Informatics and Technology in Clinical Care and Public Health
J. Mantas et al. (Eds.)

doi:10.3233/SHTI210874

Latent COVID-19 Clusters in Patients with Opioid Misuse

Fatemeh SHAH-MOHAMMADI[a,1], Wanting CUI[a], Keren BACHI[a], Yasmin HURD[a] and Joseph FINKELSTEIN[a]
[a]*Icahn School of Medicine at Mount Sinai, New York, NY, USA*

Abstract. The goal of this paper is to apply unsupervised machine learning techniques in order to discover latent clusters in patients who have opioid misuse and also undergone COVID-19 testing. Target dataset has been constructed based on COVID-19 testing results at Mount Sinai Health System and opioid treatment program (OTP) information from New York State Office of Addiction Service and Support (OASAS). The dataset was preprocessed using factor analysis for mixed data (FAMD) method and then K-means algorithm along with elbow method were used to determine the number of optimal clusters. Four patient clusters were identified among which the fourth cluster constituted the maximum percentage of positive COVID-19 test results (20%). Compared to the other clusters, this cluster has the highest percentage of African Americans. This cluster has also the highest mortality rate (16.52%), hospitalization rate after receiving the COVID-19 test result (72.17%, use of ventilator (7.83%) and ICU admission rate (47.83%). In addition, this cluster has the highest percentage of patients with at least one chronic disease (99.13%) and age-adjusted comorbidity score more than 1 (83.48%). Longer participation in OTP was associated with the highest morbidity and mortality from COVID-19.

Keywords. Cluster analysis, COVID-19, Opioid Treatment Program

1. Introduction

In recent years, opioid use disorder has become a significant public health problem in the United States, leading to thousands of death. According to CDC, prescription opioid and heroin overdose deaths have been increasing since 1999 [1]. Opioid abusers are exposed to high risk of not only contracting infectious disease, but also development of mental illness [2]. In addition to health concerns, opioid crisis also creates serious financial costs. In 2009, the annual costs of prescription and illicit opioid abuse, including lost productivity and health care costs, were estimated to be over $55 billion [3]. Methadone and buprenorphine has been reported as effective treatments for opioid dependence, and their widespread use could mitigate the negative health and societal effects of opioid use disorder [4]. Opioid Treatment Programs (OTPs) are among the licensed providers of medication for opioid abusers, and usually require patients to take medication at a clinic.

Novel Coronavirus Disease 2019 (COVID-19) caused a worldwide pandemic outbreak and resulted in large number of infections and million deaths worldwide [5].

[1] Corresponding Author, Fatemeh Shah-Mohammadi, Icahn School of Medicine at Mount Sinai, 1770 Madison Ave, 2nd Fl, New York, NY, USA, 10035; E-mail: fatemeh.shah-mohammadi@mountsinai.org.

Recent studies were able to identify latent COVID-19 clusters in general population [7] as well as in people with chronic respiratory conditions [10]. Identification of latent COVID-19 subgroups can help prioritize COVID-19 screening and treatment [7, 10]. In our previous study [11] we also demonstrated that socio-demographic and OTP characteristics can be successfully used for OTP machine learning analytics. The aim of this study was to identify the latent clusters of patient characteristics who have enrolled in OTP and also tested for COVID-19 using an unsupervised machine learning approach. This approach is also critical for comprehensive understanding of the COVID-19 risk factors for patients with opioid misuse.

2. Methods

Two datasets (AIMS and a COVID-19) were merged to generate an analytical dataset. AIMS is a unique dataset that contains patients' demographic information, employment status, admission, transfer and discharge records, and the substance usage information from the MSHS opioid treatment programs of the New York State area with the size of 158,989 records for 22,044 unique patients. COVID-19 dataset was generated by querying electronic health records at Mount Sinai Health System (MSHS) in New York to identify all patients who underwent COVID-19 testing between March 2020 and July 2021. The data in AIMS was collected from the New York State Office of Addiction Service and Supports' (OASAS), OTP. We only considered the information for the patients who received treatments at MSHS. COVID-19 dataset contains 704,223 records of 388,981 unique patients. Some patients tested multiple times for COVID-19. These patients have been considered as tested negative only if all their records show negative results. From AIMS, we considered the variables "Sex", "Race", "Admission date", "Discharge date", "Date of Birth", "Employment status" and "Secondary abuse substance" which showing the gender, race, the date that the patient gets admitted to the OTP, discharge date from the program, birth date of the patient, whether the patient is employed and what the patient secondary abuse substance is, respectively. Since primary abuse substance for majority of patients (more than 97%) in all clusters was heroin, we did not consider it as a variable in our analysis. Discharge date and the admission date from this dataset was used to calculate new variable "Length of Enrollment in OTP". To calculate this new variable, the value for both dates needs to be valid. Since unknown values for discharge date conveys the fact that the patient is still enrolled in the OTP, this value has been replaced with the most recent admission date in the AIMS. For the patients who admitted in the program multiple times with different discharge dates, we calculated the time between the first time's admission date and the most recent discharge date. Variables in COVID-19 dataset are as follow: hospitalization date, admission to ICU, ICD-10 codes, use of ventilator, alive indicator, COVID-19 test result and the date that the patient received the test result. Patient's age was calculated based on subtracting COVID-19 test result delivery date and the birth date. In addition, the age-adjusted comorbidity score was calculated based on patients' medical history using ICD-10 codes and patient's age [12]. We also considered a variable named "Hospitalized after COVID-19 test" that shows whether the patient was hospitalized after receiving COVID-19 test result. This variable was calculated based on comparing the patient's hospitalization date and the date when patient received the result for COVID-19 test. The variable "Presence of a Chronic Condition", showing whether a patients has a history of chronic disease, was also added and identified based on the ICD10 codes [13]. After merging two datasets

(finding the patients who had records in both datasets) and eliminating patients with missing values, the number of unique patients was reduced to 866 patients. Among 866 patients, 105 patients tested positive for COVID-19.

We used factor analysis for mixed data (FAMD) method for preprocessing [8]. It should be used when we have mixed variables (categorical as well as numerical variables). To discover the latent clusters, we applied K-means clustering on preprocessed data (we only selected the first 2 principal components). Optimal number of clusters were found through cluster analysis. To be more exact, we ranged the number of clusters from 2 to 20, and for each number of clusters we performed K-means and calculated within cluster sum of squares (WCSS). We plotted the WCSS against the number of clusters and used the elbow method to determine optimal number of clusters. All analysis was performed in Anaconda Jupyter Notebook, using Python 3.9.0.

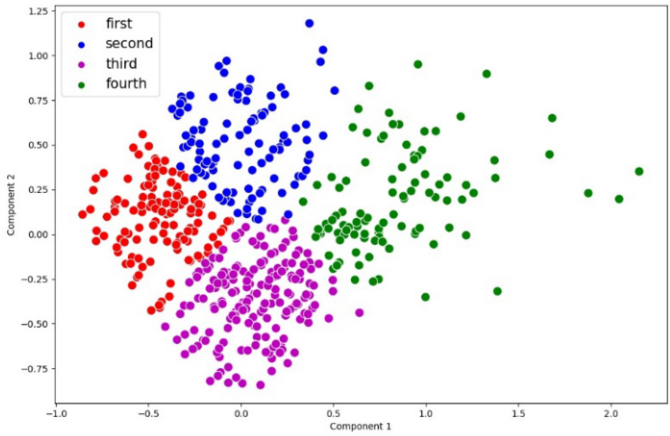

Figure 1. Visualization of the 4 clusters based on the first two principal components

3. Results

Four clusters were identified (Figure 1). According to Table. 1, there was a significantly larger amount of male patients compared to the female patients in all clusters. The largest percentage of positive COVID-19 result cases (20%) belongs to the cluster # 4, while the lowest (around 4%) comes from cluster # 2. More than 50% of patients in cluster # 2 have no comorbid conditions. Compared to the patients mostly in their adulthood in cluster # 2, patients in cluster # 4 were generally in their older adulthood with almost three times more "Length of Enrollment in OTP". More than 80% of patients in cluster # 4 have age-adjusted comorbidity score more than 1 while this value is around 7% for cluster # 2. All of the patients in both clusters have been hospitalized at least once among which around 73% of hospitalizations in cluster # 4 occurred after receiving the COVID-19 test result while this number is 40% for cluster # 2. The highest percentage of patients admitted in ICU belongs to the cluster # 4 (around 50%). This cluster has the highest percentage of patients with chronic disease (99.13%), the highest percentage of death

(16.52%) and the longest participation in OTP. Moreover, most of the patients in this cluster have African American and Hispanic race (around 70%) and are unemployed (46.15%).

4. Discussion

Among the selected variables for the analysis "Race", "Age-Adjusted Comorbidity Score", "Post-Testing Status", "COVID-19 Testing Result", "ICU Status", "Hospitalized after COVID-19 test" and "Age" were important factors in separation of the clusters. Since the highest percentage of mortality, ICU admission, hospitalization after COVID-19 test and the use of ventilator belongs to the cluster with the largest percentage of positive COVID test (i.e. cluster # 4), patients in this cluster need more rigorous testing and treatments. Percentage of patients with comorbidity score more than 1 is the highest for this cluster. On the other hand, more than 50% of patients in the cluster with the lowest percentage of positive COVID-19 cases (i.e. cluster # 2) have no comorbidities. These findings confirm that for OTP patients comorbidity condition can be considered as the COVID-19 risk factor. Moreover, the highest percentage of African Americans and Hispanics (70.44) are in cluster # 4 while the same races constitute the lowest percentage (28.24%) in cluster # 2. The cluster # 3&4 are also represented by the longest duration in OTP.

Our results are congruent with previous reports which used similar clustering techniques for patients with chronic respiratory conditions and pregnant women [7,9].

Table 1. Descriptive statistics o clusters

Clusters	**1**	**2**	**3**	**4**
Count	239	177	335	115
Age				
Young	20.08%	24.86%	0.00%	0.00%
Adult	77.41%	73.45%	13.73%	20.00%
Older adult	2.51%	1.69%	86.27%	80.00%
Sex				
Female	29.71%	20.34%	26.87%	21.74%
Male	70.29%	79.66%	73.13%	78.26%
Race				
Black / African American	7.98%	9.03%	38.51%	40.87%
White / Non-Hispanic	51.26%	54.24%	14.93%	15.65%
Hispanic	26.89%	19.21%	31.04%	29.57%
Other	13.87%	17.51%	15.52%	13.91%
Obesity				
Obese	17.99%	25.99%	18.51%	26.09%
No obese	82.01%	74.01%	81.49%	73.91%
Hospitalized (ever) - Yes	9.62%	100.00%	30.75%	100.00%
Presence of a Chronic Condition-Yes	68.20%	99.00%	82.39%	99.13%
Age Adjusted Comorbidity Score				
0	51.88%	53.67%	0.00%	0.00%
1	46.44%	39.55%	23.88%	16.52%
2	1.68%	6.78%	48.36%	37.45%
3	0.00%	0.00%	20.90%	38.34%
4	0.00%	0.00%	5.96%	6.82%
5	0.00%	0.00%	0.90%	0.87%
Length of Enrollment in OTP				
Mean number of days	981.23	612.91	2304.66	1826.76
Employment				
Employed	12.28%	12.41%	15.79%	17.31%
Not in labor force	13.16%	13.87%	29.47%	36.54%
Unemployed	74.56%	73.72%	54.74%	46.15%

Secondary Substance Use				
Alprazolam(Xanax) and Benzodiazepine	6.58%	6.57%	2.45%	1.92%
Opioid-based substance	12.73%	5.84%	5.26%	9.61%
None	71.05%	65.69%	84.21%	82.69%
COVID-19 Testing Result-Positive	14.17%	3.95%	14.03%	20.00%
Hospitalized after COVID-19 Test- Yes	0.00%	40.00%	0.60%	72.17%
On Ventilator - Yes	0.00%	0.00%	0.00%	7.83%
ICU Status -Yes	0.00%	6.00%	0.00%	47.83%
Post-Testing Vital Status				
Alive	100.00%	98.87%	100%	83.48%
Deceased	0.00%	1.13%	0.00%	16.52%

5. Conclusion

Four clusters have been identified in patients who participated in OTPs and underwent COVID-19 testing. K-means algorithm is used for cluster analysis in order to determine the number of optimal clusters. The largest percentage of COVID-19 positive cases belonged to the fourth cluster with patients mostly unemployed and in their older adulthood with the highest percentage of African American and Hispanic races. This conveys the fact that the COVID-19 pandemic's impact has more exposed race inequities. This cluster has the highest percentage of hospitalization rate after COVID test, death rate, ventilator usage, ICU admissions and comorbidity score more than 1. Due to the highest death rate, patients in this cluster need more rigorous treatments.

References

[1] Centers for Disease Control and Prevention, National Center for Health Statistics. Multiple Cause of Death 1999-2017 on CDC WONDER Online Database, 2018.
[2] Ehrich E, Turncliff R, Du Y, et. al., Evaluation of opioid modulation in major depressive disorder, Neuropsychopharmacology, 2015;40:1448-1455.
[3] Birnbaum, Howard G, et. al., Societal costs of prescription opioid abuse, dependence, and misuse in the United States, Pain medicine, 2011;12:657-667.
[4] Mattick RP, Kimber J, Breen C, Davoli M, Buprenorphine maintenance versus placebo or methadone maintenance for opioid dependence, Cochrane Database Syst Rev., 2008.
[5] WHO-China Joint Mission, Report of the WHO-China Joint Mission on Coronavirus Disease 2019 (COVID-19), 2020.
[6] Centers for Disease Control and Prevention. Coronavirus disease 2019 (COVID-19): cases in US, Accessed April 24, 2020. https://www.cdc.gov/coronavirus/2019-ncov/cases-updates/cases-in-us.html
[7] Cui W, Robins D, Finkelstein J. Unsupervised Machine Learning for the Discovery of Latent Clusters in COVID-19 Patients Using Electronic Health Records. Stud Health Technol Inform. 2020;272:1-4.
[8] Mori Y., Kuroda M., Makino N. Multiple Correspondence Analysis. In: Nonlinear Principal Component Analysis and Its Applications. Springer Briefs in Statistics. Springer, Singapore. 2016.
[9] Qi M, Li X, Liu S, Li Y, Huang W., Impact of the COVID-19 epidemic on patterns of pregnant women's perception of threat and its relationship to mental state: A latent class analysis, PLoS One. 2020; 15.
[10] Cui, Wanting et al., Latent COVID-19 Clusters in Patients with Chronic Respiratory Conditions, Studies in health technology and informatics. 2020;275:32-36.
[11] Cui W, Bachi K, Hurd Y, Finkelstein J. Using Big Data to Predict Outcomes of Opioid Treatment Programs. Stud Health Technol Inform. 2020;272:366-369.
[12] Sundararajan V, Henderson T, Perry C, Muggivan A, Quan H, Ghali WA. New ICD-10 version of the Charlson comorbidity index predicted in-hospital mortality. J Clin Epidemiol. 2004;57(12):1288-1294.
[13] Healthcare Cost and Utilization Project (HCUP). Chronic Condition Indicator for ICD-10-CM, v2021.1. Agency for Healthcare Research and Quality, October 2020.