Beyond Acute COVID-19: Identifying Pediatric Post-Acute Subphenotypes Through Topic Modeling

Yishan Shen[1,2,*], MA, Yiwen Lu[1,2,*], Yuqing Lei[1,2,*], MS, Ting Zhou[1,2],MD, Jiayi Tong[1,2], PhD, Christopher B. Forrest[3], MD, PhD, Yong Chen[1,2], PhD

[1] The Center for Health AI and Synthesis of Evidence (CHASE), University of Pennsylvania, Philadelphia, PA, USA,

[2] Department of Biostatistics, Epidemiology and Informatics, University of Pennsylvania, Philadelphia, PA 19104, USA.

[3]Applied Clinical Research Center, Children's Hospital of Philadelphia, Philadelphia, PA 19104, USA

[*]Contributed equally.

Background

Post-acute sequelae of SARS-CoV-2 infection (PASC), or long COVID, involves persistent symptoms that affect various organ systems, impacting daily functioning and increasing healthcare burdens[1–3]. This prolonged condition highlights the need for deepened understanding to guide public health responses and treatment strategies[4,5]. Although extensively studied in adults[6–8], PASC in pediatric populations remains less understood, with data scarcity hindering effective subphenotype classification and treatment. Hence in this study, we investigate the pediatric PASC subphenotypes using a large-scale cohort from the Researching COVID to Enhance Recovery (RECOVER) (https://recovercovid.org) electronic health records (EHR) database including nineteen children's hospitals across the United States. We aim to provide more insights into PASC and generate real-world evidence through large-scale analytics which is particularly relevant to the mission of the OHDSI community. Traditional clustering methods to find subphenotypes of a complex syndrome disease include latent class analysis[9] and principal component analysis (PCA) along with k-means clustering. We leverage advanced data-driven technique – topic modeling[10] to examine the co-incidence patterns of 128 diagnosis categories derived from the Clinical Classifications Software Refined (CCSR) categories that are associated with PASC within 28-179 days following COVID-19 infection.

Methods

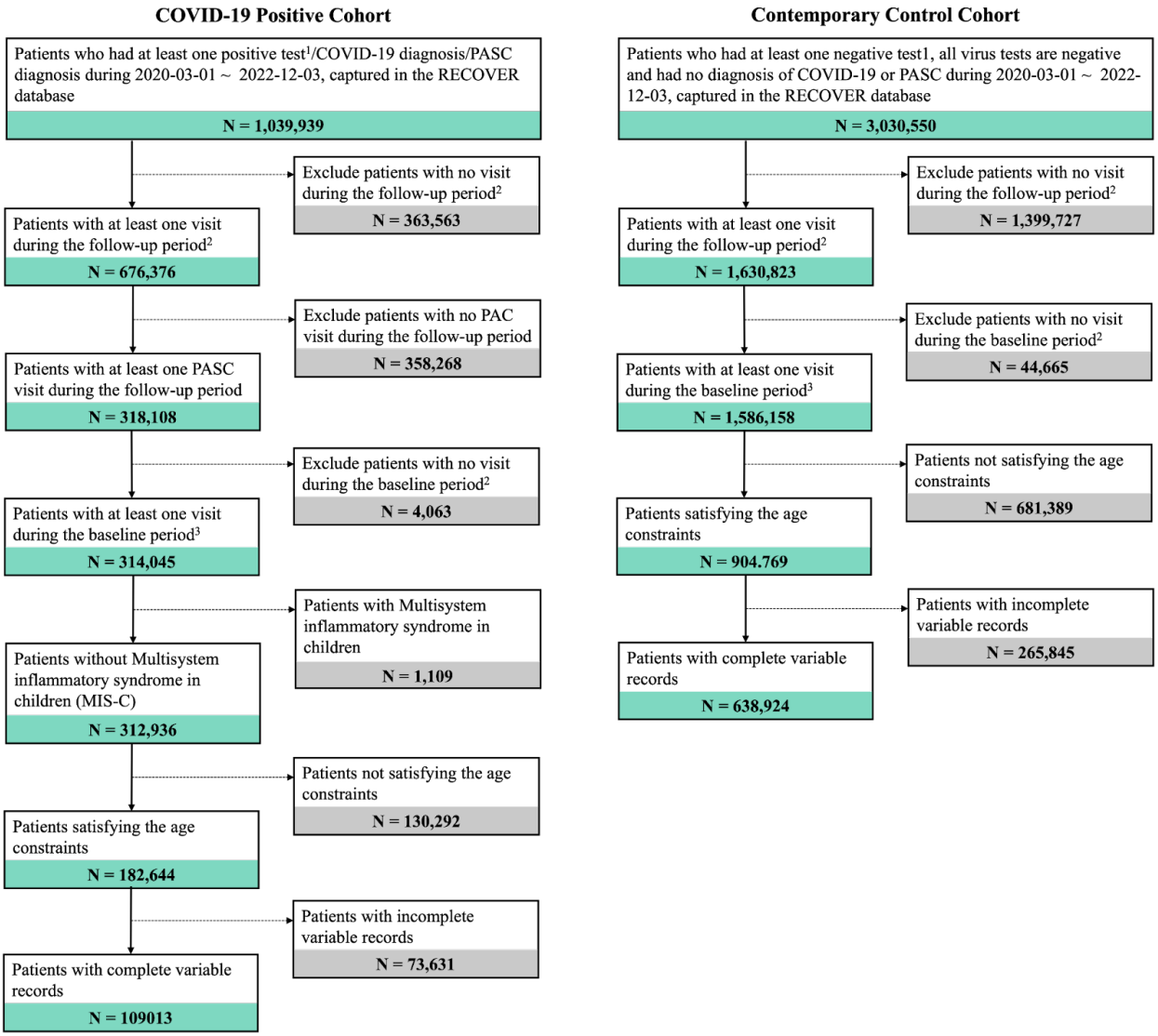
First, to ensure covariate balance between the COVID-19 positive and negative cohorts, we employed a propensity score-based stratification approach. The propensity scores were estimated using a logistic regression model, where COVID-19 positivity served as the binary outcome and patient characteristics functioned as covariates. After stratifying data into 5 strata, the confounders balances were assessed using Standardized Mean Differences (SMDs), with a threshold of less than 0.1 indicating adequate balance. Heatmap was generated to visualize similarities between the control and positive cohorts.

Second, to compare the co-incidence patterns of conditions during follow-up periods for patients within COVID-19 and contemporary Control Cohort, advanced topic modeling and clustering techniques were utilized. We used a topic modeling approach to learn the co-occurrence patterns among various semantic topics. For each patient, a binary vector was used to indicate whether the list of 128 potential non-MIS-C PASC Conditions presented in follow-up period or not. Poisson factor analysis, a specific form of topic modeling, was applied to each patient's 128-dimensional binary vector to map it onto a lower K-dimensional continuous potential PASC topic space. The Python package Pydpm v3.0.1 (available at https://pypi.org/project/pydpm/) was utilized to train the model. Each topic stands for a set of PASC that

is more likely to co-occur during a patient's follow-up period. The optimal number of topics was determined based on data likelihood and topic coherence metrics. Moreover, to delineate PASC subphenotypes, the hierarchical agglomerative clustering was employed using the previously obtained K-dimensional topic loading vectors for each patient.

Results

We conducted a retrospective study spanning from March 2020 to December 2022. The study included patients aged 21 years or younger, who had at least one healthcare visit during the 24 months to 7 days prior to the index date, which we designated as the baseline period. Additionally, participants were required to have at least one healthcare encounter between 28 and 179 days following the index date, defined as the follow-up period. The selection process for both COVID-19 positive and negative patients, drawn from real-world RECOVER data, is depicted in **Figure 1**.



[1] Including PCR, antigen, and serology tests. [2]28 days to 179 days after the index date [3] 24 months to 7 days before the index date

Figure 1. Results for the attrition table of a detailed flowchart illustrating the construction of our cohort.

Figure 2 presents a heatmap displaying the clustering of symptoms analyzed from the RECOVER pediatric cohort. Each column represents one of eleven identified topics, which depict patterns of symptom co-occurrence. Each row corresponds to a specific symptom or condition category from a collection of twenty prominently Clinical Classifications Software Refined (CCSR) categories, while additional, less frequent symptoms have been grouped together due to their lower incidence rates. The color intensity in each cell indicates the prevalence of each symptom within a particular topic. For COVID-19 positive cohort, Topic 1 is primarily associated with diseases of the respiratory system majorly including respiratory signs and symptoms, and other specified upper respiratory infections; topics 2, 5, 7 & 8 focus on the conditions of digestive system, mental, behavioral and neurodevelopmental disorders; topics 3 & 4 is characterized by a variety of respiratory system issues and allergic reactions; topics 6 mainly captures the conditions of musculoskeletal and nervous system; topics 9 include a mix of general signs and symptoms. After identifying potential PASC topics, we characterized pediatric patients with PASC using these topics and derived potential PASC subphenotypes. Specifically, four distinct subphenotypes were identified from the RECOVER cohort. **Figure 3** employs polar area charts to visually represent the distribution of disease categories within four distinct subphenotypes identified among pediatric patients. Each chart corresponds to a different cluster, with the angular width of each segment indicating the prevalence of specific disease categories within that subphenotype. Subphenotype 1 (top left) comprises 43.50% of the pediatric patients studied, Subphenotype 2 (top right) accounts for 26.91% of the patients, Subphenotype 3 (bottom left) includes 20.48%, and Subphenotype 4 (bottom right) encompasses 9.11% of the cohort.

CCSR Domain	PASC conditions	Covid-19 positive cohort									Covid-19 negative cohort								
		Topic1	Topic2	Topic3	Topic4	Topic5	Topic6	Topic7	Topic8	Topic9	Topic1	Topic2	Topic3	Topic4	Topic5	Topic6	Topic7	Topic8	Topic9
Diseases of the Digestive System	Other specified and unspecified gastrointestinal disorders	0.00%	4.67%	0.00%	0.00%	4.17%	0.00%	0.00%	10.37%	5.28%	0.00%	6.08%	0.00%	0.00%	2.22%	0.02%	0.00%	12.77%	4.83%
	Abdominal pain and other digestive abdomen signs and symptoms	4.03%	8.90%	0.00%	0.00%	8.11%	1.99%	0.00%	17.83%	1.60%	0.00%	2.09%	0.00%	2.04%	6.65%	1.77%	0.00%	21.76%	7.91%
	Nausea and vomiting	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.50%	0.01%	1.51%	4.10%	0.00%	0.00%	10.83%	5.59%
Diseases of the Musculoskeletal System and Connective Tissue	Musculoskeletal pain (not low back pain)	3.72%	2.55%	0.95%	0.01%	1.79%	28.65%	0.31%	2.39%	1.33%	0.00%	0.00%	2.07%	0.00%	2.76%	28.92%	0.29%	2.07%	3.12%
Diseases of the Nervous System	Nervous system pain and pain syndromes	0.00%	1.50%	0.00%	0.00%	1.52%	10.69%	0.00%	1.01%	0.29%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%
	Neurodevelopmental disorders	0.00%	0.00%	3.45%	1.83%	3.70%	0.00%	19.65%	1.39%	3.74%	13.08%	2.52%	1.54%	1.23%	3.20%	0.00%	18.19%	0.00%	0.00%
	Epilepsy; convulsions	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.02%	10.45%	0.00%	0.00%	0.00%	0.00%	2.00%	0.00%	0.14%
Diseases of the Respiratory System	Asthma	4.98%	2.11%	14.47%	15.37%	1.59%	0.00%	3.26%	3.67%	3.15%	16.29%	3.15%	19.11%	9.06%	1.43%	0.57%	0.02%	2.95%	0.03%
	Respiratory signs and symptoms	13.73%	3.46%	13.66%	18.07%	0.00%	0.00%	0.00%	4.63%	5.24%	3.42%	4.42%	17.38%	15.99%	1.55%	0.00%	0.00%	1.82%	3.08%
	Other specified upper respiratory infections	11.83%	0.00%	15.82%	24.89%	0.00%	2.47%	0.11%	11.31%	0.04%	2.53%	0.00%	20.20%	17.60%	1.70%	1.28%	0.00%	4.61%	0.00%
General and others	Other general signs and symptoms	2.62%	5.73%	1.87%	1.30%	5.31%	6.70%	1.76%	4.34%	10.14%	2.73%	10.75%	0.41%	2.17%	3.69%	7.58%	2.12%	4.82%	5.73%
Injury, Poisoning and Certain Other Consequences of External Causes	Allergic reactions	3.92%	1.21%	17.86%	16.70%	0.79%	0.00%	3.25%	6.41%	0.00%	16.00%	1.37%	19.43%	10.57%	0.00%	0.03%	0.00%	3.05%	0.04%
Mental, Behavioral and Neurodevelopmental Disorders	Anxiety and fear related disorders	1.84%	1.45%	0.00%	0.00%	14.26%	1.15%	21.30%	0.00%	0.93%	8.56%	0.00%	0.00%	0.00%	12.80%	0.70%	21.52%	1.96%	1.56%
	Depressive disorders	0.00%	0.00%	0.00%	0.00%	10.57%	0.00%	14.27%	0.00%	0.00%	4.82%	0.00%	0.00%	0.00%	9.50%	0.00%	15.02%	0.02%	0.00%

Figure 2. Heat map of PASC topics learned from the RECOVER cohort. Each row denotes a potential PASC diagnosis category defined by ICD-10 codes classified through CCSR, and each column denotes a particular PASC topic. Each PASC topic represents a unique post-acute incidence probability distribution over all 128 individual potential PASC diagnosis categories.

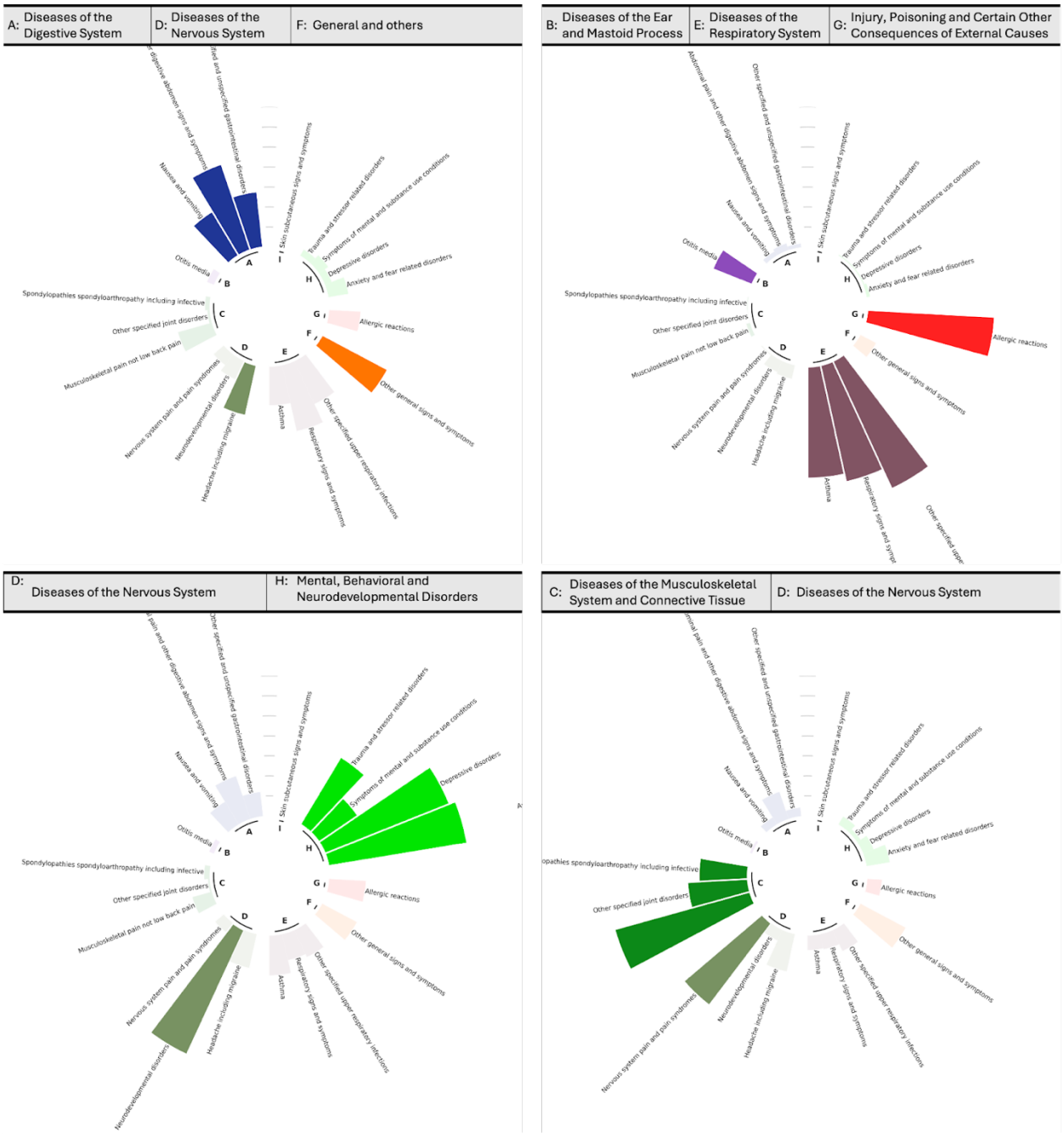


Figure 3. Incidence rates of potential PASC conditions in each subphenotype, where potential PASC conditions were grouped into different categories shown in different colored bars. A condition was highlighted from a particular subphenotype where it had the highest incidence rate compared to other subphenotypes.

Conclusion

Our research explores the diversity and complexity of conditions emerging 30–180 days post-confirmation of SARS-CoV-2 infection, identifying four consistent subphenotypes using EHR data from the extensive Clinical Research Networks through topic modeling techniques. These insights could assist clinicians and

healthcare systems in creating tailored care models for patients with PASC. The approach is readily applicable to any datasets standardized by the OMOP Common Data Model without the need for additional preprocessing or variable selection. Future work will extend the evaluation of this method across various diseases within OHDSI studies.

References

1. Bowe, B., Xie, Y. & Al-Aly, Z. Postacute sequelae of COVID-19 at 2 years. *Nat. Med. 2023 299* **29**, 2347–2357 (2023).
2. Thaweethai, T. *et al.* Development of a Definition of Postacute Sequelae of SARS-CoV-2 Infection. *JAMA* **329**, 1934–1946 (2023).
3. Soriano, J. B., Murthy, S., Marshall, J. C., Relan, P. & Diaz, J. V. A clinical case definition of post-COVID-19 condition by a Delphi consensus. *Lancet. Infect. Dis.* **22**, e102–e107 (2022).
4. Davis, H. E., McCorkell, L., Vogel, J. M. & Topol, E. J. Long COVID: major findings, mechanisms and recommendations. *Nat. Rev. Microbiol. 2023 213* **21**, 133–146 (2023).
5. Parotto, M. *et al.* Post-acute sequelae of COVID-19: understanding and addressing the burden of multisystem manifestations. *Lancet Respir. Med.* **11**, 739–754 (2023).
6. Zhang, H. *et al.* Data-driven identification of post-acute SARS-CoV-2 infection subphenotypes. *Nat. Med. 2022 291* **29**, 226–235 (2022).
7. Ozonoff, A. *et al.* Features of acute COVID-19 associated with post-acute sequelae of SARS-CoV-2 phenotypes: results from the IMPACC study. *Nat. Commun. 2024 151* **15**, 1–17 (2024).
8. Dagliati, A. *et al.* Characterization of long COVID temporal sub-phenotypes by distributed representation learning from electronic health record data: a cohort study. *eClinicalMedicine* **64**, (2023).
9. Bandeen-roche, K. *et al.* Latent Variable Regression for Multiple Discrete Outcomes. *J. Am. Stat. Assoc.* **92**, 1375–1386 (1997).
10. Blei, D. M. Surveying a suite of algorithms that offer a solution to managing large document archives. Probabilistic topic models. doi:10.1145/2133806.2133826.