Covid-19 and Diabetes: A Relational Comorbidity Cluster Analysis

Joseph Cruz
josephbcruz@lewisu.edu
DATA-51000-001, Summer 1 (2021)
Data Mining and Analytics
Lewis University

I. Introduction

In 2020, the coronavirus disease of 2019 (Covid-19) was officially declared a global pandemic. At that time, much was unknown about the virus to researchers aside from that it was caused by a coronavirus like that of the severe acute respiratory syndrome virus (SARS). Throughout the course of the last year, scientists have been able to learn a multitude of things about the Covid-19 virus such as potential transmission rates, mortality rates, and even the mechanism of infection. Scientists and doctors have also begun speculating that certain pre-existing conditions may contribute to a more serious Covid-19 infection. One of these pre-existing conditions is diabetes (either type 1 or type 2) [1]. Diabetes is a chronic condition that causes irregularities in how the body converts food into energy through how it controls glucose levels in the bloodstream. Pending upon the type of diabetes, insulin, the hormone responsible for balancing blood glucose levels, is either deficient, where the body cannot produce enough insulin, or the body has insulin resistances, where the insulin cannot effectively reduce blood glucose levels.

Since diabetes has the potential to cause more severe illness with a Covid-19 infection, it is incredibly important to understand how and why the infection becomes more severe. One way to determine this is to view comorbidities that could have led to overall mortality in patients with diabetes. Furthermore, determining relationships between these potentially developed comorbidities would allow for a deeper understanding of the interplay between them, diabetes, and the Covid-19 virus. To begin identifying these comorbidities that may be related to Covid-19 infection and diabetes, one can begin by grouping and identifying cases that led to higher mortality with these conditions. By performing a cluster analysis, it is possible to group these cases into clusters that can then be analyzed for relationships and to understand some of the complications that may arise with having both Covid-19 and diabetes.

In this paper, two different clustering methods are applied and examined in the determination of comorbidities associated with Covid-19 infection and diabetes. These clustering methods that will be explored are the *k-means* clustering method and the *hierarchical* clustering method. These clustering methods will be applied to a report containing the provisional diabetes death counts for 2020 from the Centers for Disease Control and Prevention (CDC) [2]. Throughout this report, the following will be present: description of the dataset, methodology, results and discussion, and the conclusion from the data. In Section II, a description of the dataset used in the analysis will be covered. In Section III, there will be an overview of the methodology used for the analysis at hand. Section IV will provide the reader with the results of the cluster analyses along with a discussion. Lastly, in Section V, conclusions will be presented.

II. Data Description

The dataset that will be used for this cluster analysis is a report from the CDC of the provisional diabetes death counts for 2020 [2]. There are a total of 226 rows, or instances, in the dataset with a total of 16 features. The features of this dataset have been shown in Table 1. To begin with, the dataset contains the "Data as of" attribute, which is effectively meant to describe that the data, as of this date, is said value. The purpose of this attribute is that provisional death counts are not always complete, as not all states are able to get the information to the CDC at the same time, hence they are not necessarily a complete set of data [3]. Next, we have the "Date_of_Death_Year" and "Date_of_Death_Month", where these features are the year and month in which the deaths occurred, respectively. The data also has a categorical feature of "AgeGroup", where the deaths are split amongst their respective age groups. The "Sex" feature is meant to group the deaths based upon respective biological sexes. The dataset then has the "COVID19" feature, which is a provisional death count associated with Covid-19. There are also the attributes "Diabetes.uc" and "Diabetes.mc", which represent the provisional death counts for those with associated with diabetes and either ulcerative colitis or major cardiovascular disease. Then, there are the comorbidities that occurred in conjunction with Covid-19. The first, is "C19PlusDiabetes", which is the death count where Covid-19 and diabetes were present. The next attribute is "C19PlusHypertensiveDiseases", which is the death count where Covid-19 and hypertensive diseases were present at time of death. Then, "C19PlusMajorCardiovascularDiseases" is the death count where Covid-19 and major cardiovascular diseases were present at time of death. "C19PlusHypertensiveDiseasesAndMCVD" are the death counts where Covid-19 and both hypertensive and major

TABLE I. ATTRIBUTES OF PROVISIONAL DIABETES DEATH COUNTS FOR 2020 DATASET

Attribute	Type	Example Value	Description
Data as of	Categorical	10/20/2020	Date of the updated date of the data.
Date_Of_Death_Year	Numeric	2020	Year deaths occurred.
Date_Of_Death_Month	Numeric	1	Month deaths occurred (1-12).
AgeGroup	Categorical	18-29 years	The age group the deaths belonged to.
Sex	Categorical	Female (F)	The biological sexes of the deaths.
COVID19	Numeric	5	Deaths associated with Covid-19.
Diabetes.uc	Numeric	4	Deaths associated with diabetes that had ulcerative colitis.
Diabetes.mc	Numeric	7	Deaths associated with diabetes that had major cardiovascular disease.
C19PlusDiabetes	Numeric	16	Deaths associated with Covid-19 and diabetes.
C19PlusHypertensiveDiseases	Numeric	47	Deaths associated with Covid-19 and hypertensive diseases.
C19PlusMajorCardiovascularDiseases	Numeric	81	Deaths associated with Covid 19 and cardiovascular disease.
C19PlusHypertensiveDiseasesAndMCVD	Numeric	11	Deaths associated with Covid-19 and hypertension and major cardiovascular disease.
C19PlusChronicLowerRespiratoryDisease	Numeric	6	Deaths associated with Covid-19 and chronic lower respiratory disease.
C19PlusKidneyDisease	Numeric	47	Deaths associated with Covid-19 and kidney disease.
C19PlusChronicLiverDiseaseAndCirrhosis	Numeric	29	Deaths associated with Covid-19 and chronic liver disease and cirrhosis.
C19PlusObesity	Numeric	26	Deaths associated with Covid-19 and obesity.

cardiovascular diseases were present at time of death. “C19PlusChronicLowerRespiratoryDisease” is the death count in which Covid-19 and chronic lower respiratory disease were present at time of death. “C19PlusKidneyDisease” is the death count in which Covid-19 and kidney disease was present at the time of death. “C19PlusChronicLiverDiseaseAndCirrhosis” is the death count in which Covid-19 and chronic liver disease and cirrhosis of the liver were present at the time of death. Finally, “C19PlusObesity” is the death count in which Covid-19 and obesity were present at the time of death.

Of the dataset, the following attributes were used for the cluster analysis: “AgeGroup”, “COVID19”, “Diabetes.mc”, “C19PlusDiabetes”, “C19PlusHypertensiveDiseases”, “C19PlusHypertensiveDiseasesAndMCVD”, “C19PlusChronicLowerRespiratoryDisease”, “C19PlusKidneyDisease”, “C19PlusChronicLiverDiseaseAndCirrhosis”, “C19PlusObesity”. The frequency distributions of the attributes used are found in Figures 1-9. The “AgeGroup” attribute is split into categories of age ranges (<18 years, 18-29 years, 30-39 years, 40-49 years, 50-59 years, 60-69 years, 70-79 years, 80+ years, etc.). The average, standard deviation, and the percent coefficient of variation (%CV) of the used numerical attributes are shown in Table II.

Fig. 1. Frequency distribution of COVID19 attribute.

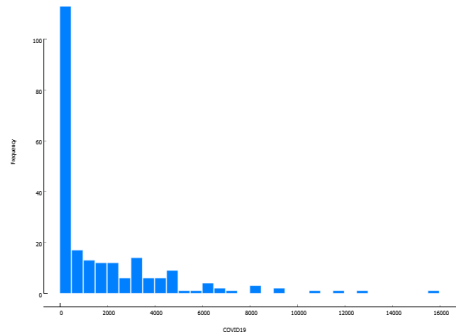


Fig 2. Frequency distribution of Diabetes.mc attribute.

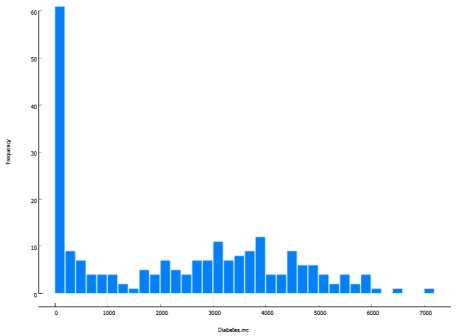


Fig 3. Frequency distribution of C19PlusDiabetes attribute.

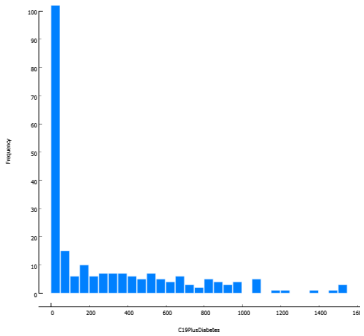


Fig 4. Frequency distribution of C19PlusHypertensiveDisesases attribute.

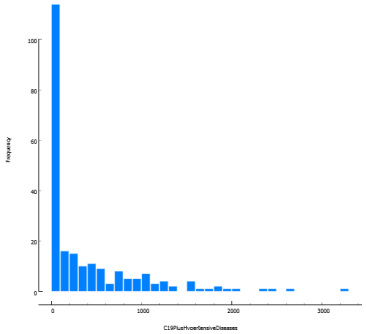


Fig 5. Frequency distribution of C19PlusChronic LowerRespiratoryDisease attribute.

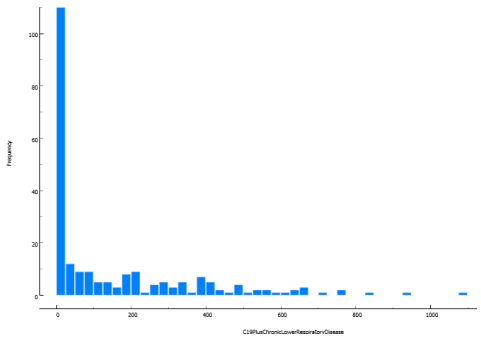


Fig 6. Frequency distribution of C19PlusHypertensiveDisesases AndMCVD attribute.

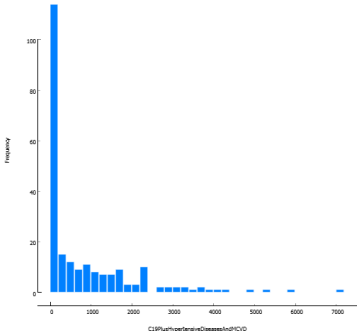


Fig 7. Frequency distribution of C19PlusChronicLiver DiseaseAndCirrhosis attribute.

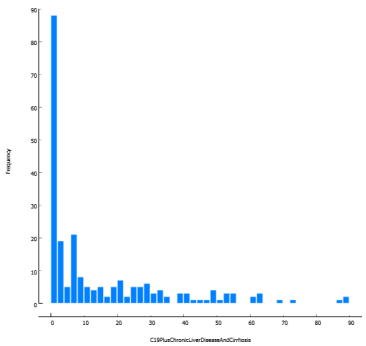


Fig 8. Frequency distribution of C19PlusKidneyDisease attribute.

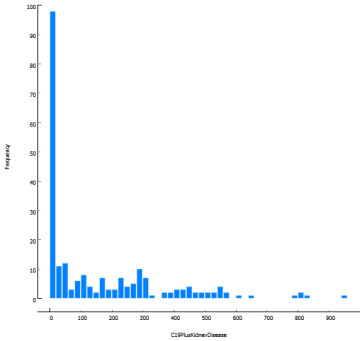


Fig 9. Frequency distribution of C19PlusObesity attribute.

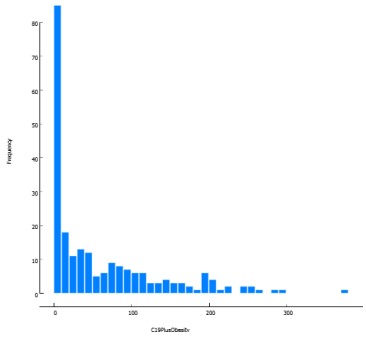


TABLE II. STATISTICAL ANALYSIS OF PROVISIONAL DIABETES DEATH COUNTS FOR 2020 ATTRIBUTES (USED FOR ANALYSIS)

	COVID19	Diabetes.mc	C19Plus Diabetes	C19Plus HypertensiveDiseases	C19Plus HypertensiveDiseases AndMCVD	C19Plus ChronicLower RespiratoryDisease	C19Plus KidneyDisease	C19Plus ChronicLiver DiseaseAndCirrhosis	C19Plus Obesity
Average (AVG)	1711.0	2306.3	280.8	373.8	796.7	147.0	150.1	13.9	58.7
Standard Deviation (SD)	2542.2	1939.9	367.5	563.6	1186.8	211.7	198.7	19.4	73.8
Percent CV (%CV)	148.6	84.1	130.9	150.8	149.0	144.1	132.4	139.1	125.6

The “COVID19” feature has an average death count of approximately 1711 deaths with a standard deviation of 2542.2 and a %CV of 148.6%. In the “Diabetes.mc” feature, the average death count is approximately 2306.3 with a standard deviation of 1939.9 and a %CV of 84.1%. The “C19PlusDiabetes” feature has an average of 280.8 and a standard deviation of 367.5 and a %CV 130.9%. The “C19PlusHypertensiveDiseases” has an average death count of 373.8 with a standard deviation of 563.6 and a %CV of 150.8%. The “C19PlusHypertensiveDiseasesAndMCVD” feature has an average death count of 796.7 with a standard deviation of 1186.8 and a %CV of 149%. The “C19PlusChronicLowerRespiratoryDisease” attribute has an average death count of approximately 147 with a standard deviation of 211.7 and a %CV of 144.1%. The “C19PlusKidneyDisease” feature has an average death count of 150.1 with a standard deviation of 198.7 with a %CV of 132.4%. The “C19PlusChronicLiverDiseaseAndCirrhosis” feature has an average death count of 13.9 with a standard deviation of 19.4 with a %CV of 139.1%. The “C19PlusObesity” attribute has an average death count of 58.7 with a standard deviation of 73.8 with a %CV of 125.6%.

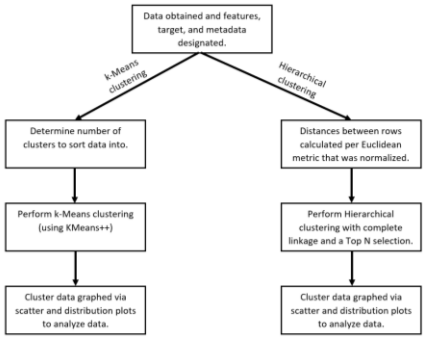
III. METHODOLOGY

To perform the cluster analysis of the data, the data mining toolkit Orange (version 3.26) was used. The flowchart present in Figure 10 briefly summarizes the steps required for these cluster analyses. First, the data was obtained and the features, metadata, and target where all designated using the file widget in Orange. The “AgeGroup” and “Diabetes.mc” attributes were marked as metadata. The “Data as of”, “Date_Of_Death_Year”, “Date_Of_Death_Month”, “Sex”, “Diabetes.uc”, and the “C19PlusMajorCardiovascularDisease” attributes were marked to skip since these values were not going to contribute to the overall analysis of the required features (“C19PlusMajorCardiovascularDisease” and the “C19PlusHypertensiveDiseasesAndMCVD” attributes were identical for each value in the data, thus it was removed and replaced with the “C19PlusHypertensiveDiseasesAndMCVD” attribute). The “C19PlusDiabetes” feature was labeled as the target. Finally, the “C19PlusHypertensiveDiseases”, “C19PlusHypertensiveDiseasesAndMCVD”, “C19PlusChronicLowerRespiratoryDisease”, “C19PlusKidneyDisease”, “C19PlusChronicLiverDiseaseAndCirrhosis”, “C19PlusObesity” were marked as features. From here, the data was then able to be split into the two clustering methods.

The first clustering method used was k-means clustering. The first step in this clustering method was to determine the number of clusters that the data will be sorted into. To do this, the file data was fed into the k-Means widget in Orange. The widget determined the silhouette scores of clustering the data into a range from 2 clusters to 8 clusters and the cluster assignment with the highest silhouette score was utilized for the clustering. Furthermore, the columns in the data were normalized and initialized with KMeans++ using 10 re-runs and a maximum of 300 iterations. At this point, the data was clustered and fed into multiple different scatter plots and distribution widgets to graph their clusters and examine the distribution of the data.

The other clustering method used was hierarchical clustering method. The first step in this clustering method was to calculate the distances between the features. To do this, the ‘Distances’ widget in Orange was used to calculate the distance between rows using the Euclidean distance metric with normalization. Once the distances were calculated, they were fed into the hierarchical clustering widget in Orange. The hierarchical clustering utilized a complete linkage with a pruning max depth of 5 and the selection was based upon a ‘Top N’ of 3. From here, the data was clustered and fed into multiple different scatter plot and distribution widgets to graph their clusters and examine the distribution of the data.

Fig 10. Flowchart of clustering methodology.



IV. Results and Discussion

A. *K-means Clustering Results*

The k-means clustering of the data yielded a maximum silhouette score of 0.614 for two clusters, hence the data was split into two clusters. The population of the data was split so that 30% (67 instances) of the population was in cluster C1 and 70% (159 instances) of the population was in cluster C2. To better visualize and understand the results in terms of the data, scatter plots were created. In the scatter plots, the attributes of the data were graphed against each other to investigate for potential relationships based off the clustering. Of the scatter plots analyzed, it was found that the scatter plots in Figures 11-13 had the most potential for correlation based upon the clustering. In these figures, blue represents the first cluster (C1), and the red represents the second cluster (C2). In Figure 11, the number of Covid-19 cases (x-axis) were graphed against the number of Covid-19 cases with diabetes (y-axis). In Figure 12, the number of deaths attributed to Covid-19 with a diabetes diagnosis (x-axis) were plotted against the number of deaths attributed to Covid-19 and hypertensive diseases and major cardiovascular disease (y-axis). Finally, in Figure 13, the number of deaths attributed to Covid-19 with a diabetes diagnosis (x-axis) were plotted against the number of deaths attributed to Covid-19 and kidney diseases (y-axis). For each of these scatter plots, it

Fig 11. K-means scatter plot of Covid-19 cases (x-axis) v.s Covid-19 with diabetes (y-axis).

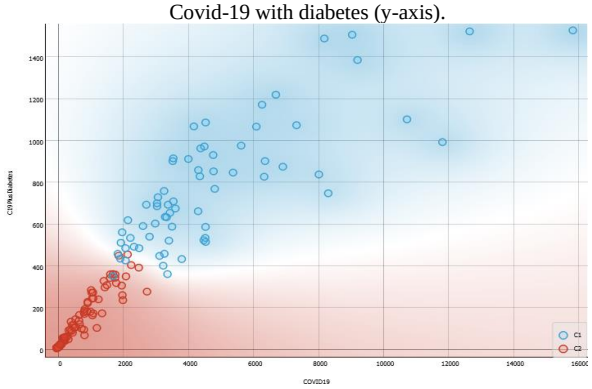


Fig 12. K-means scatter plot of Covid-19 with diabetes (x-axis) v.s. Covid-19 with HypertensiveDisesasesAndMCVD (y-axis).

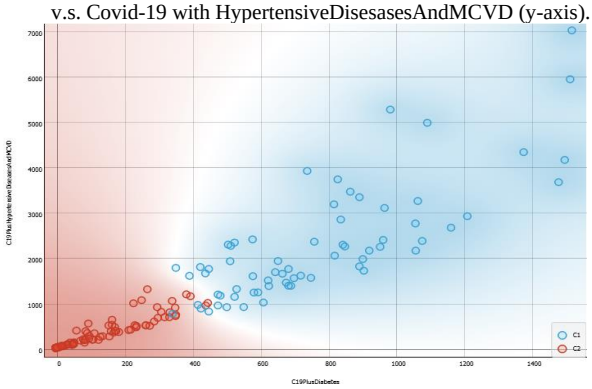
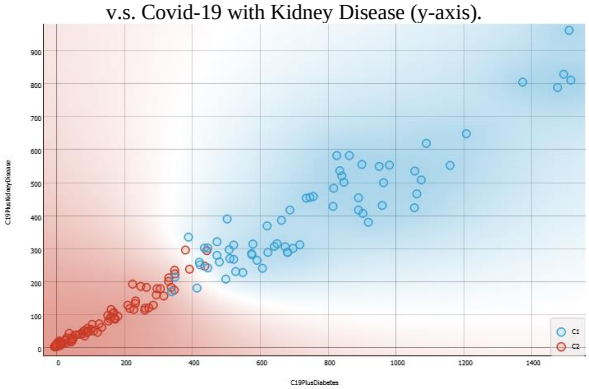


Fig 13. K-means scatter plot of Covid-19 with diabetes (x-axis) v.s. Covid-19 with Kidney Disease (y-axis).



should be noted that the higher death counts between the attributes belong to cluster C1 while the lower death counts belong to cluster C2. For instance, in Figure 11, most of the higher number of Covid-19 cases with a diabetes diagnosis belong to cluster C1, while those with less seem to belong to cluster C2. This trend is present throughout the figures for the k-means clustering scatter plots. Furthermore, there also seems to be upward linear trends between each of the pairs of attributes graphed throughout the scatter plots.

B. Hierarchical Clustering Results

The hierarchical clustering of the data was separated into a total of three clusters, due to the 'Top N' value of 3 used. The resulting dendrogram of the clustering can be shown in Figure 14. The population of the data was split so that 1.77% (4 instances) of the population was in cluster C1, 73% (165 instances) of the population was in cluster C2, and 25.2% of the population (57 instances) belonged to cluster C3. To better understand and visualize results, scatter plots were generated. The pairs of attributes graphed were the same as the k-means scatter plots since they had high potential for correlations due to clustering and it was a good way to compare the two clustering methods. The resulting scatter plots in Figures 15-17 had the most potential for correlation based upon the clustering. In these figures, the blue represents the first cluster (C1), the red represents the second cluster (C2), and the green represents the third cluster (C3). In Figure 15, the number of Covid-19 cases (x-axis) were graphed against the number of Covid-19 cases with diabetes (y-axis). In Figure 16, the number of deaths attributed to Covid-19 with a diabetes diagnosis (x-axis) were plotted against the number of deaths attributed to Covid-19 and hypertensive diseases and major cardiovascular disease (y-axis). Finally, in Figure 17, the number of deaths attributed to Covid-19 with a diabetes diagnosis (x-axis) were plotted against the number of deaths attributed to Covid-19 and kidney diseases (y-axis). As shown in the k-means plots, there is still a linear trend between each pair of the attributes.

C. Discussion

By performing two different clustering methods upon the data, two very different results were obtained. The k-means clustering designated two clusters while the hierarchical clustering utilized three clusters. Of the k-means clusters, the cluster C2 had significantly more instances of the population. On the other hand, the hierarchical clustering split into three clusters with the smallest cluster C1 only having 4 instances and the other two having similar spreads as the k-means clusters. This may be a consequence of the data being too far away from the other points in cluster C3, so they were not included and instead formed their own cluster. It does seem that given this data, for the higher number of Covid-19 deaths with diabetes there are higher numbers of the comorbidities associated deaths; thus, it is possible that these higher numbers may signify drastic spikes and/or may be associated with age. After checking the 4 instances' age attributes, the instances in cluster C1

Fig 14. Resulting dendrogram of hierarchical clustering (annotated with the number of deaths attribute to Covid-19 with a diabetes diagnosis).

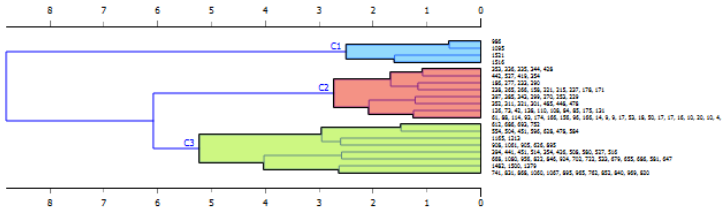


Fig 15. Hierarchical scatter plot of Covid-19 cases (x-axis) v.s. Covid-19 with diabetes (y-axis).

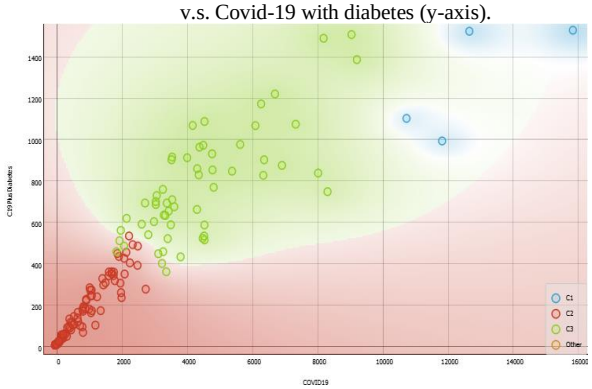


Fig 16. Hierarchical scatter plot of Covid-19 with diabetes (x-axis) v.s. Covid-19 with HypertensiveDisesasesAndMCVD(y-axis).

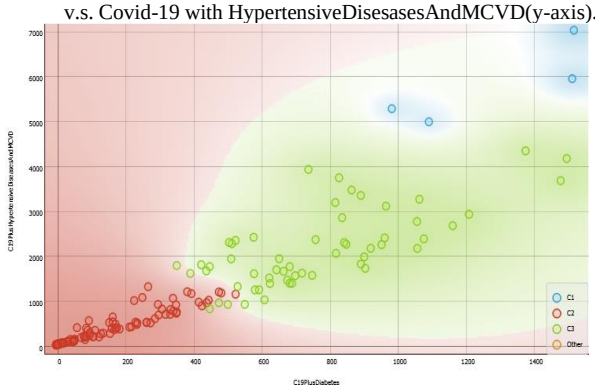
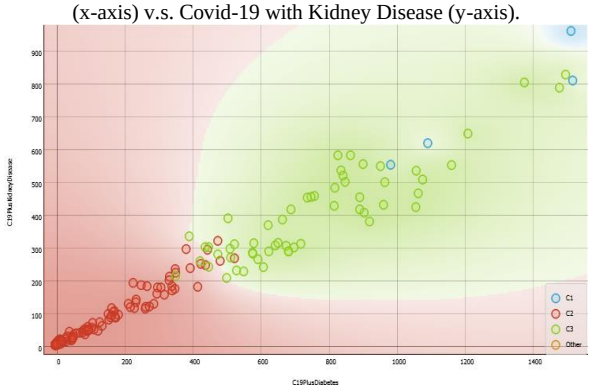


Fig 17. Hierarchical scatter plot of Covid-19 with diabetes (x-axis) v.s. Covid-19 with Kidney Disease (y-axis).



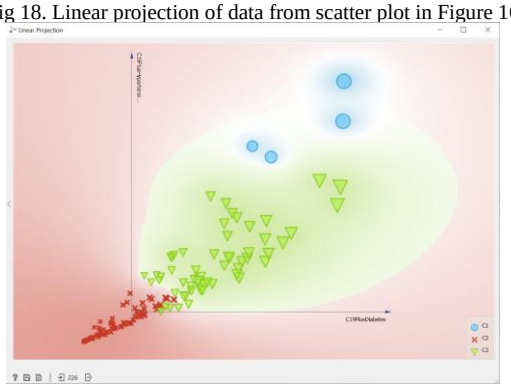
were from age groups 80+ years and 85+ years. Thus, the small cluster is likely attributed to the greatest risk of severe illness for elderly individuals [4], which also tended to be much higher in overall mortality rates.

Furthermore, some points in the clusters in the hierarchical clustering do not seem like they should be in the assigned clusters when looking at the scatter plot. For instance, some points in cluster C1 seem to be in cluster C3, however this is not actually the case. Since these scatter plots are two-dimensional representations of the clustering, it is difficult to visualize the three-dimensional space they are in. Figure 18 is an example of a linear projection of the data that was graphed in Figure 16. Notice how the cluster C1 seems to be on top of cluster C3. This may be a consequence of the complete linkage that was used for the hierarchical clustering. Since the complete linkage uses the 'farthest neighbor' method, tighter clusters are generally created but at the expense of them being incredibly close together.

In respect to comorbidities, the cluster sizes of the k-means and hierarchical clustering showed a much higher density in the clusters with lower quantities of Covid-19 deaths with diabetes and Covid-19 with comorbidities deaths. For this reason, cluster C1 (k-means) and clusters C1 and C3 (hierarchical) could be labeled as high-comorbidity Covid-19/diabetes clusters while the clusters C2 (for both k-means and hierarchical) would be labeled low-comorbidity Covid-19/diabetes clusters. This is further corroborated by the first scatter plots for both clustering methods, where they served as controls to determine the clustering of just Covid-19 with diabetes deaths against overall Covid-19 deaths. For the k-means, there was a 30% clustering of the data set with the higher quantities of Covid-19 with diabetes and Covid-19 with comorbidities while 70% of the instances fell in line with the lower quantities of both. For the hierarchical clustering, there were 73% of the instances associated with lower quantities of Covid-19 with diabetes and Covid-19 with comorbidities and a cumulative 27% of the instances that were associated with higher quantities of both. One could surmise that this may indicate slight correlation with Covid-19 with diabetes and the other comorbidities mentioned. Although not necessarily indicative that the comorbidities are in fact a consequence of diabetes and Covid-19 diagnosis, there seems to be some sort of correlation in which these factors are present in approximately 27% to 30% of the instance data. Thus, there may be some sort of relationship between Covid-19 with diabetes and hypertension/major cardiovascular and kidney disease.

One potential explanation for this relationship between diabetes, hypertension, major cardiovascular disease, and kidney disease is that hypertension affects the flow of blood through the body. Hypertension specifically can cause blood vessels to constrict and narrow causing high blood pressure [5]. Major cardiovascular disease can be a consequence of hypertension due to the damaging of the blood vessels, where blood clots may form and cause arterial and aortic damage. If blood vessel damage occurs in the kidneys, the kidneys may have difficulty performing their functions and may lead to inability to remove waste from the blood stream and further increasing blood pressure. Diabetes can also affect the blood vessels through the

Fig 18. Linear projection of data from scatter plot in Figure 16



build-up of plaque in arteries which may lead to blood clots [6]. If Covid-19 had a means of causing an increase in blood pressure in those with diabetes, it may be possible that the Covid-19 infection could potentially lead to the development of hypertension, major cardiovascular, and kidney disease. On the other hand, it is also possible that some with diabetes had hypertension before contracting Covid-19 and such predisposition ultimately led to the development of the other comorbidities. Regardless, there is possibility that Covid-19 infection with diabetes may contribute to the development of comorbidities.

V. Conclusions

Throughout the course of this paper, the k-means and hierarchical clustering methods were used to cluster and identify potential comorbidities associated with the Covid-19 infection and diabetes diagnosis. Consequentially, it was determined that 27%-30% of the instance data indicated that higher Covid-19 and diabetes deaths were associated with higher deaths of either hypertension, major cardiovascular disease, or kidney disease. This data also demonstrates that there may be some underlying relationship between having Covid-19 and diabetes and the comorbidities that caused a significant number of deaths. To fully elucidate these findings, further studies on Covid-19 and diabetes with each of the comorbidities need to be performed from a medical level. These studies would help in the understanding of whether Covid-19 and diabetes can cause other comorbidities to develop.

References

[1] U.S. Department of Health & Human Services, "Covid-19: People with Certain Medical Conditions", Centers for Disease Control and Prevention, May 13, 2021. [Online]. Available: https://www.cdc.gov/coronavirus/2019-ncov/need-extra-precautions/people-with-medical-conditions.html. [Accessed: May 29, 2021].

[2] Data.Gov, *AH Provisional Diabetes Death Counts, 2020*, Atlanta, GA: Centers for Disease Control and Prevention, 2020.[Dataset]. Available: https://catalog.data.gov/dataset/ah-provisional-diabetes-death-counts-2020. [Accessed: May 29, 2021]

[3] U.S. Department of Health & Human Services, "Public Health Surveillance and Data: Understanding Death Data", Centers for Disease Control and Prevention, August 24, 2018. [Online]. Available: https://www.cdc.gov/surveillance/projects/understanding-death-data.html. [Accessed: May 29, 2021].

[4] U.S. Department of Health & Human Services, "Covid-19: Older Adults at greater risk of requiring hospitalization or dying if diagnosed with COVID-19", Centers for Disease Control and Prevention, May 14, 2021. [Online]. Available: https://www.cdc.gov/coronavirus/2019-ncov/need-extra-precautions/older-adults.html#:~:text=The%20greatest%20risk%20for%20severe,they%20may%20even%20die. [Accessed: May 29, 2021].

[5] National Institute of Diabetes and Digestive and Kidney Diseases, "High Blood Pressure & Kidney Disease", National Institute of Diabetes and Digestive and Kidney Diseases, March, 2020. [Online]. Available: https://www.niddk.nih.gov/health-information/kidney-disease/high-blood-pressure. [Accessed: May 31, 2021].

[6] American Heart Association, "Understand Your Risk for Excessive Blood Clotting", American Heart Association, 2021. [Online]. Available: https://www.heart.org/en/health-topics/venous-thromboembolism/understand-your-risk-for-excessive-blood-clotting. [Accessed: May 31, 2021].