medRxiv preprint doi: https://doi.org/10.1101/2022.05.21.22275412; this version posted May 23, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

Machine Learning for Identifying Data-Driven Subphenotypes of Incident Post-Acute SARS-CoV-2 Infection Conditions with Large Scale Electronic Health Records: Findings from the RECOVER Initiative

Hao Zhang[1], Chengxi Zang[1], Zhenxing Xu[1], Yongkang Zhang[1], Jie Xu[2], Jiang Bian[2], Dmitry Morozyuk[1], Dhruv Khullar[1], Yiye Zhang[1], Anna S. Nordvig[3], Edward J. Schenck[4], Elizabeth A. Shenkman[2], Russel L. Rothman[5], Jason P. Block[6], Kristin Lyman[7], Mark Weiner[1], Thomas W. Carton[7], Fei Wang[1]*, Rainu Kaushal[1]

[1] Department of Population Health Sciences, Weill Cornell Medicine, New York, NY, USA.
[2] Department of Health Outcomes Biomedical Informatics, University of Florida, Gainesville, FL, USA.
[3] Department of Neurology, Weill Cornell Medicine, New York, NY, USA.
[4] Division of Pulmonary and Critical Care Medicine, Department of Medicine, Weill Cornell Medicine, New York, NY, USA.
[5] Center for Health Services Research, Vanderbilt University Medical Center, Nashville, Tennessee, USA.
[6] Department of Population Medicine, Harvard Pilgrim Health Care Institute, Harvard Medical School, Boston, MA, USA.
[7] Louisiana Public Health Institute, New Orleans, Louisiana, USA.

* Corresponding few2001@med.cornell.edu

NOTE: This preprint reports new research that has not been certified by peer review and should not be used to guide clinical practice.

medRxiv preprint doi: https://doi.org/10.1101/2022.05.21.22275412; this version posted May 23, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

Abstract

The post-acute sequelae of SARS-CoV-2 infection (PASC) refers to a broad spectrum of symptoms and signs that are persistent, exacerbated, or newly incident in the post-acute SARS-CoV-2 infection period of COVID-19 patients. Most studies have examined these conditions individually without providing concluding evidence on co-occurring conditions. To answer this question, this study leveraged electronic health records (EHRs) from two large clinical research networks from the national Patient-Centered Clinical Research Network (PCORnet) and investigated patients' newly incident diagnoses that appeared within 30 to 180 days after a documented SARS-CoV-2 infection. Through machine learning, we identified four reproducible subphenotypes of PASC dominated by blood and circulatory system, respiratory, musculoskeletal and nervous system, and digestive system problems, respectively. We also demonstrated that these subphenotypes were associated with distinct patterns of patient demographics, underlying conditions present prior to SARS-CoV-2 infection, acute infection phase severity, and use of new medications in the post-acute period. Our study provides novel insights into the heterogeneity of PASC and can inform stratified decision-making in the treatment of COVID-19 patients with PASC conditions.

Introduction

A variety of symptoms and signs involving multiple organ systems (e.g., cardiovascular[1], mental[2], metabolic[3], and renal[4]) were persistent, exacerbated or newly developed following the acute phase of SARS-CoV-2 infection. Currently, our understanding of these conditions, typically regarded as post-acute sequelae of SARS-CoV-2 infection (PASC)[5,6], remains limited. Most existing studies have investigated these PASC conditions individually (e.g., by examining the incidence[7] or excess burden[8] of each symptom or condition in the post-acute period for COVID-19 patients relative to controls). It is unclear if any of these PASC symptoms and conditions tend to co-appear together or are more likely to develop in certain patient populations.

To answer the above question, we developed a machine learning approach to derive subphenotypes of SARS-CoV-2 infected patients based on the newly incident conditions in the post-acute period (defined as 30 to 180 days after their confirmed SARS-CoV-2 infection) using the data from electronic health records (EHR). We focused on new incidences in this study because it provided a clean way of defining PASC phenotypes without complicated consideration of pre-existing conditions. We leveraged the EHR repositories from two large-scale clinical research networks (CRNs) from the national Patient-Centered Clinical Research Network (PCORnet): the INSIGHT network[9] , which includes 12 million patients in the New York City (NYC) area, and the OneFlorida+ network[10], which includes 19 million patients from Florida, Georgia, and Alabama. We examined the incidence of 137 diagnosis categories defined from the Clinical Classifications Software Refined (CCSR) categories[11]. Four distinct subphenotypes were identified from the INSIGHT CRN data and validated in the OneFlorida+ CRN data. Patients in Subphenotype 1 were older with incident blood and circulatory conditions in the post-acute phase. Subphenotype 2 included patients who are younger with incident respiratory problems. Subphenotype 3 included patients who developed musculoskeletal and nervous system conditions. Lastly, patients in Subphenotype 4 are characterized by digestive systems incident conditions. Our study dissects the heterogeneity of potential PASC conditions as different subgroups, which can inform the classification and treatment of PASC. This study is part of the NIH Researching COVID to Enhance Recovery (RECOVER) Initiative[12], which seeks to understand, treat, and prevent the post-acute sequelae of SARS-CoV-2 infection (PASC). For more information on RECOVER, visit https://recovercovid.org/

medRxiv preprint doi: https://doi.org/10.1101/2022.05.21.22275412; this version posted May 23, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

Results

Overall Pipeline

Our overall analytics pipeline is demonstrated in Figure 1. Using the two CRNs, we extracted patients who had positive nucleic acid amplification or antigen viral tests for SARS-CoV-2 from March 2020 to November 2021. A list of 137 potential PASC diagnosis categories (Methods) were compiled and only patients who had documented new incidences of these conditions in their post-acute infection period were retained. Each patient was initially represented as a 137-dimensional binary vector according to whether a particular condition appeared in the post-acute infection period or not (Step 1). Then a set of "PASC topics" was learned based on the co-incidence patterns of these conditions (Step 2) and the initial patient vectors were projected onto these learned topics to obtain their topic loading representations (Step 3), which was further used in a clustering procedure to identify the subphenotypes (Step 4). Details on the concepts and methods involved in our pipeline are presented in Methods. In the following text, we present our main results.

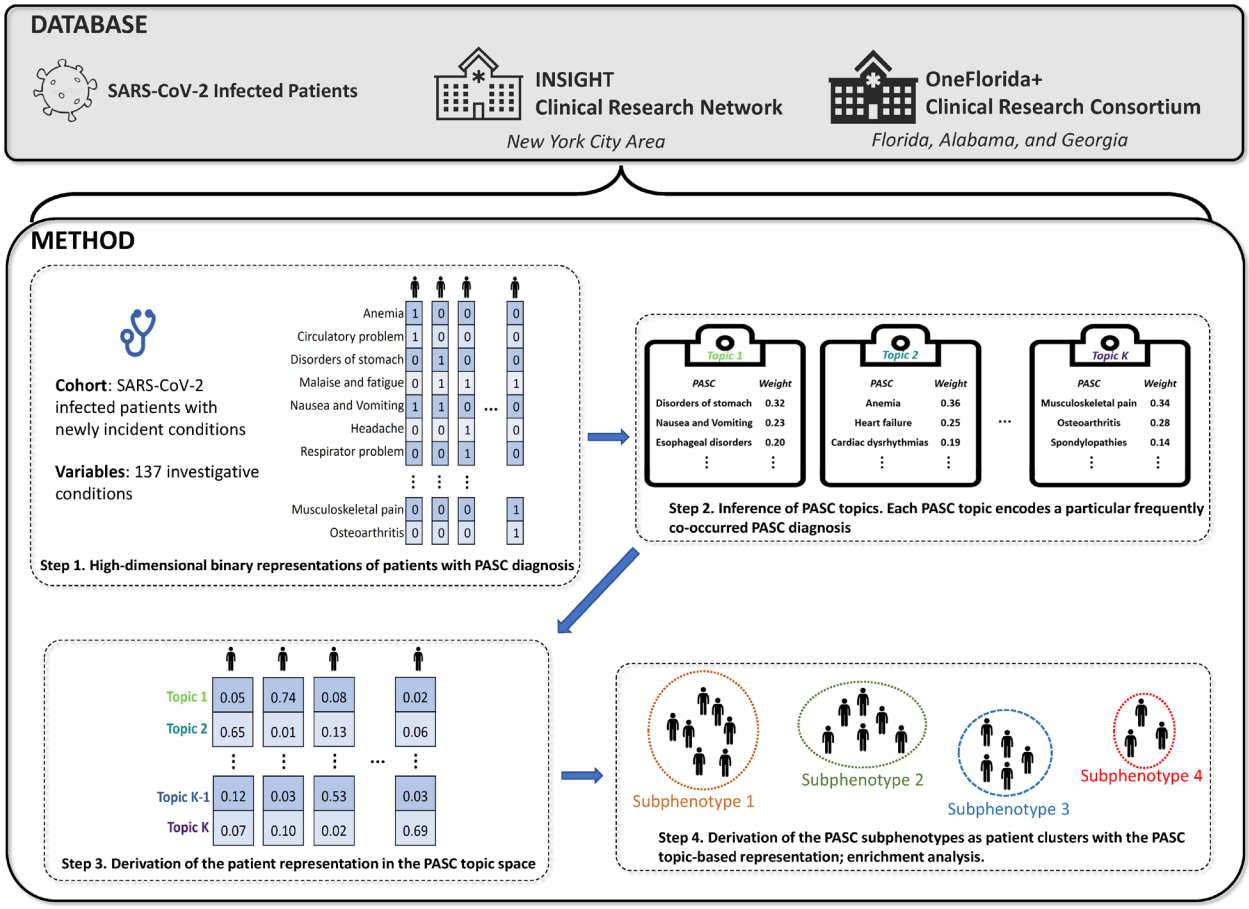


Figure 1. Data creation and subphenotype pipeline. To study the subphenotypes for patients with PASC, we constructed cohorts from the INSIGHT and OneFlorida+ clinical research networks. After obtaining high-dimensional binary representations of patients with PASC diagnoses (**step 1**), we learned PASC topics (**step 2**) and inferred the patient representations in the low-dimensional PASC topic space (**step 3**) by a topic modeling approach. Finally, we

medRxiv preprint doi: https://doi.org/10.1101/2022.05.21.22275412; this version posted May 23, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

derived PASC subphenotypes as patient clusters with the PASC topic-based representations (**step 4**).

Study Cohorts.

Our study included 20,881 patients from the INSIGHT CRN and 13,724 patients from the OneFlorida+ CRN who tested positive for SARS-CoV-2 on viral tests (see Methods for detailed inclusion-exclusion criteria). The patients within the INSIGHT cohort had a median age of 58.0 (interquartile range [IQR] [42.0-70.0]) and a median Area Deprivation Index (ADI)[13] of 15.0 (IQR [6.0-25.0]), consisting of 12,188 (58.37%) females, 7013 (33.59%) White patients, and 4771 (22.85%) Black patients. The OneFlorida+ cohort contained patients who were younger (median age of 51.0 (IQR [35.0-65.0])), with more disadvantaged social conditions (median ADI 59.0; IQR [42.0-76.0])), and more white patients (7175; 52.28%). 33.04% of the INSIGHT CRN patients had a confirmed SARS-CoV-2 infection from March to June 2020 (compared to 8.83% of the patients from OneFlorida+). This coincided with the first wave of COVID-19 in the US when NYC was the epicenter. Patients from OneFlorida+ were more likely to test positive from July to October 2020 (26% vs. 6% of patients from INSIGHT). Table 1 summarizes the summary statistics of the patients from the two cohorts.

Table 1. Characteristics of the INSIGHT cohort (for development) and the OneFlorida+ cohort (for validation).

Characteristics	INSIGHT cohort development	OneFlorida+ cohort validation
No. of patients	20881	13724
Age, y, Median (IQR)[c]	58.0 [42.0-70.0]	51.0 [35.0-65.0]
Age group – no. (%)		
20-<40 years	4603 (22.04%)	4284 (31.22%)
40-<55 years	4522 (21.66%)	3318 (24.18%)
55-<65 years	4308 (20.63%)	2528 (18.42%)
65-<75 years	3810 (18.25%)	1923 (14.01%)
75-<85 years	2553 (12.22%)	1178 (8.58%)
85+ years	1085 (5.20%)	493 (3.59%)
Sex – no. (%)		
Female	12188 (58.37%)	8468 (61.70%)
Male	8692 (41.63%)	5255 (38.29%)
Other/Missing	1 (0.00%)	1 (0.00%)
Race – no. (%)		
Asian	945 (4.52%)	144 (1.05%)
Black or African American	4771 (22.85%)	4076 (29.70%)
White	7013 (33.59%)	7175 (52.28%)
Other	6238 (29.87%)	2123 (15.47%)
Missing	1914 (9.17%)	206 (1.50%)
Ethnic group – no. (%)		
Hispanic: Yes	6838 (32.74%)	2881 (20.99%)
Hispanic: No	12248 (58.66%)	9329 (67.98%)
Hispanic: Other/Missing	1795 (8.60%)	1514 (11.03%)
Area deprivation index, Median (IQR)	15.0 [6.0-25.0]	59.0 [42.0-76.0]
Healthcare utilization in the past 3 yr – no. (%)		
Inpatient 0	14631 (70.07%)	7437 (54.19%)
Inpatient 1-2	4359 (20.88%)	3018 (21.99%)
Inpatient 3-4	1074 (5.14%)	1195 (8.71%)
Inpatient >=5	817 (3.91%)	2074 (15.11%)
Outpatient 0	1241 (5.94%)	2759 (20.10%)
Outpatient 1-2	2012 (9.64%)	1736 (12.65%)
Outpatient 3-4	1595 (7.64%)	820 (5.97%)

medRxiv preprint doi: https://doi.org/10.1101/2022.05.21.22275412; this version posted May 23, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

Outpatient >=5	16033 (76.78%)	8409 (61.27%)
Emergency 0	11321 (54.22%)	4908 (35.76%)
Emergency 1-2	5975 (28.61%)	3386 (24.67%)
Emergency 3-4	1730 (8.29%)	1465 (10.67%)
Emergency >=5	1855 (8.89%)	3965 (28.89%)
Index time period of patient – no. (%)		
03/20-06/20	6899 (33.04%)	1212 (8.83%)
07/20-10/20	1180 (5.65%)	3570 (26.01%)
11/20-02/21	8714 (41.73%)	3988 (29.06%)
03/21-06/21	3390 (16.24%)	1531 (11.16%)
07/21-11/21	698 (3.34%)	3423 (24.94%)
Acute phase severities of COVID-19 (-1~16 days)[a] – no. (%)		
Hospitalized	9076 (43.47%)	5036 (36.69%)
Ventilation	495 (2.37%)	465 (3.39%)
Critical care	1144 (5.48%)	833 (6.07%)
Underlying conditions[b] – no. (%)		
Alcohol Abuse	697 (3.34%)	503 (3.66%)
Anemia	3288 (15.75%)	2674 (19.48%)
Arrythmia	3656 (17.51%)	2106 (15.35%)
Asthma	2591 (12.41%)	1641 (11.96%)
Cancer	2406 (11.52%)	1194 (8.70%)
Chronic Kidney Disease	3223 (15.44%)	2175 (15.85%)
Chronic Pulmonary Disorders	3893 (18.64%)	2778 (20.24%)
Cirrhosis	408 (1.95%)	241 (1.76%)
Coagulopathy	1753 (8.39%)	776 (5.65%)
Congestive Heart Failure	2508 (12.01%)	2003 (14.59%)
COPD	1246 (5.97%)	1125 (8.20%)
Coronary Artery Disease	3234 (15.49%)	1924 (14.02%)
Dementia	1046 (5.01%)	688 (5.01%)
Diabetes Type 1	287 (1.37%)	265 (1.93%)
Diabetes Type 2	5173 (24.77%)	3544 (25.82%)
End Stage Renal Disease on Dialysis	1080 (5.17%)	622 (4.53%)
Hemiplegia	311 (1.49%)	242 (1.76%)
HIV[d]	376 (1.81%)	123 (0.90%)
Hypertension	9165 (43.89%)	6486 (47.26%)
Hypertension and Type 1 or 2 Diabetes	4367 (20.91%)	3064 (22.33%)
Inflammatory Bowel Disorder	215 (1.03%)	145 (1.06%)
Lupus or SLE[e]	186 (0.89%)	173 (1.26%)
Mental Health Disorders	2556 (12.24%)	2559 (18.65%)
Multiple Sclerosis	114 (0.55%)	73 (0.53%)
Parkinson's Disease	144 (0.68%)	107 (0.78%)
Peripheral vascular disorders	1791 (8.57%)	1174 (8.55%)
Pregnant	686 (3.29%)	788 (5.74%)
Pulmonary Circulation Disorder	467 (2.24%)	342 (2.49%)
Rheumatoid Arthritis	375 (1.79%)	296 (2.16%)
Seizure/Epilepsy	546 (2.61%)	511 (3.72%)
Severe Obesity (BMI>=40 kg/m2)	1600 (7.66%)	1789 (13.04%)
Weight Loss	1145 (5.48%)	783 (5.71%)
Down's Syndrome	20 (0.09%)	20 (0.15%)
Other Substance Abuse	1503 (7.20%)	1857 (13.53%)
Cystic Fibrosis	8 (0.04%)	16 (0.12%)
Autism	25 (0.12%)	46 (0.34%)
Sickle Cell	158 (0.76%)	139 (1.01%)
Corticosteroids Prescription	3193 (15.29%)	2618 (19.08%)
Immunosuppressant Prescription	1018 (4.88%)	353 (2.57%)

Note: a. Time range for acute phase severities is from one day before index date to sixteen days after index date; **b**. Coexisting conditions existed if two records in the 3-years prior to index event; **c**. IQR: inter-quartile range; **d**. HIV:

medRxiv preprint doi: https://doi.org/10.1101/2022.05.21.22275412; this version posted May 23, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

human immunodeficiency virus; **e**. SLE: systemic lupus erythematosus; f. For the healthcare utilization in the past three years, including inpatient, outpatient, and emergency visits, we binned the number of visits into five levels.

Potential PASC Topics

A list of 137 potentially PASC-related diagnoses groups defined by ICD-10 diagnosis codes and CCSR categories[11] (Supplementary Table 1) was compiled for our study. We first investigated the co-incidence patterns across different diagnoses within 30-180 days after the SARS-CoV-2 infection confirmation for COVID-19 positive patients (Methods). We achieved this goal through probabilistic topic modeling[14], originally proposed for learning word co-occurrence patterns in documents with different semantic topics. With this approach (see details in Methods), we were able to identify ten distinct "PASC topics", each of which is characterized by a unique post-acute infection incidence probability distribution across the 137 individual conditions.

Figure 2 shows the heatmap matrix of the learned topics from the INSIGHT cohort. Each column was a learned topic, and each row was a potential PASC condition category (we demonstrated 31 of them in the heatmap and aggregated the remaining 106 because none of their incident probabilities exceeded 0.1 in any of the learned topics). Each entry in the matrix corresponded to the probability of the specific PASC condition in the corresponding topic. All entry values in the same column added up to one so that each topic was characterized by a rigorous incident probability distribution over the 137 PASC conditions. Specifically, Topics T1, T2, and T5 concentrated on the conditions of the musculoskeletal system, digestive system, and nervous system, respectively. T4, T7, and T9 included respiratory conditions mixed with sleep disorder and anxiety along with symptoms such as headache and chest pain. T3 included fluid and electrolyte disorders combined with anemia and cardiac problems. T6 was musculoskeletal and skin conditions with headache and fatigue problems. T8 was anemia and digestive system problems. T10 was a mixture of circulatory problems, renal failure, fluid and electrolyte problems, and others.

medRxiv preprint doi: https://doi.org/10.1101/2022.05.21.22275412; this version posted May 23, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

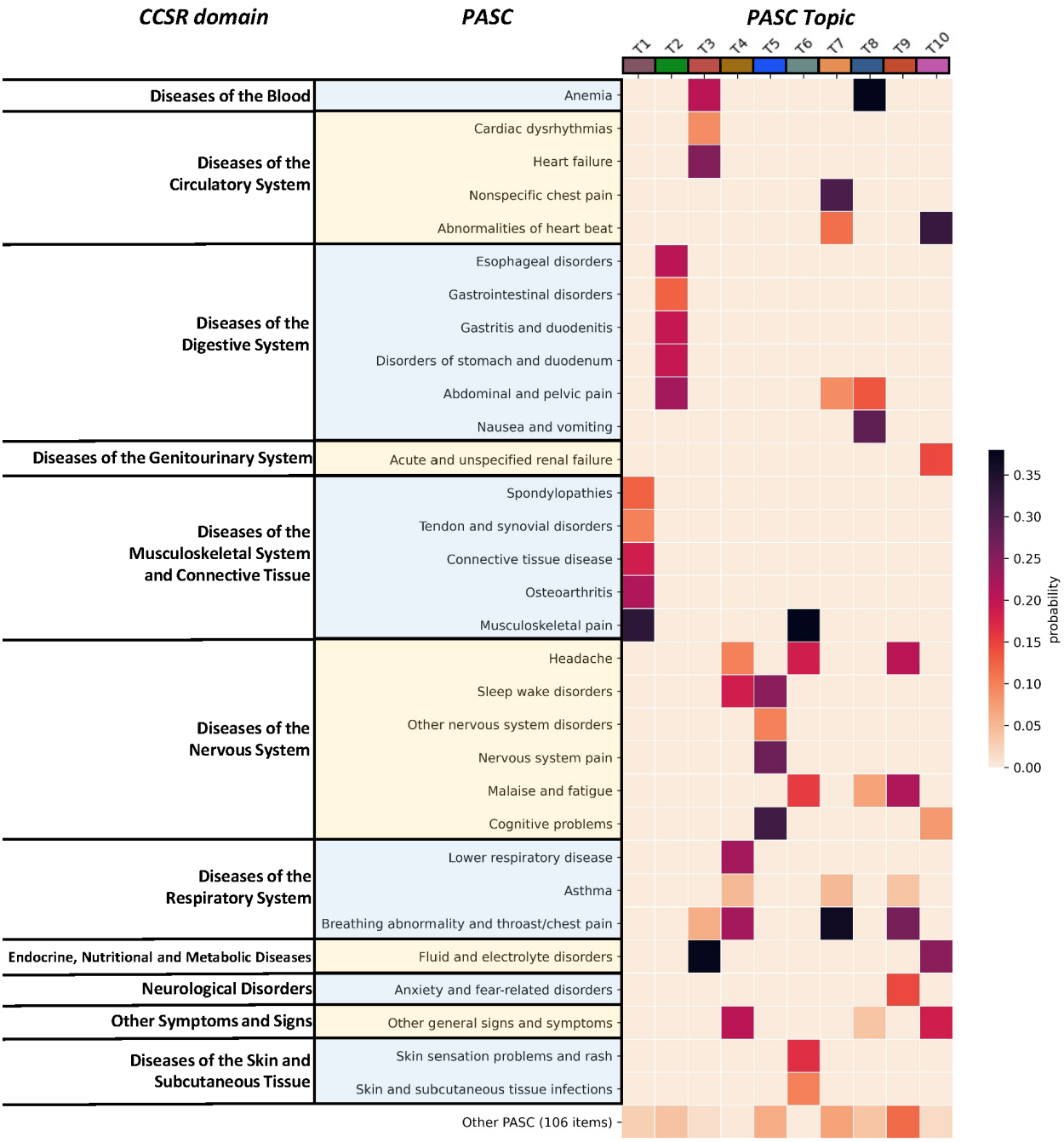


Figure 2. The heatmap of PASC topics learned on the INSIGHT cohort. Each row denotes a potential PASC category grouped by different CCSR domains, and each column denotes a particular PASC topic. Each PASC topic is characterized by a unique post-acute incidence probability distribution over all 137 individual potential PASC categories.

medRxiv preprint doi: https://doi.org/10.1101/2022.05.21.22275412; this version posted May 23, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

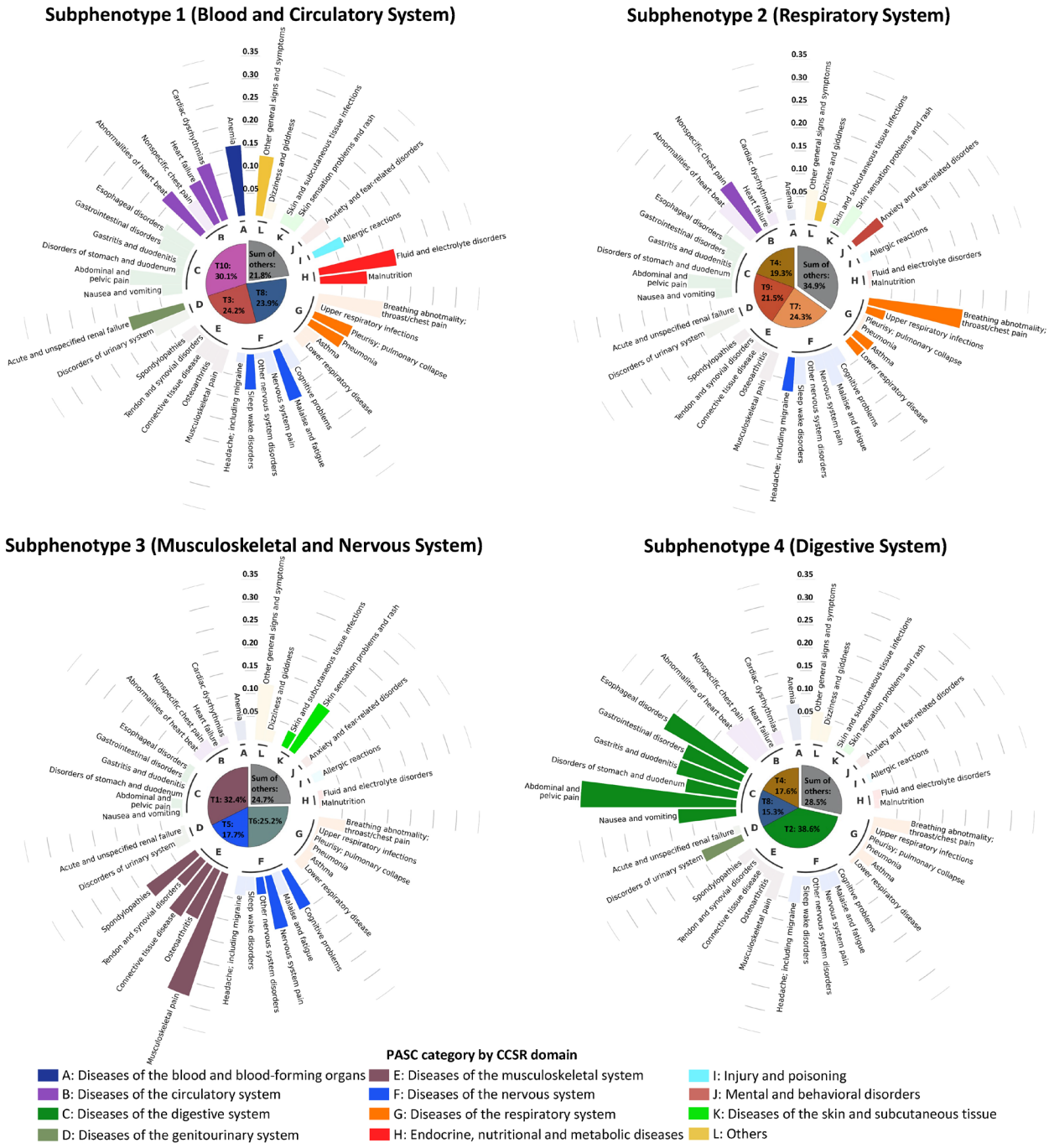


Figure 3. The incidence rates of potential PASC conditions in each subphenotype for the INSIGHT cohort, where potential PASC conditions were grouped into different categories shown in different colored bars outside the center pie chart. A condition is highlighted in the subphenotype where it has the highest incidence rate. The center pie chart of each subphenotype shows the mean topic proportions of the patients it included, and the meanings of the topic indices can be referred to Figure 2.

medRxiv preprint doi: https://doi.org/10.1101/2022.05.21.22275412; this version posted May 23, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

Potential PASC Subphenotypes

With the potential PASC topics identified, we could describe PASC-affected patients with them and derive potential PASC subphenotypes as patient clusters (Methods). In the INSIGHT cohort, four subphenotypes were identified. Table 2 summarizes their characteristics with respect to patient demographics, disease severity in the acute phase according to treatment setting, as well as the prevalence of comorbidities in the baseline period. Across the subphenotypes, we also demonstrated the prevalence of potential PASC conditions and incident prescriptions of medications in the post-acute infection period for patients in different subphenotypes in Figures 3 and 4. From Figure 3, we observed that four subphenotypes had different prevalent PASC conditions, which were consistent with the top PASC conditions in the topics with large proportions. Next, we characterized these subphenotypes in detail as follows.

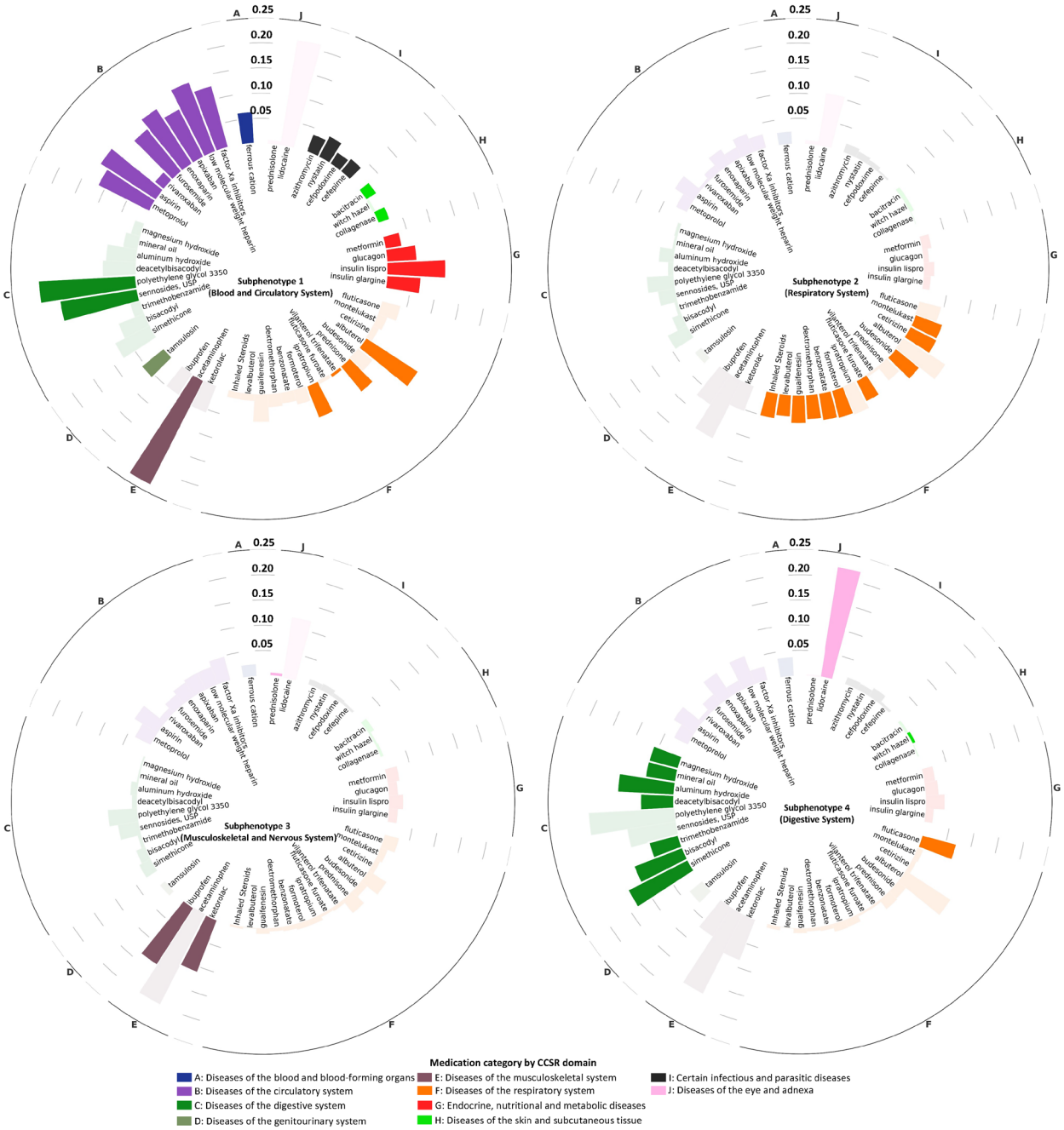
Subphenotype 1 (Blood and Circulatory System) consisted of 7,047 (33.75%) patients. It was dominated by blood- and circulation-related topics (T3, T8, T10), including anemia, fluid and electrolyte problems, and circulatory and cardiac problems. Compared to other subphenotypes, patients in this subphenotype were older (median age 65.0 years, IQR [52.0-75.0]) and had the highest proportion of males (48.53%). They also had a higher severity of COVID-19 in the acute phase (with the highest rate of hospitalization [61.15%], use of mechanical ventilation [4.81%], and critical care services [9.95%]). Furthermore, this subphenotype had the highest portion of patients (37.38%) infected with SARS-CoV-2 during the first wave of the pandemic (March to June 2020) in NYC. In addition, patients in this subphenotype had a higher prevalence of underlying conditions than other subphenotypes, especially for blood, circulation, and endocrine comorbidities. Correspondingly, patients in this subphenotype had a high incident prescription for medications to treat circulatory and endocrine problems and anemia.

Subphenotype 2 (Respiratory System) included 6,838 (32.75%) patients. It was dominated by respiratory conditions (topics T4, T7, and T9), sleep disorders, anxiety, and symptoms such as headache and chest pain. This subphenotype had the youngest patients among the four (median age 51.0 years, IQR [35.0-64.0]), the highest proportion of females (62.8%), and the lowest rate of hospitalization (31.28%) for COVID-19. It also had the largest proportion of patients who tested positive for SARS-CoV-2 from November 2020 to November 2021 (64.47%). Patients in this subphenotype had higher baseline comorbidity burdens for respiratory conditions such as upper respiratory problems and chronic obstructive pulmonary disease, breathing problems, and a higher incident prescription for a diverse set of anti-asthma, anti-allergy, and anti-inflammation medications including inhaled steroids, levalbuterol, and montelukast.

Subphenotype 3 (Musculoskeletal and Nervous System) consisted of 4,879 (23.37%) patients. It mainly contained musculoskeletal and nervous system problems (topics T1, T5, and T6) such as musculoskeletal pain, headaches, and sleep-wake problems. This subphenotype included patients with a median age of 57.0 years (IQR [42.0-69.0]), with 60.71% female. It had the highest proportion of patients with more than five outpatient visits (78.4%). Patients in this subphenotype had higher baseline comorbidity burdens of autoimmune and allergy conditions such as rheumatoid arthritis and asthma, as well as other musculoskeletal and nervous system problems including soft tissue, bone, and sleep problems. This subphenotype was also associated with a higher incident prescription risk of pain medications (ibuprofen and ketorolac) in the post-acute infection period.

medRxiv preprint doi: https://doi.org/10.1101/2022.05.21.22275412; this version posted May 23, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

Subphenotype 4 (Digestive System) included 2,117 (10.14%) patients mainly with digestive system problems such as abdominal pain, vomiting, and respiratory conditions (topics T2, T4, T8). Patients in this subphenotype had a median age of 54.0 (IQR [39.0-67.0]) with 61.64% female. Patients in this subphenotype had the highest proportion of patients without any baseline emergency visits (57.06%) and the lowest rates of mechanical ventilation (0.8%) and critical care admission (2.79%) in the acute phase of COVID-19. Compared with the other three subphenotypes, this subphenotype had an overall lower prevalence of underlying conditions, and a slightly higher prevalence of digestive problems such as hematemesis, stomach and duodenum disorders, and digestive system neoplasm. In addition, this subphenotype had higher incident prescription rates of drugs for treating the digestive system.



medRxiv preprint doi: https://doi.org/10.1101/2022.05.21.22275412; this version posted May 23, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

Figure 4. The prevalence of incident prescriptions of medications in the post-acute infection period for each subphenotype on the INSIGHT cohort, where medications are grouped into different categories shown by different colors. For one of the medications, if it is most prevalent in one subphenotype, we highlighted it in this subphenotype.

Contrast with COVID-19 Negative Patients

These potential PASC subphenotypes were derived from SARS-CoV-2 infected patient. However, it was unclear how the potential PASC diagnosis co-incidence patterns encoded in them differed from the non-infected patients. To answer this question, we compared the incidence patterns of 28 selected potential PASC conditions in 30-180 days after the COVID-19 lab test between positive and matched negative patients (Methods). The results were demonstrated in Figure 5, where the nodes in each network corresponded to a particular potential PASC condition with their sizes proportional to the incidence rate in the corresponding subphenotype or matched controls, and each line linking a pair of nodes indicated co-incidence of the corresponding potential PASC diagnoses with its thickness proportional to the co-incidence rate. From the figure, we could see that the conditions we used to characterize each subphenotype were clearly associated with larger-sized nodes, representing higher incidence rates. At the same time, we do not observe differences in node sizes on the matched COVID-19 negative networks. In addition, we observed denser connections in COVID-19 positive networks, which suggested that the potential PASC conditions did not appear independently, but collectively, and those larger nodes were network hubs associated with more lines.

medRxiv preprint doi: https://doi.org/10.1101/2022.05.21.22275412; this version posted May 23, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

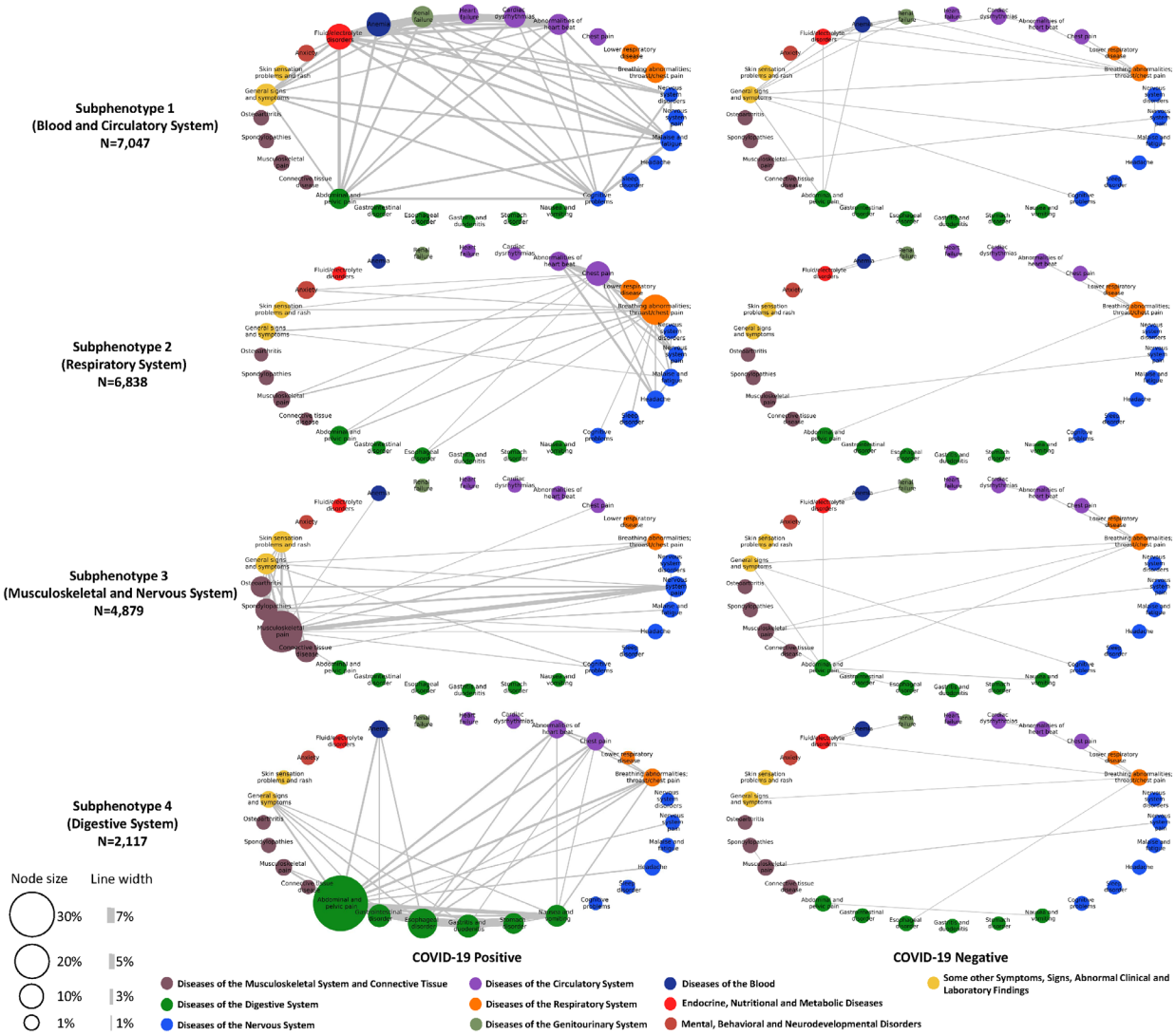


Figure 5. Difference of the incidence patterns of selected PASC conditions (grouped by CCSR domains) in 30-180 days after COVID-19 lab test between positive and matched negative patients on the INSIGHT cohort. The bubbles in each network correspond to a PASC condition with their sizes proportional to the incidences in the particular subphenotype or matched controls. The edge linking a pair of bubbles indicates co-incidence of the corresponding potential PASC conditions with its thickness proportional to the co-incidence rate, where lines are visible if the rate is larger than 1%.

Table 2. Characteristics of the identified subphenotypes on the INSIGHT cohort.

Variable	Total	Subphenotype 1 (Blood & Circulatory System)	Subphenotype 2 (Respiratory System)	Subphenotype 3 (Musculoskeletal and Nervous System)	Subphenotype 4 (Digestive System)	P-value[a]	Post hoc analysis
No. of patients (%)	20881 (100%)	7047 (33.75%)	6838 (32.75%)	4879 (23.37%)	2117 (10.14%)		
Age, y, Median (IQR[b])	58.0 (42.0-70.0)	65.0 (52.0-75.0)	51.0 (35.0-64.0)	57.0 (42.0-69.0)	54.0 (39.0-67.0)	0	2 vs. 1, 3 vs. 1, 4 vs. 1, 3 vs. 2, 4 vs. 2, 4 vs. 3
Age group – no. (%)							
20-<40 years	4603 (22.04%)	820 (11.64%)	2200 (32.17%)	1026 (21.03%)	557 (26.31%)	5.73E-190	1 vs. 2, 1 vs. 3, 1 vs. 4, 2 vs. 3, 2 vs. 4, 3 vs. 4
40-<55 years	4522 (21.66%)	1180 (16.74%)	1677 (24.52%)	1153 (23.63%)	512 (24.19%)	7.37E-33	1 vs. 2, 1 vs. 3, 1 vs. 4
55-<65 years	4308 (20.63%)	1486 (21.09%)	1336 (19.54%)	1067 (21.87%)	419 (19.79%)	0.00992	1 vs. 2, 2 vs. 3
65-<75 years	3810 (18.25%)	1657 (23.51%)	930 (13.60%)	878 (18%)	345 (16.3%)	8.48E-51	1 vs. 2, 1 vs. 3, 1 vs. 4, 2 vs. 3, 2 vs. 4
75-<85 years	2553 (12.22%)	1278 (18.14%)	518 (7.58%)	532 (10.9%)	225 (10.63%)	4.5E-82	1 vs. 2, 1 vs. 3, 1 vs. 4, 2 vs. 3, 2 vs. 4
85+ years	1085 (5.20%)	626 (8.88%)	177 (2.59%)	223 (4.57%)	59 (2.79%)	1.48E-68	1 vs. 2, 1 vs. 3, 1 vs. 4, 2 vs. 3, 3 vs. 4
Sex – no. (%)							
Female	12188 (58.37%)	3627 (51.47%)	4294 (62.80%)	2962 (60.71%)	1305 (61.64%)	4.91E-46	1 vs. 2, 1 vs. 3, 1 vs. 4, 2 vs. 3
Male	8692 (41.63%)	3420 (48.53%)	2543 (37.19%)	1917 (39.29%)	812 (38.36%)	4.08E-46	1 vs. 2, 1 vs. 3, 1 vs. 4, 2 vs. 3
Other/Missing	1 (0.00%)	0 (0.00%)	1 (0.01%)	0 (0%)	0 (0%)	0.561	-
Race – no. (%)							
Asian	945 (4.52%)	278 (3.94%)	339 (4.96%)	237 (4.86%)	91 (4.3%)	0.019	1 vs. 2, 1 vs. 3
Black or African American	4771 (22.85%)	1752 (24.86%)	1482 (21.67%)	1116 (22.87%)	421 (19.89%)	4.99E-07	1 vs. 2, 1 vs. 3, 1 vs. 4, 3 vs. 4
White	7013 (33.59%)	2407 (34.16%)	2255 (32.98%)	1670 (34.23%)	681 (32.17%)	0.174	-
Other	6238 (29.87%)	2052 (29.12%)	2041 (29.85%)	1411 (28.92%)	734 (34.67%)	0.00000509	1 vs. 4, 2 vs. 4, 3 vs. 4
Missing	1914 (9.17%)	558 (7.92%)	721 (10.54%)	445 (9.12%)	190 (8.97%)	0.00000238	1 vs. 2, 1 vs. 3, 2 vs. 3, 2 vs. 4
Ethnic group – no. (%)							
Hispanic: Yes	6838 (32.74%)	2227 (31.60%)	2277 (33.30%)	1554 (31.85%)	780 (36.84%)	0.0000392	1 vs. 2, 1 vs. 4, 2 vs. 4, 3 vs. 4
Hispanic: No	12248 (58.66%)	4354 (61.79%)	3901 (57.05%)	2856 (58.54%)	1137 (53.71%)	2.4E-12	1 vs. 2, 1 vs. 3, 1 vs. 4, 2 vs. 4, 3 vs. 4
Hispanic: Other/Missing	1795 (8.60%)	466 (6.61%)	660 (9.65%)	469 (9.61%)	200 (9.45%)	1.54E-11	1 vs. 2, 1 vs. 3, 1 vs. 4
Area deprivation index, Median (IQR)	15.0 (6.0-25.0)	16.0 (7.0-25.0)	14.0 (6.0-24.0)	15.0 (6.0-23.5)	15.0 (7.0-25.0)	4.21E-09	2 vs. 1
Healthcare utilization in the past 3 yr – no. (%)							
Inpatient 0	14631 (70.07%)	4109 (58.31%)	5299 (77.49%)	3601 (73.81%)	1622 (76.62%)	8.58E-156	1 vs. 2, 1 vs. 3, 1 vs. 4, 2 vs. 3, 3 vs. 4

Inpatient 1-2	4359 (20.88%)	1930 (27.39%)	1139 (16.66%)	929 (19.04%)	361 (17.05%)	4.15E-61	1 vs. 2, 1 vs. 3, 1 vs. 4, 2 vs. 3
Inpatient 3-4	1074 (5.14%)	553 (7.85%)	232 (3.39%)	212 (4.35%)	77 (3.64%)	1.71E-35	1 vs. 2, 1 vs. 3, 1 vs. 4, 2 vs. 3
Inpatient >=5	817 (3.91%)	455 (6.46%)	168 (2.46%)	137 (2.81%)	57 (2.69%)	1.16E-39	1 vs. 2, 1 vs. 3, 1 vs. 4
Outpatient 0	1241 (5.94%)	493 (7.00%)	378 (5.53%)	244 (5%)	126 (5.95%)	0.0000271	1 vs. 2, 1 vs. 3
Outpatient 1-2	2012 (9.64%)	694 (9.85%)	666 (9.74%)	439 (9%)	213 (10.06%)	0.366	-
Outpatient 3-4	1595 (7.64%)	508 (7.21%)	550 (8.04%)	371 (7.6%)	166 (7.84%)	0.312	-
Outpatient >=5	16033 (76.78%)	5352 (75.95%)	5244 (76.69%)	3825 (78.4%)	1612 (76.15%)	0.0154	1 vs. 3, 2 vs. 3, 3 vs. 4
Emergency 0	11321 (54.22%)	3496 (49.61%)	3884 (56.80%)	2733 (56.02%)	1208 (57.06%)	8.53E-20	1 vs. 2, 1 vs. 3, 1 vs. 4
Emergency 1-2	5975 (28.61%)	2123 (30.13%)	1880 (27.49%)	1388 (28.45%)	584 (27.59%)	0.00412	1 vs. 2, 1 vs. 3, 1 vs. 4
Emergency 3-4	1730 (8.29%)	646 (9.17%)	552 (8.07%)	361 (7.4%)	171 (8.08%)	0.00514	1 vs. 2, 1 vs. 3
Emergency >=5	1855 (8.89%)	782 (11.10%)	522 (7.63%)	397 (8.14%)	154 (7.27%)	3.1E-14	1 vs. 2, 1 vs. 3, 1 vs. 4
Index time period of patient – no. (%)							
03/20-06/20	6899 (33.04%)	2634 (37.38%)	2040 (29.83%)	1554 (31.85%)	671 (31.7%)	8.52E-21	1 vs. 2, 1 vs. 3, 1 vs. 4, 2 vs. 3
07/20-10/20	1180 (5.65%)	394 (5.59%)	390 (5.70%)	245 (5.02%)	151 (7.13%)	0.00606	1 vs. 4, 2 vs. 4, 3 vs. 4
11/20-02/21	8714 (41.73%)	2708 (38.43%)	3016 (44.11%)	2129 (43.64%)	861 (40.67%)	4.75E-12	1 vs. 2, 1 vs. 3, 2 vs. 4, 3 vs. 4
03/21-06/21	3390 (16.24%)	1112 (15.78%)	1113 (16.28%)	794 (16.27%)	371 (17.52%)	0.298	-
07/21-11/21	698 (3.34%)	199 (2.82%)	279 (4.08%)	157 (3.22%)	63 (2.98%)	0.000347	1 vs. 2, 2 vs. 3, 2 vs. 4
Acute phase severities of Covid-19 (-1~16 days) – no. (%)							
Hospitalized	9076 (43.47%)	4309 (61.15%)	2207 (32.28%)	1855 (38.02%)	705 (33.3%)	1.05E-301	1 vs. 2, 1 vs. 3, 1 vs. 4, 2 vs. 3, 3 vs. 4
Ventilation	495 (2.37%)	339 (4.81%)	85 (1.24%)	54 (1.11%)	17 (0.8%)	2.58E-59	1 vs. 2, 1 vs. 3, 1 vs. 4
Critical care	1144 (5.48%)	701 (9.95%)	210 (3.07%)	174 (3.57%)	59 (2.79%)	4.62E-89	1 vs. 2, 1 vs. 3, 1 vs. 4
Underlying Conditions (ICD10; Category[c]) – no. (%)							
Intestinal infectious diseases (A01-09; CIPD)	494 (2.37%)	197 (2.79%)	96 (1.40%)	88 (1.80%)	113 (5.34%)	5.5899E-26	1 vs. 2, 1 vs. 3, 1 vs. 4, 2 vs. 4, 3 vs. 4
Other sepsis (A41; CIPD)	1000 (4.79%)	547 (7.76%)	198 (2.9%)	197 (4.04%)	58 (2.74%)	1.531E-46	1 vs. 2, 1 vs. 3, 1 vs. 4, 2 vs. 3, 3 vs. 4
Benign neoplasm of digestive system (D13; Neoplasms)	101 (0.48%)	21 (0.30%)	11 (0.16%)	10 (0.20%)	59 (2.79%)	2.6718E-56	1 vs. 4, 2 vs. 4, 3 vs. 4
Benign neoplasm of unspecified sites (D36; Neoplasms)	115 (0.55%)	18 (0.26%)	8 (0.12%)	15 (0.31%)	74 (3.50%)	5.3098E-81	1 vs. 4, 2 vs. 3, 2 vs. 4, 3 vs. 4
Anaemia in chronic diseases (D63; DB)	1237 (5.92%)	725 (10.29%)	212 (3.1%)	217 (4.45%)	83 (3.92%)	1.5802E-80	1 vs. 2, 1 vs. 3, 1 vs. 4, 2 vs. 3

Other anaemias (D64; DB)	2178 (10.43%)	726 (10.29%)	683 (9.99%)	540 (11.07%)	229 (10.82%)	0.2598	-
Purpura and other haemorrhagic conditions (D65-D69; DB)	1477 (7.07%)	738 (10.47%)	342 (5.00%)	277 (5.67%)	120 (5.67%)	8.413E-41	1 vs. 2, 1 vs. 3, 1 vs. 4
Type 2 diabetes mellitus (E11; ENMD)	5173 (24.77%)	2384 (33.83%)	1196 (17.49%)	1131 (23.18%)	462 (21.82%)	1.1594E-112	1 vs. 2, 1 vs. 3, 1 vs. 4, 2 vs. 3, 2 vs. 4
Disorders of endocrine glands (E20-E35; ENMD)	1097 (5.25%)	447 (6.34%)	201 (2.94%)	251 (5.14%)	198 (9.35%)	6.923E-35	1 vs. 2, 1 vs. 3, 1 vs. 4, 2 vs. 3, 2 vs. 4, 3 vs. 4
Disorders of lipoprotein metabolism (E78; ENMD)	7517 (36.00%)	3090 (43.85%)	2049 (29.96%)	1724 (35.34%)	654 (30.89%)	2.3131E-69	1 vs. 2, 1 vs. 3, 1 vs. 4, 2 vs. 3, 3 vs. 4
Disorders of mineral metabolism (E83; ENMD)	1497 (7.17%)	761 (10.8%)	318 (4.65%)	294 (6.03%)	124 (5.86%)	2.2801E-47	1 vs. 2, 1 vs. 3, 1 vs. 4, 2 vs. 3, 2 vs. 4
Volume depletion (E86; ENMD)	1094 (5.24%)	573 (8.13%)	225 (3.29%)	215 (4.41%)	81 (3.83%)	3.7677E-40	1 vs. 2, 1 vs. 3, 1 vs. 4, 2 vs. 3
Disorders of fluid, electrolyte and acid-base balance (E87; ENMD)	2858 (13.69%)	1485 (21.07%)	593 (8.67%)	554 (11.35%)	226 (10.68%)	3.8121E-110	1 vs. 2, 1 vs. 3, 1 vs. 4, 2 vs. 3, 2 vs. 4
Organic mental disorders (F01-F09; MBD)	1204 (5.77%)	746 (10.58%)	140 (2.05%)	235 (4.81%)	83 (3.92%)	2.4693E-107	1 vs. 2, 1 vs. 3, 1 vs. 4, 2 vs. 3, 2 vs. 4
Depressive episode (F32; MBD)	1965 (9.41%)	839 (11.91%)	521 (7.62%)	445 (9.12%)	160 (7.56%)	1.4254E-18	1 vs. 2, 1 vs. 3, 1 vs. 4, 2 vs. 3, 3 vs. 4
Other anxiety disorders (F41; MBD)	1986 (9.51%)	624 (8.85%)	836 (12.23%)	403 (8.26%)	123 (5.81%)	1.5564E-22	1 vs. 2, 1 vs. 4, 2 vs. 3, 2 vs. 4, 3 vs. 4
Migraine (G43; DNS)	1039 (4.98%)	260 (3.69%)	453 (6.62%)	239 (4.9%)	87 (4.11%)	1.5353E-14	1 vs. 2, 1 vs. 3, 2 vs. 3, 2 vs. 4
Sleep disorders (G47; DNS)	2093 (10.02%)	665 (9.44%)	588 (8.60%)	619 (12.69%)	221 (10.44%)	2.751E-12	1 vs. 3, 2 vs. 3, 2 vs. 4, 3 vs. 4
Nerve, nerve root and plexus disorders (G50-G59; DNS)	759 (3.63%)	246 (3.49%)	170 (2.49%)	271 (5.55%)	72 (3.40%)	8.947E-17	1 vs. 2, 1 vs. 3, 2 vs. 3, 2 vs. 4, 3 vs. 4
Other disorders of brain (G93; DNS)	1389 (6.65%)	685 (9.72%)	340 (4.97%)	271 (5.55%)	93 (4.39%)	1.6964E-35	1 vs. 2, 1 vs. 3, 1 vs. 4, 3 vs. 4
Disorders of eyelid, lacrimal system and orbit (H00-H05; DE)	900 (4.31%)	296 (4.20%)	165 (2.41%)	370 (7.58%)	69 (3.26%)	1.9423E-41	1 vs. 2, 1 vs. 3, 2 vs. 3, 2 vs. 4, 3 vs. 4
Disorders of lens (H25-H28; DE)	1322 (6.33%)	567 (8.05%)	350 (5.11%)	298 (6.11%)	107(5.05%)	1.4737E-12	1 vs. 2, 1 vs. 3, 1 vs. 4, 2 vs. 3
Disorders of choroid and retina (H30-H36; DE)	638 (3.06%)	218 (3.09%)	150 (2.19%)	212 (4.35%)	58 (2.74%)	8.0234E-10	1 vs. 2, 1 vs. 3, 2 vs. 3, 3 vs. 4
Essential hypertension (I10; DCS)	8600 (41.19%)	3472 (49.27%)	2287 (33.45%)	2025 (41.5%)	816 (38.55%)	6.5003E-79	1 vs. 2, 1 vs. 3, 1 vs. 4, 2 vs. 3, 2 vs. 4, 3 vs. 4

Hypertensive heart disease (I11; DCS)	1103 (5.28%)	598 (8.49%)	199 (2.91%)	209 (4.28%)	97 (4.58%)	2.7382E-50	1 vs. 2, 1 vs. 3, 1 vs. 4, 2 vs. 3, 2 vs. 4
Hypertensive renal disease (I12; DCS)	1712 (8.20%)	995 (14.12%)	274 (4.01%)	334 (6.85%)	109 (5.15%)	1.1775E-113	1 vs. 2, 1 vs. 3, 1 vs. 4, 2 vs. 3, 2 vs. 4, 3 vs. 4
Hypertensive heart and renal disease (I13; DCS)	919 (4.40%)	552 (7.83%)	131 (1.92%)	174 (3.57%)	62 (2.93%)	2.465E-68	1 vs. 2, 1 vs. 3, 1 vs. 4, 2 vs. 3, 2 vs. 4
Ischaemic heart diseases (I20-I25; DCS)	4905 (23.49%)	2533 (35.94%)	998 (14.59%)	1027 (21.05%)	347 (16.39%)	3.6267E-213	1 vs. 2, 1 vs. 3, 1 vs. 4, 2 vs. 3, 2 vs. 4, 3 vs. 4
Other pulmonary heart diseases (I27; DCS)	853 (4.09%)	488 (6.92%)	148 (2.16%)	172 (3.53%)	45 (2.13%)	1.8174E-50	1 vs. 2, 1 vs. 3, 1 vs. 4, 2 vs. 3, 3 vs. 4
Atrial fibrillation and flutter (I48; DCS)	1607 (7.70%)	894 (12.69%)	311 (4.55%)	297 (6.09%)	105 (4.96%)	1.3607E-82	1 vs. 2, 1 vs. 3, 1 vs. 4, 2 vs. 3
Heart failure (I50; DCS)	2076 (9.94%)	1149 (16.3%)	379 (5.54%)	390 (7.99%)	158 (7.46%)	2.0673E-108	1 vs. 2, 1 vs. 3, 1 vs. 4, 2 vs. 3, 2 vs. 4
Acute upper respiratory infections (J00-J06; DRS)	3285 (15.73%)	962 (13.65%)	1330 (19.45%)	720 (14.75%)	273 (12.89%)	7.8839E-24	1 vs. 2, 2 vs. 3, 2 vs. 4, 3 vs. 4
Pneumonia, organism unspecified (J18; DRS)	1311 (6.28%)	465 (6.60%)	527 (7.71%)	235 (4.82%)	84 (3.97%)	2.3448E-13	1 vs. 2, 1 vs. 3, 1 vs. 4, 2 vs. 3, 2 vs. 4
Diseases of upper respiratory tract (J30-J39; DRS)	2790 (13.36%)	807 (11.45%)	1133 (16.57%)	601 (12.32%)	249 (11.76%)	7.2016E-20	1 vs. 2, 2 vs. 3, 2 vs. 4
Chronic obstructive pulmonary disease (J44; DRS)	1473 (7.05%)	518 (7.35%)	607 (8.88%)	257 (5.27%)	91 (4.3%)	4.5469E-18	1 vs. 2, 1 vs. 3, 1 vs. 4, 2 vs. 3, 2 vs. 4
Asthma (J45; DRS)	2272 (10.88%)	708 (10.05%)	684 (10.00%)	676 (13.86%)	204 (9.64%)	1.299E-12	1 vs. 3, 2 vs. 3, 3 vs. 4
Pleural effusion (J90; DRS)	545 (2.61%)	320 (4.54%)	96 (1.4%)	100 (2.05%)	29 (1.37%)	9.4426E-35	1 vs. 2, 1 vs. 3, 1 vs. 4, 2 vs. 3
Respiratory failure (J96; DRS)	1004 (5.81%)	465 (6.6%)	349 (5.10%)	144 (2.95%)	46 (2.17%)	9.7419E-26	1 vs. 2, 1 vs. 3, 1 vs. 4, 2 vs. 3, 2 vs. 4
Gastro-oesophageal reflux disease (K21; DDS)	3771 (18.06%)	1416 (20.09%)	1095 (16.01%)	905 (18.55%)	355 (16.77%)	3.6011E-09	1 vs. 2, 1 vs. 3, 1 vs. 4, 2 vs. 3
Diseases of stomach and duodenum (K31; DDS)	669 (3.20%)	221 (3.14%)	185 (2.71%)	106 (2.17%)	157 (7.42%)	6.8625E-31	1 vs. 3, 1 vs. 4, 2 vs. 4, 3 vs. 4
Noninfective enteritis and colitis (K50-K52; DDS)	712 (3.41%)	252 (3.57%)	165 (2.41%)	146 (2.99%)	149 (7.04%)	2.3918E-23	1 vs. 2, 1 vs. 4, 2 vs. 4, 3 vs. 4
Intestinal disorders (K59; DDS)	2556 (12.24%)	1105 (15.68%)	696 (10.18%)	556 (11.4%)	199 (9.4%)	1.1535E-26	1 vs. 2, 1 vs. 3, 1 vs. 4, 2 vs. 3, 3 vs. 4
Hematemesis (K92; DDS)	682 (3.27%)	252 (3.58%)	178 (2.6%)	128 (2.62%)	124 (5.86%)	1.3275E-13	1 vs. 2, 1 vs. 3, 1 vs. 4, 2 vs. 4, 3 vs. 4
Rheumatoid arthritis (M05-M06; DMSCT)	393 (1.88%)	95 (1.35%)	76 (1.11%)	182 (3.73%)	40 (1.89%)	1.6417E-26	1 vs. 3, 2 vs. 3, 2 vs. 4, 3 vs. 4
Gonarthrosis (M17; DMSCT)	963 (4.61%)	295 (4.19%)	206 (3.01%)	388 (7.95%)	74 (3.50%)	3.8129E-37	1 vs. 2, 1 vs. 3, 2 vs. 3, 3 vs. 4
Other arthrosis (M19; DMSCT)	1401 (6.71%)	531 (7.54%)	292 (4.27%)	477 (9.78%)	101 (4.77%)	3.4763E-34	1 vs. 2, 1 vs. 3, 1 vs. 4, 2 vs. 3, 3 vs. 4

Joint disorders (M25; DMSCT)	4074 (19.51%)	1427 (20.25%)	1053 (15.40%)	1196 (24.51%)	398 (18.80%)	2.8357E-33	1 vs. 2, 1 vs. 3, 2 vs. 3, 2 vs. 4, 3 vs. 4
Spondylopathies (M45-M49; DMSCT)	2637 (12.63%)	917 (13.01%)	717 (10.48%)	778 (15.94%)	225 (10.62%)	1.7986E-18	1 vs. 2, 1 vs. 3, 1 vs. 4, 2 vs. 3, 3 vs. 4
Soft tissue disorders (M60-M79; DMSCT)	7827 (37.48%)	2677 (37.99%)	2255 (32.98%)	2196 (45.01%)	699 (33.02%)	3.1706E-42	1 vs. 2, 1 vs. 3, 1 vs. 4, 2 vs. 3, 3 vs. 4
Other disorders of bone (M89; DMSCT)	832 (3.98%)	274 (3.89%)	223 (3.26%)	261 (5.35%)	74 (3.49%)	1.4758E-07	1 vs. 2, 1 vs. 3, 2 vs. 3, 3 vs. 4
Acute renal failure (N17; DGS)	2002 (9.59%)	1078 (15.3%)	373 (5.45%)	362 (7.42%)	189 (8.93%)	2.7279E-92	1 vs. 2, 1 vs. 3, 1 vs. 4, 2 vs. 3, 2 vs. 4, 3 vs. 4
Chronic kidney disease (N18; DGS)	2938 (14.07%)	1697 (24.08%)	557 (8.15%)	416 (8.53%)	268 (12.66%)	5.537E-197	1 vs. 2, 1 vs. 3, 1 vs. 4, 2 vs. 4, 3 vs. 4
Cystitis (N30; DGS)	667 (3.19%)	236 (3.35%)	173 (2.53%)	132 (2.71%)	126 (5.95%)	2.8725E-14	1 vs. 2, 1 vs. 3, 1 vs. 4, 2 vs. 4, 3 vs. 4
Diseases of female pelvic organs (N70-N77; DGS)	749 (3.59%)	162 (2.30%)	388 (5.67%)	130 (2.66%)	69 (3.26%)	1.4805E-28	1 vs. 2, 1 vs. 4, 2 vs. 3, 2 vs. 4
Disorders of female genital tract (N80-N98; DGS)	2717 (13.01%)	706 (10.02%)	1200 (17.55%)	534 (10.94%)	277 (13.08%)	8.5594E-43	1 vs. 2, 1 vs. 4, 2 vs. 3, 2 vs. 4, 3 vs. 4
Maternal care for pregnancy (O26; PCP)	522 (2.50%)	90 (1.28%)	274 (4.01%)	107 (2.19%)	51 (2.41%)	1.8855E-23	1 vs. 2, 1 vs. 3, 1 vs. 4, 2 vs. 3, 2 vs. 4
Other maternal diseases (O99; PCP)	566 (2.71%)	103 (1.46%)	289 (4.23%)	126 (2.58%)	48 (2.27%)	3.2785E-22	1 vs. 2, 1 vs. 3, 1 vs. 4, 2 vs. 3, 2 vs. 4
Abnormalities of heart beat (R00; Others)	3259 (15.61%)	1333 (18.92%)	917 (13.41%)	726 (14.88%)	283 (13.37%)	3.6165E-20	1 vs. 2, 1 vs. 3, 1 vs. 4, 2 vs. 3
Abnormalities of breathing (R06; Others)	2857 (13.68%)	919 (13.04%)	1246 (18.22%)	502 (10.29%)	190 (8.97%)	4.672E-45	1 vs. 2, 1 vs. 3, 1 vs. 4, 2 vs. 3, 2 vs. 4
Pain associated with micturition (R30; Others)	737 (3.53%)	257 (3.64%)	187 (2.73%)	161 (3.30%)	132 (6.24%)	8.5231E-13	1 vs. 2, 1 vs. 4, 2 vs. 4, 3 vs. 4
Abnormal findings of lung (R91; Others)	900 (4.31%)	329 (4.67%)	384 (5.62%)	132 (2.71%)	55 (2.60%)	2.2536E-16	1 vs. 2, 1 vs. 3, 1 vs. 4, 2 vs. 3, 2 vs. 4
Examination for infectious (Z11; Others)	2298 (11.00%)	693 (9.83%)	934 (13.66%)	467 (9.57%)	204 (9.64%)	8.2708E-16	1 vs. 2, 2 vs. 3, 2 vs. 4
Pregnancy examination and test (Z32; Others)	775 (3.71%)	148 (2.10%)	362 (5.29%)	182 (3.73%)	83 (3.92%)	2.1184E-21	1 vs. 2, 1 vs. 3, 1 vs. 4, 2 vs. 3, 2 vs. 4
Long term drug therapy (Z79; Others)	4177 (20.00%)	1994 (28.3%)	961 (14.05%)	868 (17.79%)	354 (16.72%)	2.0363E-104	1 vs. 2, 1 vs. 3, 1 vs. 4, 2 vs. 3, 2 vs. 4
Cardiac/vascular implants (Z95; Others)	1560 (7.47%)	856 (12.15%)	303 (4.43%)	297 (6.09%)	104 (4.91%)	4.3267E-75	1 vs. 2, 1 vs. 3, 1 vs. 4, 2 vs. 3
Weeks of gestation (Z3A; Others)	759 (3.63%)	124 (1.76%)	371 (5.43%)	176 (3.61%)	88 (4.16%)	4.5153E-29	1 vs. 2, 1 vs. 3, 1 vs. 4, 2 vs. 3, 2 vs. 4

a. **P-value**: continuous variables are tested by one-way analysis of variance (ANOVA, with Tukey HSD post hoc test) for Normal distribution or by Kruskal-Wallis test (with Dunn post hoc test) for non-Normal distribution; categorical variables are tested by Fisher's exact test with pair-wise Fisher test for post hoc analysis.
b. **IQR**: inter-quartile range.
c. **Abbreviation for ICD10 category**. We grouped ICD10 code according to the first three digits. CIPD: Certain Infectious and Parasitic Diseases; DB: Diseases of the Blood; ENMD: Endocrine, Nutritional and Metabolic Diseases; MBD: Mental and Behavioral Disorders; DNS: Diseases of the Nervous System; DE: Diseases of the ear; DCS: Diseases of the Circulatory System; DRS: Diseases

of the Respiratory System; DDS: Diseases of the Digestive System; DMSCT: Diseases of the Musculoskeletal System and Connective Tissue; DGS: Diseases of the Genitourinary System; PCP: Pregnancy, Childbirth and the Puerperium.

medRxiv preprint doi: https://doi.org/10.1101/2022.05.21.22275412; this version posted May 23, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

Replication on OneFlorida+ CRN

We repeated the same subphenotyping process on the OneFlorida+ cohort; the subphenotypes generated were highly overlapping with those from the INSIGHT cohort. Following the same analytical pipeline (Figure 1), we first identified incident diagnoses for the 137 potential PASC conditions with 30-180 days of testing positive for SARS-CoV-2; we then conducted topic modeling. Supplemental Figure 6 displays the heatmap of all learned potential PASC topics, where we see topics concentrated on problems with the musculoskeletal system (T1), digestive system (T2), nervous system (T5), and topics mixed with respiratory system problems and blood/circulatory system problems (T3), as well as headache and sleep-wake problems (T7). Some topics also include a mixture of diagnoses. For example, T9 is throat/chest pain mixed with breathing/heartbeat abnormalities; T6 is a mixture of musculoskeletal pain, headache, malaise and fatigue, and skin sensory problems; T8 and T10 are topics mixed with electrolyte/fluid disorders and anemia/arrhythmias, and T4 is a topic mixing up problems involving digestive, nervous and respiratory systems. To quantify areas of overlap with findings from the INSIGHT cohort, we evaluated the pairwise similarities between the topics learned from the two different cohorts (Methods) and visualized the results in Supplemental Figure 4, which clearly showed a one-to-one correspondence between the topics learned from the two cohorts (as identified by darker colors on the diagonal line).

With the learned topics, we also built topic loading-based representations for patients in the OneFlorida+ cohort, derived four potential PASC subphenotypes with HAC (Methods), and described the characteristics of patients classified into each subphenotype. (Supplemental Table 3, Figures 7 and 8). Subphenotype 1 was dominated by incidental blood and circulatory system problems in the post-acute infection period, which included 25.43% of the patients who were older (with a median age of 62.0 and IQR [49.0-74.0]), with the highest proportion of males (46.93%, compared to 38.29% for the overall population) and the highest rates of hospitalization (57.34%, compared to 36.69% for the overall population), mechanical ventilation (8.57%, compared to 3.39% for the overall population) and critical care admission (12.52%, compared to 6.07% for the overall population) in the acute phase of COVID-19. This subphenotype was associated with a higher prevalence of underlying conditions and more new prescriptions for medications treating circulatory system, blood, and endocrine problems. Subphenotype 2 was dominated by incidental respiratory problems and was the largest subphenotype containing 5281 (38.48%) patients with a median age of 47.0 (IQR [33.0-61.0]). They had a higher prevalence of respiratory conditions at baseline including chronic obstructive pulmonary disease, pneumonia, upper respiratory tract problems, and had a higher post-acute infection incident prescription for respiratory medications. Subphenotype 3 was dominated by incident problems with musculoskeletal and nervous problems in the post-acute infection phase. It included 3205 (23.35%) patients with a median age of 48.0 (IQR [33.0-61.0]) and had the lowest hospitalization rate in the acute phase (27.8%). This subphenotype had a higher prevalence of baseline musculoskeletal and connective tissue problems and asthma, and more new prescriptions for pain medications, including ketorolac and ibuprofen in the post-acute infection phase. Subphenotype 4 was dominated by incidental digestive problems. It was the youngest (median age 46.0 [32.0-60.0]) and smallest (including 1748 [12.74%] patients) subphenotype, with the highest proportion of females (67.11%, compared to 61.70% overall) and lowest rates of mechanical ventilation (0.97%) and critical care admission (2.8%) in the acute phase. This subphenotype was associated with a higher baseline burden of digestive system problems, and more new prescriptions for medications focused on the digestive system. These observations and characterizations were highly consistent with the subphenotypes identified from the INSIGHT cohort. In addition, the difference in the incidence patterns of 28 selected potential PASC conditions in 30-180 days after the COVID-19 lab test between positive and matched

medRxiv preprint doi: https://doi.org/10.1101/2022.05.21.22275412; this version posted May 23, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

negative patients on the OneFlorida+ cohort is shown in Supplemental Figure 9, which was also highly consistent with the results from the INSIGHT cohort.

Discussion

Many studies have pointed out the existence of a diverse set of symptoms and signs that may develop, persist, or recur in the post-acute SARS-CoV-2 infection period[7]. These conditions, typically referred to as PASC, involve a wide range of organ systems. Different from most of the existing research which studied these conditions independently, we developed a data-driven framework shown in Figure 1 to identify subphenotypes of SARS-CoV-2 infected patients based on newly incident symptoms and signs from 30 to 180 days (the post-acute period) after their infection confirmation, such that patients within the same subphenotype share a similar distribution of potential PASC condition incidences in the post-acute periods.

Within both INSIGHT and OneFlorida+ CRNs, we identified four consistent subphenotypes dominated by new conditions of the blood and circulatory systems (Subphenotype 1), respiratory system (Subphenotype 2), musculoskeletal and nervous systems (Subphenotype 3), and digestive system (Subphenotype 4).

Comparing the subphenotypes derived from both cohorts, we observed that Subphenotype 1 included older patients with higher baseline comorbidity burden, with greater severity of illness during the acute phase (e.g., higher rates of hospitalization, critical care admission and mechanical ventilation) and with higher incident prescription rates of medications for treating diseases of many different organ systems. This subphenotype had the highest proportion of males, which aligns with the finding that males had more severe acute SARS-CoV-2 infections[15]. This subphenotype had a large proportion of patients with their SARS-CoV-2 infection confirmed during the early pandemic (March to September 2020) when treatment standards were still evolving. Temporally, NYC was the epicenter for the first wave. This may explain the observation that this was the largest subphenotype for INSIGHT (containing 33.75% of the patients) but the second largest subphenotype for OneFlorida+ (containing 25.43% of the patients). Early cases had greater acute phase severity, which may explain the more severe incident conditions (e.g., heart failure and renal failure) in the post-acute infection period of these patients, which could be caused by the hyperinflammation during the acute phase[16].

Subphenotype 2 is another major subphenotype for both cohorts (the second largest for INSIGHT occupying 32.75% of the patients, and the largest for OneFlorida+ accounting for 38.48% of the population). It is the youngest subphenotype for INSIGHT and the second youngest subphenotype for OneFlorida+ and contains the highest proportion of patients who had an acute SARS-CoV-2 infection confirmed from July to November 2021. The young age and infection recency are consistent with the milder incident PASC conditions of this subphenotype. Of note, patients in this subphenotype had a high baseline rate of pulmonary comorbidities which is likely correlated with the high rate of incident respiratory medication prescriptions in the post-acute SARS-CoV-2 period.

Subphenotype 3 and 4 were two smaller subphenotypes with Subphenotype 3 associated with musculoskeletal and neurological conditions, whereas Subphenotype 4 was associated with gastrointestinal conditions. Patients in Subphenotype 3 also suffered from dermatologic conditions and had the highest rates of related conditions at baseline, including autoimmune diagnoses such as rheumatoid arthritis and allergy conditions. Conversely, patients in Subphenotype 4 had the mildest acute phase severity (e.g., lowest rates of mechanical ventilation and critical care admissions.

medRxiv preprint doi: https://doi.org/10.1101/2022.05.21.22275412; this version posted May 23, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

There are several strengths in our study. First, we adopted a topic modeling approach to derive compact patient representations based on the co-incidence patterns across different diagnoses. Unlike other dimensionality reduction techniques such as Principal Component Analysis (PCA)[17], topic modeling is designed specifically for data samples with binary or count features[18,19] and thus appropriate for our analysis. Second, INSIGHT and OneFlorida+ include patients from distinct geographic regions in the US with different characteristics, allowing us to validate the robustness of the derived subphenotypes. Third, our study period (March 2020 to November 2021) covers different COVID-19 waves associated with different SARS-CoV-2 virus variants. Our study cohorts contained robust patient populations in New York and Florida, representing the different waves of SARS-CoV-2 infected cases in US. This is important factor contributing to the different distributions of the four subphenotypes in the two cohorts.

Our study is not without limitations. First, our analysis is based on longitudinal observational patient data, which cannot explain the biological mechanisms behind PASC directly. Second, the PASC diagnoses we investigated were encoded as CCSR categories, which may not reflect the co-incidence patterns of fine-grained diagnosis conditions in the context of PASC. Third, we focused on new incidences of conditions in the post-acute infection period for COVID-19 patients and did not consider pre-existing conditions that are persistent or exacerbated due to the acute SARS-CoV-2 infection. Finally, our study period did not represent the recent wave dominated by the Omicron variants of SARS-CoV-2.

To summarize, our study dissects the complexity and heterogeneity of newly incident conditions in 30-180 days after SARS-CoV-2 infection confirmation into four reproducible subphenotypes based on the EHR repositories from two large CRNs using machine learning. These four subphenotypes included a severe one involving problems with the blood and circulatory system and associated with high baseline comorbidity burden and disease severity in its acute phase, a milder one in younger people mainly with respiratory problems, and two pain-dominated ones (musculoskeletal/nervous system pain and abdominal pain respectively). Overall, patients in each subphenotype tend to have higher rates of related conditions in the baseline period. Our study provides the first systematic study on the co-incidence patterns of conditions in the post-acute infection period of SARS-CoV-2 infected adult patients, which can inform more nuanced and tailored diagnosis and treatment plans.

Methods

EHR Data Repositories

Two large-scale de-identified real-world EHR data warehouses were utilized in our analyses. Our first cohort data was based on the EHR from INSIGHT CRN[9], which contains the longitudinal clinical information of around 12 million patients in New York City area. Our second cohort data was based on the EHR from the OneFlorida+ CRN[10], which contains the information of nearly 15 million patients majorly from Florida and selected cities in Georgia and Alabama. The use of the INSIGHT data was approved by the Institutional Review Board (IRB) of Weill Cornell Medicine following protocol 21-10-95-380 with title “Adult PCORnet-PASC Response to the Proposed Revised Milestones for the PASC EHR/ORWD Teams (RECOVER)”. The use of the OneFlorida+ data for this study was approved under the University of Florida IRB number IRB202001831.

Potential PASC Conditions

medRxiv preprint doi: https://doi.org/10.1101/2022.05.21.22275412; this version posted May 23, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

We compiled a list of 137 diagnostic categories covering near 6,500 ICD-10-CM codes as potential PASC conditions for our study. The list was built based on the Clinical Classifications Software Refined (CCSR) v2022.1 covering 66,534 ICD-10-CM Diagnoses. The codes that would not plausibly be considered post-acute sequelae of COVID-19 in the adult population (e.g., HIV, tuberculosis, infection by non-COVID causes, neoplasms, injury due to external causes, etc.) were excluded, and parent codes (e.g., the first 3-digits of ICD-10 codes) were systematically added. The full list of diagnosis codes for the PASC is provided in Supplemental Table 1.

Cohort Construction

For the both cohorts, adult patients (age ≥ 20) with at least one SARS-CoV-2 polymerase-chain-reaction (PCR) or antigen laboratory test (Supplemental Table 2) between March 01, 2020 and November 30, 2021 were selected. Then we chose the patients who had at least one positive test and had at least one potential PASC conditions in the follow-up (or post-acute infection) period defined as below. We further made sure those potential PASC conditions were new incidences in the follow-up period by excluding patients who had any of them in both baseline and follow-up periods. The overall inclusion-exclusion cascade was shown in Supplemental Figure 1, and the relevant definitions are provided below.

- Index date: the date of the first COVID-19 positive test.
- Baseline period: from 3 years to one week prior to the index date.
- Follow-up (post-acute infection) period: from 31 days after the index date to the day of documented death, last record in the database, 180 days after baseline, or the end of our observational window (Nov. 30, 2021), whichever came first.

Topic Modeling

We use binary vectors $\{\boldsymbol{x}_n\}_{n=1}^{N}$ to represent the patients, where n is the patient index the i-th element of $\boldsymbol{x}_n$, or $\boldsymbol{x}_n(i) = 1$ if the i-th potential PASC condition appears in the post-acute infection period of the n-th patient's EHR, otherwise $\boldsymbol{x}_n(i) = 0$. Therefore, each $\boldsymbol{x}_n$ is a 137-dimensional binary vector (Step 1 in Figure 1). Topic modeling (TM)[14] as then applied on these vectors to learn a set of potential PASC topics. Specifically, assume that each patient can be represented as a mixture of K latent PASC topics $\boldsymbol{\Phi} \in R^{D\times K}$, where each topic $\boldsymbol{\varphi}_k$ can be viewed as a set of PASC that are more likely to be co-incident in the post-acute infection period of a particular patient (Step 2 in Figure 1). Then for each patient, TM infers the mixture memberships $\boldsymbol{\theta}_n \in R^K$, also called topic proportions or topic loadings, as the new representation for each patient (Step 3 in Figure 1). A patient with higher loading value on a particular topic indicates that he/she suffers from more co-incident condition patterns from this topic. In other words, TM transforms the representations of each patient from the original 137-dimensional binary space $\boldsymbol{x}_n$ to the low-dimensional continuous PASC topic space $\boldsymbol{\theta}_n$, which will be leveraged further for subphenotype identification through clustering later (Step 4 in Figure 1). Specifically, we used Poisson factor analysis (PFA)[20] as the concrete TM method, which generates $\boldsymbol{x}_n$ as follows.

- Draw a topic proportion $\boldsymbol{\theta}_n$ for the n-th patient from a Gamma distribution $\boldsymbol{\theta}_n \sim Gamma(1,1)$;
- Draw the k-th PASC topic from a Dirichlet distribution $\boldsymbol{\varphi}_k \sim Dirichlet(0.01)$, $k = 1, \cdots, K$;
- Draw a binary vector $\boldsymbol{x}_n$ from the Bernoulli distribution by Bernoulli-Poisson link[21]:

$$\boldsymbol{x}_n = \mathbf{1}(\boldsymbol{u}_n \geq 1),\ \boldsymbol{u}_n \sim Poisson(\boldsymbol{\Phi}\boldsymbol{\theta}_n);$$

medRxiv preprint doi: https://doi.org/10.1101/2022.05.21.22275412; this version posted May 23, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

where $\mathbf{1}(\cdot)$ is an indicator function representing $\boldsymbol{x}_n = 1$ if $\boldsymbol{u}_n \geq 1$, and $\boldsymbol{x}_n = 0$ if $\boldsymbol{u}_n = 0$. We use Gibbs sampling[22] to infer the posterior of PASC topic $\boldsymbol{\Phi}$ and topic proportions $\{\boldsymbol{\theta}_n\}_{n=1}^{N}$.

Determining the Number of Topics

The number of topics, K, is an important parameter in TM. To determine an optimal K based on the data, we used two metrics: data likelihood and topic coherence[23]. Data likelihood is used to evaluate the fitness of the model on the current data set, where the larger value indicates better fitness. Topic coherence is used to evaluate the relevance of our learned PASC topics to the investigative condition list, where the value is from zero to one and higher value indicates better coherence. Detailed calculations of these two metrics are provided in the Supplemental methods and results are shown in Supplemental Figure 2. From this figure, we can see that more topics can provide higher data likelihood because we have more topics to represent the original PASCs. However, more topics may bring down topic coherence, which suggests the redundancy between the newly added and the old ones. With these considerations, we set the final number of topics as 10 for both INSIGHT and OneFlorida+ cohorts as it achieved the best topic coherence and reasonable data likelihood (we do not want the data likelihood to be too perfect as that may suggests overfitting).

Topic Robustness

As TM is a probabilistic process, we want to guarantee the robustness of the identified topics. To achieve this goal, we firstly did 1000 bootstrapping (randomly choose 80% patients from each cohort) to learn topics. Then according to the importance of each topic (mean topic loadings over all patients), we reordered the topics and calculated the cosine similarity among all topics from each pair of bootstrapping:

$$S_{pq} = \frac{1}{1000} \sum_{i=1}^{1000} \frac{1}{999} \sum_{j=1, j \neq i}^{1000} cos\left(\boldsymbol{\varphi}_p^{(i)}, \boldsymbol{\varphi}_q^{(j)}\right),$$

where $\boldsymbol{\varphi}_p^{(i)}$ is the p-th topic vector learned from the i-th bootstrapped samples, and S_{pq} is the similarity between the p-th topic and the q-th topic. Supplemental Figure 3 demonstrated the heatmap of the similarity matrix with S_{pq} as its (p, q)-th entry, from which we can clearly observe a darker diagonal line, which indicates a high similarity between the topics learned from different bootstrapped samples and thus implies the learned topics are robust.

Topic Consistency

We also quantitatively evaluated the consistency between the two set of topics learned from different cohorts. Specifically, denoting the topic matrices learned from the two cohorts as $\boldsymbol{\Phi}^{(1)}$ and $\boldsymbol{\Phi}^{(2)}$, then we can evaluate the consistency between the i-th topic in cohort 1 and the j-th topic in cohort 2 as the cosine similarity between their corresponding topic vector $\boldsymbol{\varphi}_i^{(1)}$ and $\boldsymbol{\varphi}_j^{(2)}$ as:

$$S_{ij} = cos\left(\boldsymbol{\varphi}_i^{(1)}, \boldsymbol{\varphi}_j^{(2)}\right),\ i, j = 1, \cdots, K,$$

medRxiv preprint doi: https://doi.org/10.1101/2022.05.21.22275412; this version posted May 23, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

Finally, the heatmap of the topic consistency matrix with S_{ij} as its (i, j)-th entry was shown in Supplemental Figure 4, from which we can clearly observe darker diagonal values, which suggests high consistency between the two sets of topics.

Subphenotyping through Clustering

With the learned K-dimensional topic loading vector for n-th patient $\boldsymbol{\theta}_n$, we applied hierarchical agglomerative clustering method with Euclidean distance calculation and Ward linkage criterion[24] to derive subphenotypes as patient clusters. For determining the optimal number of clusters (subphenotypes), we applied NbClust R package[25], which includes 21 cluster indices to evaluate the quality of clusters. With the patients from the INSIGHT and OneFlorida+ CRN, 13 and 12 out of the 21 indices agreed 4 is the optimal number of clusters. Through majority voting, we set the number of clusters as 4 in both two cohorts. Supplemental Figure 5 demonstrates the UMAP embeddings[26] and dendrogram of these clusters for both cohorts.

We have also examined the robustness of the identified clusters on both cohorts. Specifically, for each cohort, we used the subphenotypes derived from all patients as references, so that the subphenotype index for the n-th patient is denoted as y_n. Then we ran 1000 bootstrapping (randomly choose 80% patients from the cohort denoted as set $\Omega_{\mathrm{i}}, i = 1, \cdots, 1000$) to learn topic model and then derive subphenotypes with the same procedure as described above. For each bootstrapped sample set i, we can obtain another subphenotype index for n-th patient from Ω_{i} as $\hat{y}_{n,i}$. We calculated mean and 95% confidence interval of adjusted rand score (ARI) and normalized mutual information (NMI) between the clustering results on bootstrapped sample sets and reference as 0.902 (95% confidence interval (CI): $0.863 - 0.927$) and 0.937 (95% CI: $0.908 - 0.952$) for the INSIGHT cohort, and 0.914 (95% CI: $0.907 - 0.929$) and 0.950 (95% CI: $0.936 - 0.968$) for the OneFlorida+ cohort, which suggests the identified clusters are highly robust.

Comparison with SARS-CoV-2 Infection Negative Patients

We compared the co-incidence patterns of the investigative conditions in the follow-up periods for patients with SARS-CoV-2 infection testing positive and negative. The SARS-CoV-2 infection negative patients are with all negative results for their SARS-CoV-2 infection lab tests during March 2020 to November 2021, and there were no documented COVID-19 related diagnoses during this time period. The index date for each individual patient in non-infected group is defined as the date of the first (negative) lab test.

To make fair comparisons, we performed similarity matching to identify appropriate negative patients for each positive patient based on the following hypothetical confounding variables.

- Demographics: age, gender, race, and ethnicity, where age was binned into different groups (20-<40 years, 40-<55 years, 55-<65 years, 65-<75 years, 75-<85 years, 85+ years).
- The area deprivation index (10-rank bins of national ADI) for capturing socioeconomic disadvantage of patients' neighborhood[13].
- Index date for considering the effect of different stages of pandemic, which was binned into different time intervals (March 2020 – June 2020, July 2020 – October 2020, November 2020 - February 2021, March 2021 – June 2021, July 2021 – November 2021).

medRxiv preprint doi: https://doi.org/10.1101/2022.05.21.22275412; this version posted May 23, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

- Medical utilizations measured by numbers of inpatient, outpatient, and emergency encounters in the baseline period (binned into 0 visit, 1 or 2 visits, 3 or 4 visits, 5+ visits for each encounter type).
- Coexisting conditions including comorbidities and medications based on a tailored list of the Elixhauser comorbidities[27]. We defined the patient having a particular condition if he/she had at least two related records during the baseline period.

For identifying the negative controls for each patient in a particular subphenotype, we first required exact match for confounders of demographics, ADI, and index date to obtain an initial set, and then performed robust propensity score (PS) matching on other hypothetical confounders robust propensity score to rank the patients in the initial set and we finally picked the top 2. We used standardized mean difference (SMD) to quantify the goodness-of-balance of confounders between two groups $SMD(X_1, X_0) = \frac{|E[X_1]-E[X_0]|}{\sqrt[2]{(Var(X_1)+Var(X_0))/2}}$, where $SMD < 0.2$ is the threshold to examine whether this confounder is balanced[28]. On both ISNIGHT and OneFlorida cohort, we found that all confounders on all subphenotypes were balanced.

Reference

1 Xie, Y., Xu, E., Bowe, B. & Al-Aly, Z. Long-term cardiovascular outcomes of COVID-19. *Nat Med* **28**, 583-590 (2022). https://doi.org:10.1038/s41591-022-01689-3

2 Xie, Y., Xu, E. & Al-Aly, Z. Risks of mental health outcomes in people with covid-19: cohort study. *BMJ* **376**, e068993 (2022). https://doi.org:10.1136/bmj-2021-068993

3 Xie, Y. & Al-Aly, Z. Risks and burdens of incident diabetes in long COVID: a cohort study. *Lancet Diabetes Endocrinol* (2022). https://doi.org:10.1016/S2213-8587(22)00044-4

4 Bowe, B., Xie, Y., Xu, E. & Al-Aly, Z. Kidney Outcomes in Long COVID. *J Am Soc Nephrol* **32**, 2851-2862 (2021). https://doi.org:10.1681/ASN.2021060734

5 Daugherty, S. E. *et al.* Risk of clinical sequelae after the acute phase of SARS-CoV-2 infection: retrospective cohort study. *BMJ* **373**, n1098 (2021). https://doi.org:10.1136/bmj.n1098

6 Crook, H., Raza, S., Nowell, J., Young, M. & Edison, P. Long covid—mechanisms, risk factors, and management. *bmj* **374** (2021).

7 Al-Aly, Z., Xie, Y. & Bowe, B. High-dimensional characterization of post-acute sequelae of COVID-19. *Nature* **594**, 259-264 (2021). https://doi.org:10.1038/s41586-021-03553-9

8 Xie, Y., Bowe, B. & Al-Aly, Z. Burdens of post-acute sequelae of COVID-19 by severity of acute infection, demographics and health status. *Nat Commun* **12**, 6571 (2021). https://doi.org:10.1038/s41467-021-26513-3

9 Kaushal, R. *et al.* Changing the research landscape: the New York City Clinical Data Research Network. *J Am Med Inform Assoc* **21**, 587-590 (2014). https://doi.org:10.1136/amiajnl-2014-002764

10 Shenkman, E. *et al.* OneFlorida Clinical Research Consortium: Linking a Clinical and Translational Science Institute With a Community-Based Distributive Medical Education Model. *Acad Med* **93**, 451-455 (2018). https://doi.org:10.1097/ACM.0000000000002029

11 *Clinical Classifications Software Refined (CCSR)*, <https://www.hcup-us.ahrq.gov/toolssoftware/ccsr/ccs_refined.jsp> (

12 *RECOVER: Researching COVID to Enhance Recovery*, <https://recovercovid.org> (

medRxiv preprint doi: https://doi.org/10.1101/2022.05.21.22275412; this version posted May 23, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

13 Kind, A. J. & Buckingham, W. R. Making neighborhood-disadvantage metrics accessible—the neighborhood atlas. *The New England journal of medicine* **378**, 2456 (2018).
14 Blei, D. M. Probabilistic topic models. *Communications of the ACM* **55**, 77-84 (2012).
15 Peckham, H. *et al.* Male sex identified by global COVID-19 meta-analysis as a risk factor for death and ITU admission. *Nat Commun* **11**, 6317 (2020). https://doi.org:10.1038/s41467-020-19741-6
16 Zaim, S., Chong, J. H., Sankaranarayanan, V. & Harky, A. COVID-19 and Multiorgan Response. *Curr Probl Cardiol* **45**, 100618 (2020). https://doi.org:10.1016/j.cpcardiol.2020.100618
17 Abdi, H. & Williams, L. J. Principal component analysis. *Wiley interdisciplinary reviews: computational statistics* **2**, 433-459 (2010).
18 Blei, D. M., Ng, A. Y. & Jordan, M. I. Latent dirichlet allocation. *the Journal of machine Learning research* **3**, 993-1022 (2003).
19 Griffiths, T. L. & Steyvers, M. Finding scientific topics. *Proceedings of the National academy of Sciences* **101**, 5228-5235 (2004).
20 Zhou, M., Hannah, L., Dunson, D. & Carin, L. Beta-negative binomial process and Poisson factor analysis. *Artificial Intelligence and Statistics*, 1462-1471 (2012).
21 Zhang, Y., Zhao, Y., David, L., Henao, R. & Carin, L. in *2016 IEEE 16th International Conference on Data Mining (ICDM).* 1359-1364 (IEEE).
22 Gelfand, A. E. Gibbs sampling. *Journal of the American statistical Association* **95**, 1300-1304 (2000).
23 Newman, D., Lau, J. H., Grieser, K. & Baldwin, T. in *Human language technologies: The 2010 annual conference of the North American chapter of the association for computational linguistics.* 100-108.
24 Murtagh, F. & Legendre, P. Ward's hierarchical agglomerative clustering method: which algorithms implement Ward's criterion? *Journal of classification* **31**, 274-295 (2014).
25 Charrad, M., Ghazzali, N., Boiteau, V. & Niknafs, A. NbClust: an R package for determining the relevant number of clusters in a data set. *Journal of statistical software* **61**, 1-36 (2014).
26 McInnes, L., Healy, J. & Melville, J. Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426* (2018).
27 *Elixhauser Comorbidity Software Refined for ICD-10-CM*, <https://www.hcup-us.ahrq.gov/toolssoftware/comorbidityicd10/comorbidity_icd10.jsp> (
28 Ozery-Flato, M., Goldschmidt, Y., Shaham, O., Ravid, S. & Yanover, C. Framework for identifying drug repurposing candidates from observational healthcare data. *JAMIA Open* **3**, 536-544 (2020). https://doi.org:10.1093/jamiaopen/ooaa048

Acknowledgement

This research was funded by the National Institutes of Health (NIH) Agreement OTA OT2HL161847 (contract number EHR-01-21) as part of the Researching COVID to Enhance Recovery (RECOVER) research program.

medRxiv preprint doi: https://doi.org/10.1101/2022.05.21.22275412; this version posted May 23, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

Supplemental Figure and Table

Supplemental Table 1. PASC list in our study defined by CCSR domain with corresponding ICD-10-CM codes.

Supplemental Table 2. SARS-CoV-2 polymerase-chain-reaction or antigen laboratory test.

Supplemental Table 3. Characteristics of the identified subphenotypes on the OneFlorida+ cohort.

Variable	Total	Subphenotype 1 (Blood & Circulatory System)	Subphenotype 2 (Respiratory System)	Subphenotype 3 (Musculoskeletal and Nervous System)	Subphenotype 4 (Digestive System)	P-value[a]	Post hoc analysis
No. of patients (%)	13724 (100%)	3490 (25.43%)	5281 (38.48%)	3205 (23.35%)	1748 (12.74%)		
Age, y, Median (IQR[b])	51.0 (35.0-65.0)	62.0 (49.0-74.0)	47.0 (33.0-61.0)	48.0 (33.0-61.0)	46.0 (32.0-60.0)	2.14E-267	2 vs. 1, 3 vs. 1, 4 vs. 1
Age group – no. (%)							
20-<40 years	4284 (31.22%)	502 (14.38%)	1968 (37.27%)	1135 (35.41%)	679 (38.84%)	5.72E-135	1 vs. 2, 1 vs. 3, 1 vs. 4, 3 vs. 4
40-<55 years	3318 (24.18%)	636 (18.22%)	1347 (25.51%)	873 (27.24%)	462 (26.43%)	3.3402E-20	1 vs. 2, 1 vs. 3, 1 vs. 4
55-<65 years	2528 (18.42%)	789 (22.61%)	891 (16.87%)	574 (17.91%)	274 (15.68%)	1.2492E-12	1 vs. 2, 1 vs. 3, 1 vs. 4, 3 vs. 4
65-<75 years	1923 (14.01%)	756 (21.66%)	591 (11.19%)	379 (11.83%)	197 (11.27%)	3.7093E-49	1 vs. 2, 1 vs. 3, 1 vs. 4
75-<85 years	1178 (8.58%)	559 (16.02%)	347 (6.57%)	178 (5.55%)	94 (5.38%)	5.765E-72	1 vs. 2, 1 vs. 3, 1 vs. 4
85+ years	493 (3.59%)	248 (7.11%)	137 (2.59%)	66 (2.06%)	42 (2.4%)	2.6827E-36	1 vs. 2, 1 vs. 3, 1 vs. 4
Sex – no. (%)							
Female	8468 (61.70%)	1852 (53.07%)	3445 (65.23%)	1998 (62.34%)	1173 (67.11%)	1.6794E-34	1 vs. 2, 1 vs. 3, 1 vs. 4, 2 vs. 3, 3 vs. 4
Male	5255 (38.29%)	1638 (46.93%)	1835 (34.75%)	1207 (37.66%)	575 (32.89%)	1.439E-34	1 vs. 2, 1 vs. 3, 1 vs. 4, 2 vs. 3, 3 vs. 4
Other/Missing	1 (0.00%)	0 (0%)	1 (0.02%)	0 (0%)	0 (0%)	0.65965	-
Race – no. (%)							
Asian	144 (1.05%)	33 (0.95%)	57 (1.08%)	30 (0.94%)	24 (1.37%)	0.46317	-
Black or African American	4076 (29.70%)	1093 (31.32%)	1506 (28.52%)	992 (30.95%)	485 (27.75%)	0.0036448	1 vs. 2, 1 vs. 4, 2 vs. 3, 3 vs. 4
White	7175 (52.28%)	1811 (51.89%)	2813 (53.27%)	1658 (51.73%)	893 (51.09%)	0.30123	-
Other	2123 (15.47%)	510 (14.61%)	816 (15.45%)	476 (14.85%)	321 (18.36%)	0.0027825	1 vs. 4, 2 vs. 4, 3 vs. 4
Missing	206 (1.50%)	43 (1.23%)	89 (1.69%)	49 (1.53%)	25 (1.43%)	0.39223	-
Ethnic group – no. (%)							
Hispanic: Yes	2881 (20.99%)	575 (16.48%)	1173 (22.21%)	673 (21%)	460 (26.32%)	1.0397E-16	1 vs. 2, 1 vs. 3, 1 vs. 4, 2 vs. 4, 3 vs. 4
Hispanic: No	9329 (67.98%)	2459 (70.46%)	3556 (67.34%)	2216 (69.14%)	1098 (62.81%)	1.7398E-07	1 vs. 2, 1 vs. 4, 2 vs. 4, 3 vs. 4
Hispanic: Other/Missing	1514 (11.03%)	456 (13.07%)	552 (10.45%)	316 (9.86%)	190 (10.87%)	0.0001027	1 vs. 2, 1 vs. 3, 1 vs. 4
Area deprivation index, Median (IQR)	59.0 (42.0-76.0)	59.0 (43.0-76.0)	59.0 (42.0-76.0)	60.0 (43.0-77.0)	58.0 (42.0-74.0)	0.0331	-
Healthcare utilization in the past 3 yr – no. (%)							
Inpatient 0	7437 (54.19%)	1359 (38.94%)	3059 (57.92%)	1965 (61.31%)	1054 (60.3%)	7.4073E-97	1 vs. 2, 1 vs. 3, 1 vs. 4, 2 vs. 3
Inpatient 1-2	3018 (21.99%)	896 (25.67%)	1109 (21%)	650 (20.28%)	363 (20.77%)	3.4309E-08	1 vs. 2, 1 vs. 3, 1 vs. 4
Inpatient 3-4	1195 (8.71%)	398 (11.4%)	433 (8.2%)	226 (7.05%)	138 (7.89%)	5.2719E-10	1 vs. 2, 1 vs. 3, 1 vs. 4

Inpatient >=5	2074 (15.11%)	837 (23.98%)	680 (12.88%)	364 (11.36%)	193 (11.04%)	4.2633E-63	1 vs. 2, 1 vs. 3, 1 vs. 4, 2 vs. 3, 2 vs. 4
Outpatient 0	2759 (20.10%)	577 (16.53%)	1074 (20.34%)	668 (20.84%)	440 (25.17%)	2.6683E-12	1 vs. 2, 1 vs. 3, 1 vs. 4, 2 vs. 4, 3 vs. 4
Outpatient 1-2	1736 (12.65%)	426 (12.21%)	687 (13.01%)	375 (11.7%)	248 (14.19%)	0.055214	-
Outpatient 3-4	820 (5.97%)	210 (6.02%)	294 (5.57%)	204 (6.37%)	112 (6.41%)	0.38788	-
Outpatient >=5	8409 (61.27%)	2277 (65.24%)	3226 (61.09%)	1958 (61.09%)	948 (54.23%)	6.4449E-13	1 vs. 2, 1 vs. 3, 1 vs. 4, 2 vs. 4, 3 vs. 4
Emergency 0	4908 (35.76%)	1263 (36.19%)	1892 (35.83%)	1158 (36.13%)	595 (34.04%)	0.43416	-
Emergency 1-2	3386 (24.67%)	864 (24.76%)	1233 (23.35%)	799 (24.93%)	490 (28.03%)	0.0012887	1 vs. 4, 2 vs. 4, 3 vs. 4
Emergency 3-4	1465 (10.67%)	368 (10.54%)	559 (10.59%)	350 (10.92%)	188 (10.76%)	0.95595	-
Emergency >=5	3965 (28.89%)	995 (28.51%)	1597 (30.24%)	898 (28.02%)	475 (27.17%)	0.034739	2 vs. 3, 2 vs. 4
Index time period of patient – no. (%)							
03/20-06/20	1212 (8.83%)	268 (7.68%)	498 (9.43%)	297 (9.27%)	149 (8.52%)	0.028427	1 vs. 2, 1 vs. 3
07/20-10/20	3570 (26.01%)	969 (27.77%)	1290 (24.43%)	870 (27.15%)	441 (25.23%)	0.0016853	1 vs. 2, 2 vs. 3
11/20-02/21	3988 (29.06%)	1069 (30.63%)	1490 (28.21%)	904 (28.21%)	525 (30.03%)	0.047141	1 vs. 2, 1 vs. 3
03/21-06/21	1531 (11.16%)	377 (10.8%)	586 (11.1%)	344 (10.73%)	224 (12.81%)	0.11711	-
07/21-11/21	3423 (24.94%)	807 (23.12%)	1417 (26.83%)	790 (24.65%)	409 (23.4%)	0.00032816	1 vs. 2, 2 vs. 3, 2 vs. 4
Acute phase severities of Covid-19 (-1~16 days) – no. (%)							
Hospitalized	5036 (36.69%)	2001 (57.34%)	1642 (31.09%)	891 (27.8%)	502 (28.72%)	6.2272E-188	1 vs. 2, 1 vs. 3, 1 vs. 4, 2 vs. 3
Ventilation	465 (3.39%)	299 (8.57%)	101 (1.91%)	48 (1.5%)	17 (0.97%)	1.2671E-83	1 vs. 2, 1 vs. 3, 1 vs. 4, 2 vs. 4
Critical care	833 (6.07%)	437 (12.52%)	225 (4.26%)	122 (3.81%)	49 (2.8%)	7.9898E-75	1 vs. 2, 1 vs. 3, 1 vs. 4, 2 vs. 4
Baseline Coexisting conditions (ICD10; Category[c]) – no. (%)							
Intestinal infectious diseases (A01-09; CIPD)	145 (1.06%)	40 (1.15%)	43 (0.81%)	24 (0.75%)	38 (2.17%)	0.000005849	1 vs. 4, 2 vs. 4, 3 vs. 4
Other sepsis (A41; CIPD)	1368 (9.97%)	596 (17.08%)	393 (7.44%)	241 (7.52%)	138 (7.89%)	6.47E-57	1 vs. 2, 1 vs. 3, 1 vs. 4
Benign neoplasm of digestive system (D13; Neoplasms)	74 (0.54%)	20 (0.57%)	23 (0.44%)	10 (0.31%)	21 (1.2%)	0.0003453	1 vs. 4, 2 vs. 4, 3 vs. 4
Benign neoplasm of unspecified sites (D36; Neoplasms)	99 (0.72%)	30 (0.86%)	26 (0.49%)	15 (0.47%)	28 (1.6%)	7.1641E-06	1 vs. 2, 1 vs. 4, 2 vs. 4, 3 vs. 4
Anaemia in chronic diseases (D63; DB)	1238 (9.02%)	620 (17.77%)	313 (5.93%)	195 (6.08%)	110 (6.29%)	3.0714E-94	1 vs. 2, 1 vs. 3, 1 vs. 4
Other anaemias (D64; DB)	2947 (21.47%)	1062 (30.43%)	1003 (18.99%)	562 (17.54%)	320 (18.31%)	1.5268E-48	1 vs. 2, 1 vs. 3, 1 vs. 4

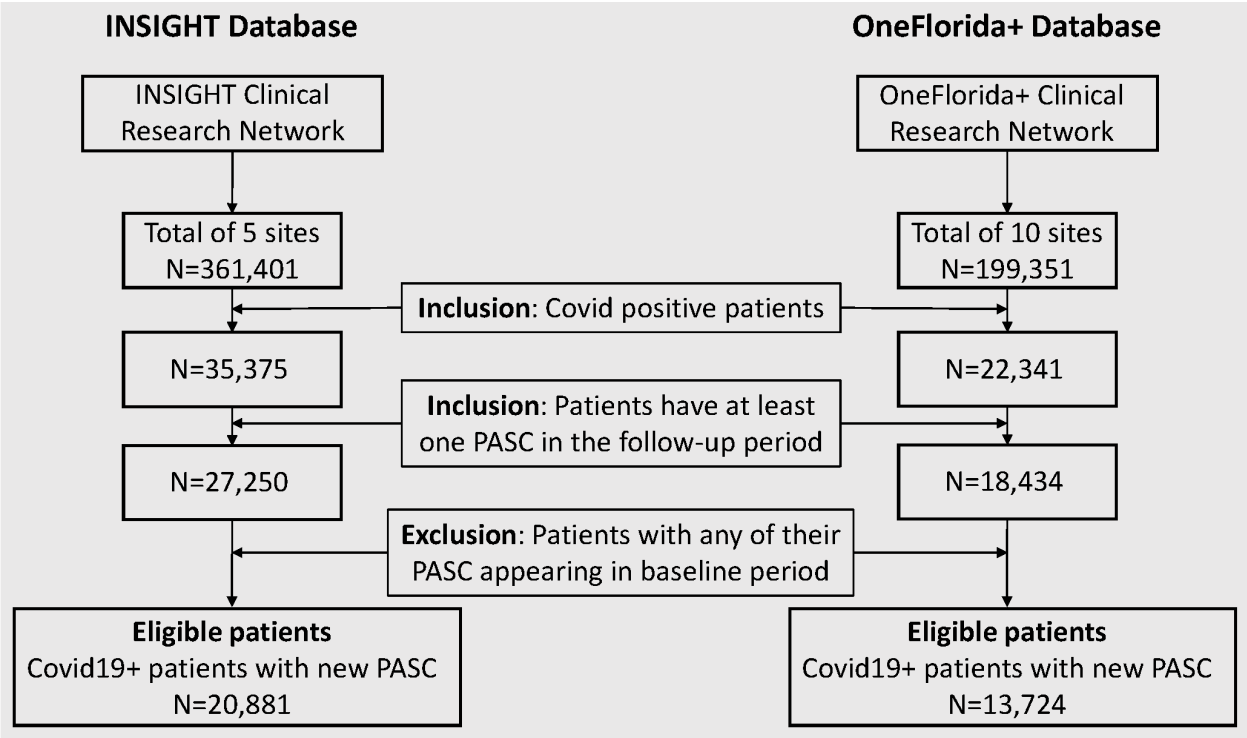
Purpura and other haemorrhagic conditions (D65-D69; DB)	1243 (9.06%)	520 (14.9%)	403 (7.63%)	213 (6.65%)	107 (6.12%)	8.5145E-43	1 vs. 2, 1 vs. 3, 1 vs. 4, 2 vs. 4
Type 2 diabetes mellitus (E11; ENMD)	3883 (28.29%)	1495 (42.84%)	1225 (23.2%)	736 (22.96%)	427 (24.43%)	1.045E-105	1 vs. 2, 1 vs. 3, 1 vs. 4
Disorders of endocrine glands (E20-E35; ENMD)	343 (2.50%)	90 (2.58%)	89 (1.69%)	66 (2.06%)	98 (5.61%)	1.3996E-18	1 vs. 2, 1 vs. 4, 2 vs. 4, 3 vs. 4
Disorders of lipoprotein metabolism (E78; ENMD)	5102 (37.18%)	1710 (49%)	1797 (34.03%)	1044 (32.57%)	551 (31.52%)	2.5933E-61	1 vs. 2, 1 vs. 3, 1 vs. 4
Disorders of mineral metabolism (E83; ENMD)	1442 (10.51%)	619 (17.74%)	432 (8.18%)	255 (7.96%)	136 (7.78%)	3.7207E-56	1 vs. 2, 1 vs. 3, 1 vs. 4
Volume depletion (E86; ENMD)	1713 (12.48%)	625 (17.91%)	593 (11.23%)	323 (10.08%)	172 (9.84%)	6.0039E-28	1 vs. 2, 1 vs. 3, 1 vs. 4
Disorders of fluid, electrolyte and acid-base balance (E87; ENMD)	3186 (23.21%)	1259 (36.07%)	1013 (19.18%)	605 (18.88%)	309 (17.68%)	3.7678E-94	1 vs. 2, 1 vs. 3, 1 vs. 4
Organic mental disorders (F01-F09; MBD)	1418 (10.33%)	654 (18.74%)	418 (7.92%)	229 (7.15%)	117 (6.69%)	1.2034E-77	1 vs. 2, 1 vs. 3, 1 vs. 4
Depressive episode (F32; MBD)	2443 (17.80%)	773 (22.15%)	932 (17.65%)	513 (16.01%)	225 (12.87%)	1.65E-17	1 vs. 2, 1 vs. 3, 1 vs. 4, 2 vs. 4, 3 vs. 4
Other anxiety disorders (F41; MBD)	2592 (18.89%)	496 (14.21%)	1297 (24.56%)	571 (17.82%)	228 (13.04%)	1.4963E-43	1 vs. 2, 1 vs. 3, 2 vs. 3, 2 vs. 4, 3 vs. 4
Migraine (G43; DNS)	804 (5.86%)	107 (3.07%)	499 (9.45%)	131 (4.09%)	67 (3.83%)	5.6955E-44	1 vs. 2, 1 vs. 3, 2 vs. 3, 2 vs. 4
Sleep disorders (G47; DNS)	2710 (19.75%)	917 (26.28%)	948 (17.95%)	573 (17.88%)	272 (15.56%)	3.3046E-28	1 vs. 2, 1 vs. 3, 1 vs. 4, 2 vs. 4, 3 vs. 4
Nerve, nerve root and plexus disorders (G50-G59; DNS)	489 (3.56%)	112 (3.21%)	138 (2.61%)	183 (5.71%)	56 (3.2%)	1.0695E-12	1 vs. 3, 2 vs. 3, 3 vs. 4
Other disorders of brain (G93; DNS)	1289 (9.39%)	566 (16.22%)	402 (7.61%)	216 (6.74%)	105 (6.01%)	3.0763E-56	1 vs. 2, 1 vs. 3, 1 vs. 4, 2 vs. 4
Disorders of eyelid, lacrimal system and orbit (H00-H05; DE)	594 (4.33%)	149 (4.27%)	153 (2.9%)	240 (7.49%)	52 (2.97%)	6.1428E-24	1 vs. 2, 1 vs. 3, 1 vs. 4, 2 vs. 3, 3 vs. 4
Disorders of lens (H25-H28; DE)	962 (7.01%)	327 (9.37%)	311 (5.89%)	223 (6.96%)	101 (5.78%)	1.4547E-09	1 vs. 2, 1 vs. 3, 1 vs. 4
Disorders of choroid and retina (H30-H36; DE)	387 (2.82%)	101 (2.89%)	106 (2.01%)	133 (4.15%)	47 (2.69%)	2.4157E-07	1 vs. 2, 1 vs. 3, 2 vs. 3, 3 vs. 4
Essential hypertension (I10; DCS)	5375 (39.17%)	1692 (48.48%)	1778 (33.67%)	1247 (38.91%)	658 (37.64%)	3.2388E-42	1 vs. 2, 1 vs. 3, 1 vs. 4, 2 vs. 3, 2 vs. 4
Hypertensive heart disease (I11; DCS)	1283 (9.35%)	597 (17.11%)	353 (6.68%)	227 (7.08%)	106 (6.06%)	4.9844E-72	1 vs. 2, 1 vs. 3, 1 vs. 4
Hypertensive renal disease (I12; DCS)	1476 (10.75%)	721 (20.66%)	371 (7.03%)	241 (7.52%)	143 (8.18%)	9.1825E-104	1 vs. 2, 1 vs. 3, 1 vs. 4
Hypertensive heart and renal disease (I13; DCS)	852 (6.21%)	450 (12.89%)	197 (3.73%)	131 (4.09%)	74 (4.23%)	9.8371E-78	1 vs. 2, 1 vs. 3, 1 vs. 4
Ischaemic heart diseases (I20-I25; DCS)	3632 (26.46%)	1562 (44.76%)	1106 (20.94%)	631 (19.69%)	333 (19.05%)	8.901E-175	1 vs. 2, 1 vs. 3, 1 vs. 4

Other pulmonary heart diseases (I27; DCS)	585 (4.26%)	274 (7.85%)	174 (3.29%)	89 (2.78%)	48 (2.75%)	3.4999E-32	1 vs. 2, 1 vs. 3, 1 vs. 4
Atrial fibrillation and flutter (I48; DCS)	1236 (9.07%)	584 (16.73%)	347 (6.57%)	211 (6.58%)	94 (5.38%)	3.7043E-74	1 vs. 2, 1 vs. 3, 1 vs. 4
Heart failure (I50; DCS)	1810 (13.19%)	867 (24.84%)	497 (9.41%)	298 (9.3%)	148 (8.47%)	3.1198E-120	1 vs. 2, 1 vs. 3, 1 vs. 4
Acute upper respiratory infections (J00-J06; DRS)	4166 (30.36%)	941 (26.96%)	2069 (39.18%)	737 (23%)	419 (23.97%)	4.5632E-71	1 vs. 2, 1 vs. 3, 1 vs. 4, 2 vs. 3, 2 vs. 4
Pneumonia, organism unspecified (J18; DRS)	814 (5.93%)	184 (5.27%)	417 (7.9%)	150 (4.68%)	63 (3.6%)	4.4997E-14	1 vs. 2, 1 vs. 4, 2 vs. 3, 2 vs. 4
Diseases of upper respiratory tract (J30-J39; DRS)	2763 (20.13%)	601 (17.22%)	1280 (24.24%)	597 (18.63%)	285 (16.3%)	2.7364E-20	1 vs. 2, 2 vs. 3, 2 vs. 4, 3 vs. 4
Chronic obstructive pulmonary disease (J44; DRS)	858 (6.25%)	221 (6.33%)	414 (7.84%)	153 (4.77%)	70 (4%)	8.9998E-11	1 vs. 2, 1 vs. 3, 1 vs. 4, 2 vs. 3, 2 vs. 4
Asthma (J45; DRS)	2196 (16.00%)	560 (16.05%)	784 (14.85%)	606 (18.91%)	246 (14.07%)	1.2327E-06	1 vs. 3, 2 vs. 3, 3 vs. 4
Pleural effusion (J90; DRS)	773 (5.63%)	353 (10.11%)	215 (4.07%)	139 (4.34%)	66 (3.78%)	2.93E-38	1 vs. 2, 1 vs. 3, 1 vs. 4
Respiratory failure (J96; DRS)	1160 (8.45%)	568 (16.28%)	303 (5.74%)	187 (5.83%)	102 (5.84%)	6.4292E-80	1 vs. 2, 1 vs. 3, 1 vs. 4
Gastro-oesophageal reflux disease (K21; DDS)	3724 (27.13%)	1195 (34.24%)	1336 (25.3%)	805 (25.12%)	388 (22.2%)	3.4009E-27	1 vs. 2, 1 vs. 3, 1 vs. 4, 2 vs. 4, 3 vs. 4
Diseases of stomach and duodenum (K31; DDS)	363 (2.65%)	94 (2.69%)	111 (2.1%)	68 (2.12%)	90 (5.15%)	2.9317E-11	1 vs. 4, 2 vs. 4, 3 vs. 4
Noninfective enteritis and colitis (K50-K52; DDS)	1220 (8.89%)	312 (8.94%)	435 (8.24%)	240 (7.49%)	233 (13.33%)	1.7437E-11	1 vs. 3, 1 vs. 4, 2 vs. 4, 3 vs. 4
Intestinal disorders (K59; DDS)	1993 (14.52%)	672 (19.26%)	717 (13.58%)	420 (13.1%)	184 (10.53%)	2.4213E-20	1 vs. 2, 1 vs. 3, 1 vs. 4, 2 vs. 4, 3 vs. 4
Hematemesis (K92; DDS)	394 (2.87%)	108 (3.09%)	122 (2.31%)	81 (2.53%)	83 (4.75%)	1.3587E-06	1 vs. 2, 1 vs. 4, 2 vs. 4, 3 vs. 4
Rheumatoid arthritis (M05-M06; DMSCT)	178 (1.30%)	36 (1.03%)	51 (0.97%)	73 (2.28%)	18 (1.03%)	6.664E-07	1 vs. 3, 2 vs. 3, 3 vs. 4
Gonarthrosis (M17; DMSCT)	554 (4.04%)	132 (3.78%)	151 (2.86%)	207 (6.46%)	64 (3.66%)	8.3063E-15	1 vs. 2, 1 vs. 3, 2 vs. 3, 3 vs. 4
Other arthrosis (M19; DMSCT)	2100 (15.30%)	776 (22.23%)	727 (13.77%)	417 (13.01%)	180 (10.3%)	4.9208E-40	1 vs. 2, 1 vs. 3, 1 vs. 4, 2 vs. 4, 3 vs. 4
Joint disorders (M25; DMSCT)	3093 (22.54%)	1003 (28.74%)	590 (11.17%)	1029 (32.11%)	471 (26.95%)	1.1035E-141	1 vs. 2, 1 vs. 3, 2 vs. 3, 2 vs. 4, 3 vs. 4
Spondylopathies (M45-M49; DMSCT)	1772 (12.91%)	520 (14.9%)	491 (9.3%)	588 (18.35%)	173 (9.9%)	4.885E-37	1 vs. 2, 1 vs. 3, 1 vs. 4, 2 vs. 3, 3 vs. 4
Soft tissue disorders (M60-M79; DMSCT)	5447 (39.69%)	1393 (39.91%)	1859 (35.2%)	1563 (48.77%)	632 (36.16%)	2.5479E-35	1 vs. 2, 1 vs. 3, 1 vs. 4, 2 vs. 3, 3 vs. 4
Other disorders of bone (M89; DMSCT)	366 (2.67%)	69 (1.98%)	152 (2.88%)	108 (3.37%)	37 (2.12%)	0.0014737	1 vs. 2, 1 vs. 3, 3 vs. 4
Acute renal failure (N17; DGS)	1950 (14.21%)	857 (24.56%)	576 (10.91%)	349 (10.89%)	168 (9.61%)	3.3151E-89	1 vs. 2, 1 vs. 3, 1 vs. 4
Chronic kidney disease (N18; DGS)	2075 (15.12%)	993 (28.45%)	543 (10.28%)	339 (10.58%)	200 (11.44%)	1.7002E-140	1 vs. 2, 1 vs. 3, 1 vs. 4

Cystitis (N30; DGS)	442 (3.22%)	113 (3.24%)	128 (2.42%)	95 (2.96%)	106 (6.06%)	2.8623E-12	1 vs. 2, 1 vs. 4, 2 vs. 4, 3 vs. 4
Diseases of female pelvic organs (N70-N77; DGS)	1088 (7.93%)	199 (5.7%)	574 (10.87%)	201 (6.27%)	114 (6.52%)	3.4114E-22	1 vs. 2, 2 vs. 3, 2 vs. 4
Disorders of female genital tract (N80-N98; DGS)	2352 (17.14%)	433 (12.41%)	1215 (23.01%)	423 (13.2%)	281 (16.08%)	2.5398E-47	1 vs. 2, 1 vs. 4, 2 vs. 3, 2 vs. 4, 3 vs. 4
Maternal care for pregnancy (O26; PCP)	647 (4.71%)	64 (1.83%)	332 (6.29%)	160 (4.99%)	91 (5.21%)	1.8282E-20	1 vs. 2, 1 vs. 3, 1 vs. 4, 2 vs. 3
Other maternal diseases (O99; PCP)	668 (4.87%)	67 (1.92%)	358 (6.78%)	158 (4.93%)	85 (4.86%)	4.417E-23	1 vs. 2, 1 vs. 3, 1 vs. 4, 2 vs. 3, 2 vs. 4
Abnormalities of heart beat (R00; Others)	2851 (20.77%)	931 (26.68%)	1017 (19.26%)	606 (18.91%)	297 (16.99%)	3.1408E-22	1 vs. 2, 1 vs. 3, 1 vs. 4, 2 vs. 4
Abnormalities of breathing (R06; Others)	3191 (23.25%)	708 (20.29%)	1517 (28.73%)	636 (19.84%)	330 (18.88%)	2.5158E-31	1 vs. 2, 2 vs. 3, 2 vs. 4
Pain associated with micturition (R30; Others)	1377 (10.03%)	301 (8.62%)	562 (10.64%)	358 (11.17%)	156 (8.92%)	0.00077407	1 vs. 2, 1 vs. 3, 2 vs. 4, 3 vs. 4
Abnormal findings of lung (R91; Others)	1113 (8.11%)	230 (6.59%)	622 (11.78%)	163 (5.09%)	98 (5.61%)	1.6597E-34	1 vs. 2, 1 vs. 3, 2 vs. 3, 2 vs. 4
Examination for infectious (Z11; Others)	3824 (27.86%)	935 (26.79%)	1497 (28.35%)	932 (29.08%)	460 (26.32%)	0.070226	-
Pregnancy examination and test (Z32; Others)	407 (2.97%)	39 (1.12%)	244 (4.62%)	76 (2.37%)	48 (2.75%)	1.1827E-20	1 vs. 2, 1 vs. 3, 1 vs. 4, 2 vs. 3, 2 vs. 4
Long term drug therapy (Z79; Others)	3373 (24.58%)	1334 (38.22%)	1072 (20.3%)	621 (19.38%)	346 (19.79%)	8.7734E-102	1 vs. 2, 1 vs. 3, 1 vs. 4
Cardiac/vascular implants (Z95; Others)	1401 (14.21%)	643 (18.42%)	415 (7.86%)	227 (7.08%)	116 (6.64%)	5.7165E-75	1 vs. 2, 1 vs. 3, 1 vs. 4
Weeks of gestation (Z3A; Others)	866 (6.31%)	88 (2.52%)	444 (8.41%)	226 (7.05%)	108 (6.18%)	2.3253E-27	1 vs. 2, 1 vs. 3, 1 vs. 4, 2 vs. 3, 2 vs. 4

a. **P-value**: continuous variables are tested by one-way analysis of variance (ANOVA, with Tukey HSD post hoc test) for Normal distribution or by Kruskal-Wallis test (with Dunn post hoc test) for non-Normal distribution; categorical variables are tested by Fisher's exact test with pair-wise Fisher test for post hoc analysis.
b. **IQR**: inter-quartile range.
c. **Abbreviation for ICD10 category**. We grouped ICD10 code according to the first three digits. CIPD: Certain Infectious and Parasitic Diseases; DB: Diseases of the Blood; ENMD: Endocrine, Nutritional and Metabolic Diseases; MBD: Mental and Behavioral Disorders; DNS: Diseases of the Nervous System; DE: Diseases of the ear; DCS: Diseases of the Circulatory System; DRS: Diseases of the Respiratory System; DDS: Diseases of the Digestive System; DMSCT: Diseases of the Musculoskeletal System and Connective Tissue; DGS: Diseases of the Genitourinary System; PCP: Pregnancy, Childbirth and the Puerperium.

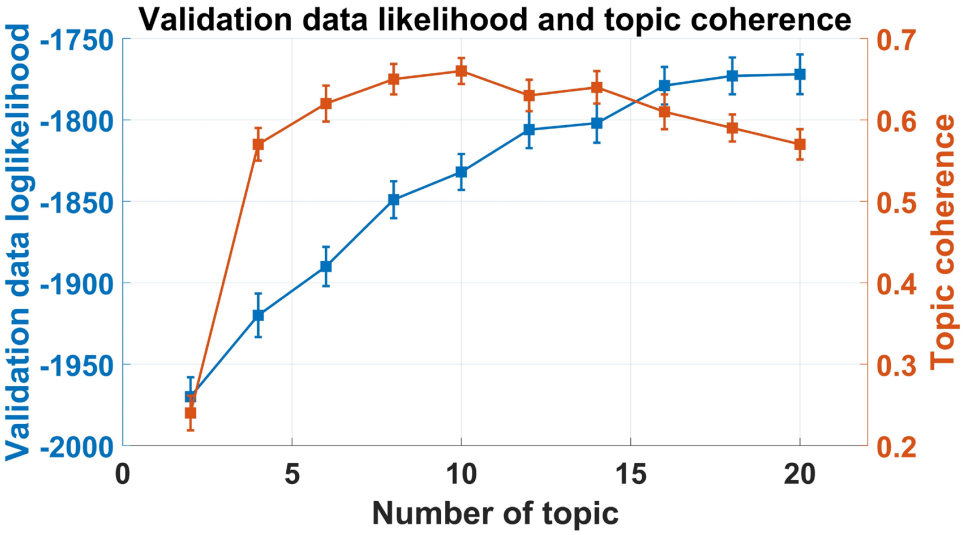
medRxiv preprint doi: https://doi.org/10.1101/2022.05.21.22275412; this version posted May 23, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.



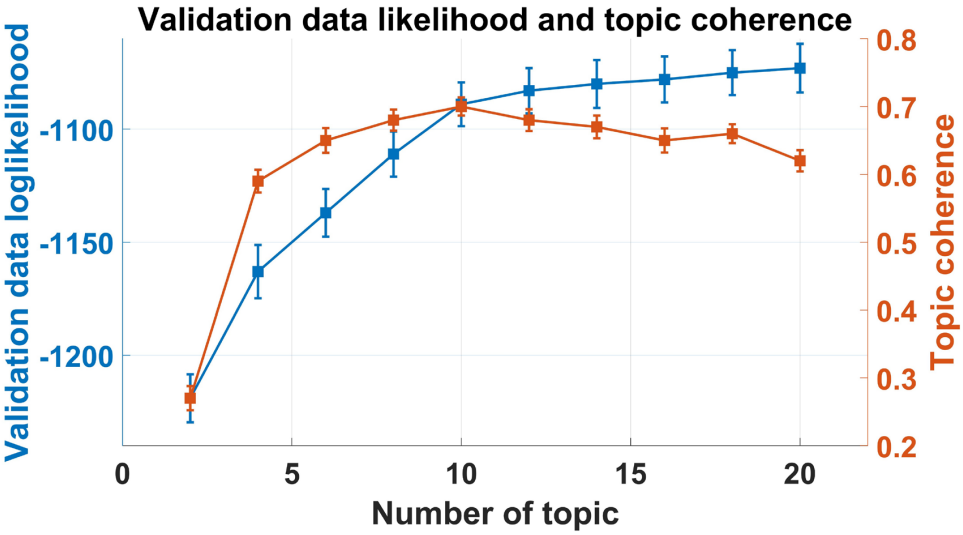
Supplemental Figure 1. Inclusion-exclusion cascade for the INSIGHT and OneFlorida+ cohorts.

medRxiv preprint doi: https://doi.org/10.1101/2022.05.21.22275412; this version posted May 23, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

a. Evaluation on the INSIGHT cohort

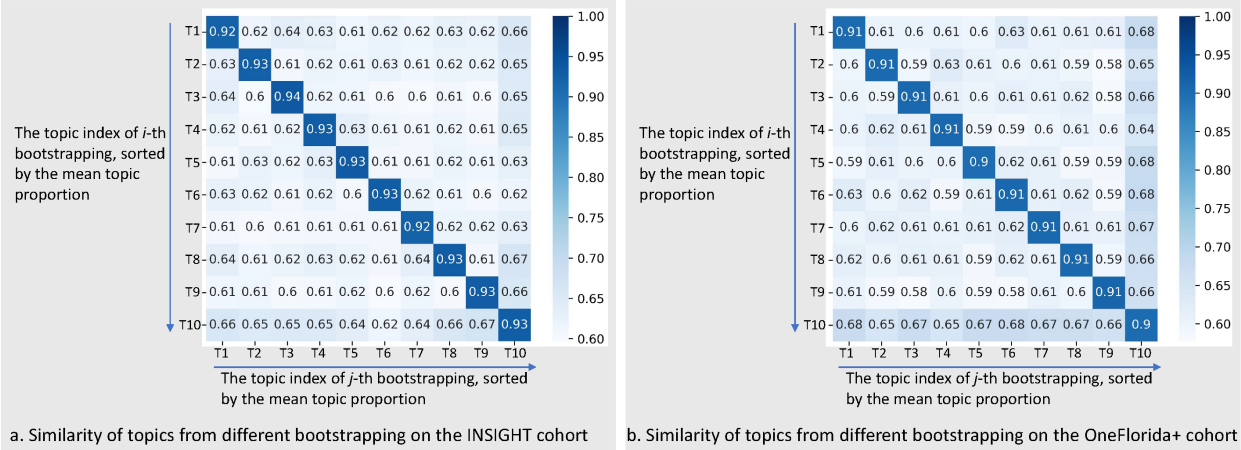


b. Evaluation on the OneFlorida+ cohort

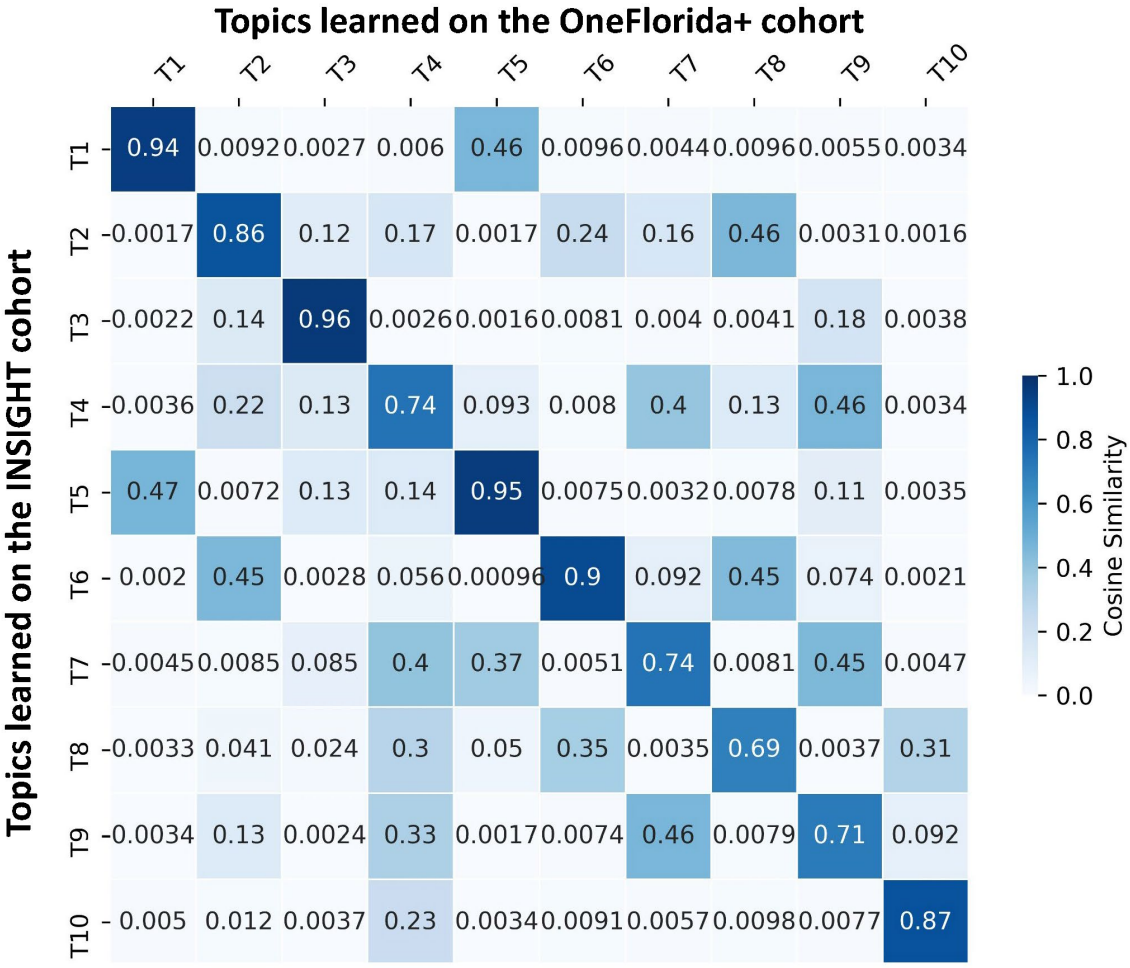


Supplemental Figure 2. The data likelihood and topic coherence conditioned on different number of topics, which were regarded as the criteria to select the optimal number of topics.

medRxiv preprint doi: https://doi.org/10.1101/2022.05.21.22275412; this version posted May 23, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.



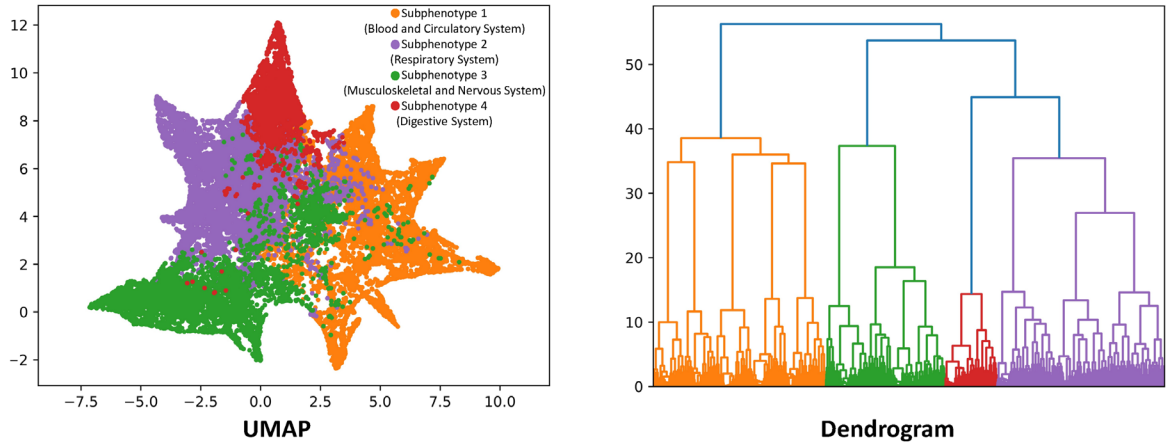
Supplemental Figure 3. The similarity of topics from different bootstrapping, which were used to evaluate the robustness of topics.



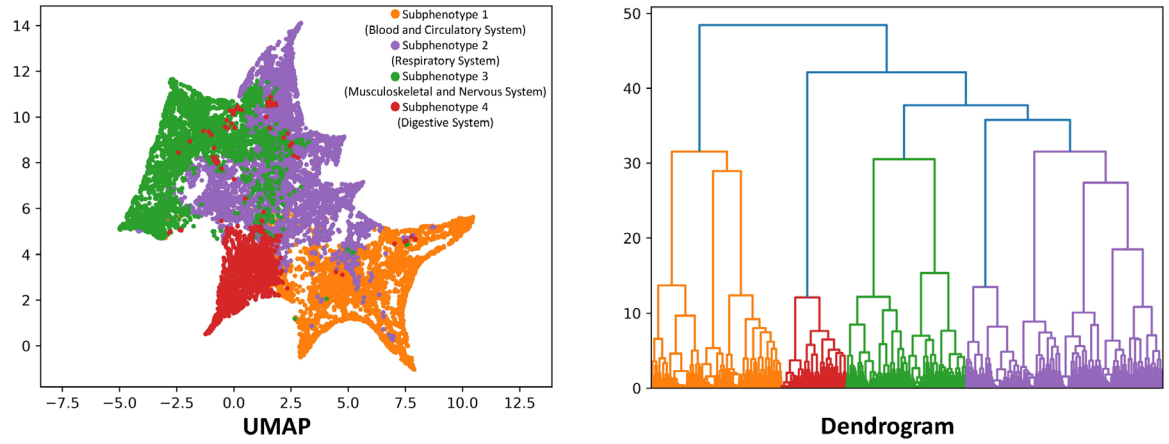
Supplemental Figure 4. The similarity of topics from two cohorts, which were used to evaluate the overlap of topics learned on two cohorts.

medRxiv preprint doi: https://doi.org/10.1101/2022.05.21.22275412; this version posted May 23, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

a. UMAP and Dendrogram of hierarchical clustering for topic proportion, evaluated on the INSIGHT cohort

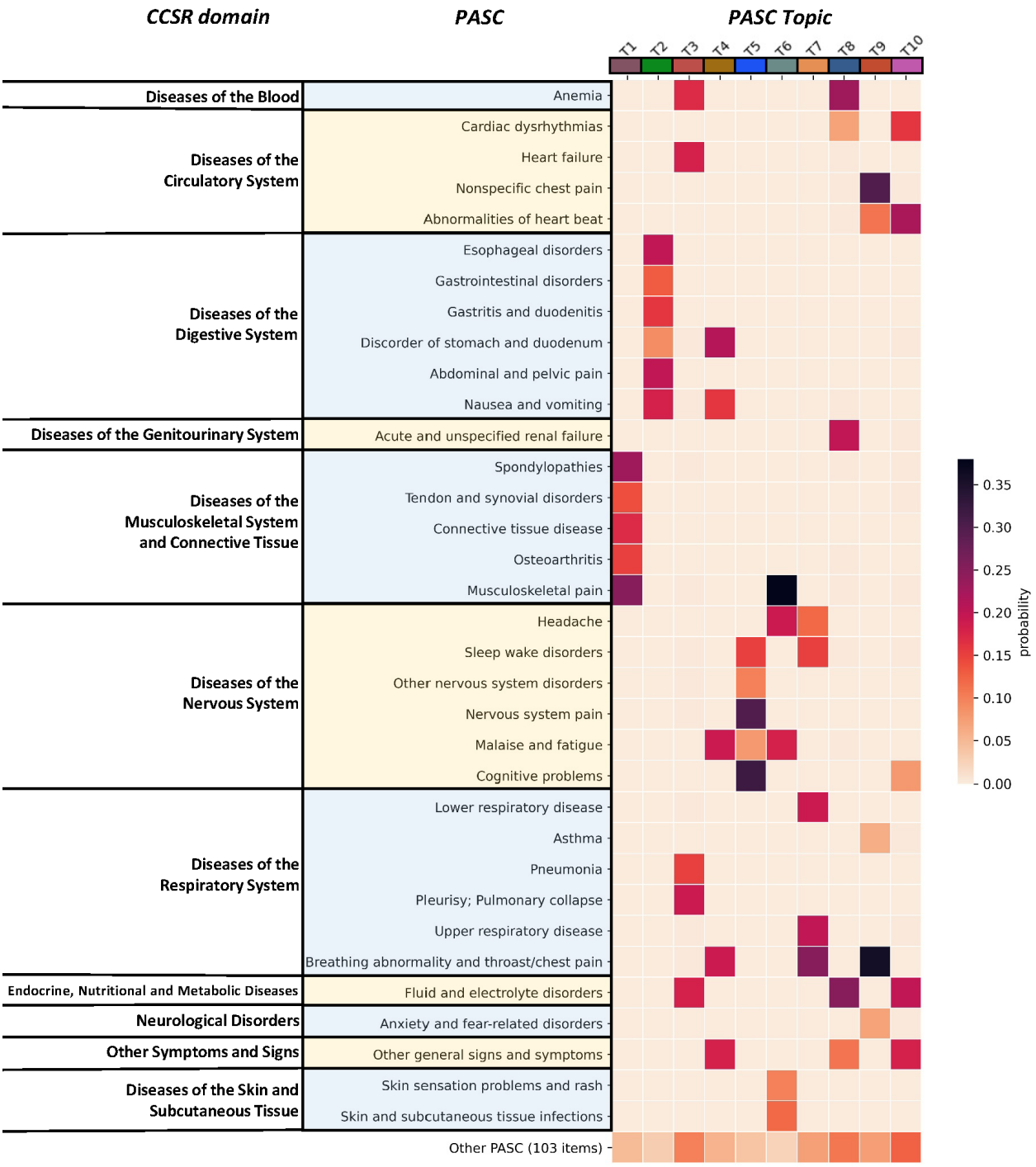


b. UMAP and Dendrogram of hierarchical clustering for topic proportion, evaluated on the OneFlorida+ cohort



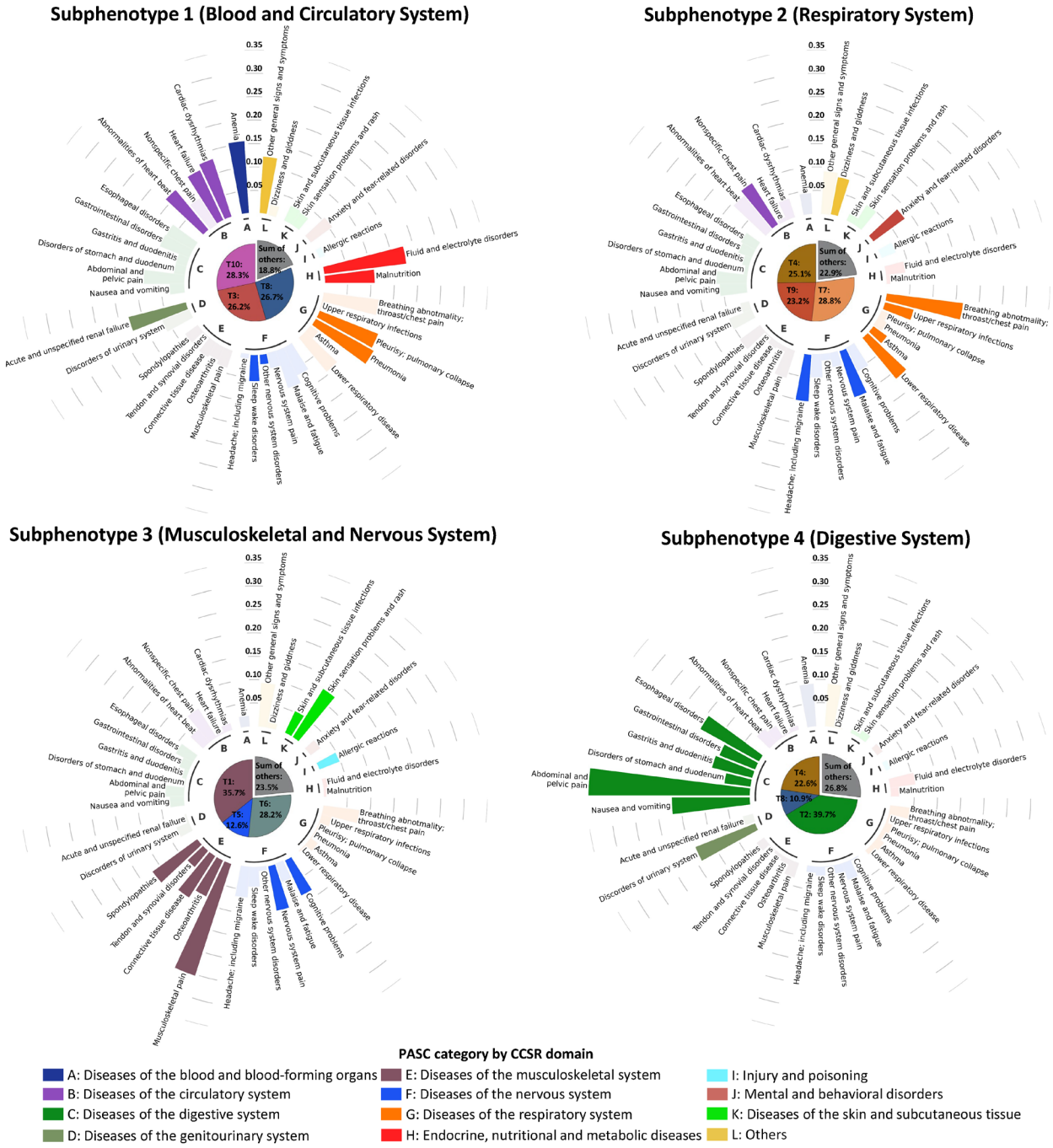
Supplemental Figure 5. UMAP and dendrogram for two cohorts.

medRxiv preprint doi: https://doi.org/10.1101/2022.05.21.22275412; this version posted May 23, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.



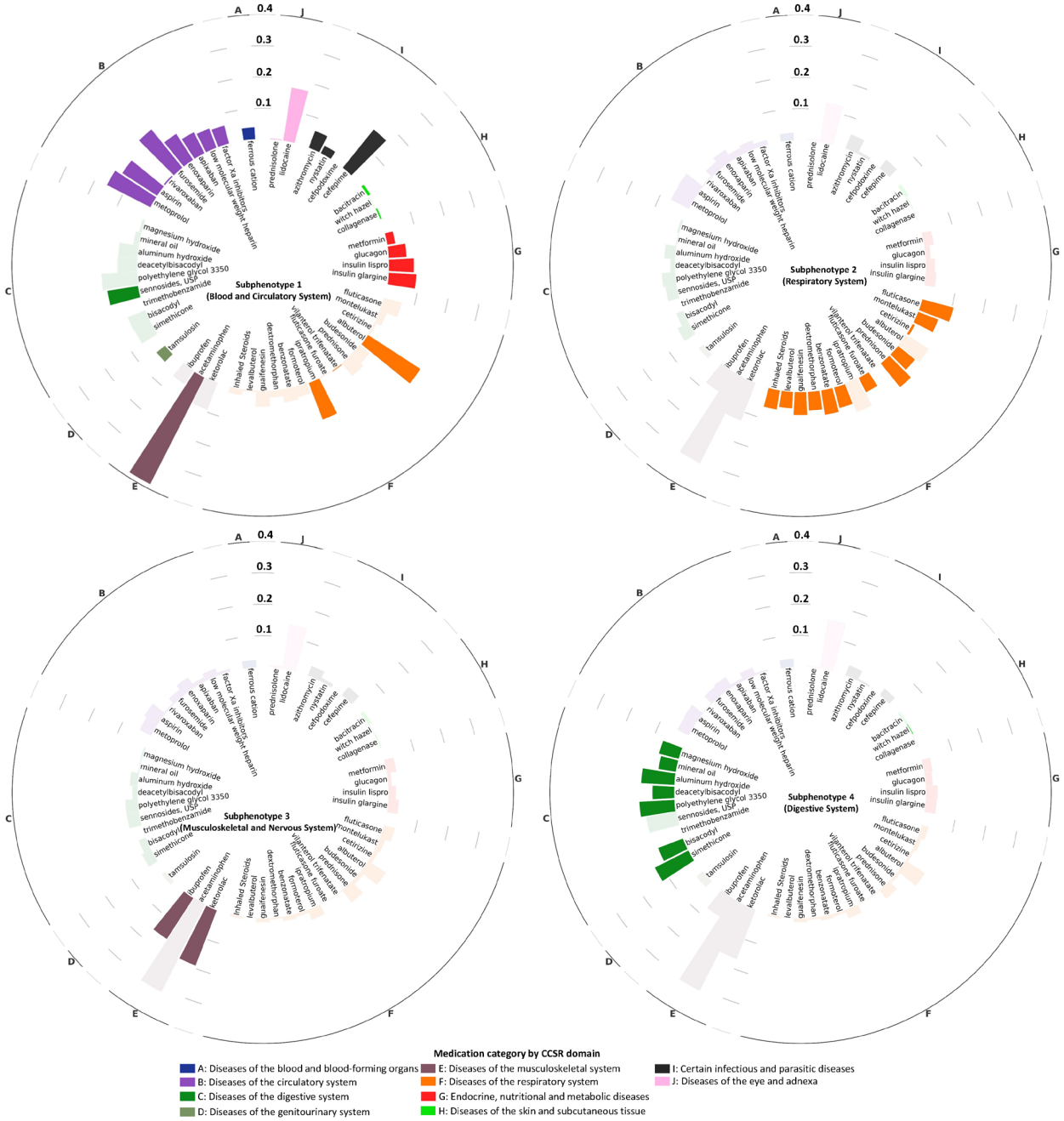
Supplemental Figure 6. The heatmap of PASC topics learned on the OneFlorida+ cohort. Each row denotes a potential PASC category grouped by different CCSR domains, and each column denotes a particular PASC topic. Each PASC topic is characterized by a unique post-acute incidence probability distribution over all 137 individual potential PASC categories.

medRxiv preprint doi: https://doi.org/10.1101/2022.05.21.22275412; this version posted May 23, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.



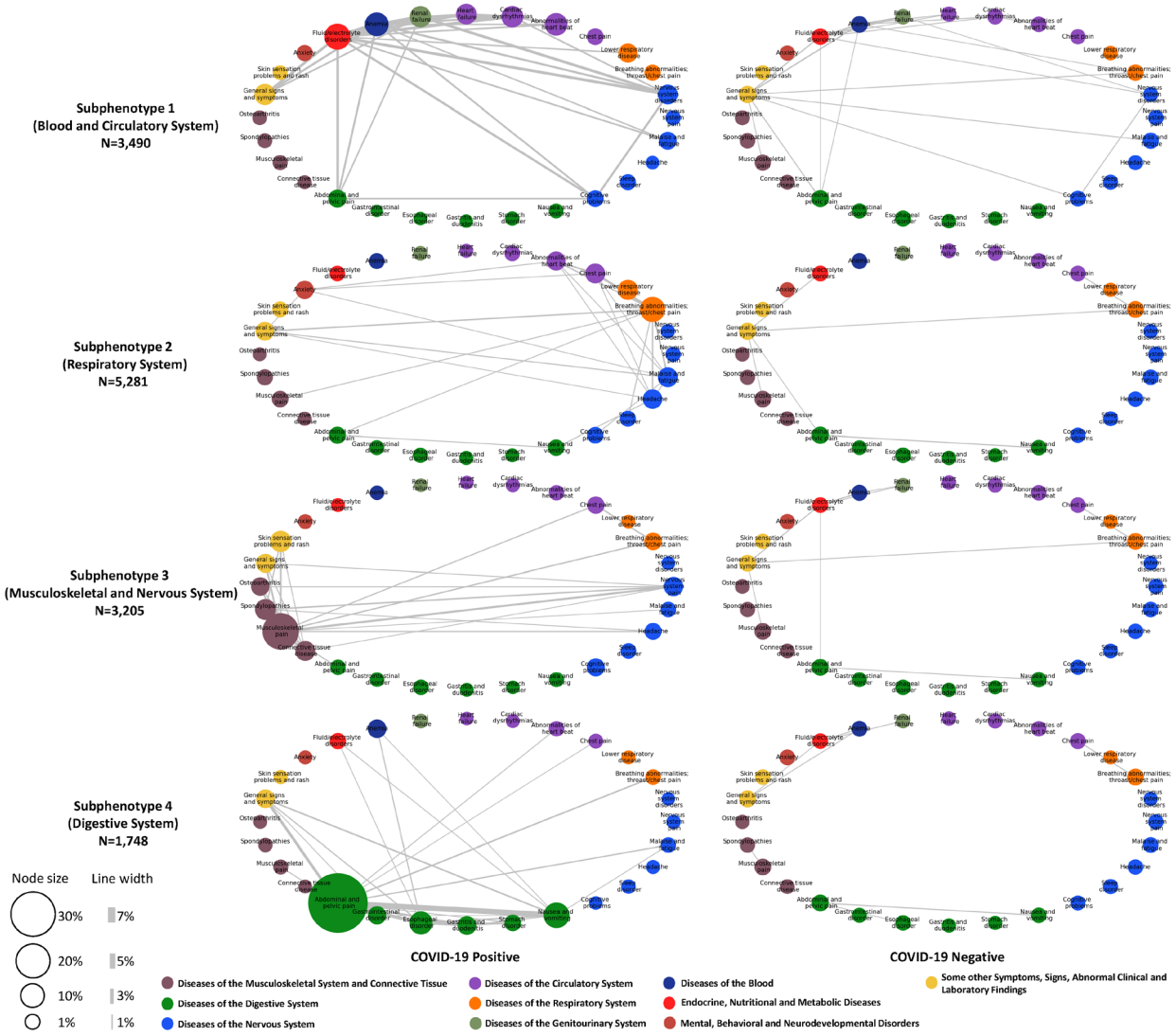
Supplement Figure 7. The prevalence (denoted by the percentage of patients in this subphenotype having each PASC condition) of PASC conditions for each subphenotype on the OneFlorida+ cohort, where PASC conditions were grouped into different categories shown by different colors. For one PASC condition, if it is most prevalent in one subphenotype, we highlighted it in this subphenotype. In the center of each plot, we used pie chart to represent the mean topic proportions on this subphenotype.

medRxiv preprint doi: https://doi.org/10.1101/2022.05.21.22275412; this version posted May 23, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.



Supplemental Figure 8. The prevalence of incident prescriptions of medications in the post-acute infection period for each subphenotype on the OneFlorida+ cohort, where medications are grouped into different categories shown by different colors. For one of the medications, if it is most prevalent in one subphenotype, we highlighted it in this subphenotype.

medRxiv preprint doi: https://doi.org/10.1101/2022.05.21.22275412; this version posted May 23, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.



Supplemental Figure 9. Difference of the incidence patterns of selected PASC conditions (grouped by CCSR domains) in 30-180 days after COVID-19 lab test between positive and matched negative patients on the OneFlorida+ cohort. The bubbles in each network correspond to a PASC condition with their sizes proportional to the incidence rate in the particular subphenotype or matched controls. The edge linking a pair of bubbles indicates co-incidence of the corresponding investigative conditions with its thickness proportional to the co-incidence rate, where we showed the line if the rate is larger than 1%.

medRxiv preprint doi: https://doi.org/10.1101/2022.05.21.22275412; this version posted May 23, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

Supplemental Methods

Evaluate the data likelihood and topic coherence

To select the optimal number of topics, based on different number of topics, we learned topic model and calculated the data likelihood and topic coherence (Supplemental Figure 2) according to the following methods.

Data likelihood. In PFA, to model the binary PASC vector, we used the Bernoulli-Poisson link as:

$$\boldsymbol{x}_n = \mathbf{1}(\boldsymbol{u}_n \geq 1),\ \boldsymbol{u}_{\boldsymbol{n}} \sim Poisson(\boldsymbol{\Phi}\boldsymbol{\theta}_n);$$

where, $\boldsymbol{x}_n$ is the binary PASC vector, $\boldsymbol{\Phi}$ is the topic matrix, $\boldsymbol{\theta}_n$ is the topic proportion vector, $\boldsymbol{u}_{\boldsymbol{n}}$ is the latent variables which links the binary observation and topic representation. According to the property of Bernoulli-Poisson link, $\boldsymbol{u}_{\boldsymbol{n}}$ can be marginalized out and then one can obtain a Bernoulli likelihood as

$$\boldsymbol{x}_n \sim Ber\left(\mathbf{1} - \exp^{-\boldsymbol{\Phi}\boldsymbol{\theta}_n}\right).$$

Given $\{\boldsymbol{x}_n\}_{n=1}^{N}$, after learning $\boldsymbol{\Phi}$ and $\{\boldsymbol{\theta}_n\}_{n=1}^{N}$, we can calculate the Bernoulli data likelihood.

Topic coherence. Topic coherence is an important matric to evaluate the quality of topics based on the input data. It is measured based on a sliding window (in our case, the length of the window is the total number of PASCs), and then one can calculate the normalized pointwise mutual information (NPMI) between input data and the learned topics. We used the python package GENSIM (https://radimrehurek.com/gensim/) to calculate the topic coherence.