The Importance of Health Informatics in Public Health during a Pandemic
J. Mantas et al. (Eds.)

doi:10.3233/SHTI200478

Unsupervised Machine Learning for the Discovery of Latent Clusters in COVID-19 Patients Using Electronic Health Records

Wanting CUI [a,1], Daniel ROBINS [a] and Joseph FINKELSTEIN [a]
[a] *Icahn School of Medicine at Mount Sinai, New York, NY, USA*

Abstract. The goal of this paper was to apply unsupervised machine learning techniques towards the discovery of latent clusters in COVID-19 patients. Over 6,000 adult patients tested positive for the SARS-CoV-2 infection at the Mount Sinai Health System in New York, USA met the inclusion criteria for analysis. Patients' diagnoses were mapped onto chronicity and one of the 18 body systems, and the optimal number of clusters was determined using K-means algorithm and the elbow method. 4 clusters were identified; the most frequently associated comorbidities involved infectious, respiratory, cardiovascular, endocrine, and genitourinary disorders, as well as socioeconomic factors that influence health status and contact with health services. These results offer a strong direction for future research and more granular analysis.

Keywords. Big Data Analytics, Unsupervised Machine Learning

1. Introduction

Coronavirus Disease (COVID-19) has spread rapidly throughout North America, and so observations and data about this illness on the American population are now available. New York, in particular, has become the largest single focus of confirmed cases in the USA [1], and it is effectively the American epicenter of the pandemic since March 2020. The large, high-density population of NY has led to the growth of an even larger medical record dataset, which must be methodically analyzed to produce actionable information. Early efforts to find patient characteristics and risk factors are already underway; one case series recently demonstrated that hypertension, obesity, and diabetes are the most frequent comorbidities found in laboratory confirmed COVID patients that require hospitalization [2]. However, a broad approach to discover *latent* clusters is critical for a comprehensive understanding of the course of illness. The goal of this paper is to identify these latent clusters of patient characteristics by using an unsupervised machine learning approach.

[1] Corresponding Author, 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. Methods

The dataset was pulled from Epic, an electronic health record at Mount Sinai Health System between January 2020 and April 2020, and de-identified for further analysis. This analysis aimed to identify latent clusters from patients who contracted coronavirus, and thus only records of patients who tested positive for SARS-CoV-2 were included. The initial analytical dataset comprised 6220 eligible patients. After deletion of records with missing values the final dataset included 6101patients.

Since each patient could have multiple underlying comorbidities, we mapped diagnoses into chronicity indicator and one of the 18 body systems, based on ICD-10 codes [3]. If patients had more than one diagnosis, chronicity and body systems would be the aggregated information of all diagnoses. Chronicity was positive if one or more chronic diseases were detected and negative if no chronic diseases were detected. The same rule applied to all 18 body systems. In addition, the age-adjusted comorbidity index was calculated using patient's age and ICD-10 code of diagnoses [4]. In the end, age, comorbidity, ICU status, alive, sex, race, ethnicity, chronicity and body systems were the variables we used for clustering analysis. All numerical variables were normalized between 0 and 1; and all categorical variables were changed into dummy variables. We performed PCA to address the multi-colinearity issues between variables. 15 components were selected, as they explained over 80% of variance. Important variables were chronicity, body systems (3, 7, 10, and 18), age and race. K-means algorithm was used for clustering and the elbow method was used to determine optimal number of clusters. All analyses were performed in Anaconda Jupyter Notebook, using Python 3.7.3. T-test, proportion testing were used. All tests were two-sided, with $p<0.05$ being considered statistically significant.

3. Results

4 clusters were identified (Figure 1). Clusters 0 and 2 constituted patients who were sicker and had more conditions (Table 1). Clusters 1 and 3, constituted patients with less comorbidities. Over 97% of patients in clusters 0 and 2 had one or more chronic diseases. In contrast, only 15% of patients from cluster 1 and 3 had chronic diseases. The average ages of clusters 0 and 2 were 10 years older than clusters 1 and 3. In addition, patients from cluster 2 had a 20% death rate, whereas the death rate of cluster 3 was around 5%. Race was also an important factor in forming the clusters.

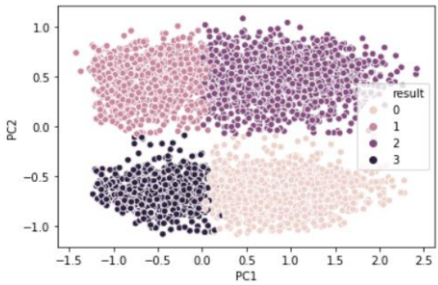

Figure 1. Visualization of the 4 clusters based on the first 2 principle components

Age and comorbidity index were significantly different between clusters 0, 2 and 1; but not between clusters 1 and 3. ICU length of stay were all significantly different across all four clusters. Patients in clusters 0 and 2 had high comorbidity index, whereas patients in cluster 1 and 3 had low comorbidity index. Although patients in cluster 2 had a shorter ICU length of stay and percentage use of ventilators, comparing to cluster 0, these patients had higher comorbidity index and had 3% higher death rate.

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

Clusters	0	1	2	3
count	1153	1959	1572	1417
Numeric Variables				
Age				
Mean	64.54	51.51	67.95	50.29
SD	15.62	17.97	14.68	17.20
Comorbidity				
Mean	3.12	1.20	3.45	1.09
SD	2.17	1.43	2.12	1.34
ICU Length				
Mean	1.73	0.15	1.34	0.38
SD	3.79	1.03	3.36	1.54
Categorical Variables				
STATUS				
Alive	82.74%	94.38%	79.71%	94.71%
Deceased	17.26%	5.62%	20.29%	5.29%
SEX				
Female	44.32%	49.52%	44.59%	42.48%
Male	55.68%	50.48%	55.41%	57.52%
RACE				
American Indian or Alaskan	0.00%	0.05%	0.13%	0.00%
Asian	0.69%	7.55%	7.25%	0.85%
Black	0.00%	38.74%	44.78%	0.00%
Pacific Islander	0.26%	6.43%	4.45%	0.14%
Other	99.05%	0.00%	0.00%	99.01%
White	0.00%	47.22%	43.38%	0.00%
ETHNICITY				
Hispanic	3.56%	0.05%	0.25%	0.71%
Not Hispanic	96.44%	99.95%	99.75%	99.29%
On Ventilator	16.22%	1.79%	10.50%	3.53%
ICU	31.40%	4.44%	23.47%	10.16%
Chronicity	96.96%	14.04%	97.77%	15.74%

In terms of body systems, all patients had relatively higher proportion of heath conditions in body systems 1, 8, 16 and 18 (Table 2). However, patients from cluster 0 and 2 exhibited higher proportion of conditions in body systems 3, 7, 10 and 18. There were significant differences between cluster 0 and cluster 2 in representation of health conditions in body systems3, 16 and body system "none."

Table 2. Percentage affected of body systems

	Cluster			
Body System	0	1	2	3
1. Infectious and parasitic disease	74.41%	40.07%	75.51%	46.93%
2.Neoplasms	12.40%	2.65%	11.83%	1.62%
3. Endocrine, nutritional, and metabolic diseases and immunity disorders	70.42%	3.52%	66.28%	4.80%
4. Diseases of blood and blood-forming organs	20.21%	1.28%	18.70%	1.91%
5. Mental disorders	13.79%	2.76%	15.20%	2.54%

6. Diseases of the nervous system and sense	30.27%	2.19%	29.01%	2.33%
7. Diseases of the circulatory system	65.05%	1.53%	68.19%	2.40%
8. Diseases of the respiratory system	83.87%	57.58%	85.05%	60.69%
9. Diseases of the digestive system	20.03%	3.06%	18.26%	3.60%
10. Diseases of the genitourinary system	47.88%	5.00%	45.29%	4.45%
11. Complications of pregnancy, childbirth, and the puerperium	0.43%	3.62%	0.57%	2.12%
12. Diseases of the skin and subcutaneous	6.68%	1.53%	7.32%	1.13%
13. Diseases of the musculoskeletal system	20.21%	4.34%	20.99%	4.16%
14. Congenital anomalies	1.39%	0.05%	1.46%	0.07%
15. Certain conditions originating in the	0.17%	0.26%	0.13%	0.00%
16. Symptoms, signs, and ill-defined conditions	87.94%	54.77%	90.71%	59.14%
17. Injury and poisoning	12.75%	2.40%	12.53%	2.33%
18. Factors influencing health status and contact with health services	48.40%	21.34%	46.31%	16.51%
Body System "None"	2.17%	0.92%	3.94%	0.21%

4. Discussion

The majority of patients who tested positive for COVID-19 had respiratory symptoms. Patients with high comorbidity scores and chronic diseases were affected most by COVID-19. Age was also an important factor, but it is highly correlated to comorbidity and chronicity. In addition, patients with a history of metabolic diseases and immune system disorders, or medical conditions in the circulatory system or genitourinary system were more susceptible to serious complications when infected.

The discovery of these 4 clusters represents an important milestone towards understanding the course of illness, and subsequently optimizing disease prevention and patient care. Future investigations and a more granular analysis should be directed towards these clusters.

5. Conclusions

4 clusters were identified. Clusters 0 and 2 contained patients with the most serious conditions. Age, comorbidity index, chronicity, race and body systems (3, 7, 10 and 18) were important variables in separating these clusters.

References

[1] 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.

[2] Richardson S, et al, Presenting Characteristics, Comorbidities, and Outcomes Among 5700 Patients Hospitalized With COVID-19 in the New York City Area. JAMA; 2020. [Epub ahead of print].

[3] 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): 1288-94.

[4] 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;96(36): e8040.