Identifying Clinical Subtypes of Long COVID: An Unsupervised Machine Learning Approach

Michael Munsell, PhD[1], Mark Friedman, MD[1], Joseph Menzin, PhD[1]

[1] Panalgo, Boston MA, USA

Background & Objective

- Nearly half of COVID-19 survivors experience at least one unresolved symptom within four months after acute infection, independent of hospitalization status.[1]
- The symptoms of patients with post-acute COVID-19 syndrome, colloquially known as long COVID, are heterogenous, impact multiple systems and are often non-specific and difficult to diagnose.
- This study used real-world data and unsupervised machine learning techniques to identify distinct groups of long COVID patients and describe their underlying symptom profile.

Methods

Data Source

- Optum's de-identified Clinformatics® Data Mart Database (January 1, 2000 through March 31, 2022).

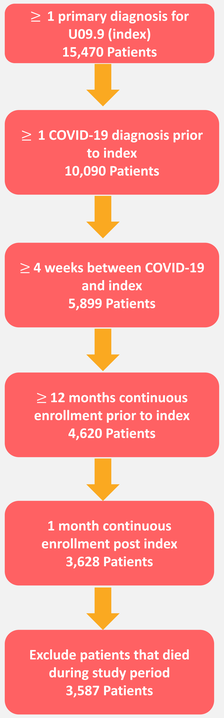
Inclusion and Exclusion Criteria

- Children/adolescents (age 0-17), and adults (18-64 and ≥65 years) were included in the analysis (i.e., no exclusion was made based on age).
- The first primary diagnosis for U09.9 "Post-COVID-19 condition" identified during the time period from 10/01/2021 (ICD-10 code introduction date) through 02/28/2022 was deemed index.
- Patients had ≥1 diagnosis for COVID-19 at least four weeks before index.
- Continuous enrollment during the 12 months prior to index (baseline) and one month post index (follow up) were required.
- Patients that died during the study period were excluded.

Analysis

- ICD-10-CM diagnosis codes recorded ±2 weeks from index that were not present prior to the initial COVID-19 diagnosis were captured. The 100 most prevalent diagnoses were used as patient features for *k*-means clustering.
- The *k*-means algorithm by Hartigan and Wong[2] was used for clustering, where *k* was the number yielding the minimum within-cluster sum of squares for values 2 to 7 within cluster centers randomly chosen for 10 trials.
- Final cluster assignments for further analysis were selected based on silhouette coefficient and clinical relevancy of groupings.
- Diagnosis codes used for clustering were reviewed and grouped based on COVID-19 symptom domain. The prevalence of symptom domains was analyzed descriptively and stratified by cluster assignment.
- All analyses were performed using the Instant Health Data (IHD) software (Panalgo LLC, Boston MA, USA).

Figure 1. Study Cohort Attrition



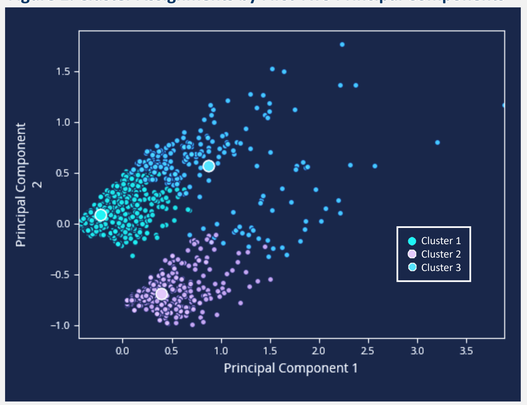
Results

- 3,587 patients met eligibility criteria and were included in the unsupervised clustering analysis.
- A model with three clusters resulted in a silhouette coefficient of 0.146 and was selected for further analysis. Baseline demographics stratified by cluster assignment are presented in Table 1.

Table 1. Baseline Patient Demographics by Clustering Assignment

	Cluster 1	Cluster 2	Cluster 3
N	2,578	651	358
Age at Index			
Mean (SD)	58.26 (17.08)	57.67 (16.05)	66.94 (13.26)
Female Sex, N (%)	1,505 (58.40%)	378 (58.06%)	181 (50.56%)
Region, N (%)			
Midwest	683 (26.52%)	172 (26.42%)	94 (26.26%)
Northeast	283 (10.99%)	65 (9.98%)	28 (7.82%)
South	1,039 (40.35%)	279 (42.86%)	158 (44.13%)
West	570 (22.14%)	135 (20.74%)	78 (21.79%)
Race, N (%)			
Asian	55 (2.29%)	11 (1.82%)	5 (1.48%)
Black	253 (10.53%)	62 (10.25%)	45 (13.35%)
Caucasian	1,770 (73.66%)	446 (73.72%)	238 (70.62%)
Hispanic	325 (13.52%)	86 (14.21%)	49 (14.54%)
Baseline CCI			
Mean (SD)	1.45 (1.83)	1.37 (1.64)	2.68 (2.17)

Figure 2. Cluster Assignments by First Two Principal Components



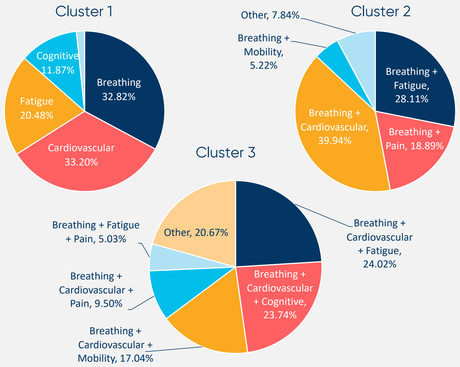
- The majority of patients were grouped into a single cluster (Cluster 1, N=2,578 [71.9%]), with the remaining patients belonging to two smaller groupings (Cluster 2, N=651 [18.1%]; Cluster 3, N=358 [9.9%]; Figure 2).
- Clusters 1 and 2 were similar in terms of age and baseline comorbidity profile, while Cluster 3 was slightly older and had a higher prevalence of Charlson comorbidities prior to index.
- Concurrent conditions recorded at the time of long COVID diagnosis were grouped into six domains based on underlying symptom presentation: symptoms related to breathing, cardiovascular symptoms, neurological symptoms and issues with cognitive health, fatigue, issues with mobility, and pain. The most common ICD-10 code-level diagnoses observed in each domain are listed in Table 2.
- The median number of symptom domains present at the time of index differed across clusters, with the first cluster being characterized by patients with a single symptom domain, the second cluster with two domains and the third cluster with three domains.

Results Cont'd.

Table 2. Identified Symptom Domains Among Patients Receiving a Long COVID Diagnosis

Symptom Domain	Most common ICD-10-CM diagnoses ± 2 weeks of index
Breathing	Pneumonia, respiratory infection, respiratory failure, dyspnea, shortness of breath, dependence on respirator/supplemental oxygen
Cardiovascular	Hypertension, abnormalities of heart-beat, tachycardia, bradycardia, palpitations
Cognitive Health	Anxiety disorders, disorientation, amnesia, attention and concertation deficit, psychomotor deficit, unspecified symptoms involving cognitive function and awareness
Fatigue	Malaise and fatigue, weakness, chronic fatigue
Mobility	Abnormalities of gait and mobility, ataxic gait, paralytic gait, difficulty in walking, unsteadiness on feet
Pain	Pain in throat, chest pain on breathing, precordial pain, pleurodynia, intercostal pain

Figure 3. Prevalence of Symptom Domain Groupings By Cluster



- Two thirds of patients in Cluster 1 had either breathing or cardiovascular symptoms alone; 20% presented with only fatigue and 11% with cognitive symptoms alone.
- All patients in Cluster 2 presented with breathing symptoms accompanied by one additional domain (cardiovascular 40%, fatigue 28%, pain 18%).
- Patients in Cluster 3 experienced breathing symptoms accompanied by two additional domains (fatigue and cardiovascular 24%, cardiovascular and cognitive 24%).

Conclusion

- Unsupervised machine learning identified distinct groups of long COVID patients based on the frequency and combination of symptom domains, which may help inform multidisciplinary care needs.
- Our analysis suggest that many patients with long COVID may experience symptoms from only a single domain, and multisystem illness may generally include breathing complications accompanied by fatigue and/or cardiovascular complications.

References

1. O'Mahoney LL, Routen A, Gillies C, Ekezie W, Welford A, Zhang A, Karamchandani U, Simms-Williams N, Cassambai S, Ardavani A, Wilkinson TJ, Hawthorne G, Curtis F, Kingsnorth AP, Almaqhawi A, Ward T, Ayoubkhani D, Banerjee A, Calvert M, Shafran R, Stephenson T, Sterne J, Ward H, Evans RA, Zaccardi F, Wright S, Khunti K. The prevalence and long-term health effects of Long Covid among hospitalised and non-hospitalised populations: A systematic review and meta-analysis. EClinicalMedicine. 2022 Dec 1;55:101762. doi: 10.1016/j.eclinm.2022.101762. PMID: 36474804; PMCID: PMC9714474.
2. Hartigan, J.A. & Wong, M.A. (1979) Algorithm AS 136 A K Means Clustering Algorithm, Applied Statistics, 28(1), 100-108.

Presented at ISPOR, May 2023, Boston, MA

Funding Source: None