Integrated Citizen Centered Digital Health and Social Care
A. Värri et al. (Eds.)

doi:10.3233/SHTI200689

Latent COVID-19 Clusters in Patients with Chronic Respiratory Conditions

Wanting CUI[1a], Manuel CABRERA[b] and Joseph FINKELSTEIN[a]
[a] *Icahn School of Medicine at Mount Sinai, New York, NY, USA*
[b] *Columbia University Irving Medical Center, NY, USA*

Abstract. The goal of this paper was to apply unsupervised machine learning techniques towards the discovery of latent COVID-19 clusters in patients with chronic lower respiratory diseases (CLRD). Patients who underwent testing for SARS-CoV-2 were identified from electronic medical records. The analytical dataset comprised 2,328 CLRD patients of whom 1,029 were tested COVID-19 positive. We used the factor analysis for mixed data method for preprocessing. It performed principle component analysis on numeric values and multiple correspondence analysis on categorical values which helped convert categorical data into numeric. Cluster analysis was an effective means to both distinguish subgroups of CLRD patients with COVID-19 as well as identify patient clusters which were adversely affected by the infection. Age, comorbidity index and race were important factors for cluster separations. Furthermore, diseases of the circulatory system, the nervous system and sense organs, digestive system, genitourinary system, metabolic diseases and immunity disorders were also important criteria in the resulting cluster analyses.

Keywords. Chronic lower respiratory diseases, cluster analysis, COVID-19

1. Introduction

Chronic lower respiratory diseases (CLRD) comprise heterogeneous chronic airway disorders that consist of multiple phenotypes with diverse clinical characteristics [1, 2]. Unsupervised machine learning has been successfully used in CLRD to identify latent clusters in such conditions as asthma [1] and chronic obstructive pulmonary disease [2]. Cluster analysis allowed to identify subgroups of COVID-19 patients with differing risk factors, comorbidities, and prognosis using electronic health records (EHR) [3]. No unsupervised learning approach has been undertaken to identify COVID-19 latent clusters in patients with CLRD. The goal of this study is to conduct cluster analysis of EHR data of CLRD patients who were tested for presence of severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2). A broad approach to discover latent clusters is critical for a comprehensive understanding of the COVID-19 risk factors in CLRD.

[1] Wanting Cui, Icahn School of Medicine at Mount Sinai, 1770 Madison Ave, 2nd Fl, New York, NY, USA, 10035, E-mail: wanting.cui@mssm.edu

2. Method

The initial dataset was generated by querying electronic health records at Mount Sinai Health System in New York to identify all patients who underwent SARS-CoV-2 testing between January 2020 and April 2020. The initial dataset contained 19,588 patients with 8,559 tested positive and 11,029 tested negative. The analytical dataset for this study was de-identified and comprised 2,422 patients over 18 years old with CLRD based on presence of ICD-10 codes in the range of J40 – J47. We further eliminated patients with missing values, the final dataset included 2,328 CLRD patients of whom 1,029 were tested positive. Variables in the dataset included age, sex, race, ethnicity, ICU status, alive indicator and COVID-19 status. We added comorbidity index and 18 body systems based on patients' medical history using ICD-10 codes [4]. A body system was positive if a patient has one or more diagnoses related to this system and was negative if a patient has no diagnosis of this system. In addition, the age-adjusted comorbidity index was calculated using patient's age and ICD-10 code of diagnoses [5].

We divided our study into 2 subsets: all CLRD patients and CLRD patients who tested positive for SARS-CoV-2. For each subset of patients we performed data processing and cluster analysis.

We used the factor analysis for mixed data (FAMD) method for preprocessing. It performed principle component analysis (PCA) on numeric values and multiple correspondence analysis (MCA) on categorical values which would help convert categorical data into numeric [6]. In PCA, we scaled all numeric variables between 0 and 1. In MCA, all categorical variables were converted into dummy variables. A dummy variable was a numeric variable that represents categorical data. If a variable had n levels, we expanded the one variable into (n-1) new variables and used a Boolean value to indicate this. FAMD was also good at reducing multi-collinearity issues between variables and achieving dimension reduction. It extracted features, emphasized variation and combined input variables in specific ways. It allowed us to drop the least important information, while still retaining trends and patterns.

The K-means algorithm was used for clustering. Cluster analysis ranged from 2 clusters to 18 clusters, because the number of clusters needed to be determined prior to running the algorithm. We calculated the Within Cluster Sum of Squares (WCSS) which was the sum of squares of the distances of each data point in all clusters to their respective centroids. We plotted WCSS against the number of clusters and used the elbow method to determine optimal number of clusters.

All analyses were performed in Anaconda Jupyter Notebook, using Python 3.7.3.

3. Results

First of all, in the subset of all CLRD patients, 2,328 patients were included and 3 clusters were found (Figure 1). The number of patients distributed evenly among the 3 clusters. According to Table 1, there was a significantly greater amount of female CLRD patients than male CLRD patients in all groups. Over 99% of patients in cluster 0 tested positive for COVID-19, while almost all patients in cluster 2 tested negative for COVID-19. Cluster 1 was a mixture of COVID-19 positive and negative patients; however, patients in this group were generally younger with less comorbidities. In contrast, patients in cluster 0 were the oldest and had the highest comorbidity index. The average age of this group was 66 years old and the comorbidity index was 5.3. In addition, these patients

had the most severe symptoms as they had the highest death rate (23.3%), ICU admission rate (16.35%), percent use of ventilator (8.39%), hospitalization rate (60.35%) and the longest ICU length of stay (1.14 days).

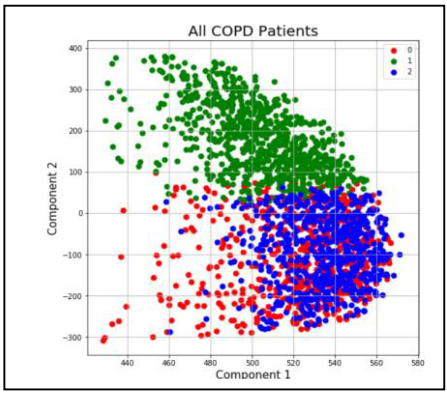

Figure 1. Clustering of all COPD patients.

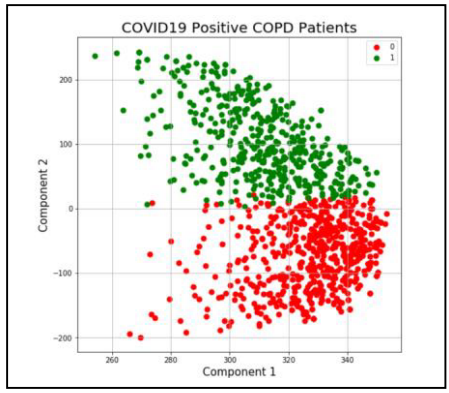

Figure 2. Clustering of COVID-19 positive patients

To further analyze the effects of COVID-19 on CLRD patients, we performed clustering analysis on CLRD patients who tested positive for COVID-19. 1,029 patients were included and 2 clusters were found (Figure 2). Patients in cluster 0 had more serious conditions when infected compared to those in cluster 1. They were older (65.78 years) and had significantly more comorbidities (5.41). In addition, there were significantly more African American patients and significantly less White patients in cluster 0 than those in cluster 1.

Table 1. Descriptive statistics of clusters (SD – standard deviation)

Subsets	**All COPD Patients**				**COVID19 Positive COPD Patients**	
Clusters	0	1	2		0	1
Count	691	881	756		583	446
Numeric Variables						
AGE						
Mean	65.99	52.47	60.25		65.78	58.85
SD	15.62	19.15	16.87		15.72	18.62
Comorbidity Index						
Mean	5.30	2.74	5.14		5.41	3.39
SD	3.03	2.08	3.53		3.10	2.27
ICU Length						
Mean	1.14	0.24	0.02		1.13	0.74
SD	3.63	1.80	0.32		3.51	3.25
Categorical Variables						
Status						
Alive	76.70%	93.42%	97.62%		78.73%	82.74%
Deceased	23.30%	6.58%	2.38%		21.27%	17.26%
Sex						
Female	59.91%	55.16%	67.72%		62.95%	48.88%
Male	40.09%	44.84%	32.28%		37.05%	51.12%
Race						

American Indian or Alaskan	0.00%	0.00%	0.13%		0.00%	0.00%
Asian	4.05%	3.29%	3.57%		3.60%	4.71%
Black	30.25%	27.47%	27.91%		31.56%	22.65%
Islander	1.45%	2.16%	1.19%		1.37%	2.02%
Other	44.14%	37.80%	37.70%		43.74%	43.05%
White	20.12%	29.28%	29.50%		19.73%	27.58%
Ethnicity						
Hispanic	2.89%	0.68%	4.10%		3.09%	0.67%
Not Hispanic	97.11%	99.32%	95.90%		96.91%	99.33%
On Ventilator	8.39%	3.41%	1.59%		7.89%	6.95%
COVID19 Positive	99.42%	38.37%	0.53%		100.00%	100.00%
ICU	16.35%	3.97%	0.40%		16.47%	10.99%
HOSPITAL	60.35%	34.17%	30.03%		61.06%	38.57%

In body systems (Table 2), over 95% of COVID-19 positive patients had endocrine, nutritional and metabolic diseases and immunity disorders. In addition, around 90% of CLRD patients with COVID-19 had diseases of the circulatory system. Furthermore, patients with diagnoses in sense organs, digestive system and genitourinary system were more likely to have serious complications when infected.

Table 2. Percentage affected based on body systems

	All COPD Patients				**COVID19 Positive COPD Patients**	
Body System	0	1	2		0	1
1. Infectious and parasitic disease	92.33%	42.34%	65.87%		92.97%	70.40%
2.Neoplasms	41.39%	11.24%	51.85%		46.31%	11.21%
3. Endocrine, nutritional, and metabolic diseases and immunity disorders	95.22%	41.20%	91.67%		95.71%	57.85%
4. Diseases of blood and blood-forming organs	57.60%	12.94%	55.42%		62.09%	16.37%
5. Mental disorders	55.86%	26.22%	63.89%		63.46%	17.49%
6. Diseases of the nervous system and sense organs	85.96%	28.83%	87.83%		90.22%	37.89%
7. Diseases of the circulatory system	89.00%	37.46%	83.73%		88.51%	56.28%
8. Diseases of the respiratory system	100%	100%	100%		100%	100%
9. Diseases of the digestive system	75.98%	24.63%	82.94%		83.88%	23.09%
10. Diseases of the genitourinary system	79.16%	29.63%	77.65%		81.99%	38.34%
11. Complications of pregnancy, childbirth, and the puerperium	2.60%	10.22%	8.60%		2.74%	4.04%
12. Diseases of the skin and subcutaneous tissue	56.15%	14.98%	67.20%		63.81%	13.23%
13. Diseases of the musculoskeletal system	86.54%	32.92%	90.61%		91.77%	37.44%
14. Congenital anomalies	9.99%	2.27%	8.20%		10.46%	2.91%
15. Certain conditions originating in the perinatal period	0.58%	0.34%	0.66%		0.51%	0.22%
16. Symptoms, signs, and ill-defined conditions	99.57%	83.09%	99.21%		99.83%	90.36%
17. Injury and poisoning	63.24%	19.86%	67.59%		69.64%	17.94%
18. Factors influencing health status and contact with health services	95.66%	62.09%	98.81%		97.60%	62.11%
Body System “None”	22.00%	5.33%	23.28%		24.53%	6.05%

4. Discussion

Around 44% of CLRD patients tested positive for COVID-19, which was similar to the statistic of all patients (45%) at Mount Sinai Health System. In the first part of this study, we found 2 cluster of patients who were older and had higher comorbidity index. However, those patients who had COVID-19 had a 23% death rate, compared to the 2% death rate in the non COVID-19 cluster. In addition, patients with immunity disorders or diseases of the circulatory system were more likely to be subjected to the illness. The second part of the study confirmed that age and comorbidities were crucial factors. Race also emerged as an important part to differentiate seriously ill patients. Patients who developed severe symptoms had significant history of concurrent conditions of the nervous system, digestive system and genitourinary system.

Cluster analysis provided initial insights of COVID-19 subgroups and risk factors in patients with CLRD. This methodology could be applied in the future towards similar studies. Our results are congruent with previous reports which used similar clustering techniques for CLRD phenotyping [7-8].

5. Conclusion

Cluster analysis was an effective means to both distinguish subgroups of CLRD patients with COVID-19 as well as identify patient clusters which were adversely affected by the infection. Age, comorbidity index and race were important factors for cluster separations. Furthermore, diseases of the circulatory system, the nervous system and sense organs, digestive system, genitourinary system, metabolic diseases and immunity disorders were also important criteria in the resulting cluster analyses.

References

[1] Horne E, Tibble H, Sheikh A, Tsanas A. Challenges of Clustering Multimodal Clinical Data: Review of Applications in Asthma Subtyping. JMIR Med Inform. 2020 May;8(5):e16452.
[2] Pikoula M, Quint JK, Nissen F, Hemingway H, Smeeth L, Denaxas S. Identifying clinically important COPD sub-types using data-driven approaches in primary care population based electronic health records. BMC Med Inform Decis Mak. 2019 Apr;19(1):86.
[3] 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 May;272:1-4.
[4] 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 Dec;57(12):1288-94.
[5] Ho CH, Chen YC, Chu CC, Wang JJ, Liao KM. Age-adjusted Charlson comorbidity score is associated with the risk of empyema in patients with COPD. Medicine (Baltimore). 2017 Sep;96(36),e8040.
[6] Mori Y., Kuroda M., Makino N. Multiple Correspondence Analysis. In: Nonlinear Principal Component Analysis and Its Applications. Springer Briefs in Statistics. Springer, Singapore. 2016.
[7] Weatherall M, Shirtcliffe P, Travers J, Beasley R. Use of cluster analysis to define COPD phenotypes. Eur Respir J. 2010 Sep;36(3):472-474.
[8] Bourbeau J, Pinto LM, Benedetti A. Phenotyping of COPD: challenges and next steps. Lancet Respir Med. 2014 Mar;2(3):172-174.