clinically relevant summarization showed the presence of distinct clusters with a high proportion of patients with (Table 2) : cancer (cluster 1), musculoskeletal diseases (cluster 2), substance abuse (cluster 3), female population with arthritis and post-menopausal conditions (cluster 4), metabolic syndrome related conditions (cluster 5), thyroid related conditions (cluster 6), females with migraines and depression (cluster 7), elderly population with multiple conditions (cluster 8) and a diabetes cohort (cluster 9). Income, medical utilization, inpatient visits were also calculated but are not shown in the current analysis.

Table 2. A summary of clinical findings from each cluster

Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5	Cluster 6	Cluster 7	Cluster 8	Cluster 9
Cancer	Musculoskeletal disorders	Substance Abuse	Menopause & Rheumatoid Arthritis	Obesity and Hypertension	Thyroid and Osteoporosis	Migraine and Depression	Elderly population with multiple conditions	Diabetes

4. Discussion

In the current paper, we use a clustering approach to identify groups of patients with similar co/multimorbid conditions in the Texas population. The clusters identified were homogenous and clinically relevant, and the real-world applications of our findings provide actionable insights for the fields of public health & healthcare provision. We present a fast and easy approach to explore patient data to better understand co/multimorbidities.

Co/multimorbidities are illnesses that coexist with a condition of interest and often lead to delayed treatment or misdiagnosis and have been shown to increase mortality in multiple populations (Song et al., 2018). They are a major source of economic burden on healthcare systems, with co/multimorbid patients experiencing worse health, economic and social outcomes compared to patients with singular health issues. Indeed, it has been shown that having multiple health conditions significantly increases the probability of reporting a diminished quality of life (Pisinger, Toft, Aadahl, Glümer, & Jørgensen, 2009; Song et al., 2018, 2018). Co/multimorbidity indices are frequently used to summarize the overall health of a population but often suffer from errors of manual data curation ("Comorbidity indices," n.d.; Sharabiani, Aylin, & Bottle, 2012). Our analysis provides a quantitative, data driven approach to exploring multimorbid patient data with the

possibility of real-time analysis.

Clustering on the Texas patient health data to isolate multimorbidity patients revealed nine well-defined clusters. An analysis of the most prevalent conditions in every cluster revealed broad groupings within each cluster (Table 2). Our first cluster was also the largest with 4,532 patients. It contained middle aged patients (median age 53.5 years) and had a slightly higher ratio of males (56%) compared to females (44%). The cluster was characterized by the highest incidence of cancers of different organ systems including colorectal cancer (2.8%), prostate cancer (3.9%), ovarian cancer (0.6%), Multiple myeloma (1.5%), malignant melanoma (1.8%), pancreatic cancer (0.5%), esophageal cancer (0.4%), stomach cancer (0.4%), skin cancer (7.2%), oral cancer (0.8%) and other cancers (8.9%). Interestingly, this cluster also had the highest incidences of kidney stones (7.7%), inflammatory bowel disease (8.6%) and sickle cell anemia (0.2%) indicating a potential causal role of these conditions in certain cancers. Indeed, a 2015 meta-analysis revealed an increased risk for kidney stone formation and renal cell carcinoma (Cheungpasitporn et al., 2015). Further corroborating our cluster-derived corollary relationship, inflammatory bowel disease patients are known to be at an increased risk of colorectal cancer and it has also been recently identified as a risk factor for oral cancer (Katsanos, Roda, Brygo, Delaporte, & Colombel, 2015; Kim & Chang, 2014).

Our second cluster had a higher ratio of older females (53% females with median age of 59.6 years) and had the highest incidences of conditions like osteoarthritis (69%) and lower back pain (71%). Interestingly, this cluster in general tends to be more expensive compared to other clusters (results not shown). Cluster three was our youngest cluster (median age 42.9 years), with more males compared to females (64% males). This cluster contained the highest percentage of members with substance abuse (80%) and related disorders including hepatitis (15%), pancreatitis (16%), neurosis (1.1%), and psychoses (9.3%). Remarkably, we also found the highest rate of post-partum neurosis disorders in this cohort, which raised the possibility of post-partum substance abuse or addiction. Indeed it has been suggested that pharmacological agents used to treat post-partum depression often lead to long term addiction and need more federal & clinical regulation (Chapman & Wu, 2013; Ross & Dennis, 2009).

Cluster 4 contained a high proportion of middle aged females (median age 56.1 years; 62.5 % females) with a relatively high proportion reporting menopause (17.2%). Further, this cluster also reported the highest incidences for conditions like rheumatoid arthritis (9.4%) and fibromyalgia (8.8%). This correlative evidence further backs the previous similar observations of links between fibromyalgia, rheumatoid arthritis and menopause (Martínez-Jauand et al., 2013; Pines, 2014). Martinez-Jauand and colleagues, have previously shown that an early menopause can reduce estrogen exposure and this causes an increased sensitivity to pain which magnifies the fibromyalgia symptoms (Martínez-Jauand et al., 2013). Cluster 5, although relatively smaller in size (895 patients), contained a very high proportion of patients with metabolic syndrome (95%) patients. As expected this cluster had the highest rates of hypertension (98%) and obesity (61%). This cluster also contained a high proportion of cervical cancer patients (0.4%) and this link has previously been demonstrated in other populations (Bussière, Sicsic, & Pelletier-Fleury, 2014; Lee, So, Piyathilake, & Kim, 2013). Cluster 6, was our smallest cluster cohort (765 patients) with the highest proportion of females (77.3% females) had the highest incidences of osteoporosis (22.4%), chronic thyroid disorders (86.4%), chronic fatigue syndrome (1.2%). A common theme related to these conditions is the interleukin-6 pathway, dysregulations of which are known to play a central role in osteoporosis, thyroid disorders and neck cancer (Guerrera et al., 2014; Lumachi, Basso, & Orlando, 2010; Papanicolaou, Wilder, Manolagas, & Chrousos, 1998; Roy, Curtis, Fears, Nahashon, & Fentress, 2016). Our results suggest that this cytokine molecular pathway may be responsible for more disorders than previously identified.

Cluster 7 also had a high proportion of young females (59% females; median age 49.1 years) who reported a high proportion of neurological and psychological disorders, including psychosis (2.8%), depression (48%), migraines/headaches (35.4%), epilepsy (12.1%), and eating disorders (0.6%). This group also had the highest proportion of fertility issues (0.7%) and gave birth to babies with low birth weight (0.1%). We also observed this group reporting an increased incidence of having a disrupted childhood (0.6%), posing a possible origin of these psychological issues. Similar co/multimorbidity associations have extensively been studied in childhood post-traumatic stress disorders (Gekker et al., 2018; Lecei et al., 2018; Nordin, Olsson, & Tomson, 2018). Cluster 8, our oldest cluster (median age 60.5 years) with the highest proportion of males (66.8% males) suffered from a combination of cardiovascular and respiratory diseases commonly seen in the elderly population. This cluster had a high incidence rates for heart failure (49%), cerebrovascular disease (80.9%), COPD (16.7%), congenital heart disease (1.7%) and ventricular arrhythmia (14.2%). Further, consistent with our expectations, we also found the highest incidences of Parkinson's (0.7%) and dementia (2.7%) in this cluster. This cohort of patients were seen to have the highest frequency of in-patient visits and the highest total cost associated with them (data not shown). Our final cluster, number 9, was dominated by middle aged males (61.8% males; median age 59.2 years) with the highest incidence of diabetes mellitus (68.8%). We also saw highest incidences of diabetes related chronic disorders like chronic renal failure (29.5%), cataract (20.8%) and glaucoma (20.6%). Relationships between these conditions have been extensively reported (Harding, Egerton, van Heyningen, & Harding,

1993; Lipton & Decker, 2015; Yoshimoto & Kato, 2016).

Overall, our clustering approach has identified cohorts of patients with similar multimorbid diseases with actionable insights that can be used to reduce disease incidence, treatment & management costs as well as the overall burden on today's healthcare system.

References

Anderson, G. F. (2003). Physician, Public, and Policymaker Perspectives on Chronic Conditions. *Archives of Internal Medicine*, *163*(4), 437–442.

Becker, H. (2005). *A Survey of Correlation Clustering*.

Bussi ère, C., Sicsic, J., & Pelletier-Fleury, N. (2014). The effects of obesity and mobility disability in access to breast and cervical cancer screening in france: results from the national health and disability survey. *PloS One*, *9*(8), e104901.

CDC - NCHS - National Center for Health Statistics. (2018, June 18). Retrieved June 27, 2018, from https://www.cdc.gov/nchs/index.htm

Chapman, S. L. C., & Wu, L.-T. (2013). Postpartum substance use and depressive symptoms: a review. *Women & Health*, *53*(5), 479–503.

Cheng, S. Y., Dy, S., Fang, P. H., Chen, C. Y., & Chiu, T. Y. (2013). Evaluation of inpatient multidisciplinary palliative care unit on terminally ill cancer patients from providers' perspectives: a propensity score analysis. *Japanese Journal of Clinical Oncology*, *43*(2), 161–169.

Cheungpasitporn, W., Thongprayoon, C., O'Corragain, O. A., Edmonds, P. J., Ungprasert, P., Kittanamongkolchai, W., & Erickson, S. B. (2015). The risk of kidney cancer in patients with kidney stones: a systematic review and meta-analysis. *QJM: monthly journal of the Association of Physicians*, *108*(3), 205–212.

Clifton, D. A., Niehaus, K. E., Charlton, P., & Colopy, G. W. (2015). Health Informatics via Machine Learning for the Clinical Management of Patients. *Yearbook of Medical Informatics*, *10*(1), 38–43.

Comorbidity indices. (n.d.). . Retrieved June 19, 2018, from http://www.clinexprheumatol.org/abstract.asp?a=8618

Deo, R. C. (2015). Machine Learning in Medicine. *Circulation*, *132*(20), 1920–1930.

Fleming, S. M., Santiago, N. A., Mullin, E. J., Pamphile, S., Karkare, S., Lemkuhl, A., & Ekhator, O. R., et al. (2018). The effect of manganese exposure in Atp13a2-deficient mice. *Neurotoxicology*, *64*, 256–266.

Fried, T. R., O'Leary, J., Towle, V., Goldstein, M. K., Trentelange, M., & Martin, D. K. (2014). The effects of comorbidity on the benefits and harms of treatment for chronic disease: a systematic review. *PloS One*, *9*(11), e112593.

Gekker, M., Coutinho, E. S. F., Berger, W., Luz, M. P. da, Ara újo, A. X. G. de, Pagotto, L. F. A. da C., & Marques-Portella, C., et al. (2018). Early scars are forever: Childhood abuse in patients with adult-onset PTSD is associated with increased prevalence and severity of psychiatric comorbidity. *Psychiatry Research*, *267*, 1–6.

Gower, J. C. (1971). A General Coefficient of Similarity and Some of Its Properties. *Biometrics*, *27*(4), 857.

Guerrera, I. C., Quetier, I., Fetouchi, R., Moreau, F., Vauloup-Fellous, C., Lekbaby, B., & Rousselot, C., et al. (2014). Regulation of interleukin-6 in head and neck squamous cell carcinoma is related to papillomavirus infection. *Journal of Proteome Research*, *13*(2), 1002–1011.

Harding, J. J., Egerton, M., van Heyningen, R., & Harding, R. S. (1993). Diabetes, glaucoma, sex, and cataract: analysis of combined data from two case control studies. *The British Journal of Ophthalmology*, *77*(1), 2–6.

Hester, M. S., Hosford, B. E., Santos, V. R., Singh, S. P., Rolle, I. J., LaSarge, C. L., & Liska, J. P., et al. (2016). Impact of rapamycin on status epilepticus induced hippocampal pathology and weight gain. *Experimental Neurology*, *280*, 1–12.

Hoffman, C., Rice, D., & Sung, H. Y. (1996). Persons With Chronic Conditions: Their Prevalence and Costs. *JAMA*, *276*(18), 1473–1479.

Hofstetter, H., Dusseldorp, E., van Empelen, P., & Paulussen, T. W. G. M. (2014). A primer on the use of cluster analysis or factor analysis to assess co-occurrence of risk behaviors. *Preventive Medicine*, *67*, 141–146.

Ilesanmi, O. S., & Fatiregun, A. A. (2014). The direct cost of care among surgical inpatients at a tertiary hospital in south west Nigeria. *The Pan African Medical Journal*, *18*, 3.

Karkare, S., Chhipa, R. R., Anderson, J., Liu, X., Henry, H., Gasilina, A., & Nassar, N., et al. (2014). Direct inhibition of retinoblastoma phosphorylation by nimbolide causes cell-cycle arrest and suppresses glioblastoma growth. *Clinical Cancer Research: An Official Journal of the American Association for Cancer Research*, *20*(1), 199–212.

Katsanos, K. H., Roda, G., Brygo, A., Delaporte, E., & Colombel, J.-F. (2015). Oral Cancer and Oral Precancerous Lesions in Inflammatory Bowel Diseases: A Systematic Review. *Journal of Crohn's & Colitis*, *9*(11), 1043–1052.

Kim, E. R., & Chang, D. K. (2014). Colorectal cancer in inflammatory bowel disease: the risk, pathogenesis, prevention and diagnosis. *World Journal of Gastroenterology*, *20*(29), 9872–9881.

Lecei, A., Decoster, J., De Hert, M., Derom, C., Jacobs, N., Menne-Lothmann, C., & van Os, J., et al. (2018). Evidence that the association of childhood trauma with psychosis and related psychopathology is not explained by gene-environment correlation: A monozygotic twin differences approach. *Schizophrenia Research*.

Lee, J. K., So, K. A., Piyathilake, C. J., & Kim, M. K. (2013). Mild obesity, physical activity, calorie intake, and the risks of cervical intraepithelial neoplasia and cervical cancer. *PloS One*, *8*(6), e66555.

Lipton, B. J., & Decker, S. L. (2015). Association between diagnosed diabetes and trouble seeing, National Health Interview Survey, 2011-13. *Journal of Diabetes*, *7*(5), 743–746.

Lumachi, F., Basso, S. M. M., & Orlando, R. (2010). Cytokines, thyroid diseases and thyroid cancer. *Cytokine*, *50*(3), 229–233.

Martínez-Jauand, M., Sitges, C., Femenia, J., Cifre, I., González, S., Chialvo, D., & Montoya, P. (2013). Age-of-onset of menopause is associated with enhanced painful and non-painful sensitivity in fibromyalgia. *Clinical Rheumatology*, *32*(7), 975–981.

Nordin, V., Olsson, I. B., & Tomson, T. (2018). [Epilepsy and comorbid neurodevelopmental disorders]. *Lakartidningen*, *115*.

Papanicolaou, D. A., Wilder, R. L., Manolagas, S. C., & Chrousos, G. P. (1998). The pathophysiologic roles of interleukin-6 in human disease. *Annals of Internal Medicine*, *128*(2), 127–137.

Pines, A. (2014). Fibromyalgia and menopause: any link? *Climacteric: The Journal of the International Menopause Society*, *17*(4), 514–515.

Pisinger, C., Toft, U., Aadahl, M., Glümer, C., & Jørgensen, T. (2009). The relationship between lifestyle and self-reported health in a general population: the Inter99 study. *Preventive Medicine*, *49*(5), 418–423.

Ross, L. E., & Dennis, C. L. (2009). The prevalence of postpartum depression among women with substance use, an abuse history, or chronic illness: a systematic review. *Journal of Women's Health (2002)*, *18*(4), 475–486.

Roy, B., Curtis, M. E., Fears, L. S., Nahashon, S. N., & Fentress, H. M. (2016). Molecular Mechanisms of Obesity-Induced Osteoporosis and Muscle Atrophy. *Frontiers in Physiology*, *7*, 439.

Sharabiani, M. T. A., Aylin, P., & Bottle, A. (2012). Systematic review of comorbidity indices for administrative data. *Medical Care*, *50*(12), 1109–1118.

Singh, S. P. (2015). *Quantitative analysis on the origins of morphologically abnormal cells in temporal lobe epilepsy*. University of Cincinnati. Retrieved November 25, 2017, from https://etd.ohiolink.edu/pg_10?0::NO:10:P10_ACCESSION_NUM:ucin1446547280

Singh, S. P. (2016). Advances in Epilepsy: A data science perspective, *58*(2), 89–92.

Singh, S. P., Chhunchha, B., Fatma, N., Kubo, E., Singh, S. P., & Singh, D. P. (2016). Delivery of a protein transduction domain-mediated Prdx6 protein ameliorates oxidative stress-induced injury in human and mouse neuronal cells. *American Journal of Physiology. Cell Physiology*, *310*(1), C1-16.

Singh, S. P., He, X., McNamara, J. O., & Danzer, S. C. (2013). Morphological changes among hippocampal dentate granule cells exposed to early kindling-epileptogenesis. *Hippocampus*, *23*(12), 1309–1320.

Singh, S. P., & Karkare, S. (2017). Stress, Depression and Neuroplasticity.

Singh, S. P., Karkare, S., Baswan, S. M., & Singh, V. P. (2018). The Application of Text Mining Algorithms In Summarizing Trends in Anti-Epileptic Drug Research. *International Journal of Statistics and Probability*, *7*(4), 11.

Singh, S. P., LaSarge, C. L., An, A., McAuliffe, J. J., & Danzer, S. C. (2015). Clonal Analysis of Newborn Hippocampal Dentate Granule Cell Proliferation and Development in Temporal Lobe Epilepsy. *eNeuro*, *2*(6).

Singh, S. P., Singh, S. P., Fatima, N., Kubo, E., & Singh, D. P. (2008). Peroxiredoxin 6-A novel antioxidant neuroprotective agent. *Neurology*, *70*(11), A480–A481.

Singh, S. P., & Singh, V. P. (2017). Quantitative Analysis on the role of Raffinose Synthase in Hippocampal Neurons. *bioRxiv*. Retrieved from http://biorxiv.org/content/early/2017/12/27/240192.abstract

Song, X., Wu, J., Yu, C., Dong, W., Lv, J., Guo, Y., Bian, Z., et al. (2018). Association between multiple comorbidities and

self-rated health status in middle-aged and elderly Chinese: the China Kadoorie Biobank study. *BMC Public Health*, *18*, 744.

Whiteman, M., & Whiteman, D. B. (1949). The application of cluster analysis to the Wechsler-Bellevue scale. *Delaware Medical Journal*, *21*(8), 174–176.

WHO | International Classification of Diseases, 11th Revision (ICD-11). (n.d.). *WHO*. Retrieved June 19, 2018, from http://www.who.int/classifications/icd/en/

Yoshimoto, M., & Kato, S. (2016). [Diagonosis and treatment of cataract and glaucoma caused by diabetes mellitus]. *Nihon Rinsho. Japanese Journal of Clinical Medicine*, *74 Suppl 2*, 148–152.