

Data-driven identification of post-acute SARS-CoV-2 infection subphenotypes

Received: 8 June 2022

Accepted: 2 November 2022

Published online: 1 December 2022

 Check for updates

Hao Zhang¹, Chengxi Zang¹, Zhenxing Xu¹, Yongkang Zhang¹, Jie Xu², Jiang Bian², Dmitry Morozuk¹, Dhruv Khullar¹, Yiye Zhang¹, Anna S. Nordvig³, Edward J. Schenck⁴, Elizabeth A. Shenkman², Russell L. Rothman⁵, Jason P. Block⁶, Kristin Lyman⁷, Mark G. Weiner¹, Thomas W. Carton⁷, Fei Wang¹ ✉ & Rainu Kaushal¹

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 period after acute SARS-CoV-2 infection. Most studies have examined these conditions individually without providing evidence on co-occurring conditions. In this study, we leveraged the electronic health record data of two large cohorts, INSIGHT and OneFlorida+, from the national Patient-Centered Clinical Research Network. We created a development cohort from INSIGHT and a validation cohort from OneFlorida+ including 20,881 and 13,724 patients, respectively, who were SARS-CoV-2 infected, and we investigated their newly incident diagnoses 30–180 days after a documented SARS-CoV-2 infection. Through machine learning analysis of over 137 symptoms and conditions, we identified four reproducible PASC subphenotypes, dominated by cardiac and renal (including 33.75% and 25.43% of the patients in the development and validation cohorts); respiratory, sleep and anxiety (32.75% and 38.48%); musculoskeletal and nervous system (23.37% and 23.35%); and digestive and respiratory system (10.14% and 12.74%) sequelae. These subphenotypes were associated with distinct patient demographics, underlying conditions before SARS-CoV-2 infection and acute infection phase severity. Our study provides insights into the heterogeneity of PASC and may inform stratified decision-making in the management of PASC conditions.

The ongoing global pandemic of Coronavirus Disease 2019 (COVID-19) caused by severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) infection has impacted hundreds of millions of people's lives. Existing studies have provided evidence that many symptoms and signs could be persistent, exacerbated or newly present after the acute phase of SARS-CoV-2 infection, referred to as post-acute sequelae of

SARS-CoV-2 infection (PASC)^{1,2}, which involve multiple organ systems, including cardiovascular³, mental⁴, metabolic⁵, renal⁶ and others. There have been various ongoing efforts into investigating the underlying biological mechanisms of PASC^{7–9}, which have typically been conducted in small patient cohorts. Large-scale clinical observational cohort studies can provide useful insights into PASC that may help develop

¹Department of Population Health Sciences, Weill Cornell Medicine, New York, NY, USA. ²Department of Health Outcomes Biomedical Informatics, University of Florida, Gainesville, FL, USA. ³Department of Neurology, Weill Cornell Medicine, New York, NY, USA. ⁴Department of Medicine, Division of Pulmonary and Critical Care Medicine, Weill Cornell Medicine, New York, NY, USA. ⁵Center for Health Services Research, Vanderbilt University Medical Center, Nashville, TN, USA. ⁶Department of Population Medicine, Harvard Pilgrim Health Care Institute, Harvard Medical School, Boston, MA, USA. ⁷Louisiana Public Health Institute, New Orleans, LA, USA. ✉e-mail: few2001@med.cornell.edu

effective mechanistic hypotheses and inform targeted treatments. Most existing observational studies investigated the PASC conditions individually (for example, by examining the incidence¹⁰, excess burden¹¹ or prevalence¹² of each symptom or condition in the post-acute period for patients infected with SARS-CoV-2 relative to non-infected individuals). Their co-appearance patterns (or subphenotypes)—that is, to what extent PASC symptoms and conditions co-appear or develop disproportionately in certain patient populations—can potentially help reveal the pathophysiology behind PASC by disentangling their phenotypical heterogeneities. Existing studies on this topic have been limited, with one study that identified PASC subphenotypes from the reported symptoms of a cohort including 233 patients with COVID-19 (ref. 13). More comprehensive analysis involving broader sets of PASC conditions with larger patient cohorts is needed.

We developed a machine learning approach to derive PASC subphenotypes based on newly incident conditions in the post-acute SARS-CoV-2 infection period (defined as 30–180 days after the confirmed infection) of patients with COVID-19. We leveraged the electronic health record (EHR) repositories of two large clinical research networks (CRNs) from the national Patient-Centered Clinical Research Network (PCORnet): the INSIGHT network¹⁴, which includes 12 million patients in the New York City (NYC) area, and the OneFlorida+ network¹⁵, which includes 19 million patients from Florida, Georgia and Alabama. We examined the incidence of 137 diagnosis categories derived from the Clinical Classifications Software Refined (CCSR) categories¹⁶ that were potentially related to PASC, and we leveraged a topic modeling (TM) approach¹⁷ to learn the co-incidence patterns of these diagnosis conditions, based on which the PASC subphenotypes were derived by clustering.

Results

Overall pipeline

Figure 1 shows our overall analytics pipeline. Using the two CRNs, we extracted the EHR of 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) was compiled, and only patients who have had documented new incidences of these conditions in the post-acute infection period were retained. Each patient was initially characterized as a 137-dimensional binary vector encoding whether a particular condition appeared in his/her post-acute infection period or not (step 1). Next, a set of ‘PASC topics’ was learned from these vectors; PASC topics in this context refer to a group of conditions that present together based on their incident probabilities (step 2). Then, the patients are further characterized by these topics according to how much each topic was represented in their post-acute period records (step 3). Finally, a clustering procedure was performed from topic-induced characterizations to identify the subphenotypes (step 4).

Study cohorts

Our study included 20,881 and 13,724 patients from the INSIGHT and OneFlorida+ CRNs who were virally tested positive for SARS-CoV-2, respectively (Methods). The diagnosis information in International Classification of Diseases (ICD)-10 codes of these patients in the post-SARS-CoV-2 infection period was analyzed. The INSIGHT cohort had a median age of 58.0 years (interquartile range (IQR) (42.0–70.0)) and a median Area Deprivation Index (ADI)¹⁸ of 15.0 (IQR (6.0–25.0)). It consisted of 58.37% female ($n = 12,188$), 33.59% White ($n = 7,013$) and 22.85% Black ($n = 4,771$). The OneFlorida+ cohort included patients with a median age of 51.0 years (IQR (35.0–65.0)) and is more socioeconomically deprived, with a median ADI of 59.0 (IQR (42.0–76.0)). It had a greater proportion of White patients (52.28%, $n = 7,175$).

Within the INSIGHT cohort, 33.04% of the patients had a confirmed SARS-CoV-2 infection from March to June 2020 (compared to 8.83% from the OneFlorida+ cohort). This reflected the centrality of NYC during

the first wave of COVID-19 in the United States. More patients from OneFlorida+ were tested positive from July to October 2020 (26% versus 6% for INSIGHT). The summary statistics of the two cohorts are available in Table 1 (with more details provided in Supplementary Table 1).

Potential PASC co-incidence patterns

We compiled a list of 137 potentially PASC-related diagnosis groups defined by ICD-10 diagnosis codes and CCSR categories¹⁶ (Supplementary Table 2). We first investigated the co-incidence patterns across different diagnoses 30–180 days after confirmed SARS-CoV-2 infection through probabilistic TM¹⁹, which was originally proposed for learning word co-occurrence patterns in documents with different semantic topics. With this approach (Methods), we identified 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 heat map matrix of topics learned from the INSIGHT cohort. Each column is a learned topic, and each row is a potential PASC condition category (we demonstrated 31 in the heat map 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 corresponds to the incident probability of the specific PASC condition in the corresponding topic. Topics T1, T2 and T5 were concentrated on the conditions of the musculoskeletal, digestive and nervous systems, respectively. Topics T4, T7 and T9 included respiratory conditions mixed with sleep disorder and anxiety along with symptoms such as headache and chest pain. Topic T3 included fluid and electrolyte disorders combined with anemia and cardiac complications. Topic T6 was dominated by musculoskeletal and skin conditions as well as headache and fatigue. Topic T8 mainly comprised anemia and digestive system conditions. Topic T10 contained a mix of circulatory conditions, renal failure, fluid and electrolyte disorders and others.

Potential PASC subphenotypes

After identifying potential PASC topics, we could describe patients with PASC using these topics and derive potential PASC subphenotypes as patient clusters (Methods). Specifically, four subphenotypes were identified from the INSIGHT cohort. Table 2 summarizes their characteristics with respect to patient demographics, medical utilizations and pre-existing conditions in the baseline period and disease severity in the acute phase according to medical utilization (more details provided in Supplementary Table 3). Across different subphenotypes, we demonstrated the incidence rates of specific potential PASC conditions (Fig. 3, with numerical results in Supplementary Table 4) and medication prescriptions (Extended Data Fig. 1). We characterized these subphenotypes in detail as follows.

Subphenotype 1 (cardiac and renal) consisted of 7,047 (33.75%) patients. It was dominated by cardiac-related, renal-related and circulation-related topics (T3, T8 and T10), including cardiac and circulatory conditions, renal failure, anemia and fluid and electrolyte disorders. 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%). The patients also had a higher acute severity of COVID-19 (with the highest rate of hospitalization (61.15%), use of mechanical ventilation (4.81%) and critical care services (9.95%)). This subphenotype had the highest proportion of patients (37.38%) infected with SARS-CoV-2 during the first wave of the pandemic (March–June 2020). Patients in this subphenotype had a higher burden of pre-existing conditions than other subphenotypes, especially for blood, circulation and endocrine comorbidities, as well as a high level of incident prescriptions of medications for treating circulatory and endocrine conditions and anemia.

Subphenotype 2 (respiratory, sleep and anxiety) 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 patients with a median

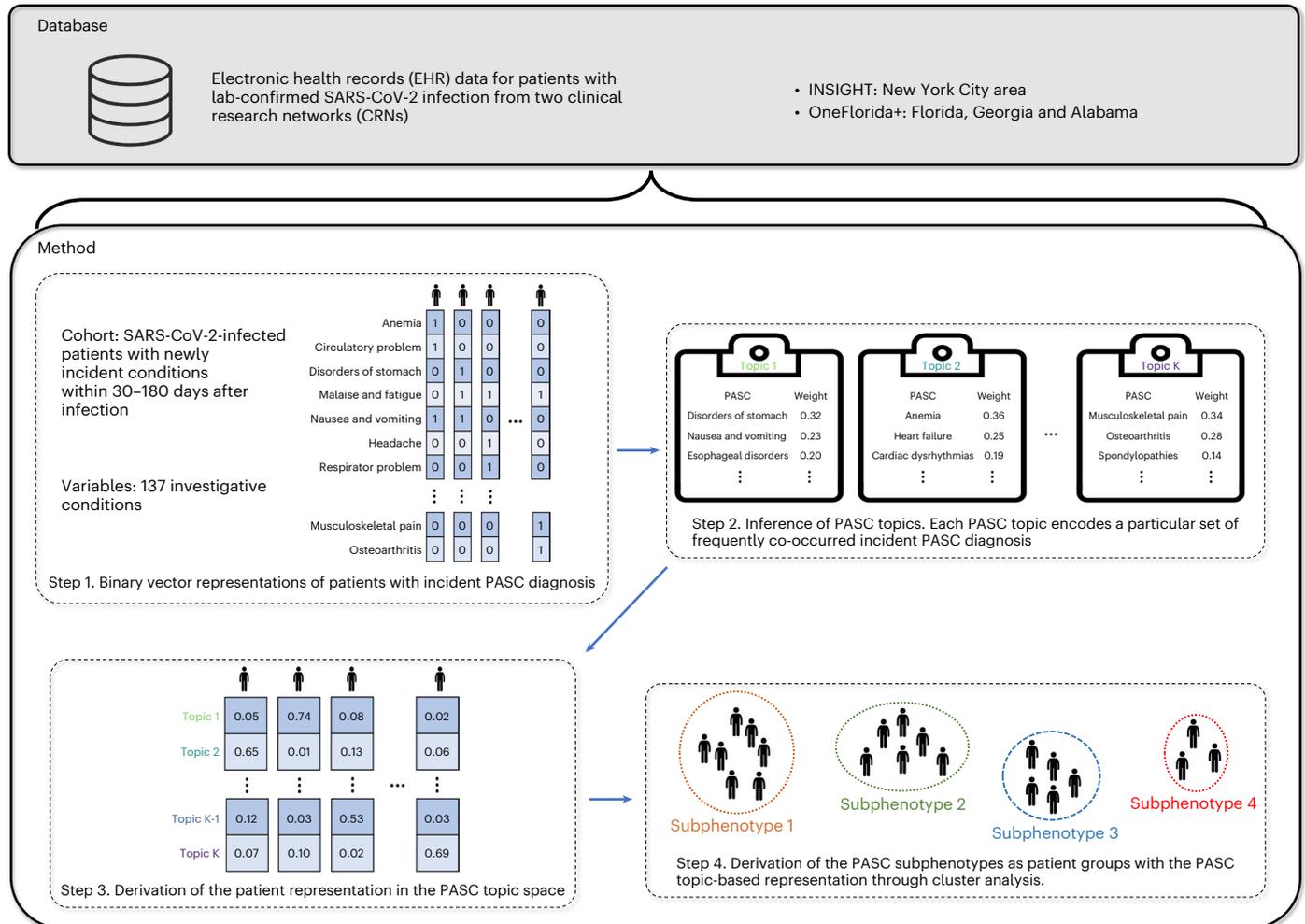


Fig. 1 | Data curation and the subphenotyping pipeline. To study the subphenotypes for patients with PASC, we constructed cohorts of SARS-CoV-2-positive and SARS-CoV-2-negative individuals from the INSIGHT and OneFlorida+ CRNs. After obtaining high-dimensional binary representations of patients with

PASC diagnoses (step 1), our algorithm learned PASC topics (step 2) and inferred the patient representations in the low-dimensional PASC topic space (step 3) through a TM approach. Finally, PASC subphenotypes were derived as patient clusters based on representations by PASC topics (step 4).

age of 51.0 years (IQR (35.0–64.0)), female representation of 62.8% and hospitalization rate of 31.28% during the acute SARS-CoV-2 infection phase. It 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 sequelae and chronic obstructive pulmonary disease. These patients were associated with higher incident prescription rates for anti-asthma, anti-allergy and anti-inflammatory medications, including inhaled steroids, levalbuterol and montelukast.

Subphenotype 3 (musculoskeletal and nervous) consisted of 4,879 (23.37%) patients who mainly had musculoskeletal and nervous system sequelae (topics T1, T5 and T6), such as musculoskeletal pain, headaches and sleep-wake disorders. This subphenotype included patients with a median age of 57.0 years (IQR (42.0–69.0)), and 60.71% were female. It had the highest proportion of patients with more than five outpatient visits before SARS-CoV-2 infection (78.4%) and higher baseline comorbidity burdens of autoimmune and allergy conditions, such as rheumatoid arthritis and asthma, as well as other musculoskeletal and nervous system conditions, including soft tissue, bone and sleep disorders. This subphenotype was associated with more incident prescriptions for pain medications (for example, ibuprofen and ketorolac).

Subphenotype 4 (digestive and respiratory) included 2,117 (10.14%) patients mainly with digestive system and respiratory conditions (topics T2, T4 and T8). Patients in this subphenotype had a median age of 54.0 years (IQR (39.0–67.0)), 61.64% female, the highest rates of zero baseline emergency visits (57.06%) as well as the lowest rates of mechanical ventilation (0.8%) and critical care admission (2.79%) in the acute phase of SARS-CoV-2 infection. Compared to other subphenotypes, Subphenotype 4 had a lower overall burden of underlying conditions but a slightly higher prevalence of digestive system conditions, such as hematemesis, stomach and duodenum disorders and digestive system neoplasm, and more incident prescription digestive system medications.

For each subphenotype, we also provide the sex-disaggregated analysis results with respect to patient demographics, baseline and acute infection phase medical utilizations, prior condition prevalence and newly PASC incidence rates in Supplementary Table 5, which can help trigger effective mechanistic hypotheses of the different PASC subphenotypes.

Comparison with patients without SARS-CoV-2 infection

To investigate how potential PASC co-incidence patterns in patients with confirmed SARS-CoV-2 infection differed from those without infection, we constructed a cohort of patients who visited the hospitals

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

Characteristics	INSIGHT cohort	OneFlorida+ cohort
	Development	Validation
No. of patients	20,881	13,724
Age in years, median (IQR) ^a	58.0 (42.0–70.0)	51.0 (35.0–65.0)
Age group, no. (%)		
20 to <40 years	4,603 (22.04%)	4,284 (31.22%)
40 to <55 years	4,522 (21.66%)	3,318 (24.18%)
55 to <65 years	4,308 (20.63%)	2,528 (18.42%)
65 to <75 years	3,810 (18.25%)	1,923 (14.01%)
75 to <85 years	2,553 (12.22%)	1,178 (8.58%)
85+ years	1,085 (5.20%)	493 (3.59%)
Sex, no. (%)		
Female	12,188 (58.37%)	8,468 (61.70%)
Male	8,692 (41.63%)	5,255 (38.29%)
Other/Missing	1 (0.00%)	1 (0.00%)
Race, no. (%)		
Asian	945 (4.52%)	144 (1.05%)
Black or African American	4,771 (22.85%)	4,076 (29.70%)
White	7,013 (33.59%)	7,175 (52.28%)
Other	6,238 (29.87%)	2,123 (15.47%)
Missing	1,914 (9.17%)	206 (1.50%)
Ethnic group, no. (%)		
Hispanic: Yes	6,838 (32.74%)	2,881 (20.99%)
Hispanic: No	12,248 (58.66%)	9,329 (67.98%)
Hispanic: Other/Missing	1,795 (8.60%)	1,514 (11.03%)
ADI, median (IQR)	15.0 (6.0–25.0)	59.0 (42.0–76.0)
Healthcare utilization ^b in the past 3 years, no. (%)		
Inpatient=0	14,631 (70.07%)	7,437 (54.19%)
0 < Inpatient < 5	5,433 (26.03%)	4,213 (30.70%)
Inpatient ≥ 5	817 (3.91%)	2,074 (15.11%)
Outpatient=0	1,241 (5.94%)	2,759 (20.10%)
0 < Outpatient < 5	3,607 (17.28%)	2,556 (18.63%)
Outpatient ≥ 5	16,033 (76.78%)	8,409 (61.27%)
Emergency=0	11,321 (54.22%)	4,908 (35.76%)
0 < Emergency < 5	7,705 (36.90%)	4,851 (35.35%)
Emergency ≥ 5	1,855 (8.89%)	3,965 (28.89%)
Index time period of patient, no. (%)		
3/20–6/20	6,899 (33.04%)	1,212 (8.83%)
7/20–10/20	1,180 (5.65%)	3,570 (26.01%)
11/20–2/21	8,714 (41.73%)	3,988 (29.06%)
3/21–6/21	3,390 (16.24%)	1,531 (11.16%)
7/21–11/21	698 (3.34%)	3,423 (24.94%)
Acute phase severities of COVID-19 (–1 to –16 days) ^c , no. (%)		
Hospitalized	9,076 (43.47%)	5,036 (36.69%)
Ventilation	495 (2.37%)	465 (3.39%)
Critical care	1,144 (5.48%)	833 (6.07%)

^aIQR, interquartile range. ^bFor the healthcare utilization in the past 3 years, including inpatient, outpatient and emergency visits, we binned the number of visits into five levels. ^cTime range for acute phase severities is from 1 day before index date to 16 days after index date.

but were tested negative for SARS-CoV-2 with a rigorous matching process considering patient demographics and baseline characteristics. Extended Data Fig. 2a shows quantitatively how well the data can be modeled with respect to different numbers of topics for patients who tested positive versus negative for SARS-CoV-2 infection within the INSIGHT cohort. We observed that a larger number of topics was needed for characterizing patients who tested negative for SARS-CoV-2. In addition, we calculated the similarities between the topics learned from the patients who tested positive versus negative for SARS-CoV-2 infection and showed them as heat maps in Extended Data Fig. 3c, from which we could observe that the topics learned from COVID-19-positive and COVID-19-negative patients were not similar to each other. To further examine the concentration patterns of the learned topics on individual conditions, we calculated the entropy of each topic vector, such that higher entropy value indicated less concentration patterns—that is, the probability values comprising each topic vector were more evenly distributed across individual conditions. Supplementary Table 6 demonstrates the detailed entropy values of each topic learned from patients who tested positive for SARS-CoV-2 infection or not, from which we could observe that the topics learned from patients who tested negative displayed higher entropy values compared to patients who tested positive. These investigations suggest that the condition incidence patterns in the follow-up period for matched patients who tested negative were much less clear—that is, these conditions appeared more randomly without clear co-incidence patterns. This further strengthened the validity of our identified potential PASC subphenotypes as they were unique to patients who tested positive for SARS-CoV-2 infection.

We also plotted heat maps of the post-acute co-incidence rate matrix of the 137 potential PASC conditions for patients in each subphenotype and their matched controls in Supplementary Fig. 1. We observed that patients in these PASC subphenotypes are associated with higher co-incidence rates of PASC conditions compared to their matched controls. We further visualized in Fig. 4 the network patterns of 28 selected potential PASC conditions whose incidence rates were larger than 1% in any of the PASC subphenotypes, where the nodes in each network are particular potential PASC conditions with their sizes proportional to the incidence rate in the records of patients from the corresponding group. Each line linking a pair of nodes indicates a co-incidence of that pair of potential PASC conditions, with its thickness proportional to the co-incidence rate in the corresponding group. Figure 4 shows that the conditions used to characterize each PASC subphenotype were clearly associated with larger-sized nodes, representing higher incidence rates. There were no clear differences in node sizes for the groups with matched controls. We also observed denser connections in PASC subphenotypes, which suggested that the potential PASC conditions did not appear independently, but rather collectively, and those larger nodes included more interconnected network hubs.

Sensitivity analysis

Our analysis so far was based on a comprehensive list of 137 potential PASC conditions compiled from existing literature and clinician input. With respect to a specific patient cohort, it was challenging to guarantee that all of these conditions would be associated with excessive incidence risk in the follow-up period for patients who tested positive versus negative for SARS-CoV-2 infection due to population heterogeneity and incomplete information capture in the EHR. One key characteristic for TM was that it could effectively suppress the impact of conditions with low incidence rates in the EHR and drew more focus on the prevalent conditions (Fig. 2). However, it was still unclear if the subphenotypes would change if only the conditions that were associated with statistically significant excessive risks in the follow-up period for patients who tested positive for SARS-CoV-2 infection compared to patients who tested negative were considered.

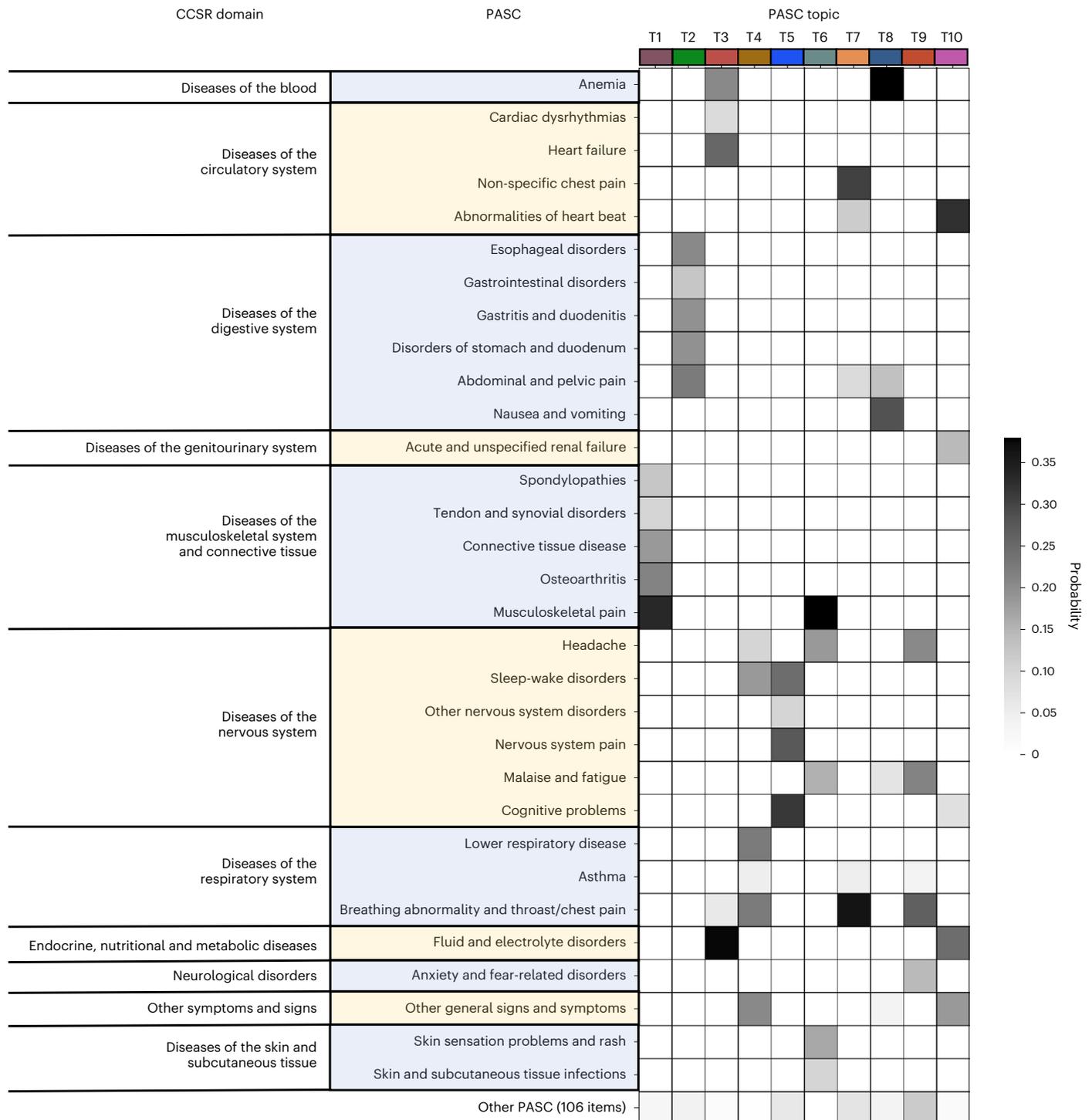


Fig. 2 | Heat map of PASC topics learned from the INSIGHT cohort. Each row denotes a potential PASC diagnosis category defined by ICD-10 codes classified through CCSR, and each column denotes a particular PASC topic. Each PASC topic is characterized by a unique post-acute incidence probability distribution over all 137 individual potential PASC diagnosis categories.

We examined the robustness of the identified PASC subphenotypes in a more restricted set of PASC conditions. Specifically, with a high-dimensional propensity score (PS) adjustment pipeline^{20,21} and existing research into PASC^{10,21,22}, we identified 44 PASC conditions (Supplementary Table 7) in the INSIGHT cohort that demonstrated statistically significant higher risk in the follow-up period for patients who tested positive versus negative for SARS-CoV-2 infection. Then, we implemented the same subphenotyping process based on these

44 PASC conditions. Finally, we quantified the overlap between these newly derived subphenotypes with the subphenotypes shown in Fig. 3.

Supplementary Fig. 2a demonstrates the learned PASC topics from this restricted condition set, where the optimal number of topics is determined according to Supplementary Fig. 2b. We further quantitatively compared these topics and the original set of topics learned from the 137 PASC conditions shown in Fig. 2 with cosine similarity, which is shown in Supplementary Fig. 2e as a heat map, suggesting

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

Variable	Total	Subphenotype 1 (cardiac and renal)	Subphenotype 2 (respiratory, sleep and anxiety)	Subphenotype 3 (musculoskeletal and nervous)	Subphenotype 4 (digestive and respiratory)	P value ^a	Post hoc analysis ^a
No. of patients (%)	20,881 (100%)	7,047 (33.75%)	6,838 (32.75%)	4,879 (23.37%)	2,117 (10.14%)		
Age in years, 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
Female, no. (%)	12,188 (58.37%)	3,627 (51.47%)	4,294 (62.80%)	2,962 (60.71%)	1,305 (61.64%)	4.91E×46	1 vs. 2, 1 vs. 3, 1 vs. 4, 2 vs. 3
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	4,771 (22.85%)	1,752 (24.86%)	1,482 (21.67%)	1,116 (22.87%)	421 (19.89%)	4.99E×7	1 vs. 2, 1 vs. 3, 1 vs. 4, 3 vs. 4
White	7,013 (33.59%)	2,407 (34.16%)	2,255 (32.98%)	1,670 (34.23%)	681 (32.17%)	0.174	–
Other	6,238 (29.87%)	2,052 (29.12%)	2,041 (29.85%)	1,411 (28.92%)	734 (34.67%)	5.13E×6	1 vs. 4, 2 vs. 4, 3 vs. 4
Missing	1,914 (9.17%)	558 (7.92%)	721 (10.54%)	445 (9.12%)	190 (8.97%)	2.43E×6	1 vs. 2, 1 vs. 3, 2 vs. 3, 2 vs. 4
Hispanic, no. (%)	6,838 (32.74%)	2,227 (31.60%)	2,277 (33.30%)	1,554 (31.85%)	780 (36.84%)	3.92E×5	1 vs. 2, 1 vs. 4, 2 vs. 4, 3 vs. 4
ADI, 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×9	2 vs. 1
Index time period of patient, no. (%)							
3/20–6/20	6,899 (33.04%)	2,634 (37.38%)	2,040 (29.83%)	1,554 (31.85%)	671 (31.7%)	8.52E×21	1 vs. 2, 1 vs. 3, 1 vs. 4, 2 vs. 3
7/20–10/20	1,180 (5.65%)	394 (5.59%)	390 (5.70%)	245 (5.02%)	151 (7.13%)	6.06E×3	1 vs. 4, 2 vs. 4, 3 vs. 4
11/20–2/21	8,714 (41.73%)	2,708 (38.43%)	3,016 (44.11%)	2,129 (43.64%)	861 (40.67%)	4.75E×12	1 vs. 2, 1 vs. 3, 2 vs. 4, 3 vs. 4
3/21–6/21	3,390 (16.24%)	1,112 (15.78%)	1,113 (16.28%)	794 (16.27%)	371 (17.52%)	0.298	–
7/21–11/21	698 (3.34%)	199 (2.82%)	279 (4.08%)	157 (3.22%)	63 (2.98%)	3.47E×4	1 vs. 2, 2 vs. 3, 2 vs. 4
Acute phase severities of COVID-19 (–1 to –16 days), no. (%)							
Hospitalized	9,076 (43.47%)	4,309 (61.15%)	2,207 (32.28%)	1,855 (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

^aP value with post hoc analysis: continuous variables were tested by one-way ANOVA, with Tukey's honestly significant difference post hoc test) for normal distribution or by Kruskal–Wallis test (with Dunn post hoc test) for non-normal distribution; categorical variables were tested by one-sided Fisher's exact test with pairwise Fisher's test for post hoc analysis. Post hoc analysis is used to uncover specific differences among subphenotypes—for example, 1 vs. 2: subphenotype 1 and 2 has significant difference. ^bIQR, interquartile range.

that they are highly similar to each other. Supplementary Fig. 2c shows the topic distributions within these newly derived subphenotypes. Finally, Supplementary Fig. 2d quantifies the overlap between the subphenotypes identified from the 44 and original 137 PASC conditions, which showed that more than 90% of the patients would remain in the same subphenotype. This demonstrated the robustness of the PASC subphenotype classification.

Subphenotype validation on the OneFlorida+ cohort

We repeated the same subphenotyping process shown in Fig. 1 on the OneFlorida+ cohort. Extended Data Fig. 4 displays the heat map of all learned potential PASC topics, where we see topics concentrated in conditions of the musculoskeletal system (T1), digestive system (T2), nervous system (T5) and topics mixed with respiratory system conditions and blood/circulatory system conditions (T3) as well as headache and sleep-wake conditions (T7). Some topics also include a mixture of diagnoses. For example, T9 includes throat/chest pain as well as

breathing/heartbeat abnormalities; T6 is a mixture of musculoskeletal pain, headache, malaise and fatigue and skin sensory conditions; T8 and T10 include electrolyte/fluid disorders and anemia/arrhythmias; and T4 includes a combination of sequelae involving digestive, nervous and respiratory systems. We quantitatively evaluated the pairwise similarities between the topics learned from the INSIGHT and OneFlorida+ cohorts (Methods) and visualized the results in Extended Data Fig. 5, which showed a one-to-one correspondence between the topics learned from the two cohorts.

Extended Data Figs. 6 and 7 plot the incidence rates of PASC conditions and medication prescriptions across different subphenotypes in the post-acute infection period in the OneFlorida+ cohort, with mode details shown in Supplementary Tables 8 and 9. The results show that Subphenotype 1 was dominated by incidental cardiac and renal conditions, which included 25.43% of the patients who were older (with a median age of 62.0 years and IQR (49.0–74.0)), with highest proportion of males (46.93%, compared to 38.29% for the overall

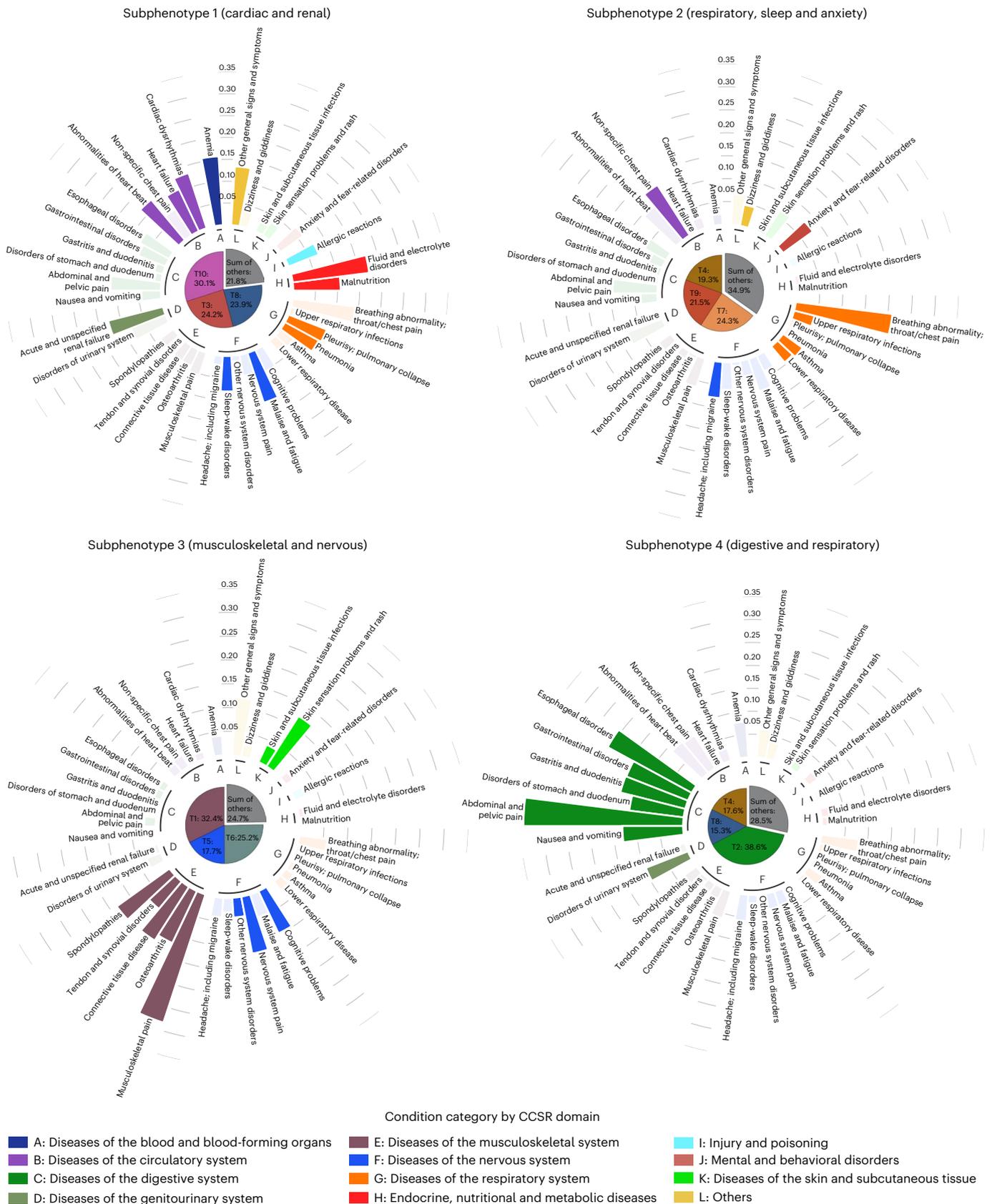


Fig. 3 | Incidence rates of potential PASC conditions in each subphenotype for the INSIGHT cohort (more numerical results are in Supplementary Table 4), where potential PASC conditions were grouped into different categories shown in different colored bars outside the center pie chart. A condition was highlighted from a particular subphenotype where it had the highest incidence rate compared to other subphenotypes. The center pie chart of each

subphenotype showed the mean topic proportions (we showed the three most dominant topics and aggregated the rest) of the patients belonging to each subphenotype, and the colors and definitions of the topics were consistent with Fig. 2. Larger percentage of a particular topic shows more representations of this topic within this subphenotype.

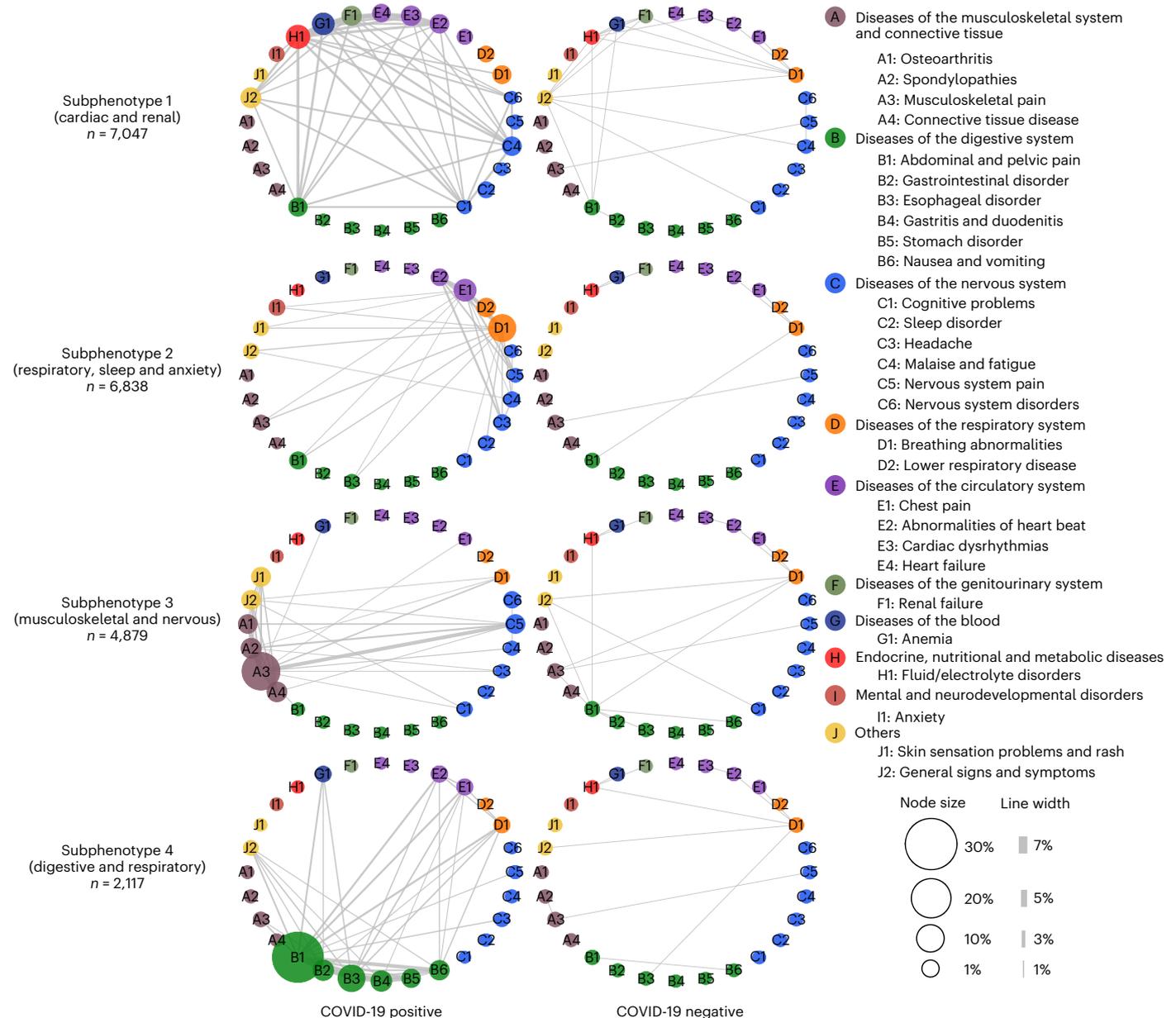


Fig. 4 | Differences in incidence patterns of selected PASC conditions (grouped by CCSR domains) in 30–180 days after COVID-19 lab test between positive and matched negative patients on the INSIGHT cohort. The bubbles in each network correspond to a PASC condition with their sizes proportional to the

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%.

population) and highest rates of hospitalization (57.34%, compared to 36.69% overall), mechanical ventilation (8.57%, compared to 3.39% overall) and critical care admission (12.52%, compared to 6.07% overall) 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 conditions. Subphenotype 2 was dominated by incidental respiratory conditions, sleep disorders and anxiety and was the largest subphenotype, containing 5,281 (38.48%) patients with a median age of 47.0 years (IQR (33.0–61.0)). This subphenotype had a higher prevalence of respiratory conditions at baseline, including chronic obstructive pulmonary disease, pneumonia and upper respiratory tract conditions, and had higher post-acute infection incident prescriptions for respiratory medications. Subphenotype 3 was dominated by incident conditions with musculoskeletal and nervous systems. It

included 3,205 (23.35%) patients with a median age of 48.0 years (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 conditions 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 and respiratory conditions. It was the youngest (median age 46.0 years (IQR (32.0–60.0))) and the subphenotype with the fewest number of patients (1,748 (12.74%) patients), with the highest proportion of females (67.11%, compared to 61.70% overall) and the 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 conditions 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, we also provide the sex-disaggregated analysis results for each subphenotype with respect to patient demographics, baseline and acute infection phase medical utilizations, underlying condition prevalence and newly PASC incidence rates in Supplementary Table 10.

We also replicated the contrast analysis with matched patients who tested negative for SARS-CoV-2 infection in the OneFlorida+ cohort. Similarly to the observations in the INSIGHT cohort, more topics were needed for characterizing the patients who tested negative for SARS-CoV-2 infection (Extended Data Fig. 2b). These topics were not similar to the topics learned from the patients who tested positive for SARS-CoV-2 infection (Extended Data Fig. 3d) and had less concentration patterns (Supplementary Table 6). The contrast of co-incidence patterns of the 137 PASC conditions and 28 selected conditions are shown in Supplementary Fig. 3 and Extended Data Fig. 8, which were also highly consistent with the observations from the INSIGHT cohort.

Discussion

Several studies have found that PASC could include a diverse set of symptoms and signs involving many organ systems^{10–12}. Unlike existing research that has studied these conditions independently, we developed a data-driven framework to identify subphenotypes of SARS-CoV-2-infected patients based on newly incident signs and symptoms 30–180 days after the date of confirmed infection. With the EHR from INSIGHT and OneFlorida+ CRNs, we identified four subphenotypes dominated by new conditions of the cardiac and renal systems (Subphenotype 1); respiratory system, sleep and anxiety problems (Subphenotype 2); musculoskeletal and nervous systems (Subphenotype 3); and digestive and respiratory systems (Subphenotype 4).

In both cohorts, Subphenotypes 1 and 2 are the largest two subphenotypes. Subphenotype 1 contains 33.75% and 25.43% of the patients in INSIGHT and OneFlorida+, respectively. It includes older patients with more baseline comorbidities, greater severity of acute illness in medical utilizations and higher proportion of males²³. Many patients in this subphenotype were confirmed with SARS-CoV-2 infection during the early pandemic (March–September 2020) when NYC was the epicenter, which may explain the observation that its size is larger for INSIGHT than OneFlorida+. Early cases had greater acute phase severity as treatment protocols were still evolving, which may explain more severe incident conditions in the post-acute infection period of these patients, possibly caused by hyperinflammation²⁴. Subphenotype 2 occupies 32.75% and 38.48% of the patients in INSIGHT and OneFlorida+. It includes younger patients who had SARS-CoV-2 infection confirmed mostly during July–November 2021.

Subphenotypes 3 and 4 were less prevalent. Subphenotype 3 included musculoskeletal and neurological conditions, whereas Subphenotype 4 was associated with gastrointestinal conditions. Patients in Subphenotype 3 also displayed dermatologic conditions and had the highest rates of related conditions at baseline, including autoimmune diagnoses such as rheumatoid arthritis and allergy conditions. Patients in Subphenotype 4 had the mildest acute phase severity (for example, lowest rates of mechanical ventilation and critical care admissions).

Our results suggest that the identified subphenotypes are highly consistent across the two cohorts with distinct patient populations and geographical characteristics. These four subphenotypes also covered the major PASC conditions that have been reported from existing independent studies, such as cardiovascular³, respiratory²⁵, neurological²⁶ and gastrointestinal²⁷ conditions. Our study verified the co-existence of these dominant subphenotypes and can inform focused disease areas of treatment development for PASC.

There is also an existing study on identifying Long-COVID symptom clusters with information reported from 233 patients enrolled in the All-Ireland Infectious Disease cohort¹³, whereas our study is based on the diagnosis information from the EHR of large general civilian

patient populations. Some of these diagnoses were with a clear diagnostic criterion, whereas others were not. For example, the conditions in Subphenotype 1, such as heart failure, pneumonia and renal failure, were mostly with objective diagnostic criteria according to underlying disease etiologies. Many conditions in Subphenotype 2 (such as breathing abnormality and non-specific chest pain), Subphenotype 3 (such as musculoskeletal and nervous system pain) and Subphenotype 4 (such as abdominal and pelvic pain, nausea and vomiting) were more subjective to diagnose. In addition, the diagnosis of certain conditions, such as esophageal and gastrointestinal disorders in Subphenotype 4, was likely to encompass functional disorders rather than clearly defined disease etiologies. This meant that our identified subphenotypes, which separated severe COVID-19 complications (Subphenotype 1) and milder PASC conditions (Subphenotypes 2, 3 and 4) that could not be explained by alternative disease etiologies and were closer to those patient-reported symptoms. These subphenotypes would help tease out the heterogeneity of these conditions and provide guidance on patient management in practice.

Our study has several strengths. First, we adopted a TM approach to derive compact patient representations based on their co-incidence patterns across different diagnoses. Unlike other dimensionality reduction techniques, such as principal component analysis (PCA)²⁸, TM is designed specifically for data samples with binary or count features^{29,30} and, thus, would be appropriate for our analysis. Second, INSIGHT and OneFlorida+ include patients from distinct geographic regions in the United States 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 also has limitations. First, our analysis is based on longitudinal observational patient data, which cannot explain the biological mechanisms behind PASC. Second, the PASC diagnoses that we investigated were encoded as CCSR categories, which may not reflect the co-incidence patterns of fine-grained diagnosis conditions. Third, we focused on new incidences of conditions in the post-acute infection period for patients with COVID-19 and did not evaluate pre-existing conditions that may persist or worsen due to acute SARS-CoV-2 infection. Fourth, the goal of our study was to identify potential PASC subphenotypes, and we did not conduct rigorous analysis on the predictability of these subphenotypes, which was left as a future research topic. Finally, our study period did not include the COVID-19 wave dominated by the Omicron SARS-CoV-2 variant.

In conclusion, our study dissects the complexity and heterogeneity of newly incident conditions 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 findings could be useful for clinicians and health systems in developing care models to meet the needs of patients with PASC.

Online content

Any methods, additional references, Nature Portfolio reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at <https://doi.org/10.1038/s41591-022-02116-3>.

References

1. 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).
2. Crook, H., Raza, S., Nowell, J., Young, M. & Edison, P. Long covid—mechanisms, risk factors, and management. *BMJ* **374**, n1648 (2021).
3. Xie, Y., Xu, E., Bowe, B. & Al-Aly, Z. Long-term cardiovascular outcomes of COVID-19. *Nat. Med.* **28**, 583–590 (2022).

4. Xie, Y., Xu, E. & Al-Aly, Z. Risks of mental health outcomes in people with covid-19: cohort study. *BMJ* **376**, e068993 (2022).
 5. Xie, Y. & Al-Aly, Z. Risks and burdens of incident diabetes in long COVID: a cohort study. *Lancet Diabetes Endocrinol.* **10**, 311–321 (2022).
 6. Bowe, B., Xie, Y., Xu, E. & Al-Aly, Z. Kidney outcomes in long COVID. *J. Am. Soc. Nephrol.* **32**, 2851–2862 (2021).
 7. Su, Y. et al. Multiple early factors anticipate post-acute COVID-19 sequelae. *Cell* **185**, 881–895 (2022).
 8. Brodin, P. et al. Studying severe long COVID to understand post-infectious disorders beyond COVID-19. *Nat. Med.* **28**, 879–882 (2022).
 9. Mehandru, S. & Merad, M. Pathological sequelae of long-haul COVID. *Nat. Immunol.* **23**, 194–202 (2022).
 10. Al-Aly, Z., Xie, Y. & Bowe, B. High-dimensional characterization of post-acute sequelae of COVID-19. *Nature* **594**, 259–264 (2021).
 11. 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).
 12. Sudre, C. H. et al. Attributes and predictors of long COVID. *Nat. Med.* **27**, 626–631 (2021).
 13. Kenny, G. et al. Identification of distinct long COVID clinical phenotypes through cluster analysis of self-reported symptoms. *Open Forum Infect. Dis.* **9**, ofac060 (2022).
 14. Kaushal, R. et al. Changing the research landscape: the New York City Clinical Data Research Network. *J. Am. Med. Inf. Assoc.* **21**, 587–590 (2014).
 15. 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).
 16. Clinical Classifications Software Refined (CCSR). https://www.hcup-us.ahrq.gov/toolssoftware/ccsr/ccs_refined.jsp
 17. Zhou, M., Hannah, L., Dunson, D. & Carin, L. Beta-negative binomial process and Poisson factor analysis. In *Proceedings of the Fifteenth International Conference on Artificial Intelligence and Statistics* 1462–1471 (2012).
 18. Kind, A. J. & Buckingham, W. R. Making neighborhood-disadvantage metrics accessible—the neighborhood atlas. *N. Engl. J. Med.* **378**, 2456 (2018).
 19. Blei, D. M. Probabilistic topic models. *Commun. ACM* **55**, 77–84 (2012).
 20. Zang, C. et al. High-throughput clinical trial emulation with real world data and machine learning: a case study of drug repurposing for Alzheimer’s disease. Preprint at <https://www.medrxiv.org/content/10.1101/2022.01.31.22270132v1> (2022).
 21. Zang, C. et al. Understanding post-acute sequelae of SARS-CoV-2 infection through data-driven analysis with longitudinal electronic health records: findings from the RECOVER Initiative. Preprint at <https://www.medrxiv.org/content/10.1101/2022.05.21.2275420v2> (2022).
 22. Nalbandian, A. et al. Post-acute COVID-19 syndrome. *Nat. Med.* **27**, 601–615 (2021).
 23. Peckham, H. et al. Male sex identified by global COVID-19 meta-analysis as a risk factor for death and ICU admission. *Nat. Commun.* **11**, 6317 (2020).
 24. Zaim, S., Chong, J. H., Sankaranarayanan, V. & Harky, A. COVID-19 and multiorgan response. *Curr. Probl. Cardiol.* **45**, 100618 (2020).
 25. Fraser, E. Long term respiratory complications of COVID-19. *BMJ* **370**, m3001 (2020).
 26. Xu, E., Xie, Y. & Al-Aly, Z. Long-term neurologic outcomes of COVID-19. *Nat. Med.* <https://doi.org/10.1038/s41591-022-02001-z> (2022).
 27. Natarajan, A. et al. Gastrointestinal symptoms and fecal shedding of SARS-CoV-2 RNA suggest prolonged gastrointestinal infection. *Med (N Y)* **3**, 371–387 (2022).
 28. Abdi, H. & Williams, L. J. Principal component analysis. *Wiley Interdiscip. Rev. Comput. Stat.* **2**, 433–459 (2010).
 29. Blei, D. M., Ng, A. Y. & Jordan, M. I. Latent Dirichlet allocation. *J. Mach. Learn. Res.* **3**, 993–1022 (2003).
 30. Griffiths, T. L. & Steyvers, M. Finding scientific topics. *Proc. Natl Acad. Sci.* **101**, 5228–5235 (2004).
- Publisher’s note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.
- Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The images or other third party material in this article are included in the article’s Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article’s Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit <http://creativecommons.org/licenses/by/4.0/>.

© The Author(s) 2022

Methods

EHR data repositories

Two large-scale, de-identified, real-world EHR data warehouses were used in our analyses. Our first cohort data were based on the EHR from the INSIGHT CRN¹⁴, which contains the longitudinal clinical information of approximately 12 million patients in the NYC area. Our second cohort data were based on the EHR from the OneFlorida+ CRN¹⁵, which contains the information of nearly 15 million patients mainly from Florida and selected cities in Georgia and Alabama.

This study is part of the National Institutes of Health (NIH) Researching COVID to Enhance Recovery (RECOVER) Initiative³¹, which seeks to understand, treat and prevent the post-acute sequelae of SARS-CoV-2 infection (PASC).

Potential PASC conditions

We systematically reviewed the literature on studies related to PASC or Long-COVID and compiled a comprehensive list of potential PASC conditions in ICD-10 codes based on the one provided in Al-Aly et al.¹⁰.

Then, our clinician team carefully reviewed the list and removed the codes that would not plausibly be considered as PASC in the adult population (for example, HIV, tuberculosis, infection by non-COVID causes, neoplasms, injury due to external causes, etc.) and systematically added the parent codes (for example, the first three digits of ICD-10 codes). Finally, these ICD-10 codes were mapped to 137 condition categories according to CCSR version 2022.1 as Al-Aly et al.¹⁰ did. We added these clarifications and provide the detailed code list in Supplementary Table 2.

Cohort construction

For both cohorts, adult patients (age ≥ 20 years) with at least one SARS-CoV-2 polymerase chain reaction (PCR) or antigen laboratory test (Supplementary Table 11) between 1 March 2020 and 30 November 2021 were included. We further required these patients to have at least one diagnosis code in the baseline period and at least one diagnosis 30–180 days after index date to confirm healthcare utilization capture. Then, we chose the patients who had at least one positive test and had at least one potential PASC condition in the follow-up (or post-acute infection) period. We further made sure that 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 is shown in Extended Data Fig. 9, and the relevant definitions are provided below.

- Index date: the date of the first COVID-19-positive test.
- Baseline period: from 3 years to 1 week before the index date.
- Follow-up (post-acute infection) period: from 31 days after the index date to the day of documented death, the last record in the database, 180 days after baseline (we followed the definition provided by the NIH and the Centers for Disease Control and Prevention (CDC)^{10,22,31,32}) or the end of our observational window (30 November 2021), whichever came first.

Statistical analysis

TM. We use binary vectors $\{\mathbf{x}_n\}_{n=1}^N$ to represent the patients, where n is the patient index the i -th element of \mathbf{x}_n , or $x_n(i) = 1$ if the i -th potential PASC condition appears in the post-acute infection period of the n -th patient's EHR, otherwise $x_n(i) = 0$. Therefore, each \mathbf{x}_n is a 137-dimensional binary vector (step 1 in Fig. 1). TM¹⁹ was then applied on these vectors to learn a set of potential PASC topics. PASC topics in this context refers to a group of conditions that present together based on their incident probabilities. Specifically, assume that each patient can be represented as a mixture of K latent PASC topics $\Phi \in R^{D \times K}$, where $D = 137$ is the number of unique PASC conditions and K is the total number of topics. Each topic ϕ_k ($k : 1 \leq k \leq K$ is the index of the topic) 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 Fig. 1). Then, for each patient, TM infers the mixture memberships $\theta_n \in R^K$, also called topic proportions or topic loadings, as the new representation for each patient (step 3 in Fig. 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 \mathbf{x}_n to the low-dimensional continuous PASC topic space θ_n , which will be leveraged further for subphenotype identification through clustering later (step 4 in Fig. 1). Specifically, we used Poisson factor analysis (PFA)¹⁷ as the concrete TM method, which generates \mathbf{x}_n as follows.

- Draw a topic proportion θ_n for the n -th patient from a Gamma distribution $\theta_n \sim \text{Gamma}(1,1)$;
- Draw the k -th PASC topic from a Dirichlet distribution $\phi_k \sim \text{Dirichlet}(0.01)$, $k = 1, \dots, K$;
- Draw a binary vector x_n from the Bernoulli distribution by Bernoulli–Poisson link^{33–36}:

$$x_n = 1(\mathbf{u}_n \geq 1), \mathbf{u}_n \sim \text{Poisson}(\Phi\theta_n);$$

where $1(\cdot)$ is an indicator function representing $x_n = 1$ if $\mathbf{u}_n \geq 1$ and $x_n = 0$ if $\mathbf{u}_n = 0$. The input of the PFA model is the observed patient binary vector $\{\mathbf{x}_n\}_{n=1}^N$. The learnable parameters of the PFA model are the topic proportion vectors $\{\theta_n\}_{n=1}^N$ and the topic matrix Φ . Because the PFA is a fully conjugated probabilistic model, we trained it by Gibbs sampling with the conditional posterior of the learnable parameters, whose specific equations can be found in Zhou et al.¹⁷. We used the Python package Pydpm version 3.0.1 (<https://pypi.org/project/pydpm/>) to train the model and released the code.

For PFA, we imposed Dirichlet as the prior distribution for k -th PASC topic so that it lies in a probability simplex space (elements in ϕ_k represent the probability of each PASC appearing in k -th topic). Compared to another popular TM method—latent Dirichlet allocation (LDA)²⁹ that imposes Dirichlet distribution on the topic proportions θ_n —PFA uses Gamma distribution that can obtain more separable compressed representations^{17,35}, which is useful for identifying more meaningful subphenotypes.

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 coherence³⁷. Data likelihood is used to evaluate the fitness of the model on the current dataset, 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 0 to 1, and higher value indicates better coherence. Detailed calculations of these two metrics are provided below.

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

$$x_n = 1(\mathbf{u}_n \geq 1), \mathbf{u}_n \sim \text{Poisson}(\Phi\theta_n);$$

where \mathbf{x}_n is the binary PASC vector, Φ is the topic matrix, θ_n is the topic proportion vector and \mathbf{u}_n is the latent variable that links the binary observation and topic representation. According to the property of Bernoulli–Poisson link, \mathbf{u}_n can be marginalized out, and then one can obtain a Bernoulli likelihood as

$$x_n \sim \text{Ber}(1 - \exp^{-\Phi\theta_n}).$$

Given $\{\mathbf{x}_n\}_{n=1}^N$, after learning Φ and $\{\theta_n\}_{n=1}^N$, we can calculate the Bernoulli data likelihood. In Extended Data Fig. 2, we show the mean (average of the number of patients) data log-likelihood:

$$\frac{1}{N} \sum_{n=1}^N \sum_{v=1}^V \hat{x}_{n,v} \log(x_{n,v}) + (1 - \hat{x}_{n,v}) \log(1 - x_{n,v}),$$

where N is the number of patients and V is the total number of PASC, and $\hat{x}_{n,v}$ is the averaged reconstruction item. Because each iteration of Gibbs sampling provides only a point estimation, to obtain the likelihood of the distribution estimation in practice we collect the parameters of every iteration after burn-in step (500 burn-in steps in our experiment) and then take the average of them (Zhou et al.¹⁷), defined as $\hat{x}_n = \frac{1}{Q} \sum_{q=1}^Q 1 - \exp^{-\langle \theta_n^{(q)} \rangle}$, where Q is the total number of collected parameters ($Q = 500$ in our experiment).

Topic coherence. Topic coherence is an important metric 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.

The results are shown in Extended Data Fig. 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 ones 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 suggest overfitting).

As TM is a probabilistic process, we want to guarantee the robustness of the identified topics. To achieve this goal, we first did 1,000 bootstrapping (randomly sample the same number of patients with replacement) 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:

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

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. Extended Data Fig. 3a,b demonstrates the heat map 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 that the learned topics are robust.

We also quantitatively evaluated the consistency between the two sets of topics learned from different cohorts. Specifically, denoting the topic matrices learned from the two cohorts as $\Phi^{(1)}$ and $\Phi^{(2)}$, 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:

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

Finally, the heat map of the topic consistency matrix with S_{ij} as its (i, j) -th entry was shown in Extended Data Fig. 5, from which we can clearly observe darker diagonal values, which suggests high consistency between the two sets of topics. In addition, the topic consistency learned from the positive and negative cohorts is provided as results in Extended Data Fig. 3c,d, which demonstrates low consistency between two sets of topics from the positive and negative cohorts.

We calculated the entropy of each topic to quantitatively evaluate the concentration patterns of each topic. Specifically, because each

topic is a probabilistic distribution over the PASC conditions, we can use the following definition of entropy to evaluate whether each topic is closer to a uniform distribution or concentrates on some related PASC.

$$H[\boldsymbol{\varphi}_k] = - \sum_{i=1}^{137} \varphi_{ki} \times \log_2 \varphi_{ki},$$

where φ_{ki} is the i -th element in $\boldsymbol{\varphi}_k$. In our case, $H[\boldsymbol{\varphi}_k]$ should be in a range of (0, 7.0980). Smaller $H[\boldsymbol{\varphi}_k]$ denotes that the k -th topic more likely concentrates on particular PASC conditions (capturing stronger co-occurrence pattern of PASC), whereas larger $H[\boldsymbol{\varphi}_k]$ denotes that the k -th topic is closer to a uniform distribution. From Supplementary Table 6, we observed that those topics learned from the positive cohort had lower entropy values than those learned from the negative cohort.

Subphenotyping through clustering. With the learned K -dimensional topic loading vector for n -th patient $\boldsymbol{\theta}_n$, we applied the hierarchical agglomerative clustering method with Euclidean distance calculation and Ward linkage criterion³⁸ to derive subphenotypes as patient clusters. For determining the optimal number of clusters (subphenotypes), according to Su et al.³⁹ and Xu et al.⁴⁰, we applied the NbClust R package⁴¹, 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 that four is the optimal number of clusters (Supplementary Table 12). Through majority voting, we set the number of clusters as four in both two cohorts. Extended Data Fig. 10 demonstrates the uniform manifold approximation and projection (UMAP) embeddings⁴² and dendrogram of these clusters for both cohorts.

We 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 1,000 bootstrapping (randomly choose 80% patients from the cohort denoted as set $\Omega_i, i = 1, \dots, 1000$) to learn the 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 the mean and 95% confidence interval (CI) of Adjusted Rand Index (ARI) score and normalized mutual information (NMI) between the clustering results on bootstrapped sample sets and reference as 0.902 (95% 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 that the identified clusters are highly robust.

Comparison with patients without SARS-CoV-2 infection. 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 patients who tested negative for SARS-CoV-2 all had negative results from lab tests between March 2020 and November 2021, and there were no documented COVID-19-related diagnoses during this time period. The index date for each individual patient in the 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 and 85+ years).
- The ADI (ten rank bins of national ADI) for capturing the socio-economic disadvantage of patients' neighborhood⁴⁸.
- Index date for considering the effect of different stages of the pandemic, which was binned into different time intervals (March–June 2020, July–October 2020, November

2020 to February 2021, March–June 2021 and July–November 2021).

- 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 and 5+ visits for each encounter type).
- Coexisting conditions, including comorbidities and medications based on a tailored list of the Elixhauser comorbidities⁴³. 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 we performed robust PS matching on other hypothetical confounders to rank the patients in the initial set, and we finally picked the top two. We used standardized mean difference (SMD) to quantify the goodness-of-balance of confounders between two groups $SMD(X_1, X_0) = \frac{|E[X_1] - E[X_0]|}{\sqrt{(\text{Var}(X_1) + \text{Var}(X_0))/2}}$ where $SMD < 0.2$ is the threshold to examine whether this confounder is balanced⁴⁴. On both INSIGHT and OneFlorida+ cohorts, we found that all confounders on all subphenotypes were balanced.

Shortening the PASC condition list with high-dimensional PS adjustment. We leveraged our high-dimensional PS adjustment pipeline^{20,21} to identify the potential PASC conditions that showed significantly higher risk during 30–180 days after lab-confirmed SARS-CoV-2 infection compared to patients who tested negative for SARS-CoV-2, after adjusting for a comprehensive list of hypothetical confounders, including demographics, baseline medical utilizations and comorbidities, and severity during the acute infection phase. More details can be found in Zang et al.²¹, and the results on different PASC conditions are provided in Supplementary Table 7. We selected a particular PASC condition if either of the following two conditions was satisfied.

- Hazard ratio (positive over negative) was larger than 1, and P value was smaller than 0.05/137.
- Hazard ratio is larger than 1, and P value is smaller than 0.05, and it is reported or suspected in the PASC-related works^{10,22}.

With these two rules, we were able to identify 44 potential PASC conditions, which are provided in Supplementary Table 2.

All statistical analysis was performed with Python 3.7–Python package scikit-learn-0.23.2, numpy-1.16.5, umap-learn-0.5.1, Pydpm-3.0 and scipy-1.7.3 for machine learning models.

Ethics approval

Use of the INSIGHT data was approved by the institutional review board (IRB) of Weill Cornell Medicine following protocol 21–10–95–380 with the title ‘Adult PCORnet-PASC Response to the Proposed Revised Milestones for the PASC EHR/ORWD Teams (RECOVER)’. Use of the OneFlorida+ data for this study was approved under University of Florida IRB number IRB202001831. All EHRs used in this study were appropriately de-identified, and, thus, no informed consents from patients were obtained.

Reporting summary

Further information on research design is available in the Nature Portfolio Reporting Summary linked to this article.

Data availability

Information of the INSIGHT CRN is provided at <https://insightcrn.org/>, and INSIGHT data (<https://onefloridaconsortium.org/>) are made available to researchers with an approved study protocol at <https://nyc-cdrn.atlassian.net/servicedesk/customer/portal/2/group/6/create/16>. Information of the OneFlorida+ CRN is provided at

<https://onefloridaconsortium.org/>, and OneFlorida+ data are made available to researchers with an approved study protocol at <https://onefloridaconsortium.org/front-door/prep-to-research-data-query/>. For questions regarding INSIGHT, email: insightcrn@med.cornell.edu. For questions regarding OneFlorida+, email: OneFloridaOperations@health.ufl.edu. Source data are provided with this paper.

Code availability

Our codes are available at <https://github.com/haozhangWCM/Subphenotyping-for-PASC>.

References

1. RECOVER: Researching COVID to Enhance Recovery (NIH, 2022); <https://recovercovid.org>
2. Centers for Disease Control and Prevention. Post-COVID conditions among adult COVID-19 survivors aged 18–64 and ≥65 years—United States, March 2020–November 2021. <https://www.cdc.gov/mmwr/volumes/71/wr/mm7121e1.htm> (2022).
3. Zhang, Y., Zhao, Y., David, L., Henao, R. & Carin, L. Dynamic Poisson factor analysis. In *Proceedings of the IEEE 16th International Conference On Data Mining (ICDM)*. <https://scholars.duke.edu/display/pub1241512> 1359–1364 (IEEE, 2017).
34. Zhang, H., Chen, B., Guo, D. & Zhou, M. WHAI: Weibull hybrid autoencoding inference for deep topic modeling. In *Proceedings of the International Conference on Learning Representations*. Available at <https://arxiv.org/abs/1803.01328> (2018).
35. Zhang, H. et al. Deep autoencoding topic model with scalable hybrid Bayesian inference. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **43**, 4306–4322 (2021).
36. Guo, D., Chen, B., Zhang, H. & Zhou, M. Deep Poisson gamma dynamical systems. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, pp. 8451–8461. 2018.
37. Newman, D., Lau, J. H., Grieser, K. & Baldwin, T. Automatic evaluation of topic coherence. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics* 100–108 (2010).
38. Murtagh, F. & Legendre, P. Ward’s hierarchical agglomerative clustering method: which algorithms implement Ward’s criterion? *J. Classification* **31**, 274–295 (2014).
39. Su, C. et al. Clinical subphenotypes in COVID-19: derivation, validation, prediction, temporal patterns, and interaction with social determinants of health. *NPJ Digit. Med.* **4**, 110 (2021).
40. Xu, Z. et al. Sepsis subphenotyping based on organ dysfunction trajectory. *Crit. Care* **26**, 197 (2022).
41. Charrad, M., Ghazzali, N., Boiteau, V. & Niknafs, A. NbClust: an R package for determining the relevant number of clusters in a data set. *J. Stat. Softw.* **61**, 1–36 (2014).
42. McInnes, L., Healy, J. & Melville, J. UMAP: uniform manifold approximation and projection for dimension reduction. Preprint at <https://arxiv.org/abs/1802.03426> (2018).
43. Elixhauser Comorbidity Software Refined for ICD-10-CM (AHRQ, 2021); https://www.hcup-us.ahrq.gov/toolssoftware/comorbidityicd10/comorbidity_icd10.jsp
44. 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).

Acknowledgements

This research was funded by NIH Agreement Other Transactions Authority (OTA). OT2HL161847 (contract no. EHR-01–21) is part of the Researching COVID to Enhance Recovery (RECOVER) research program. The PCORnet study reported in this work was conducted using PCORnet, the National Patient-Centered Clinical Research Network. PCORnet was developed with funding from the

Patient-Centered Outcomes Research Institute (PCORI). This work was conducted through use of data from the INSIGHT Clinical Research Network and supported, in part, by a PCORI PCORnet grant to the INSIGHT Clinical Research Network (grant no. RI-CORNELL-01-MC). The statements presented in this work are solely the responsibility of the author(s) and do not necessarily represent the views of other organizations participating in, collaborating with or funding PCORnet, or of PCORI.

Author contributions

H.Z. and F.W. proposed the initial idea. H.Z. designed and implemented the framework and analyzed the results. D.M. and M.W. set up the data infrastructure and analytics environment for INSIGHT. J.B. set up the data infrastructure and analytics environment for OneFlorida+. C.Z. and J.X. pre-processed the INSIGHT and OneFlorida+ data and helped with the analysis. Z.X. and Y.Z. helped with the statistical analysis. A.N., E.J.S., M.W., J.B., R.R. and R.K. provided clinical input on data, study design and results interpretation. H.Z. drafted the initial manuscript. F.W. and R.K. made critical revisions. All authors provided feedback to and proofread the final version of the paper.

Competing interests

The authors declare no competing interests.

Additional information

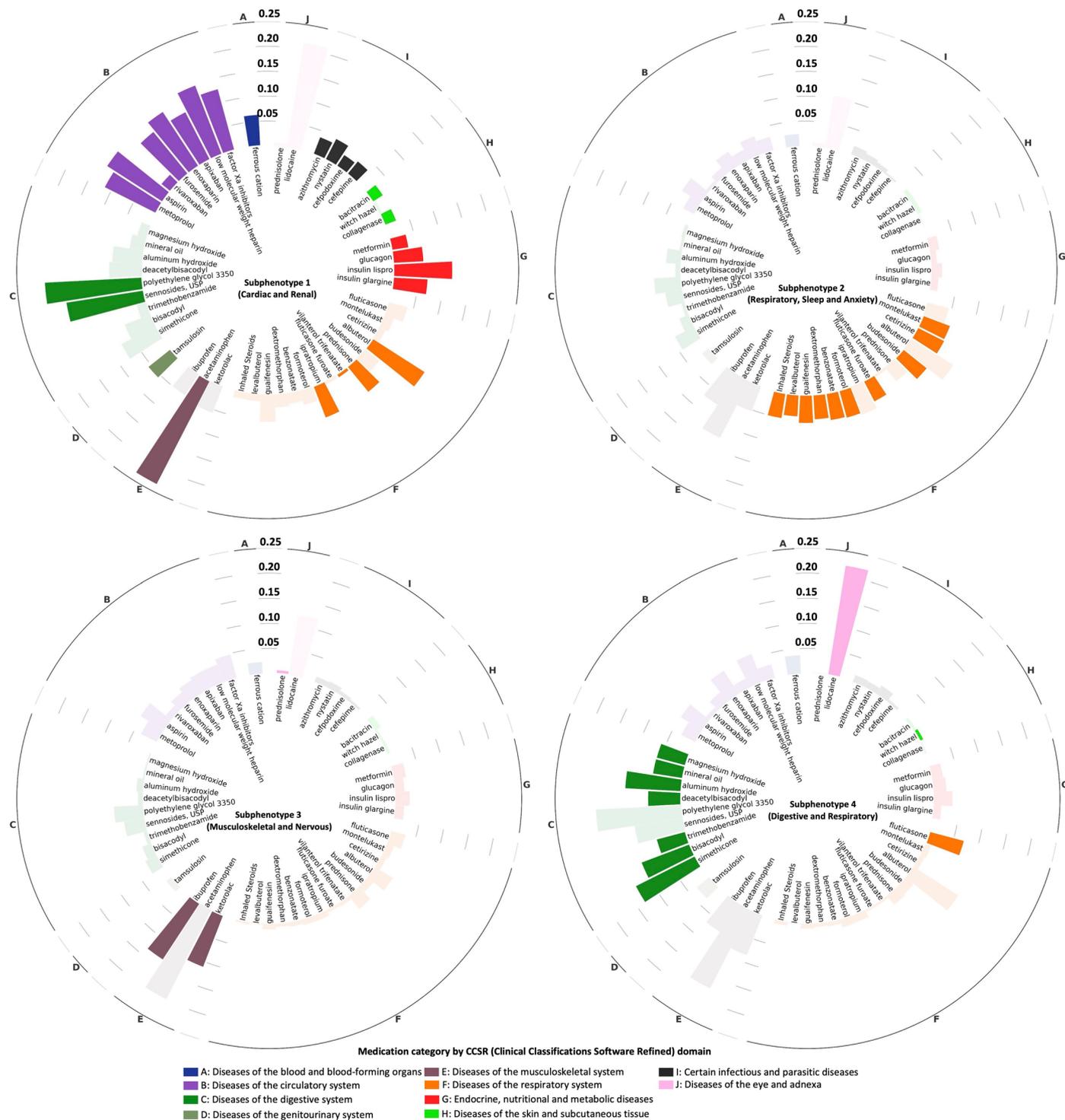
Extended data is available for this paper at <https://doi.org/10.1038/s41591-022-02116-3>.

Supplementary information The online version contains supplementary material available at <https://doi.org/10.1038/s41591-022-02116-3>.

Correspondence and requests for materials should be addressed to Fei Wang.

Peer review information *Nature Medicine* thanks the anonymous reviewers for their contribution to the peer review of this work. Primary handling editor: Ming Yang, in collaboration with the *Nature Medicine* team.

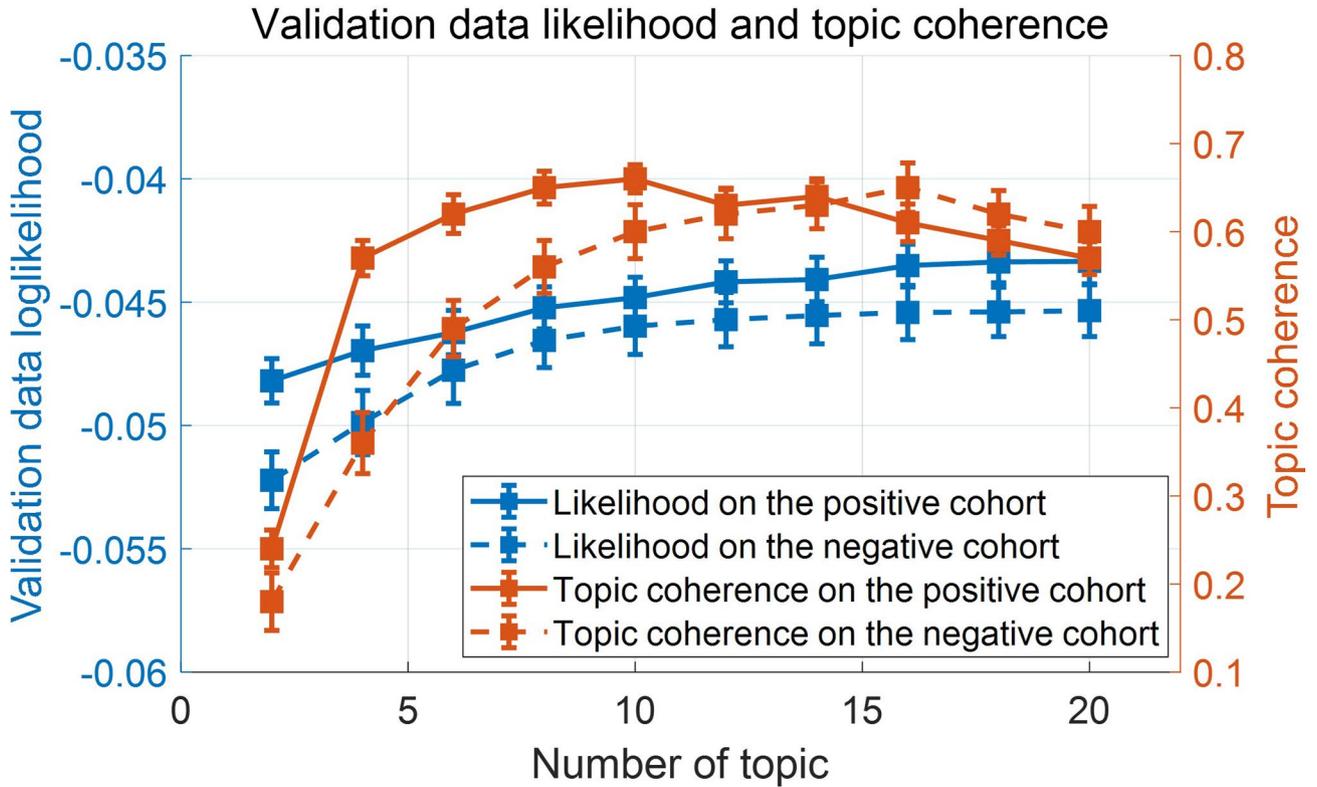
Reprints and permissions information is available at www.nature.com/reprints.



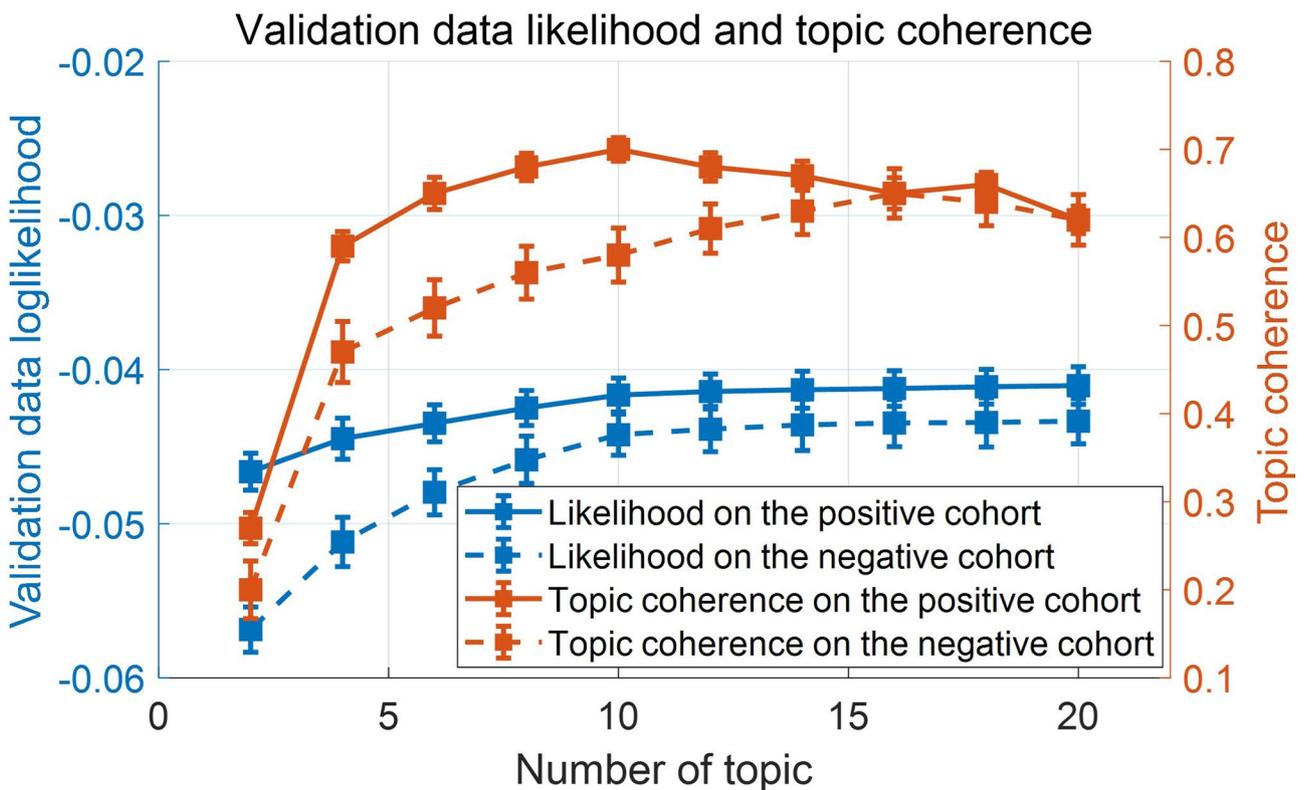
Extended Data Fig. 1 | Circos plot for the INSIGHT subphenotypes with newly incident medications. The prevalence of incident prescriptions of medications in the post-acute infection period for each subphenotype on the INSIGHT cohort,

where medications are grouped into different categories shown by different colors. The most prevalent medication in each subphenotype is highlighted.

a. Evaluation on the INSIGHT cohort



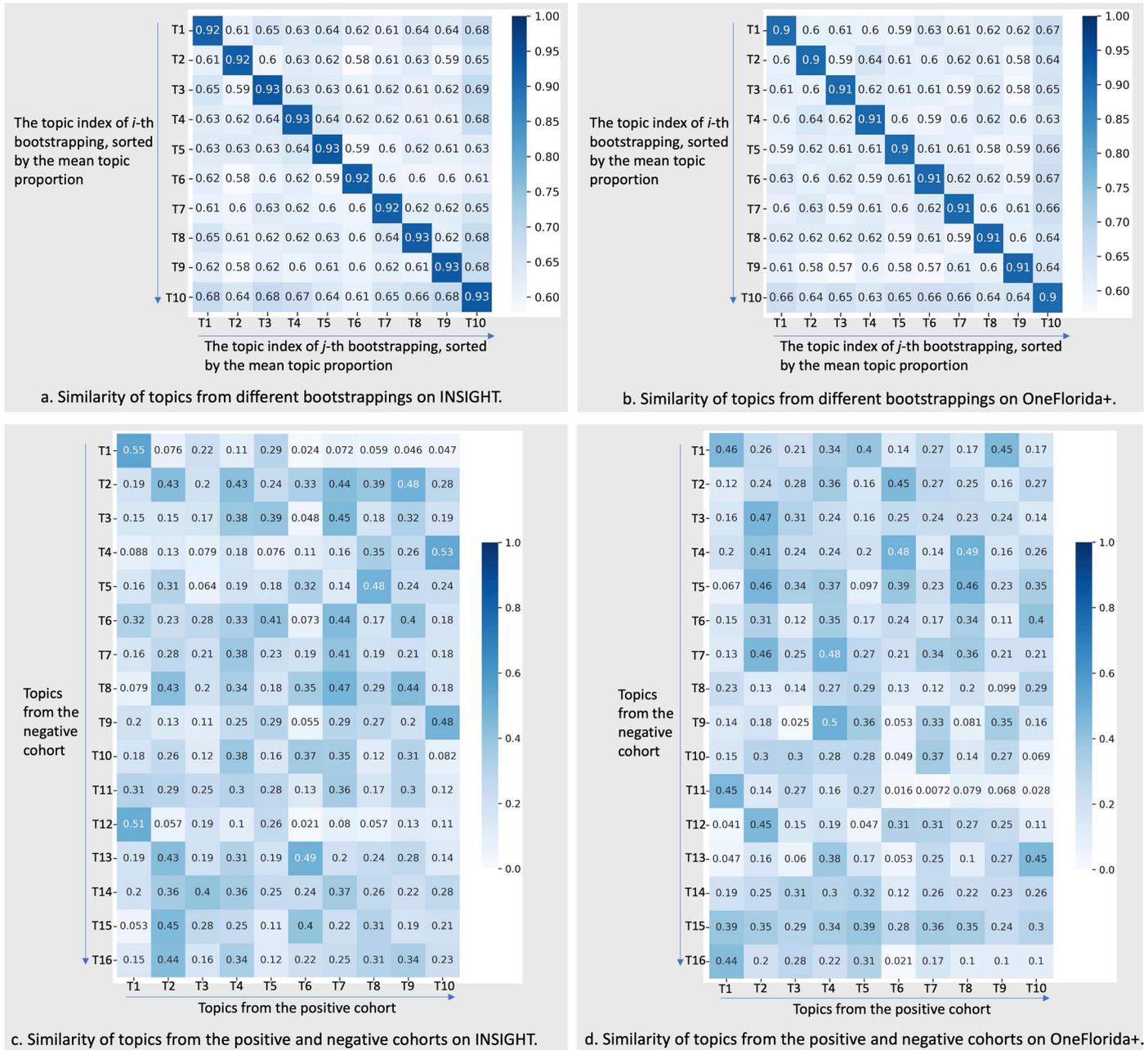
b. Evaluation on the OneFlorida+ cohort



Extended Data Fig. 2 | See next page for caption.

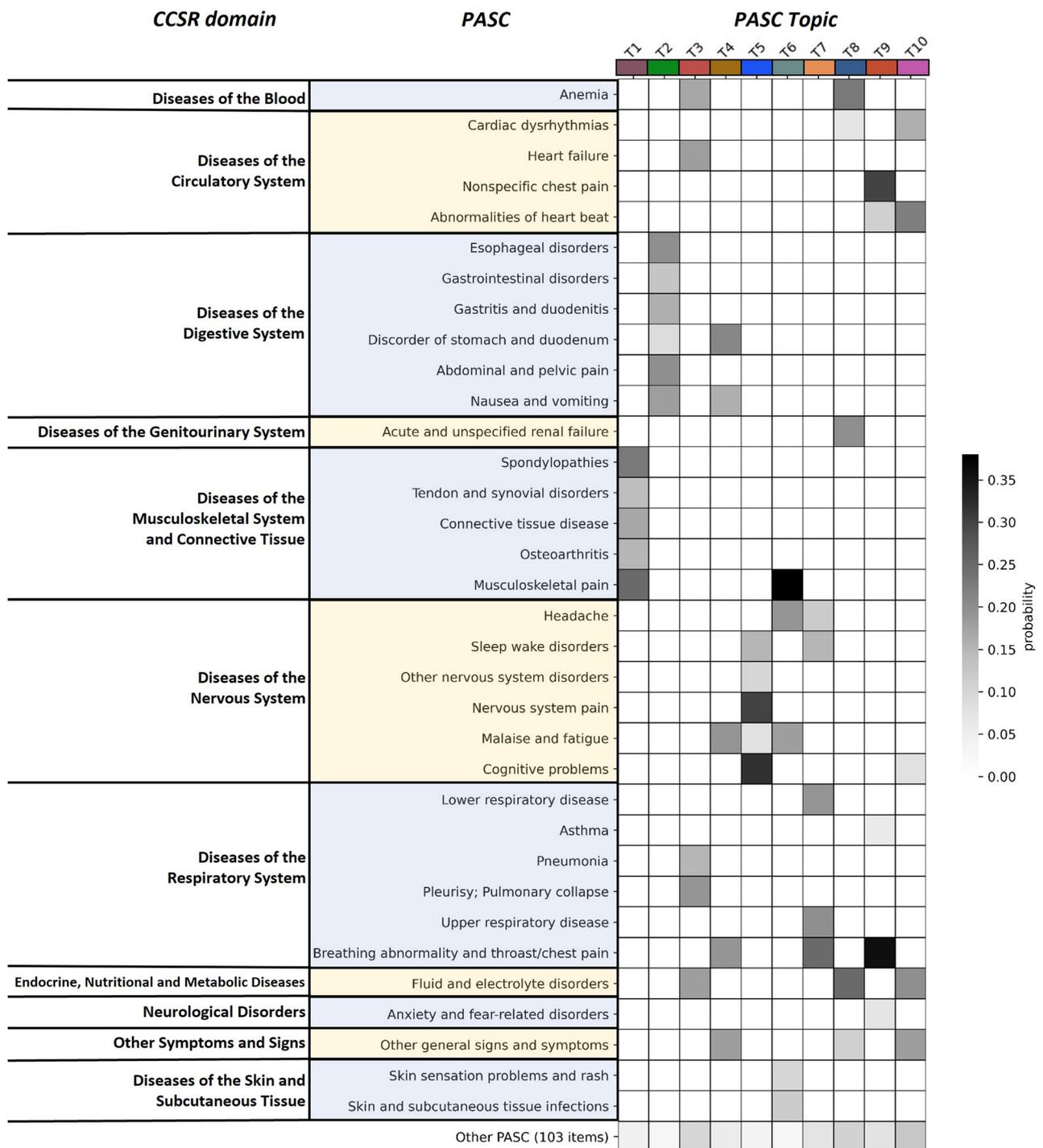
Extended Data Fig. 2 | PASC topic curve for the INSIGHT and OneFlorida+ cohorts. Data likelihood and topic coherence with respect to different number of topics on a) the INSIGHT cohort and b) the OneFlorida+ cohort. For both criteria, larger values indicate better quality of the learned topics. We evaluated them on the group of patients who tested positive for SARS-CoV-2 infection (solid curves),

and the group of patients who tested negative for SARS-CoV-2 infection (dashed curves). For each evaluation, we randomly split the corresponding patient group into 70% for training and the rest 30% for evaluation, and repeated this process for 100 times. The average evaluation results are shown as centers in the curves and standard deviations are shown as error bars.



Extended Data Fig. 3 | Topic robustness test for the INSIGHT and OneFlorida+ cohorts. The similarity of topics from different bootstrapping on a) the INSIGHT cohort and b) the OneFlorida+ cohort, which were used to evaluate the robustness of topics. c and d: the similarity of topics from the patients who tested

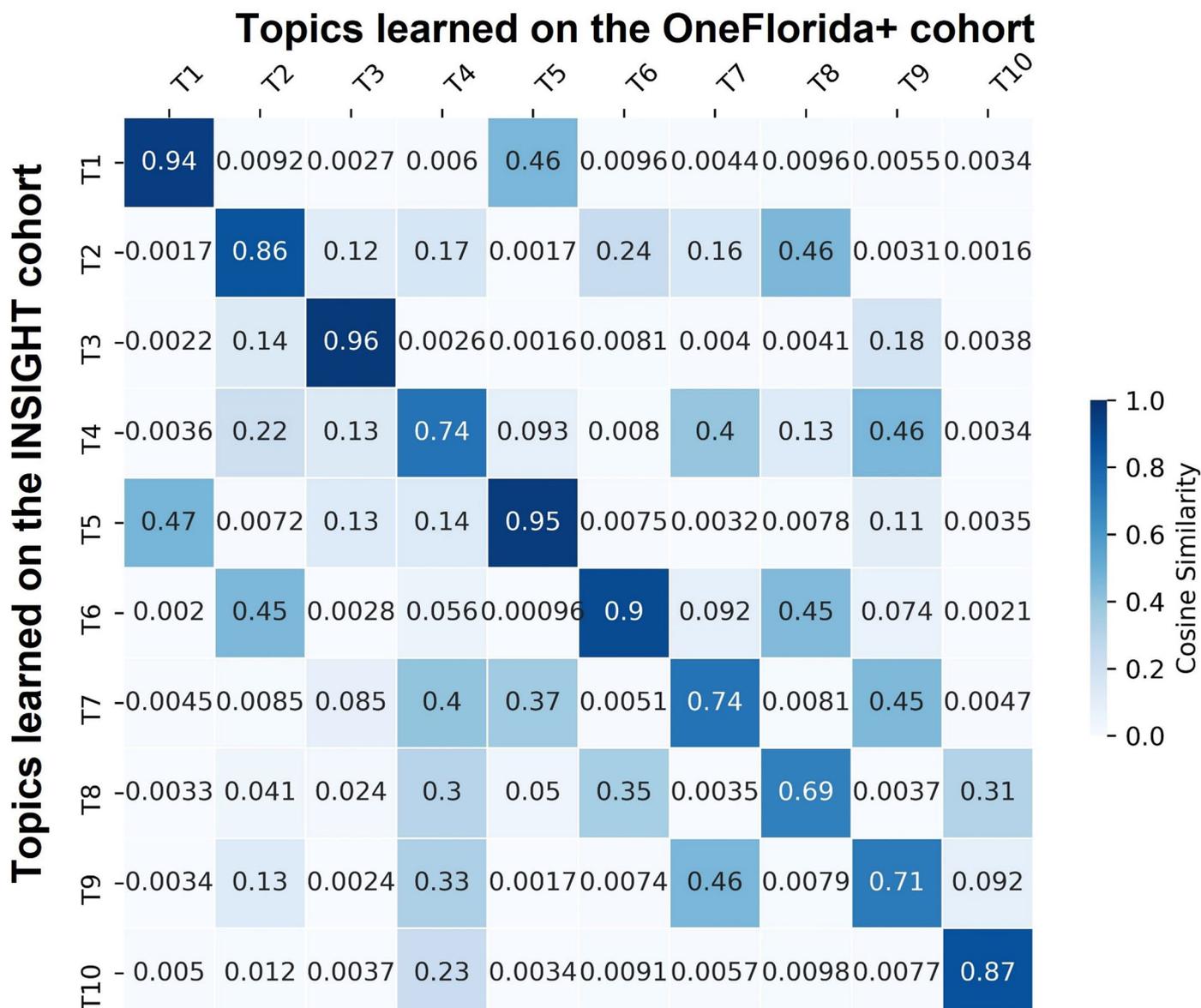
positive and negative for SARS-CoV-2 infection on INSIGHT and OneFlorida+ cohorts, which were used to evaluate the consistency of topics learned from these two cohorts.



Extended Data Fig. 4 | The PASC topic heatmap for the OneFlorida+ cohorts.

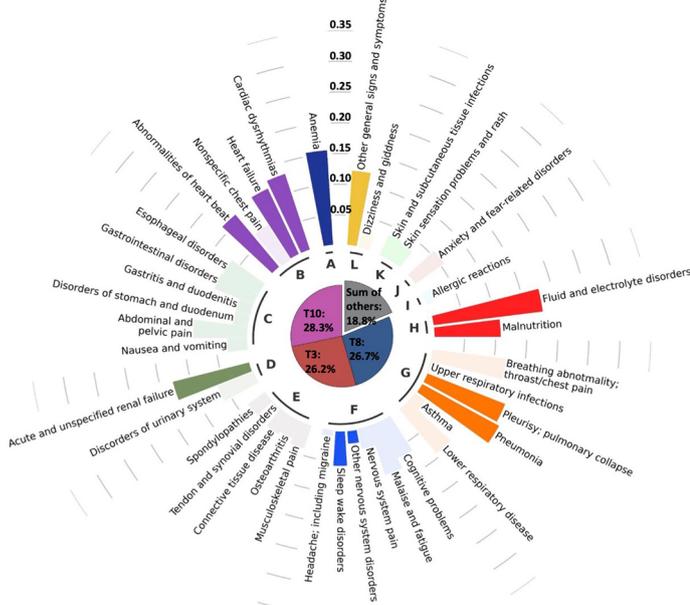
The heatmap of PASC topics learned from the OneFlorida+ cohort. Each row denotes a potential PASC category grouped by different CCSR domains, and

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.

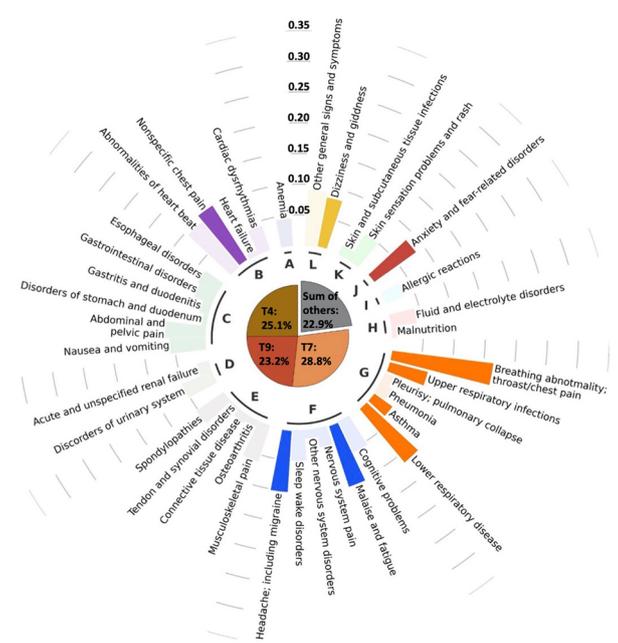


Extended Data Fig. 5 | The overlap of topics between the INSIGHT and OneFlorida+ cohorts. The similarity of topics between INSIGHT and OneFlorida+ cohorts, which were used to evaluate the overlap of topics learned on two cohorts.

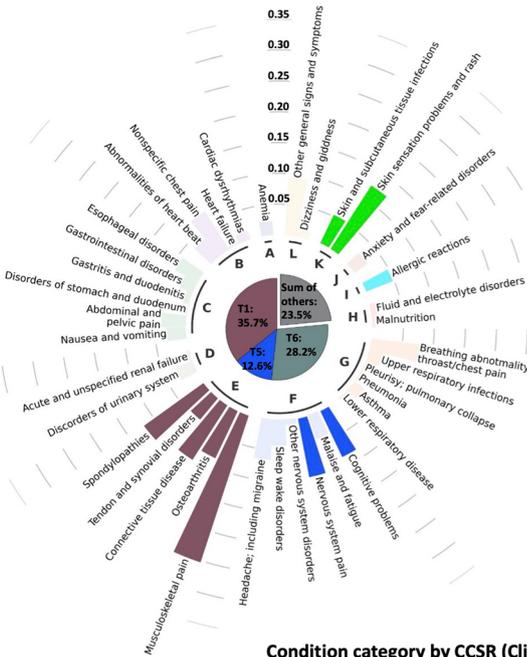
Subphenotype 1 (Cardiac and Renal)



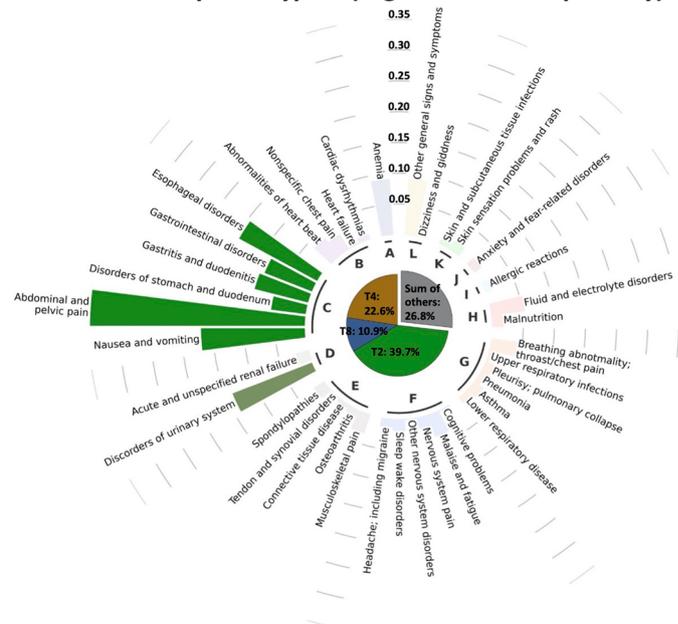
Subphenotype 2 (Respiratory, Sleep and Anxiety)



Subphenotype 3 (Musculoskeletal and Nervous)



Subphenotype 4 (Digestive and Respiratory)

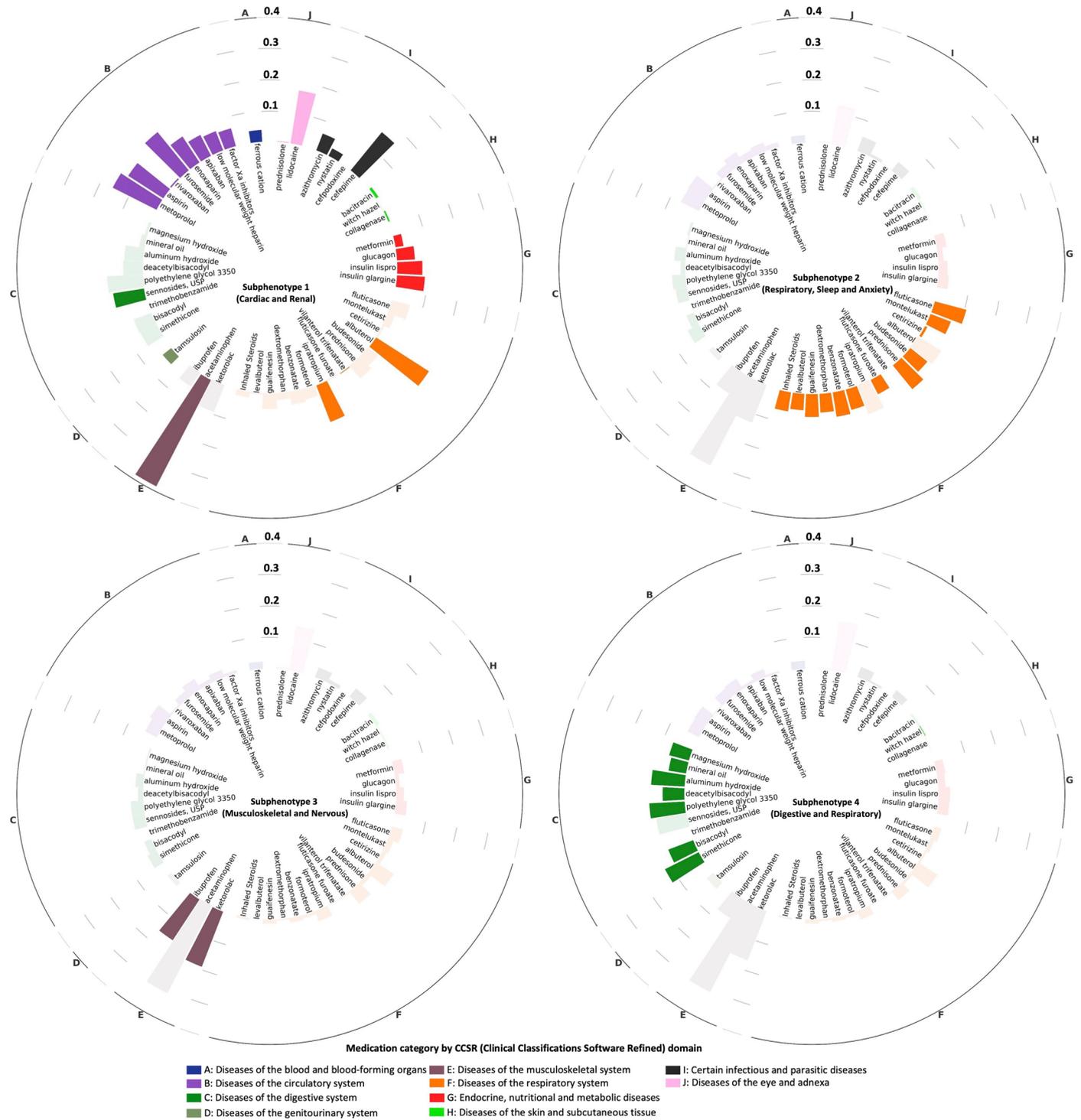


Condition category by CCSR (Clinical Classifications Software Refined) domain

- A: Diseases of the blood and blood-forming organs
- B: Diseases of the circulatory system
- C: Diseases of the digestive system
- D: Diseases of the genitourinary system
- E: Diseases of the musculoskeletal system
- F: Diseases of the nervous system
- G: Diseases of the respiratory system
- H: Endocrine, nutritional and metabolic diseases
- I: Injury and poisoning
- J: Mental and behavioral disorders
- K: Diseases of the skin and subcutaneous tissue
- L: Others

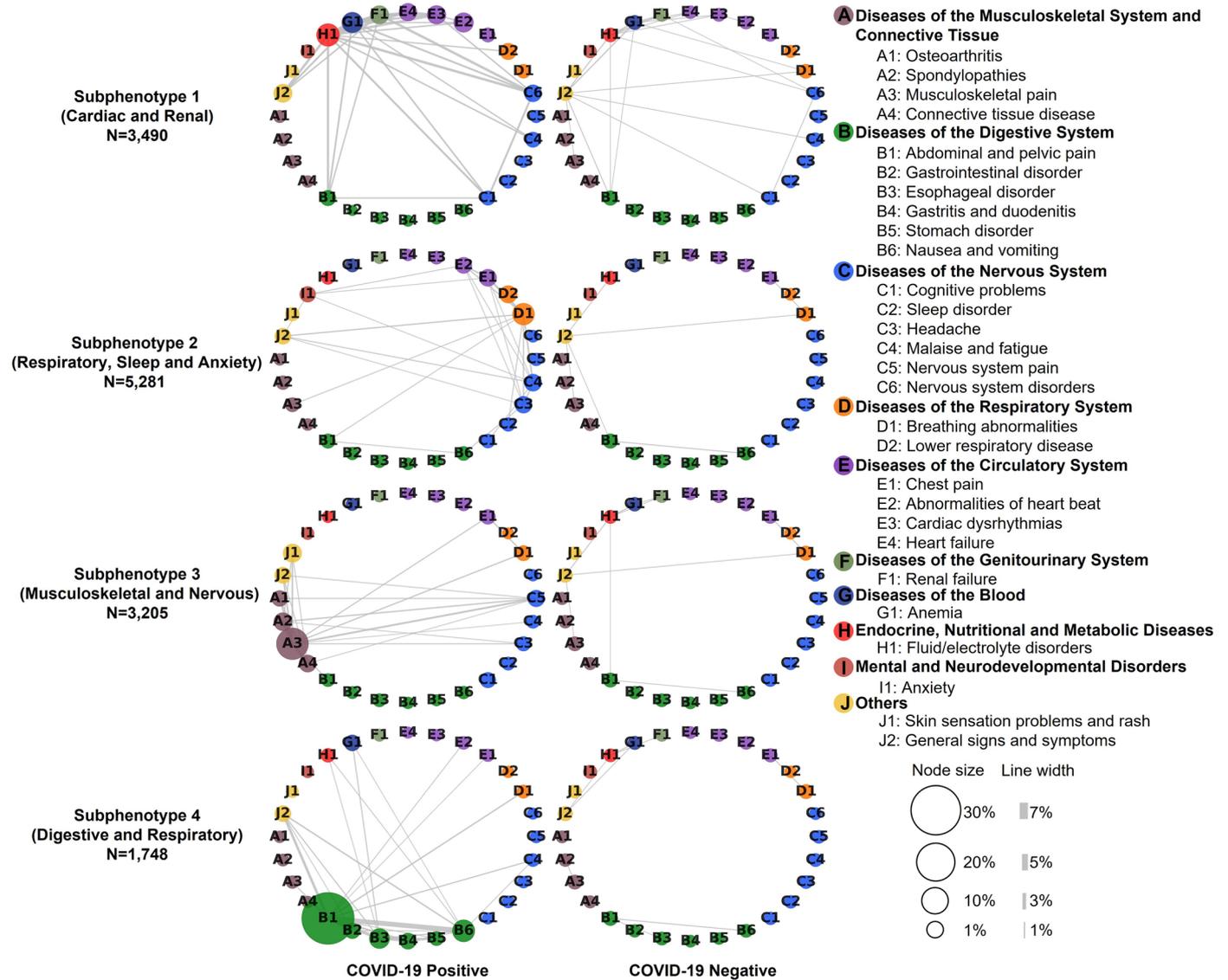
Extended Data Fig. 6 | Circos plot for the OneFlorida+ subphenotypes with PASC. The incidence rates of potential PASC conditions in each subphenotype for the OneFlorida+ cohort (more numerical results are in Supplementary Table 9), where potential PASC conditions were grouped into different categories shown in different colored bars outside the center pie chart. A condition with

the highest incidence rate is highlighted in the subphenotype. The center pie chart of each subphenotype shows the mean topic proportions of the patients it included, and the colors and meanings of the topic indices can be referred to Extended Data Fig. 4.



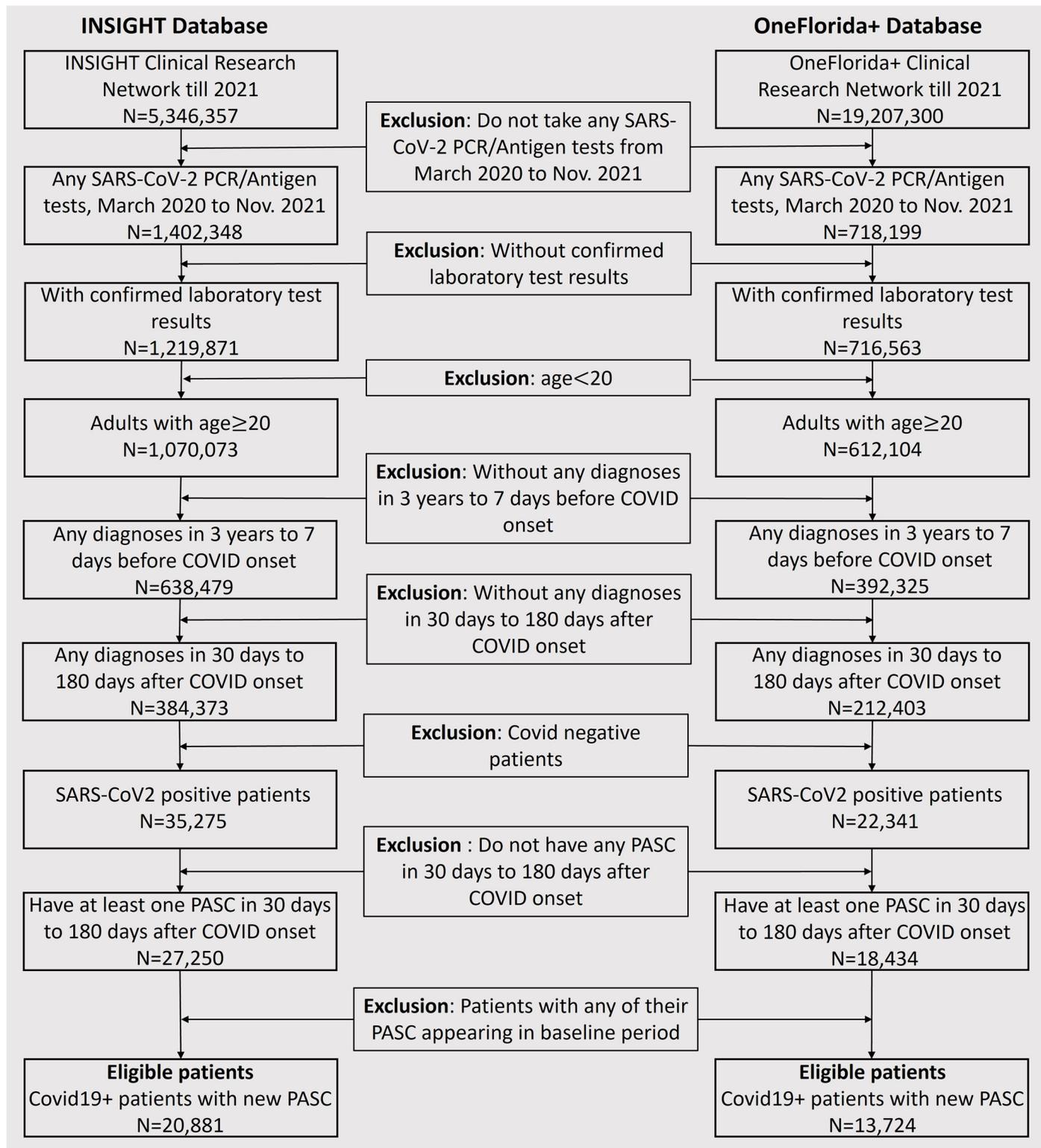
Extended Data Fig. 7 | Circos plot for the OneFlorida+ subphenotypes with newly incident medications. The prevalence of incident prescriptions of medications in the post-acute infection period for each subphenotype on the

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.



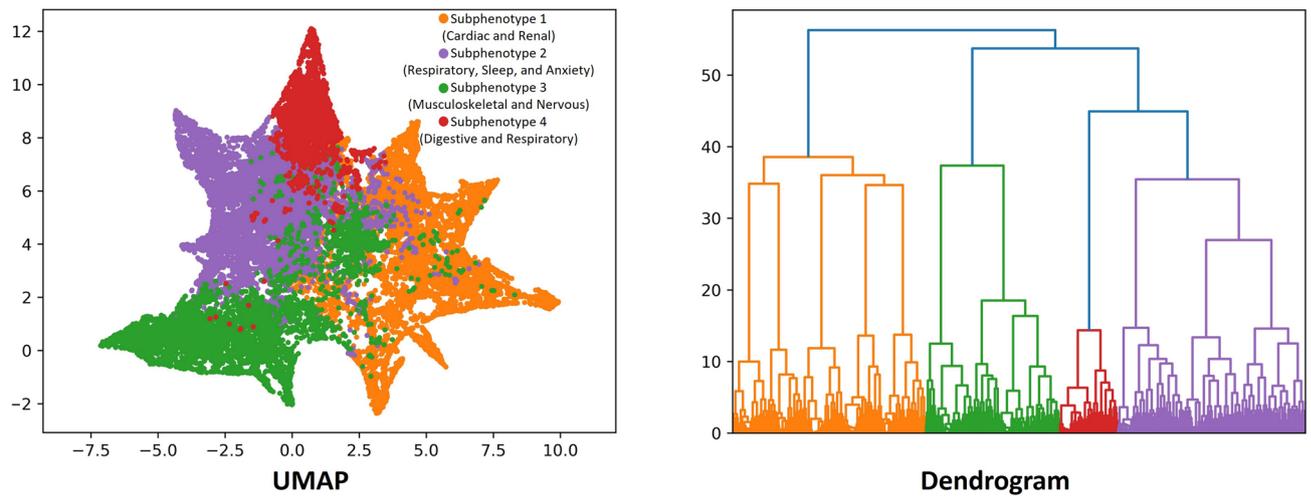
Extended Data Fig. 8 | Network contrast for the OneFlorida+ cohort. Difference of the incidence patterns of selected PASC conditions (grouped by CCSR domains) in 30–180 days after COVID-19 lab test between positive and matched negative patients on the OneFlorida+ cohort. The bubbles in each network correspond 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%.

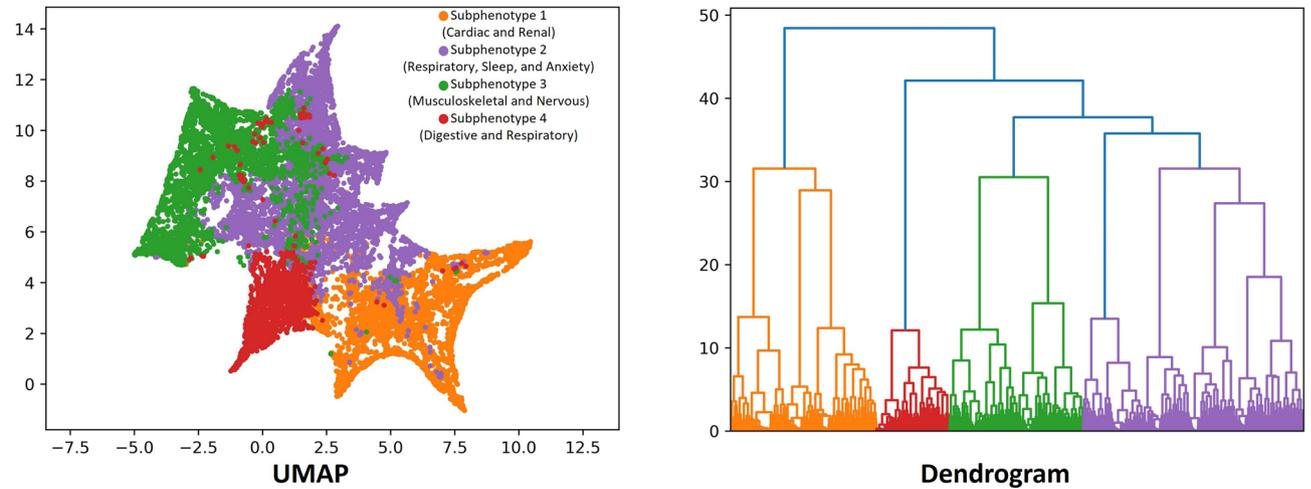


Extended Data Fig. 9 | Inclusion-exclusion cascade for the INSIGHT and OneFlorida+ cohorts. Inclusion-exclusion cascades for the INSIGHT and OneFlorida+ cohorts.

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



Extended Data Fig. 10 | UMAP and dendrogram for the INSIGHT and OneFlorida+ cohorts. Uniform Manifold Approximation and Projection (UMAP) embedding results and dendrograms for the INSIGHT and OneFlorida+ cohorts.

Reporting Summary

Nature Portfolio wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Portfolio policies, see our [Editorial Policies](#) and the [Editorial Policy Checklist](#).

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

- | | |
|-------------------------------------|--|
| n/a | Confirmed |
| <input type="checkbox"/> | <input checked="" type="checkbox"/> The exact sample size (n) for each experimental group/condition, given as a discrete number and unit of measurement |
| <input type="checkbox"/> | <input checked="" type="checkbox"/> A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly |
| <input type="checkbox"/> | <input checked="" type="checkbox"/> The statistical test(s) used AND whether they are one- or two-sided
<i>Only common tests should be described solely by name; describe more complex techniques in the Methods section.</i> |
| <input type="checkbox"/> | <input checked="" type="checkbox"/> A description of all covariates tested |
| <input type="checkbox"/> | <input checked="" type="checkbox"/> A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons |
| <input type="checkbox"/> | <input checked="" type="checkbox"/> A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals) |
| <input type="checkbox"/> | <input checked="" type="checkbox"/> For null hypothesis testing, the test statistic (e.g. F , t , r) with confidence intervals, effect sizes, degrees of freedom and P value noted
<i>Give P values as exact values whenever suitable.</i> |
| <input checked="" type="checkbox"/> | <input type="checkbox"/> For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings |
| <input checked="" type="checkbox"/> | <input type="checkbox"/> For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes |
| <input checked="" type="checkbox"/> | <input type="checkbox"/> Estimates of effect sizes (e.g. Cohen's d , Pearson's r), indicating how they were calculated |

Our web collection on [statistics for biologists](#) contains articles on many of the points above.

Software and code

Policy information about [availability of computer code](#)

Data collection SAS Enterprise Guide Version 7.1 was used to collect the data for this study. ICD-10 diagnosis codes were classified into 137 diagnostic categories based on the Clinical Classification Software Refined (CCSR) version 2021.1 (Healthcare Cost and Utilization Project, Agency for Healthcare Research and Quality).

Data analysis For reproducibility, our codes are available at <https://github.com/haozhangWCM/Subphenotyping-for-PASC>. We used Python 3.7, python package scikit-learn-0.23.2, numpy-1.16.5, umap-learn-0.5.1, and scipy-1.7.3 for machine learning models.

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors and reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Portfolio [guidelines for submitting code & software](#) for further information.

Data

Policy information about [availability of data](#)

All manuscripts must include a [data availability statement](#). This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A description of any restrictions on data availability
- For clinical datasets or third party data, please ensure that the statement adheres to our [policy](#)

The information of the INSIGHT CRN is provided at <https://insightcrn.org/> and INSIGHT data <https://onefloridaconsortium.org/> are made available to researchers

with approved study protocol at <https://nyc-cdrn.atlassian.net/servicedesk/customer/portal/2/group/6/create/16>. The information of the OneFlorida+ CRN is provided at <https://onefloridaconsortium.org/> and OneFlorida+ data are made available to researchers with approved study protocol at <https://onefloridaconsortium.org/front-door/prep-to-research-data-query/>. Questions regarding INSIGHT can email insightcrn@med.cornell.edu. Questions regarding OneFlorida+ can email OneFloridaOperations@health.ufl.edu.

Human research participants

Policy information about [studies involving human research participants and Sex and Gender in Research](#).

Reporting on sex and gender

This is a retrospective secondary analysis of two electronic health record (EHR) cohorts. Summary statistics on sex distributions within different cohorts were reported in Table 1. Summary statistics on sex distributions within different subphenotypes were reported in Table 2.

Population characteristics

Our study included 20,881 patients from the INSIGHT clinical research network (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) 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 in the manuscript summarized the summary statistics of the patients from the two cohorts.

Recruitment

This is a retrospective secondary analysis of EHR data and no patient recruitment activities are involved.

Ethics oversight

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.

Note that full information on the approval of the study protocol must also be provided in the manuscript.

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

Life sciences Behavioural & social sciences Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size

This is a retrospective secondary EHR cohort analysis leveraging two large regional clinical research networks (CRN) within US. To achieve better precision of the study results, we have enrolled all eligible patients from the two cohorts, which included 20,881 patients from the CRN and 13,724 patients from the OneFlorida+ CRN who tested positive for SARS-CoV-2 on viral tests and had at least one new incidental diagnosis in the period of 30-180 days after SARS-CoV-2 infection. To our knowledge, this is the largest and most comprehensive study focusing on general civilian adult population.

Data exclusions

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 Extended Data Figure 2, 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.

Replication

To provide results with better accuracy and precision, we have included all eligible patients in both INSIGHT and OneFlorida+ cohorts. The main analysis was conducted on the INSIGHT cohort and a replication analysis was done on the OneFlorida+ cohort once.

Randomization

This is a retrospective analysis based on clustering and no randomization procedure was involved as no treatment effect was assessed. For the contrast analysis with COVID-19 patients, a similarity based matching procedure was performed for identifying appropriate COVID-19 patients to make fair comparisons. The matching covariates include the following

- 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' neighborhood¹³.
- 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).
- 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. 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.

The detailed description of this procedure was provided in Methods.

Blinding

This is an observational study and thus blinding is not possible.

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

Methods

n/a	Involved in the study
<input checked="" type="checkbox"/>	<input type="checkbox"/> Antibodies
<input checked="" type="checkbox"/>	<input type="checkbox"/> Eukaryotic cell lines
<input checked="" type="checkbox"/>	<input type="checkbox"/> Palaeontology and archaeology
<input checked="" type="checkbox"/>	<input type="checkbox"/> Animals and other organisms
<input type="checkbox"/>	<input checked="" type="checkbox"/> Clinical data
<input checked="" type="checkbox"/>	<input type="checkbox"/> Dual use research of concern

n/a	Involved in the study
<input checked="" type="checkbox"/>	<input type="checkbox"/> ChIP-seq
<input checked="" type="checkbox"/>	<input type="checkbox"/> Flow cytometry
<input checked="" type="checkbox"/>	<input type="checkbox"/> MRI-based neuroimaging

Clinical data

Policy information about [clinical studies](#)

All manuscripts should comply with the ICMJE [guidelines for publication of clinical research](#) and a completed [CONSORT checklist](#) must be included with all submissions.

Clinical trial registration	Not a clinical trial
Study protocol	n/a, not a clinical trial
Data collection	Our study included 20,881 adult patients from the INSIGHT clinical research network (CRN) and 13,724 adult patients from the OneFlorida+ CRN who tested positive for SARS-CoV-2 on viral tests between March 01, 2020 and November 30, 2021
Outcomes	Newly incident diagnosis in the post-acute SARS-CoV-2 infection period (30-180 days after COVID confirmation)