Finding Long-COVID: Temporal Topic Modeling of Electronic Health Records from the N3C and RECOVER Programs

Shawn T. O'Neil[1], Charisse Madlock-Brown[2], Kenneth J. Wilkins[3], Brenda M. McGrath[4], Hannah E. Davis[5], Gina S. Assaf[5], Hannah Wei[5], Parya Zareie[6], Evan T. French[7], Johanna Loomba[8], Julie A. McMurry[1], Andrea Zhou[8], Christopher G. Chute[9], Richard A. Moffitt[10], Emily R Pfaff[11], Yun Jae Yoo[10], Peter Leese[11], Robert F. Chew[12], Michael Lieberman[4, 13], Melissa A. Haendel[1], on behalf of the N3C Consortium and the RECOVER Consortium

1. Department of Biomedical Informatics, University of Colorado Anschutz Medical Campus, Aurora, CO, USA
2. Health Informatics and Information Management Program, University of Tennessee Health Science Center, Memphis, TN, USA
3. Biostatistics Program, National Institute of Diabetes and Digestive and Kidney Diseases, National Institutes of Health, Bethesda, MD, USA
4. OCHIN, Inc. Portland, OR, USA
5. Patient-Led Research Collaborative
6. University of California Davis Health, Sacramento, CA, USA
7. Wright Center for Clinical and Translational Research, Virginia Commonwealth University, Richmond, VA, USA
8. The Integrated Translational Health Research Institute of Virginia (iTHRIV), University of Virginia, Charlottesville, VA, USA
9. Schools of Medicine, Public Health, and Nursing; Johns Hopkins University, Baltimore, MD, USA
10. Department of Hematology and Medical Oncology, Emory University, Atlanta, GA, USA
11. NC TraCS Institute, UNC-School of Medicine, Chapel Hill, NC, USA
12. Center for Data Science and AI, RTI International, Research Triangle Park, NC, USA
13. Department of Medical Informatics and Clinical Epidemiology, Oregon Health and Science University, Portland, OR, USA

Abstract

Post-Acute Sequelae of SARS-CoV-2 infection (PASC), also known as Long-COVID, encompasses a variety of complex and varied outcomes following COVID-19 infection that are still poorly understood. We clustered over 600 million condition diagnoses from 14 million patients available through the National COVID Cohort Collaborative (N3C), generating hundreds of highly detailed clinical phenotypes. Assessing patient clinical trajectories using these clusters allowed us to identify individual conditions and phenotypes strongly increased after acute infection. We found many conditions increased in COVID-19 patients compared to controls, and using a novel method to predict patient/cluster assignment over time, we additionally found phenotypes specific to patient sex, age, wave of infection, and PASC diagnosis status. While many of these results reflect known PASC symptoms, the resolution provided by this

unprecedented data scale suggests avenues for improved diagnostics and mechanistic understanding of this multifaceted disease.

Introduction

The long-term health consequences of SARS-CoV-2 are not fully understood.[1] Research suggests that between 7% and 30% of infected patients experience persistent symptoms,[2] known as Post-Acute Sequelae of SARS-CoV-2 infection (PASC; also known as Long-COVID), affecting multiple organ systems, including pulmonary, cardiovascular, hematological, neurological, and renal systems.[3] The disruption of the host immune response is suspected to play a role in various PASC-associated conditions, including reactivation of dormant persistent infections,[4] autoimmune responses,[5] and multi-inflammatory syndrome of children (MIS-C).[6] Emerging evidence indicates that PASC has clinical sub-phenotypes, comprising clusters of symptoms.[7–10]

Various methods have been used to identify potential PASC sub-phenotypes, including rule-based and machine learning approaches. Reese et al. combined k-means clustering with disease ontology data to identify six clusters of patients representing unique clinical manifestations.[8] Kenny et al. applied multiple correspondence analysis with hierarchical clustering of self-reported symptoms, identifying three clusters of patients.[7] Fischer et al. utilized hierarchical ascendent classification to cluster PASC patients based on select symptoms and demographics, producing three clusters stratified by severity.[11] Bowyer et al. used Latent Class Analysis (LCA) to compare patterns across PASC and non-PASC cohorts using 21 PASC-associated symptoms, revealing two clusters.[12] Overall, these studies suggest that PASC is a complex and heterogeneous condition, and further research is needed to fully understand its clinical sub-phenotypes and implications for patient care.

Topic modeling is a natural language processing technique that aims to identify commonly co-occurring terms or words, called ‘topics,’ in a corpus of documents. The objective is to associate each document in the corpus with one or more topics, based on the distribution of words and their frequency across the corpus.[13] In this study, we treat each patient’s clinical record (or phase of their clinical history, such as pre-infection, acute, or post-acute) as a ‘document,’ and their associated medical codes as ‘terms,’ resulting in topics representing clinical phenotypes. Topic modeling approaches have been previously applied to electronic health record (EHR) data, with Latent Dirichlet Allocation (LDA) or its variants being the most commonly used. Bhattacharya et al. applied LDA to SNOMED medical codes and identified clusters of co-occurring conditions that strongly match co-occurrence of symptoms reported in medical literature.[14] Pivovarov et al.[15] and Mustakim et al.[16] both provide useful reviews of LDA applications to EHR data. LDA is a probabilistic model that represents topics as probability distributions over terms, and documents as probability distributions over topics.

Topic modeling has also been used in various studies of COVID-19 cohorts. For example, Humphreys et al. compared two LDA models trained on sets of ICD10 codes assigned prior to COVID-19 infection and during the last 21 days of life, respectively.[17] Their results identified

pre-existing topics that lead to mortality topics, such as a pre-existing cancer topic leading to a pulmonary mortality topic, indicating an immune-mediated response triggered by chemotherapy.[17] However, due to a small sample size, only six mortality topics were identified, and the application of multiple models complicates comparisons. Scarpino et al. applied LDA to free-text patient narratives, comparing those of PASC and non-PASC patients, and identified three topics that were differentially associated with PASC patients.[18] Zhang et al. used Poisson factor analysis to generate topics based on presence or absence of 137 putative PASC conditions in PASC patients, identifying ten topics as potential disease phenotypes.[19] Huang et al. applied non-negative matrix factorization to EHR data and identified five acute and five post-acute symptom clusters.[20] Additionally, the authors generated a predictive model for PASC indication from select symptoms and evaluated feature importance, similar to our work in Pfaff et al.[21] However, these studies showed similarities and differences in the major features of the clusters, indicating that there is still much to learn about PASC sub-phenotypes.

To gain a better understanding of long-term COVID-19 outcomes, we utilized topic modeling on electronic health record (EHR) data from the National COVID Cohort Collaborative (N3C), consisting of over 14 million patients from 63 clinical sites representing >230 healthcare locations.[22] This approach allowed us to identify condition clusters, or topics, that are specific to PASC-labeled patients and non-PASC labeled patients, compared to COVID-naïve patients over a similar timeframe. We also investigated the potential relationship between key demographics of age and sex, as well as other covariates, such as the wave of infection, and the development of PASC sub-phenotypes. The significant range of potential post-acute sequelae presents a considerable challenge in defining, identifying, and characterizing PASC, and our study provides important insights into this complex and heterogeneous condition.

Methods

This study involved topic modeling and predictive analyses using a dataset of 13,998,246 patient condition diagnoses from the OMOP common data model's Condition Era domain, from 63 contributing sites in the N3C[23] with record dates ranging from Jan. 2018 to Aug. 2022. This dataset includes patients both with and without COVID-19 for comparative purposes. To ensure data quality, we applied a minimal quality filter to the site inclusion criteria (see Suppl. Methods). The patient dataset was randomly split into three sets (see Suppl. Figure 1): *training* for LDA model training (N=8,959,498), *validation* for LDA model validation (N=2,240,842), and *assessment* (N=2,797,906) for downstream statistical analyses. Some N3C-contributing sites reported no U09.9 PASC diagnosis codes, possibly due to lack of implementation in their EHR software. Therefore, the assessment set was limited to patients from the 39 sites with at least one record (N=1,648,012).

Model training and validation

A summary of our modeling efforts is illustrated in Figure 1. We applied the online Latent Dirichlet Allocation (LDA) method of Hoffman et al.,[24] which was implemented in Apache Spark

version 3.2.1[25] to train our topic model. To prevent significant bias towards the COVID-19 condition code and other problematic condition terms, we removed them from the corpus prior to training (Suppl. Methods). This resulted in a corpus of 48,372 unique condition identifiers with an average of 48.4 condition eras recorded per patient.

We employed the UCI Coherence metric[26] on the validation set to determine the number of topics K and measure the model's quality. This metric assesses how frequently the top-weighted terms co-occur in patient records relative to their expected occurrence if they were unrelated (see Suppl. Methods). We found that the mean topic coherence increased from K=150 to K=300, after which coherence began to decline (Suppl. Figure 2). Therefore, we selected K=300 as the final number of topics. We generated all other model and topic summary statistics using both the training and validation sets to ensure completeness.

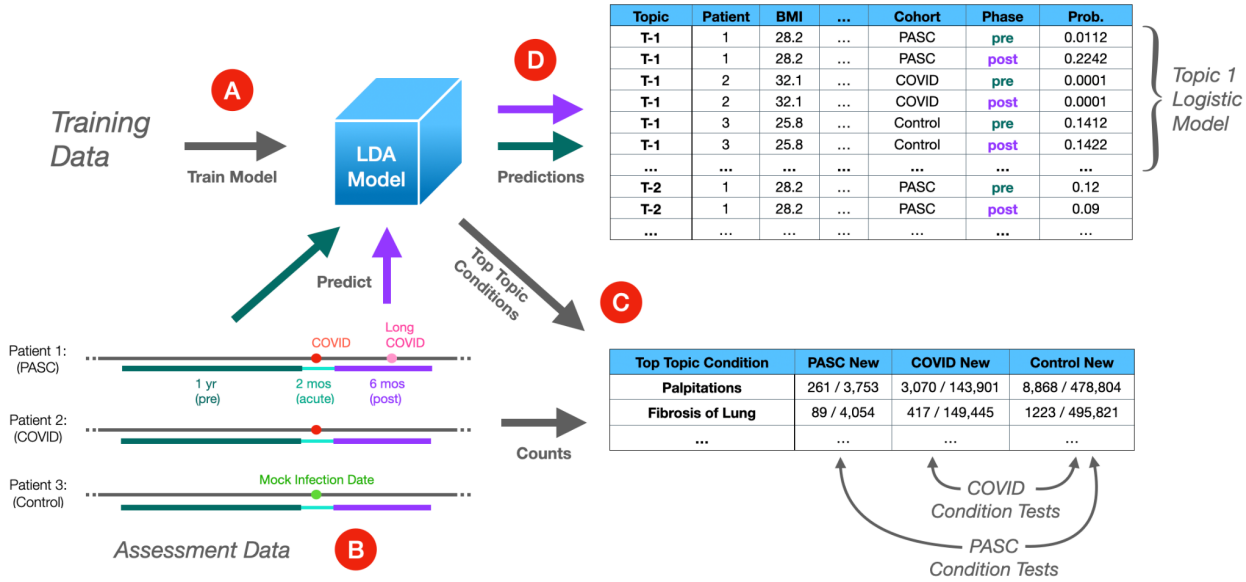


Figure 1: Experimental design summary. (A) The LDA topic model was trained on all training patient data (8.9M patients from 63 sites) using held-out validation data, resulting in the identification of K=300 clusters of co-occurring conditions (see Methods). (B) Held-out assessment patient histories were separated into pre- and post-infection phases. (C) The top 20 conditions with relevance score greater than zero (see Results) were analyzed for new incidence pre-to-post for both COVID and PASC patients compared to Controls. (D) Topic distributions were predicted independently for pre and post phases, and these predictions were modeled per-topic via logistic regression to assess demographic-specific increases in the PASC and COVID cohorts compared to Controls.

Test cohorts and clinical phases

To generate meaningful results, it is necessary to control for normal patient trajectories when studying PASC, which are defined as new-onset, persistent conditions and symptoms occurring after a COVID-19 infection. For instance, adolescents may have a tendency to be diagnosed

with ADHD and related conditions over time,[27] so a topic containing these conditions would be expected to increase over time.

To address this, we divided the ~1.6M assessment patients into three cohorts: PASC, COVID (without PASC), and Control. We excluded all other patients (N=469,325**)** from analysis. We defined an index date delineating pre- and post-infection "phases". Control patients were given a mock index date to allow longitudinal comparisons, while the index dates were uniquely defined for each cohort (see below). In all analyses, we disregarded the period from 15 days prior through 45 days after the index date as a likely acute infection phase. We only considered conditions occurring during what we define as the 'pre' and 'post' phases: 1 year prior to and 6 months after the acute phase, respectively. Patients without this observation history or those without at least 2 weeks of recorded active conditions in both the pre and post phases were excluded from our analyses. Additionally, patients whose only indication of COVID-19 or PASC were one or more visits to a Long-COVID specialty clinic were excluded, as this distinct information was provided by only six sites.

PASC (N=6,481): This cohort includes patients with the U09.9 PASC diagnosis code on or after Oct. 1, 2021, when this code was released, and/or those labeled with B94.8 *Sequelae of other specified infectious and parasitic diseases* prior to this date (the CDC-recommended label until U09.9 became available[28]). For 36.4% of this cohort, we assigned the index date as the date of their first strong COVID-19 indicator (defined below). For another 30.9% patients who do not have a strong COVID-19 indicator, we used the first PASC indicator as the index date. The remaining patients in this cohort had a strong COVID-19 indicator, but it occurred less than 45 days prior to their PASC indicator. In this case, we assumed the COVID-19 indicator date was unreliable and used the PASC indicator as the index date.

COVID (N=340,096): These patients have a confirmed primary COVID-19 infection indicated by a positive SARS-CoV-2 Polymerase Chain Reaction (PCR) or Antigen (Ag) test (Suppl. Table 1), or a U07.1 diagnosis. Their index date was the first of any of these indicators.

Control (N=832,110): Control patients are those without any indication of COVID-19 in their available data, including positive antigen, antibody, or PCR test, COVID-19 or PASC diagnosis (Suppl. Table 2), or a visit to PASC specialty clinic (unique information provided by only six sites). We also excluded patients with a diagnosis of M35.81 *Multisystem inflammatory syndrome* as potential confounders. Control patients were assigned a mock infection date, chosen uniformly at random to simulate pre-acute, acute, and post-acute phases contained entirely within their longest continuous observation period. Mock infection dates were additionally constrained to be after March 1, 2020; patients not meeting these criteria were excluded.

Topic-specific repeated-measures models

Logistic models are useful for assessing incidence rates stratified by demographic factors such as age and gender. However, in condition-specific analyses, the context provided by the topic model is lacking, and statistical power is often insufficient due to the rarity of individual conditions. To address these issues, we utilized an LDA model to assign each patient a probability distribution over topics and to assign patient histories, such as a patient's pre- or post-infection phase, a topic distribution. The generative model assumed by LDA supposes that an assigned topic weight represents the probability that a new condition will be sourced from that topic rather than some other, suggesting a logistic regression model with unobserved binary outcomes but a known positive rate.

We conducted a logistic regression analysis for each topic, including phase (pre or post) as a repeated-measure covariate, as well as patient cohort, sex (Male, Female), race (White, Black or African American, Asian or Pacific Islander, Native Hawaiian or Other Pacific Islander, Other or Unknown), BMI, life stage (Pediatric 0-10, Adolescent 11-18, Adult 19-65, Senior 66+), Quan-based Charleson comorbidity index,[29] site CDM (PCORnet, ACT, OMOP, TrinetX, and OMOP (PedsNet)), and date-based "wave" of infection as covariates. Site CDM represents the Common Data Model used by the contributing N3C site, a known source of data variation.[30] We defined infection wave based on CDC surveillance data,[31] categorizing them as Early (prior to March 1, 2021), Alpha (March 1, 2021 to June 30, 2021), or Delta (July 1, 2021 to Dec. 31, 2021); patients during the Omicron wave were excluded due to limited data across covariates. Patients without complete information were excluded from analyses. The cohorts included 2,859 PASC patients, 89,374 COVID patients, and 303,017 Control patients (Suppl. Figure 1).

Contrast analyses were used to assess differences in the estimated marginal means for different subgroups and phases within each topic regression model. Specifically, we evaluated changes in topic probability between the pre- and post-infection phases for each subgroup, such as females in the PASC cohort compared to females in the Control cohort. This resulted in an estimated odds ratio for the increase or decrease in the propensity for PASC-indicated females to experience conditions from the topic compared to Control females over a 6-month period. We ran similar contrasts for each life stage, gender, and index wave within both the COVID and PASC cohorts compared to the Control group.

To evaluate models' effectiveness, we conducted additional contrasts with the same models to evaluate expected topic propensity for females vs males, and pediatric, adolescent, and senior patients vs adults. Overall, we conducted 22 contrasts for each of the 300 topic regression models, resulting in a total of 6,600 tests that were multiply-corrected using Holm's method.

Results

Topics vary by usage and coherence across sites

Figure 2 displays selected topics as word clouds, illustrating the top conditions by their probability in the topic. The font size of each term in the cloud is proportional to its probability in the topic, while color indicates term *relevance*, with values greater than zero indicating terms more specific to the topic than overall (see Suppl. Methods). For instance, Essential Hypertension is a common and highly weighted condition across many topics, resulting in its low relevance in most topics. Conversely, High-Risk Pregnancy is highly weighted in topic T-2 compared to others, indicating its high relevance to that topic. The topics are named T-1 to T-300 in order of the relative probability mass of patients assigned to the topics. Each topic is annotated with three values: U, representing the relative usage of the topic by total weight assigned to patients in the training set (range 0-100%); H, a measure of how uniformly the topic is used by N3C-contributing sites (range 0-1, with values closer to 0 being site-specific); and C, a measure of each topics' coherence (see below) compared to the mean over all topics (as a z-score). Supplementary materials contain word clouds for all topics (Suppl. Figure 3).

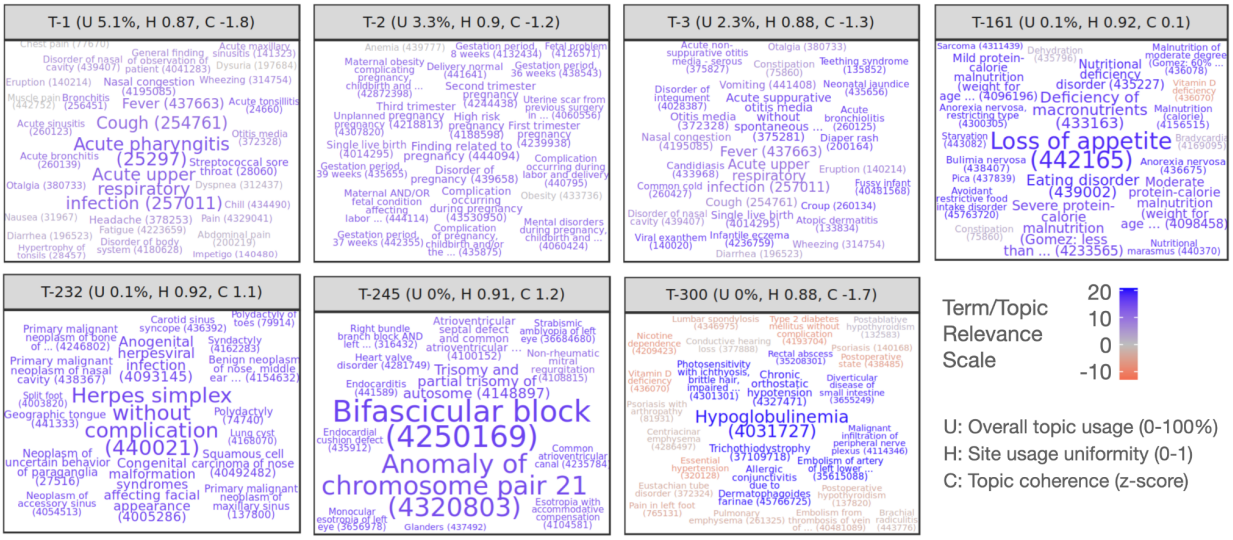


Figure 2: Word clouds illustrating top-weighted terms (conditions) for a subset of topics. *Terms are sized according to probability within their topic and colored according to relevance, with positive relevance indicating terms more probable in the topic than overall. Each term displays the numeric OMOP concept ID encoding the relevant medical code, as well as the first few words of each corresponding OMOP concept name. Per-topic statistics are also shown, summarizing overall usage of each of each topic across sites (U), how uniformly the topic is used across contributing sites (H), and relative topic quality as a normalized coherence score (C).*

Topic coherence is an important measure in addition to model validation. Coherence scores follow a roughly normal distribution (Suppl. Figure 4), and coherence tends to increase with

rarer, more specific topics except for the last 10 topics. Although we chose 300 topics to optimize the overall mean topic coherence, Suppl. Figure 5 shows that mean topic coherence varies by topic and site. All sites exhibit low coherence for the final 10 topics, and most of the final ~35 topics are low coherence for most sites, except for one site that shows high coherence for these rarer topics. Two sites report low coherence for most topics. N3C sites contribute data from one of several source CDMs (Suppl. Figure 5). The source CDM used by the sites generally do not correlate with coherence, except for two sites in the PEDSnet network specializing in pediatric care. These two sites, along with one other using the TriNetX CDM, exhibited distinctive patterns in topic coherence, such as having low coherence and usage for T-153, which pertains to Gout.

We also investigated topic usage per-site (Suppl. Figure 6). Most sites and topics follow similar patterns of usage, with a few notable exceptions. For example, T-4 was used almost exclusively by a single site and has very low coherence with only a few high-relevance terms, although this site uses other topics similarly to other sites. As with coherence, topic usage was generally uncorrelated with the site's CDM, except for the PEDSnet sites and one TriNetX site that again used topics in distinct ways. For instance, T-127 was heavily used by these sites and not others, and pertains to male-pediatric-associated conditions such as *Phimosis* and *Undescended testicle*. We evaluated topics' similarity to other topics using the Jensen-Shannon Distance metric over their term-weight distributions (Suppl. Figure 7). Overall topics had little overlap, with a median distance of 0.82 (range 0.39–0.83). The last 10 topics, T-290 to T-300, form a cluster with increased co-similarity in addition to previously mentioned low coherence and usage.

Topics surface PASC conditions

Topics reveal potential clinical sub-phenotypes in PASC patients by surfacing the top-weighted terms of each topic, which are the most commonly used and representative terms. After selecting the top 20 conditions from each topic with positive relevance scores to identify conditions of potential significance, we evaluated the new-onset incidence of each in the post-infection phase of PASC patients as compared to Controls' mock post-infection. This resulted in 4,794 unique conditions, and 9,588 two-sided Fishers' exact tests for the two cohorts, due to 533 conditions present in the top 20 terms of more than one topic. After multiple test correction (Bonferroni), we identified 213 individual conditions that were significant for the PASC cohort, 208 for the COVID cohort, and 89 for both with $p < 0.05$. The complete list of significant results is available in Suppl. Table 3.

Figure 3 labels a subset of significant conditions identified from the analysis. The PASC cohort showed larger effect sizes compared to controls for most conditions, although several prominent conditions were represented in the COVID cohort as well, such as *Pneumonia caused by SARS-CoV-2*, *Viral pneumonia*, and *Postviral fatigue syndrome.* Additionally, the following conditions had significant estimated odds ratios (ORs) greater than 2 in both cohorts: *Loss of sense of smell*, *Disorder of respiratory system*, *Acute lower respiratory tract infection*, *Upper respiratory tract infection due to Influenza*, *Telogen effluvium*, and *Non-scarring alopecia*.

Figure 3: Over-represented conditions in PASC and COVID patients compared to Controls post-infection. *The x-axis shows estimated odds ratios and the y-axis shows the adjusted p-values for new incidence of top-weighted, positive-relevance terms from all topics amongst COVID (left) and PASC (right) cohorts compared to Controls over a six month post-acute period. Many known PASC-associated conditions increased in both cohorts, while some conditions are cohort-specific. Additionally, in the COVID cohort, incidence of many conditions associated with regular care or screening is reduced compared to controls.*

Several conditions are more common in the PASC cohort, including well-known conditions such as *Chronic fatigue syndrome*, *Malaise*, *Finding related to attentiveness*, *Headache*, *Migraine* (with and without aura), and *Anxiety disorder*. *Neurosis* is also present, but it should be noted that site-labeled source codes for this were almost entirely ICD-10-CM F48.9, *Non-psychotic mental disorder, unspecified* or similar (F48.8 and ICD-9 300.9). Notably, *Impaired cognition* was more common in PASC patients (OR 4.26) but less common in COVID patients (OR 0.53) compared to Controls. Other neurological conditions more common in PASC include *Inflammatory disease of the central nervous system*, *Disorder of autonomic nervous system*, *Polyneuropathy*, *Orthostatic hypotension*, and *Familial dysautonomia* (see Discussion).

The significant results for PASC also highlight a variety of symptoms related to the cardiovascular, pulmonary, and immune systems. Cardiac conditions such as *Tachycardia*, *Palpitations*, *Congestive heart failure*, *Myocarditis*, *Cardiomyopathy*, and *Cardiomegaly* were observed. Pulmonary issues are well represented with *Pulmonary embolism*, *Bronchiectasis*,

Fibrosis of lung, and various generic labels for respiratory failure or disorder. Amongst immunological conditions are *Reactive arthritis triad*, *Elevated C-reactive protein*, *Lymphocytopenia*, *Hypogammaglobulinemia*, *Systemic mast cell disease*, and generic *Immunodeficiency disorder*. In addition, bacterial, viral, and fungal infections were common, including *Bacterial infection due to Pseudomonas*, *Aspergillosis*, and *Pneumocystosis*. Other common themes include musculoskeletal issues (*Fibromyalgia*, *Muscle weakness*, various types of pain) and hematological issues (*Blood coagulation disorder*, *Anemia*, *Hypocalcemia*, *Hypokalemia*).

Out of the 83 conditions significantly increased in the COVID cohort, 36 were not found to be significantly increased in the PASC cohort. Examples include *Intestinal infectious disease*, various forms of *sinusitis*, and a number of female reproductive health conditions such as *Irregular periods*, *Excessive and frequent menstruation*, *Human papillomavirus positive*, and *Acute vaginitis*. *Abnormal menstrual cycle* was significantly increased for both COVID and PASC patients.

The analysis also revealed estimated odds ratios less than 1, indicating decreased incidence, for 219 conditions in one or both cohorts. Most of these conditions (174) were significant only for the COVID cohort, and several related to routine screening or elective procedures potentially disrupted by a COVID-19 infection or lack of care access during the pandemic, such as *Pre-operative state*, *Nicotine dependence*, *Radiological finding*, *Gonarthrosis*, and *Hypertensive disorder*.[32] *Preoperative state* was largely coded as SNOMED CT 72077002 or ICD-10-CM Z01.818, both widely used across sites and indicative of pre-surgical examination. *Unable to Assess Risk* appears to be a custom code used by a single site, mapped to OMOP concept ID 42690761 by N3C. Other conditions may be more difficult to identify in the six months after a COVID-19 infection due to symptom masking or altered care-seeking behavior. Examples include *Diverticulosis of large intestine* and *Esophageal dysphagia*.[33,34] In addition to *Pre-operative state*, five conditions were significantly decreased for PASC patients, all related to late-term pregnancy, while *Third trimester pregnancy* was increased in COVID patients (see Discussion).

Patient demographics predict topic migration

As described in the Methods section, we used LDA models to assign probabilities of generating new conditions for each patient from specific topics. These assignments were interpreted as success rates for logistic regression models. For each topic, we fitted a regression model and used contrast analyses to identify increases or decreases in topic assignment post-infection relative to the control group, for sub-cohorts defined by sex, life-stage, and pandemic wave. The models were subjected to effectiveness tests (see Methods) to ensure their quality, resulting in expected trends specific to gender and life stage across topic (Suppl. Figure 3). For example, T-2 pertains to pregnancy, and the regression model estimated an odds ratio of 45 for female/male, 0.06 for pediatric/adult, 0.2 for adolescent/adult, and 0.03 for senior/adult. Similarly, T-3 has a high weight for neonatal conditions and an estimated pediatric/adult ratio of 43, but no significant female/male trend was observed. While unsurprising, these results

validate regression modeling and interpretation of LDA topic probabilities, themselves dense representations of patient cohort characteristics.

We tested 5,400 sex, life-stage, and wave-specific contrasts for the PASC and COVID cohorts, and 314 were significant after multiple corrections, representing 68 distinct topics. Most of these topics had small effect sizes, with only 30 contrasts across 9 topics having an OR of 2 or higher, and the large majority of strong effects were found for the PASC cohort. Figure 4 illustrates results for the subset of topics with significant odds ratio estimates >2 for more than one demographic group. All effectiveness and contrast results are listed in Suppl. Table 4 and visualized in Suppl. Figure 3.

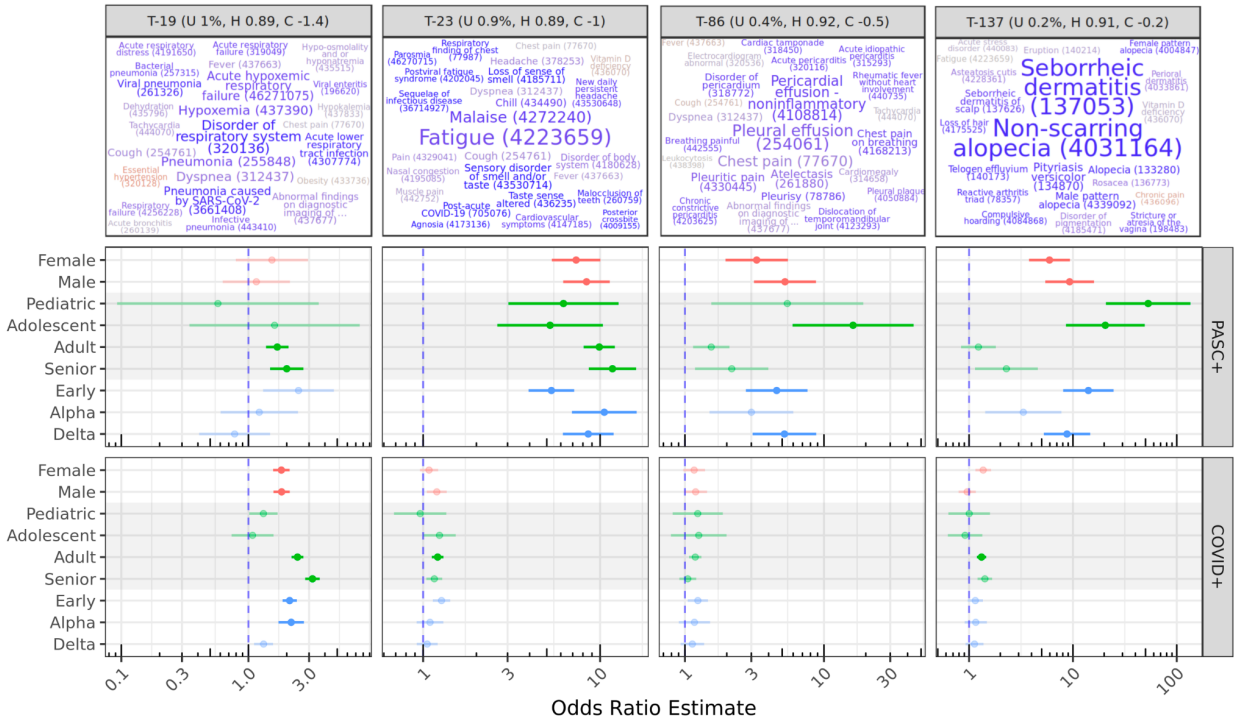


Figure 4: Topics with significant odds ratio estimates >2 for PASC or COVID cohorts, broken out by patient demographics. *The top row shows the topics evaluated for increased probability of condition generation for PASC (middle row) and COVID (bottom row) patients compared to Controls. For example, female PASC patients were found to have a ~7 times higher odds ratio of generating terms from T-23 than Control females over a similar timeframe post-acute-infection. Lines show 95% confidence intervals for estimates; semi-transparent estimates are shown for context but were not significant after multiple-test correction.*

T-23 stands out as a topic with strong migration among PASC patients, with all subgroups having a significant estimated ORs of 5-10 for generating conditions from this topic compared to controls. High-weight, high-relevance conditions in T-23 include *Fatigue*, *Malaise*, *Loss of sense of smell*, and other well-known PASC symptoms, as well as the diagnosis code for PASC itself (Post-acute COVID-19). By contrast, COVID patients do not show statistically significant migration to this topic, with the exception of Adults with a small OR of 1.2.

T-19 shows significant OR estimates for several PASC and COVID groups with similar magnitudes. This topic includes several variants of pneumonia and acute respiratory infection symptoms (*Disorder of respiratory system*, *Dyspnea*, *Hypoxemia*, *Cough*), suggesting significant long-term COVID-19 or secondary infections at least 45 days post-primary-infection. For both PASC and COVID cohorts, these increases are most associated with early-wave infections.

Topics 86 and 137 show increases for several PASC groups, especially pediatric and adolescent patients. While T-86 is characterized by *Pleural* and *Pericardial effusion* and related pain, T-137 describes skin conditions, particularly hair loss, including *Non-scarring alopecia* and *Telogen effluvium*, both identified individually above. While effusion is a known factor for severe COVID-19 pneumonia, especially in older patients[35], these results suggest differential long-term effects. A systematic review of alopecia in COVID-19 patients by Nguyen and Tosti found that *Anagen effluvium* was associated with younger patients compared to other types of alopecia, but few of the reviewed studies included young patients.[36]

Figure 5 displays additional results for selected topics with cohort or demographic-specific patterns. T-8 represents cardiovascular conditions, and shows a mild but significant increase for adult COVID patients compared to controls. T-43 (not shown) is also significant for PASC adult patients, and encompasses pulmonary conditions. Several of the top-weighted conditions within these topics were individually significant, such as *Palpitations*, *Cardiac arrhythmia*, *Chronic obstructive lung disease*, and *Pulmonary emphysema* for both cohorts, and for PASC *Dizziness and giddiness* and *Tachycardia*. While all of these were individually increased in the PASC cohort, *Cardiac arrhythmia, Chronic obstructive lung disease*, and *Pulmonary emphysema* were decreased in the COVID cohort relative to controls.

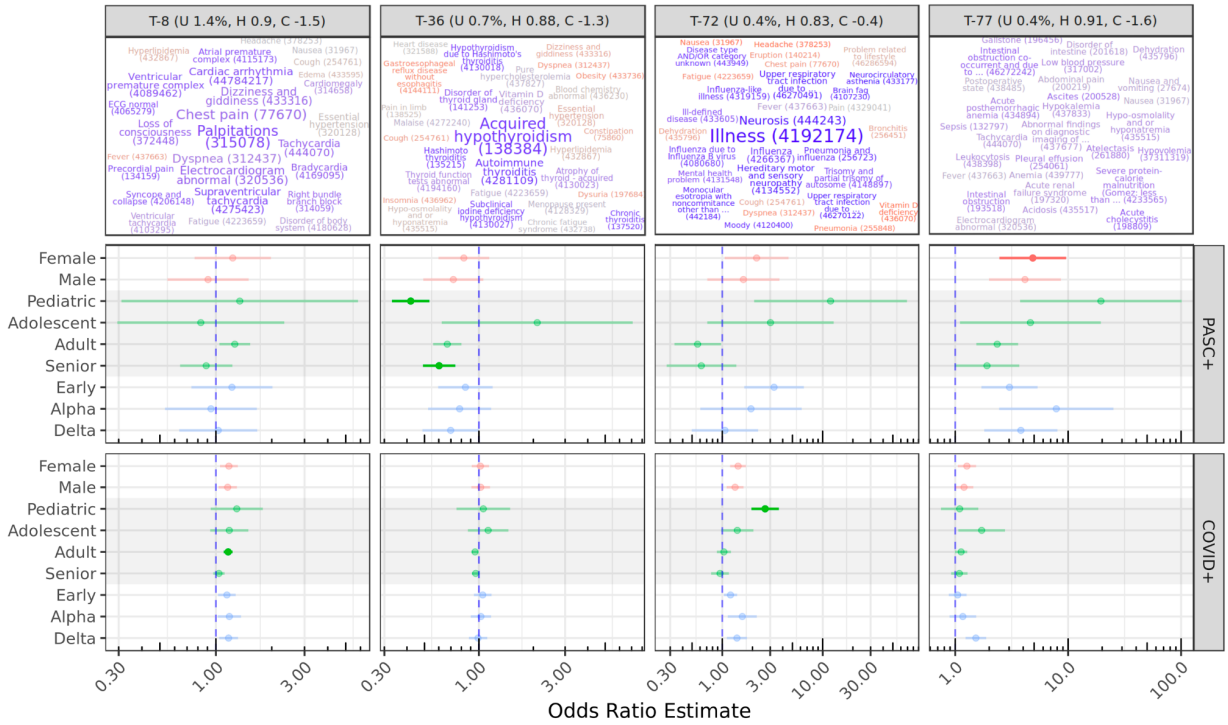


Figure 5: Other select topics with demographic or cohort-specific trends. *T-8 is statistically significant only for COVID adults compared to controls. Topics 72 and 77 include diffuse sets of conditions, while T-36 is reduced for PASC pediatric and senior patients, despite representing known PASC outcomes (see text).*

T-72 shows a significant increase in COVID pediatric patients but not PASC. It covers a range of non-specific PASC-like conditions, including *Illness*, *Neurosis* (also discussed above), *Ill-defined disease*, *Mental health problem*, and *Disease type and/or category unknown*. *Brain fog* and *Neurocirculatory asthenia* are additionally found in this topic. The fact this topic is distinct from T-23 and significant for COVID but not PASC cohorts suggests that these conditions represent a cluster of symptoms and diagnoses separate from PASC.

T-77 is increased in female PASC patients compared to controls. This topic is diffuse and has no particularly highly weighted conditions, although many had high relevance scores to the topic. Several of these are laboratory-based, such as *Hypokalemia*, *Anemia*, and *Hyponatremia*. *Tachycardia*, *Pleural effusion*, *Deficiency of macronutrients*, and *Adult failure to thrive syndrome* are also present. The low specificity and coherence of T-77 make it difficult to interpret, although many of these conditions were individually significant above. T-20 (not shown) was increased for COVID adults and COVID delta-wave patients, and also has few high-weight terms, but relevant conditions include *Acute renal failure syndrome*, *Sepsis*, and *Acidosis*.

T-36 strongly decreased for both pediatric and senior PASC patients, and covers only a few conditions with high weights and relevance scores, including *Acquired hypothyroidism* and *Autoimmune thyroiditis*. This result is paradoxical, as these conditions are common long-term outcomes of COVID-19 infection.[37] Another paradoxical result is a strong (OR 11.7) increase in

T-92 for adolescent PASC patients, which covers a variety of physical contusions, lacerations, and abrasions. The highest-weighted condition in this topic however is *Traumatic and/or non-traumatic injury*, all of which were originally coded as ICD-10 T14.8 *Other injury of unspecified body region* for these patients.

Ignoring the last 10, low quality, high co-similarity topics, adolescent PASC patients are increased in four topics: T-23, T-86, and T-137 already discussed, and T-174 which highly weights *Thyrotoxicosis*, *C-reactive protein abnormal*, and *Polymyalgia rheumatica*. PASC pediatric patients increase significantly in T-23 and T-137 already discussed, as well as T-57 covering a variety of pulmonary issues such as *Chronic cough*, *Bronchiectasis*, and *Hemoptysis*. On the other hand, PASC adolescent patients were reduced in seven topics and PASC pediatric patients showed a reduction in sixteen, covering a broad range of conditions. These assessment cohorts are small, with 49 pediatric and 66 adolescent patients. Chart reviews revealed that they were distributed across 18 and 20 sites, respectively, and had a similar mean number of conditions recorded in the year prior to infection as other cohorts in the same life stages. However, mean condition counts for these PASC patients were nearly 50% higher in the 6-month post-infection phase (Suppl. Table 5).

Discussion

While an ICD-10-CM diagnosis code (U09.9) and specialty clinics exist to treat Long COVID, there is still work to be done identifying PASC conditions and how these new diagnoses and referrals are being used in practice.[21,28] Topic modeling is a data-driven investigation method that is well-suited for exploring new diseases, as an unsupervised machine learning technique. Our model, which was trained on 387 million condition records of 8.9 million patients in the N3C Enclave, is one of the most extensive applications of topic modeling to EHR data to date, generating hundreds of diverse and clinically-relevant topics. Only a handful of topics were of low quality, and those in the middle by usage tended to have the highest coherence scores. We hypothesize that common topics are encumbered by a larger diversity in coding options and practices, while rare topics support only a few relevant conditions on top of more common and unrelated background conditions. We found these trends across our validation tests on models with different topic counts, potentially driven by the use of Dirichlet distributions initialized with sparse uniform priors. Topic usage and coherence varied across contributing sites, with notable patterns of usage at PEDsnet sites in particular. As such, topic modeling may provide insights into site differences in coding practices or data quality, which are issues of concern in federated and centralized data repositories.[30]

Investigating top-weighted topic terms for increased new-onset rates in PASC and COVID cohorts compared to Controls identified many significant conditions, including neurocognitive, cardiovascular, pulmonary, and immune-related ones. Most of these were significant for both cohorts or only the PASC cohort, but the few significant only for COVID patients included *Irregular periods*, *Excessive and frequent menstruation*, *Human papillomavirus positive*, and *Acute vaginitis*. Reproductive issues have been documented with PASC,[38] and these symptoms may thus be under-recognized. A number of conditions had a lower new incidence in COVID

patients compared to Controls, possibly due to decreased access to routine care (as in the case of breast cancer[39]) or behavioral changes (as for diverticulosis[33]) through the pandemic. Similar trends in the PASC cohort were significant only for late-term pregnancy observations (and *Pre-operative state)*, suggesting difficulties in identifying PASC during late pregnancy.

Modeling patient-topic assignment supports queries across patient demographics at a topic level. This approach identified several topics increasing in usage in PASC and COVID sub-cohorts relative to Controls. T-23 stands out as the clearest PASC-related topic across demographics and includes many conditions commonly associated with Long COVID, such as fatigue, malaise, new daily headache, and dyspnea. Other topics are demographic specific, such as T-86 covering *Pleural* and *Pericardial effusion*, T-137 with *Non-scarring alopecia* and *Seborrheic dermatitis*, and T-57 covering other pulmonary issues for younger PASC patients.

While most effects were larger for PASC patients, T-19 shows similar effect sizes for COVID adults and seniors. This topic largely represents secondary pneumonias and related symptoms, suggesting that while these are not used as indicators for PASC, they are nevertheless long-term issues for COVID-19 patients. It was also most significant for the earliest waves of the pandemic, reflecting severity of illness and lack of effective treatment protocols during this period.[40] Few such wave effects were significant overall; T-20 with *Acute renal failure syndrome*, *Acidosis*, and *Sepsis* is an exception showing increases for COVID delta-wave patients.

Despite the few young PASC pediatric patients and wide confidence interval ranges, several topics were significant for this group indicating a unique cohort with significant long-term COVID-19 health outcomes. On the other hand, estimates for COVID-only pediatric patients for most topics, including T-23, T-57, and T-137, are non-significant despite representing many more patients. Long-term COVID-19 symptoms thus appear to be rare amongst young patients, but severe and diverse when they occur.

While this study replicates many known PASC trends, a few individual results are worthy of follow-up. There was a significant increase for female PASC patients in T-77, which is diffuse, multisystem, and covers many conditions identified in other tests. More targeted analyses of this set may reveal a unique phenotype or mix of phenotypes experienced by a unique population. Additionally, T-72 represents a unique cluster of ill-defined conditions; its increase for COVID pediatric patients may reflect difficulties in PASC identification for this group. For example, the highest-weighted term, *Illness*, was originally coded as ICD-10 R69 *Illness, unspecified* in the vast majority of cases. Amongst individual conditions, *Impaired cognition* was one of the few that increased in PASC patients but decreased in COVID patients. Many of these were originally coded as R41.844, *Frontal lobe and executive function disorder*. Executive dysfunction has been linked to COVID-19, particularly for patients with acute respiratory distress syndrome.[41] This diagnosis is distinct from those typical for ADHD (F90), so it is unclear whether the reduction observed in COVID patients was a result of reduced healthcare access.

In contrast to other studies,[42,43] we found few gastrointestinal conditions increased in PASC or COVID patients, though *Abdominal pain, Viral gastroenteritis*, and *Dysphagia* were increased in

PASC patients (Suppl. Table 4). Neither did we find statistically significant sex differences, despite a known increased risk for PASC in female patients. Our experiments, however, evaluate cohorts defined by PASC diagnosis. While female patients are more likely to develop PASC, our results suggest minimal sex differences amongst patients who have been positively identified. Still, other work has suggested sex differences,[44] and similar non-significant trends in our results may be worthy of followup.The apparent reduction in late-term pregnancy conditions for PASC patients and simultaneous increase for COVID patients is notable. We hypothesize that pregnant patients are less likely to be diagnosed with PASC given the similarity of presentation, but more likely to be monitored if infected during pregnancy.

A high incidence of postural orthostatic tachycardia syndrome (POTS) has been identified in PASC clinical research,[45] but a POTS-specific ICD-10 code did not exist prior to October 1, 2022, and therefore POTS is not present in our dataset. The closest available term in the SNOMED hierarchy, *Orthostatic hypotension*, was found to be significantly elevated in PASC, as were *Disorder of the autonomic nervous system* and *Familial dysautonomia*. Many symptoms significant for the PASC cohort, such as *Tachycardia*, *Palpitations*, *Dizziness and giddiness*, *Fatigue*, and *Finding related to attentiveness* are suggestive of POTS or similar forms of dysautonomia. The presence of *Familial dysautonomia* (ICD-10-CM G90.1), a rare genetic disorder, is unlikely to be due to increased screening given that we saw no corresponding uptake in genetic testing. Rather, we suspect that frequent mis-coding may occur because the ICD-10-CM catalog has only one match for the term "dysautonomia" (G90.1 *Familial dysautonomia*), which when used alone encompasses multiple PASC-related conditions.[46] Such errors are not uncommon when using medical record software.[47]

Many of our results are immune-related, including conditions (*Lymphocytopenia*, *Hypogammaglobulinemia*, *Systemic mast cell disease*) and infections more common in immunocompromised patients (*Aspergillosis*, *Pneumocystosis*). Topic 36 highly weighting *Hypothyroidism* and *Thyroiditis* shows reductions for PASC pediatric and senior patients, a paradoxical result given that these are known post-acute sequelae.[37] It may be that patients with pre-existing thyroid disorders are underdiagnosed for PASC, while new thyroid disorders after COVID-19 infection are identified as PASC and related symptoms alone. Together these results suggest an important role for thyroid-mediated dysfunction in PASC patients, and we recommend investigation into how these related diseases are diagnosed and treated.

Limitations of this study should be considered when interpreting results. Our observational EHR data is large, but not a random sample. Indeed, records are sourced from many healthcare organizations across the United States, representing heterogeneity in specialty, coding practices, and other factors. Topic usage and coherence differ across sites; results for topics primarily used by PEDSnet sites for example may be driven by trends at these sites specifically. Unfortunately, our data are too sparse for some sites to fully control for site-level effects. To mitigate these possibilities our statistical analyses excluded patients from sites that did not have any U09.9 PASC diagnoses, and our logistic models include covariates to adjust for site U09.9 usage and relative topic usage. However, to maximize the number and specificity of testable

topics, we trained our model on a broad cohort of patients with and without COVID-19 and PASC, and condition comorbidity within a topic should be interpreted accordingly.

Like predictive models, topic models generally benefit from larger datasets.[48] Topic modeling on a large EHR dataset proves to be highly effective for assessing the progression of post-acute sequelae of SARS-CoV-2 infection. Interpreting the probabilistic assignment of patients to topics through logistic regression is a novel and flexible method, supported by effectiveness tests based on known demographic/topic relationships. Future investigations may consider other covariates such as acute disease severity or vaccination status, contrast other cohorts, consider multiple temporal phases, or investigate migration patterns between topics to reveal sub-phenotype-specific risk factors.

Ultimately, a finer understanding of presentations across populations can inform research, diagnostics, treatment, and health equity for multi-faceted diseases such as PASC. By applying topic modeling to a large-scale EHR dataset we identified hundreds of fine-grain condition clusters, or phenotypes, in the data. Tracking patient clinical trajectories over time in light of these revealed post-acute sequelae of SARS-CoV-2 infection, several of which were associated with patient sex, age, wave of infection, or presence of a PASC diagnosis. Some results, such as those highlighting immune dysfunction, thyroid involvement, and secondary infections improve our understanding of potential mechanisms for PASC. Others, such as those highlighting non-specific phenotypes in the COVID cohort, may lead to improved diagnostics and support for patients suffering from Long-COVID but yet to receive a PASC diagnosis.

Author Contributions

Concept and modeling: S.O., C.M. Statistical design: S.O., K.W., B.M., J.L. Results interpretation: S.O., H.D., G.A., H.W., M.L., M.H. Data curation: S.O., P.Z., E.F., J.L., A.Z., R.M., E.P., Y.Y., P.L., R.C. Manuscript drafting: S.O., C.M., B.M., J.M., M.H. Material and administrative support: J.M., M.H.

Competing Interests

The authors declare no competing interests.

Data Sharing

The N3C Data Enclave is managed under the authority of the NIH; information can be found at https://ncats.nih.gov/n3c/resources. The N3C data transfer to NCATS is performed under a Johns Hopkins University Reliance Protocol # IRB00249128 or individual site agreements with NIH. Enclave data is protected, and can be accessed for COVID-related research with an approved (1) IRB protocol and (2) Data Use Request (DUR). Enclave and data access instructions can be found at https://covid.cd2h.org/for-researchers.

Acknowledgements

The analyses described in this publication were conducted with data and tools accessed through the NCATS N3C Data Enclave https://covid.cd2h.org and N3C Attribution & Publication Policy v 1.2-2020-08-25b supported by NCATS U24 TR002306, Axle Informatics Subcontract: NCATS-P00438-B, National Institutes of Health grant NHLBI RECOVER Agreement OT2HL161847-01, and CTSA award No. UM1TR004360 from the National Center for Advancing Translational Sciences. This research was possible because of the patients whose information is included within the data and the organizations (https://ncats.nih.gov/n3c/resources/data-contribution/data-transfer-agreement-signatories) and scientists who have contributed to the on-going development of this community resource [https://doi.org/10.1093/jamia/ocaa196]. This study is part of the NIH Researching COVID to Enhance Recovery (RECOVER) Initiative (https://recovercovid.org/), which seeks to understand, treat, and prevent the post-acute sequelae of SARS-CoV-2 infection (PASC), and was conducted under the N3C DUR RP-5677B5. We would like to thank the National Community Engagement Group (NCEG), all patient, caregiver and community Representatives, and all the participants enrolled in the RECOVER Initiative. Authorship was determined using ICMJE recommendations.

The N3C Publication committee confirmed that this manuscript is in accordance with N3C data use and attribution policies; however, this content is solely the responsibility of the authors and does not necessarily represent the official views of the RECOVER or N3C programs or the National Institutes of Health.

Use of the N3C data for this study is authorized under the following IRB Protocols:

Site	IRB Name	Exempted vs Approved	Protocol Number
University of Colorado	Colorado Multiple Institutional Review Board	approved	21-2759
Johns Hopkins University	Johns Hopkins Office of Human Subjects Research - Institutional	approved	IRB00249128

	Review Board		
University of North Carolina	University of North Carolina Chapel Hill Institutional Review Board	exempt	21-0309
Stony Brook University	Office of Research Compliance, Division of Human Subject Protections, Stony Brook University	exempt	IRB2021-00098
OCHIN	Advarra IRB	approved	Pro00060719
RTI International	RTI Office of Research Protection	exempt	MOD10001700

We gratefully acknowledge the following core contributors to N3C:
Adam B. Wilcox, Adam M. Lee, Alexis Graves, Alfred (Jerrod) Anzalone, Amin Manna, Amit Saha, Amy Olex, Andrea Zhou, Andrew E. Williams, Andrew Southerland, Andrew T. Girvin, Anita Walden, Anjali A. Sharathkumar, Benjamin Amor, Benjamin Bates, Brian Hendricks, Brijesh Patel, Caleb Alexander, Carolyn Bramante, Cavin Ward-Caviness, Charisse Madlock-Brown, Christine Suver, Christopher Chute, Christopher Dillon, Chunlei Wu, Clare Schmitt, Cliff Takemoto, Dan Housman, Davera Gabriel, David A. Eichmann, Diego Mazzotti, Don Brown, Eilis Boudreau, Elaine Hill, Elizabeth Zampino, Emily Carlson Marti, Emily R. Pfaff, Evan French, Farrukh M Koraishy, Federico Mariona, Fred Prior, George Sokos, Greg Martin, Harold Lehmann, Heidi Spratt, Hemalkumar Mehta, Hongfang Liu, Hythem Sidky, J.W. Awori Hayanga, Jami Pincavitch, Jaylyn Clark, Jeremy Richard Harper, Jessica Islam, Jin Ge, Joel Gagnier, Joel H. Saltz, Joel Saltz, Johanna Loomba, John Buse, Jomol Mathew, Joni L. Rutter, Julie A. McMurry, Justin Guinney, Justin Starren, Karen Crowley, Katie Rebecca Bradwell, Kellie M. Walters, Ken Wilkins, Kenneth R. Gersing, Kenrick Dwain Cato, Kimberly Murray, Kristin Kostka, Lavance Northington, Lee Allan Pyles, Leonie Misquitta, Lesley Cottrell, Lili Portilla, Mariam Deacy, Mark M. Bissell, Marshall Clark, Mary Emmett, Mary Morrison Saltz, Matvey B. Palchuk, Melissa A. Haendel, Meredith Adams, Meredith Temple-O'Connor, Michael G. Kurilla, Michele Morris, Nabeel Qureshi, Nasia Safdar, Nicole Garbarini, Noha Sharafeldin, Ofer Sadan, Patricia A. Francis, Penny Wung Burgoon, Peter Robinson, Philip R.O. Payne, Rafael Fuentes, Randeep Jawa, Rebecca Erwin-Cohen, Rena Patel, Richard A. Moffitt, Richard L. Zhu, Rishi Kamaleswaran, Robert Hurley, Robert T. Miller, Saiju Pyarajan, Sam G. Michael, Samuel Bozzette, Sandeep Mallipattu, Satyanarayana Vedula, Scott Chapman, Shawn T. O'Neil, Soko Setoguchi, Stephanie S. Hong, Steve Johnson, Tellen D. Bennett, Tiffany Callahan, Umit Topaloglu, Usman Sheikh, Valery Gordon, Vignesh Subbian, Warren A. Kibbe, Wenndy Hernandez, Will Beasley, Will Cooper, William Hillegass, Xiaohan Tanner Zhang. Details of contributions available at covid.cd2h.org/core-contributors

The following institutions whose data is released or pending:
Available: Advocate Health Care Network — UL1TR002389: The Institute for Translational Medicine (ITM) • Boston University Medical Campus — UL1TR001430: Boston University

Clinical and Translational Science Institute • Brown University — U54GM115677: Advance Clinical Translational Research (Advance-CTR) • Carilion Clinic — UL1TR003015: iTHRIV Integrated Translational health Research Institute of Virginia • Charleston Area Medical Center — U54GM104942: West Virginia Clinical and Translational Science Institute (WVCTSI) • Children's Hospital Colorado — UL1TR002535: Colorado Clinical and Translational Sciences Institute • Columbia University Irving Medical Center — UL1TR001873: Irving Institute for Clinical and Translational Research • Duke University — UL1TR002553: Duke Clinical and Translational Science Institute • George Washington Children's Research Institute — UL1TR001876: Clinical and Translational Science Institute at Children's National (CTSA-CN) • George Washington University — UL1TR001876: Clinical and Translational Science Institute at Children's National (CTSA-CN) • Indiana University School of Medicine — UL1TR002529: Indiana Clinical and Translational Science Institute • Johns Hopkins University — UL1TR003098: Johns Hopkins Institute for Clinical and Translational Research • Loyola Medicine — Loyola University Medical Center • Loyola University Medical Center — UL1TR002389: The Institute for Translational Medicine (ITM) • Maine Medical Center — U54GM115516: Northern New England Clinical & Translational Research (NNE-CTR) Network • Massachusetts General Brigham — UL1TR002541: Harvard Catalyst • Mayo Clinic Rochester — UL1TR002377: Mayo Clinic Center for Clinical and Translational Science (CCaTS) • Medical University of South Carolina — UL1TR001450: South Carolina Clinical & Translational Research Institute (SCTR) • Montefiore Medical Center — UL1TR002556: Institute for Clinical and Translational Research at Einstein and Montefiore • Nemours — U54GM104941: Delaware CTR ACCEL Program • NorthShore University HealthSystem — UL1TR002389: The Institute for Translational Medicine (ITM) • Northwestern University at Chicago — UL1TR001422: Northwestern University Clinical and Translational Science Institute (NUCATS) • OCHIN — INV-018455: Bill and Melinda Gates Foundation grant to Sage Bionetworks • Oregon Health & Science University — UL1TR002369: Oregon Clinical and Translational Research Institute • Penn State Health Milton S. Hershey Medical Center — UL1TR002014: Penn State Clinical and Translational Science Institute • Rush University Medical Center — UL1TR002389: The Institute for Translational Medicine (ITM) • Rutgers, The State University of New Jersey — UL1TR003017: New Jersey Alliance for Clinical and Translational Science • Stony Brook University — U24TR002306 • The Ohio State University — UL1TR002733: Center for Clinical and Translational Science • The State University of New York at Buffalo — UL1TR001412: Clinical and Translational Science Institute • The University of Chicago — UL1TR002389: The Institute for Translational Medicine (ITM) • The University of Iowa — UL1TR002537: Institute for Clinical and Translational Science • The University of Miami Leonard M. Miller School of Medicine — UL1TR002736: University of Miami Clinical and Translational Science Institute • The University of Michigan at Ann Arbor — UL1TR002240: Michigan Institute for Clinical and Health Research • The University of Texas Health Science Center at Houston — UL1TR003167: Center for Clinical and Translational Sciences (CCTS) • The University of Texas Medical Branch at Galveston — UL1TR001439: The Institute for Translational Sciences • The University of Utah — UL1TR002538: Uhealth Center for Clinical and Translational Science • Tufts Medical Center — UL1TR002544: Tufts Clinical and Translational Science Institute • Tulane University — UL1TR003096: Center for Clinical and Translational Science • University Medical Center New Orleans — U54GM104940: Louisiana Clinical and Translational Science (LA CaTS) Center •

University of Alabama at Birmingham — UL1TR003096: Center for Clinical and Translational Science • University of Arkansas for Medical Sciences — UL1TR003107: UAMS Translational Research Institute • University of Cincinnati — UL1TR001425: Center for Clinical and Translational Science and Training • University of Colorado Denver, Anschutz Medical Campus — UL1TR002535: Colorado Clinical and Translational Sciences Institute • University of Illinois at Chicago — UL1TR002003: UIC Center for Clinical and Translational Science • University of Kansas Medical Center — UL1TR002366: Frontiers: University of Kansas Clinical and Translational Science Institute • University of Kentucky — UL1TR001998: UK Center for Clinical and Translational Science • University of Massachusetts Medical School Worcester — UL1TR001453: The UMass Center for Clinical and Translational Science (UMCCTS) • University of Minnesota — UL1TR002494: Clinical and Translational Science Institute • University of Mississippi Medical Center — U54GM115428: Mississippi Center for Clinical and Translational Research (CCTR) • University of Nebraska Medical Center — U54GM115458: Great Plains IDeA-Clinical & Translational Research • University of North Carolina at Chapel Hill — UL1TR002489: North Carolina Translational and Clinical Science Institute • University of Oklahoma Health Sciences Center — U54GM104938: Oklahoma Clinical and Translational Science Institute (OCTSI) • University of Rochester — UL1TR002001: UR Clinical & Translational Science Institute • University of Southern California — UL1TR001855: The Southern California Clinical and Translational Science Institute (SC CTSI) • University of Vermont — U54GM115516: Northern New England Clinical & Translational Research (NNE-CTR) Network • University of Virginia — UL1TR003015: iTHRIV Integrated Translational health Research Institute of Virginia • University of Washington — UL1TR002319: Institute of Translational Health Sciences • University of Wisconsin-Madison — UL1TR002373: UW Institute for Clinical and Translational Research • Vanderbilt University Medical Center — UL1TR002243: Vanderbilt Institute for Clinical and Translational Research • Virginia Commonwealth University — UL1TR002649: C. Kenneth and Dianne Wright Center for Clinical and Translational Research • Wake Forest University Health Sciences — UL1TR001420: Wake Forest Clinical and Translational Science Institute • Washington University in St. Louis — UL1TR002345: Institute of Clinical and Translational Sciences • Weill Medical College of Cornell University — UL1TR002384: Weill Cornell Medicine Clinical and Translational Science Center • West Virginia University — U54GM104942: West Virginia Clinical and Translational Science Institute (WVCTSI)

Submitted: Icahn School of Medicine at Mount Sinai — UL1TR001433: ConduITS Institute for Translational Sciences • The University of Texas Health Science Center at Tyler — UL1TR003167: Center for Clinical and Translational Sciences (CCTS) • University of California, Davis — UL1TR001860: UCDavis Health Clinical and Translational Science Center • University of California, Irvine — UL1TR001414: The UC Irvine Institute for Clinical and Translational Science (ICTS) • University of California, Los Angeles — UL1TR001881: UCLA Clinical Translational Science Institute • University of California, San Diego — UL1TR001442: Altman Clinical and Translational Research Institute • University of California, San Francisco — UL1TR001872: UCSF Clinical and Translational Science Institute

Pending: Arkansas Children's Hospital — UL1TR003107: UAMS Translational Research Institute • Baylor College of Medicine — None (Voluntary) • Children's Hospital of Philadelphia — UL1TR001878: Institute for Translational Medicine and Therapeutics • Cincinnati Children's

Hospital Medical Center — UL1TR001425: Center for Clinical and Translational Science and Training • Emory University — UL1TR002378: Georgia Clinical and Translational Science Alliance • HonorHealth — None (Voluntary) • Loyola University Chicago — UL1TR002389: The Institute for Translational Medicine (ITM) • Medical College of Wisconsin — UL1TR001436: Clinical and Translational Science Institute of Southeast Wisconsin • MedStar Health Research Institute — UL1TR001409: The Georgetown-Howard Universities Center for Clinical and Translational Science (GHUCCTS) • MetroHealth — None (Voluntary) • Montana State University — U54GM115371: American Indian/Alaska Native CTR • NYU Langone Medical Center — UL1TR001445: Langone Health's Clinical and Translational Science Institute • Ochsner Medical Center — U54GM104940: Louisiana Clinical and Translational Science (LA CaTS) Center • Regenstrief Institute — UL1TR002529: Indiana Clinical and Translational Science Institute • Sanford Research — None (Voluntary) • Stanford University — UL1TR003142: Spectrum: The Stanford Center for Clinical and Translational Research and Education • The Rockefeller University — UL1TR001866: Center for Clinical and Translational Science • The Scripps Research Institute — UL1TR002550: Scripps Research Translational Institute • University of Florida — UL1TR001427: UF Clinical and Translational Science Institute • University of New Mexico Health Sciences Center — UL1TR001449: University of New Mexico Clinical and Translational Science Center • University of Texas Health Science Center at San Antonio — UL1TR002645: Institute for Integration of Medicine and Science • Yale New Haven Hospital — UL1TR001863: Yale Center for Clinical Investigation

References

1. Brüssow, H. & Timmis, K. COVID-19: long covid and its societal consequences. *Environ. Microbiol.* **23**, 4077–4091 (2021).
2. Reardon, S. Long COVID risk falls only slightly after vaccination, huge study shows. *Nature Publishing Group UK* http://dx.doi.org/10.1038/d41586-022-01453-0 (2022) doi:10.1038/d41586-022-01453-0.
3. Nalbandian, A. *et al.* Post-acute COVID-19 syndrome. *Nat. Med.* **27**, 601–615 (2021).
4. Proal, A. D. & VanElzakker, M. B. Long COVID or Post-acute Sequelae of COVID-19 (PASC): An Overview of Biological Factors That May Contribute to Persistent Symptoms. *Frontiers in Microbiology* vol. 12 Preprint at https://doi.org/10.3389/fmicb.2021.698169 (2021).
5. Knight, J. S. *et al.* The intersection of COVID-19 and autoimmunity. *J. Clin. Invest.* **131**, (12 2021).
6. Hageman, J. R. Long COVID-19 or post-acute sequelae of SARS-CoV-2 infection in children, adolescents, and young adults. *Pediatr. Ann.* (2021).
7. 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).
8. Reese, J. T. *et al.* Generalizable Long COVID Subtypes: Findings from the NIH N3C and RECOVER Programs. *medRxiv* (2022) doi:10.1101/2022.05.24.22275398.
9. Ståhlberg, M. *et al.* Post-COVID-19 Tachycardia Syndrome: A Distinct Phenotype of Post-Acute COVID-19 Syndrome. *Am. J. Med.* **134**, 1451–1456 (2021).
10. Durstenfeld, M. S., Hsue, P. Y., Peluso, M. J. & Deeks, S. G. Findings from mayo clinic's post-COVID clinic: PASC phenotypes vary by sex and degree of IL-6 elevation. *Mayo Clin. Proc.* **97**, 430–432 (2022).

11. Fischer, A. *et al.* Long COVID Classification: Findings from a Clustering Analysis in the Predi-COVID Cohort Study. *Int. J. Environ. Res. Public Health* **19**, (2022).
12. Bowyer, R. C. E. *et al.* Characterising patterns of COVID-19 and long COVID symptoms: evidence from nine UK longitudinal studies. *Eur. J. Epidemiol.* **38**, 199–210 (2023).
13. Liu, L., Tang, L., Dong, W., Yao, S. & Zhou, W. An overview of topic modeling and its current applications in bioinformatics. *Springerplus* **5**, 1608 (2016).
14. Bhattacharya, M., Jurkovitz, C. & Shatkay, H. Co-occurrence of medical conditions: Exposing patterns through probabilistic topic modeling of snomed codes. *J. Biomed. Inform.* (2018).
15. Pivovarov, R. *et al.* Learning probabilistic phenotypes from heterogeneous EHR data. *J. Biomed. Inform.* **58**, 156–165 (2015).
16. Mustakim, M., Wardoyo, R., Mustofa, K., Rahayu, G. R. & Rosyidah, I. Latent Dirichlet Allocation for Medical Records Topic Modeling: Systematic Literature Review. in *2021 Sixth International Conference on Informatics and Computing (ICIC)* 1–7 (2021).
17. Humpherys, J. *et al.* Topic-to-Topic Modeling for COVID-19 Mortality. in *2021 IEEE 9th International Conference on Healthcare Informatics (ICHI)* 258–264 (2021).
18. Scarpino, I., Zucco, C., Vallelunga, R., Luzza, F. & Cannataro, M. Investigating Topic Modeling Techniques to Extract Meaningful Insights in Italian Long COVID Narration. *BioTech (Basel)* **11**, (2022).
19. Zhang, H. *et al.* Data-driven identification of post-acute SARS-CoV-2 infection subphenotypes. *Nat. Med.* (2022) doi:10.1038/s41591-022-02116-3.
20. Huang, Y. *et al.* COVID Symptoms, Symptom Clusters, and Predictors for Becoming a Long-Hauler Looking for Clarity in the Haze of the Pandemic. *Clin. Nurs. Res.* **31**, 1390–1398 (2022).
21. Pfaff, E. R. *et al.* Identifying who has long COVID in the USA: a machine learning approach using N3C data. *Lancet Digit Health* **4**, e532–e541 (2022).

22. Haendel, M. A. *et al.* The National COVID Cohort Collaborative (N3C): Rationale, design, infrastructure, and deployment. *J. Am. Med. Inform. Assoc.* **28**, 427–443 (2021).
23. *N3C Phenotype 4.0*. (Github).
24. Hoffman, M. D., Blei, D. M. & Bach, F. Online learning for Latent Dirichlet Allocation. https://papers.nips.cc/paper/2010/file/71f6278d140af599e06ad9bf1ba03cb0-Paper.pdf.
25. Meng, X. *et al.* MLlib: Machine Learning in Apache Spark. *arXiv [cs.LG]* (2015).
26. 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 (Association for Computational Linguistics, 2010).
27. Danielson, M. L. *et al.* Prevalence of Parent-Reported ADHD Diagnosis and Associated Treatment Among U.S. Children and Adolescents, 2016. *J. Clin. Child Adolesc. Psychol.* **47**, 199–212 (2018).
28. Pfaff, E. R. *et al.* Coding Long COVID: Characterizing a new disease through an ICD-10 lens. *medRxiv* (2022) doi:10.1101/2022.04.18.22273968.
29. Quan, H. *et al.* Updating and validating the Charlson comorbidity index and score for risk adjustment in hospital discharge abstracts using data from 6 countries. *Am. J. Epidemiol.* **173**, 676–682 (2011).
30. Pfaff, E. R. *et al.* Synergies between centralized and federated approaches to data quality: a report from the national COVID cohort collaborative. *J. Am. Med. Inform. Assoc.* **29**, 609–618 (2022).
31. Lambrou, A. S. *et al.* Genomic Surveillance for SARS-CoV-2 Variants: Predominance of the Delta (B.1.617.2) and Omicron (B.1.1.529) Variants - United States, June 2021-January 2022. *MMWR Morb. Mortal. Wkly. Rep.* **71**, 206–211 (2022).
32. Sisó-Almirall, A., Kostov, B., Sánchez, E., Benavent-Àreu, J. & González de Paz, L. Impact of the COVID-19 Pandemic on Primary Health Care Disease Incidence Rates: 2017 to

2020. *Ann. Fam. Med.* **20**, 63–68 (2022).

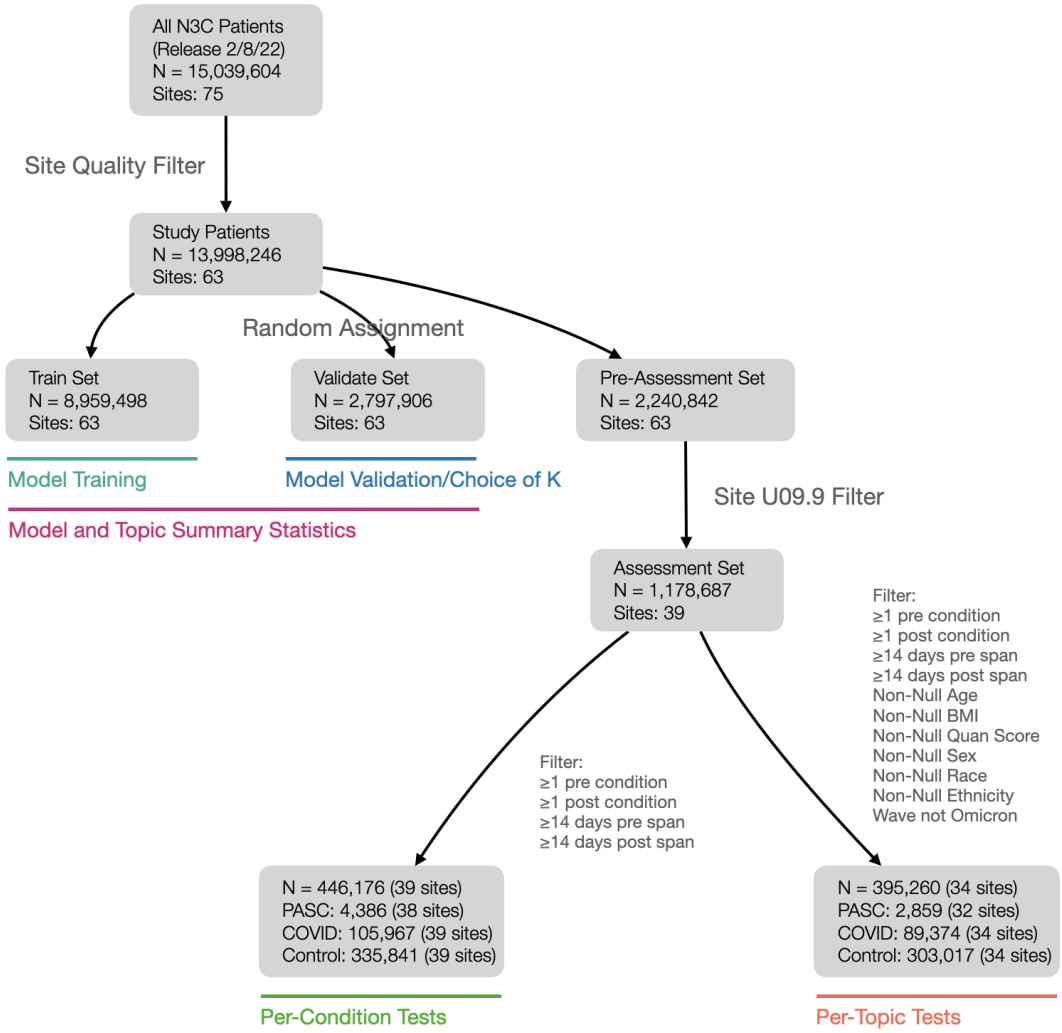
33. Pj, W., Tv, V. & Whiley, P. J. The impact of delayed acute diverticulitis presentations during the COVID-19 pandemic on acuity and surgical complexity in the long-term. *Glob. Surg.* (2022) doi:10.15761/GOS.1000239.
34. Miles, A. *et al.* An International Commentary on Dysphagia and Dysphonia During the COVID-19 Pandemic. *Dysphagia* **37**, 1349–1374 (2022).
35. Li, K. *et al.* The Clinical and Chest CT Features Associated With Severe and Critical COVID-19 Pneumonia. *Invest. Radiol.* **55**, 327–331 (2020).
36. Nguyen, B. & Tosti, A. Alopecia in patients with COVID-19: A systematic review and meta-analysis. *JAAD Int* **7**, 67–77 (2022).
37. Naguib, R. Potential relationships between COVID-19 and the thyroid gland: an update. *J. Int. Med. Res.* **50**, 3000605221082898 (2022).
38. Davis, H. E., McCorkell, L., Vogel, J. M. & Topol, E. J. Long COVID: major findings, mechanisms and recommendations. *Nat. Rev. Microbiol.* **21**, 133–146 (2023).
39. Lowry, K. P. *et al.* Breast Biopsy Recommendations and Breast Cancers Diagnosed during the COVID-19 Pandemic. *Radiology* **303**, 287–294 (2022).
40. Kuriakose, S. *et al.* Developing Treatment Guidelines During a Pandemic Health Crisis: Lessons Learned From COVID-19. *Ann. Intern. Med.* **174**, 1151–1158 (2021).
41. Ali Awan, H. *et al.* SARS-CoV-2 and the Brain: What Do We Know about the Causality of 'Cognitive COVID? *J. Clin. Med. Res.* **10**, (2021).
42. Norouzi Masir, M. & Shirvaliloo, M. Symptomatology and microbiology of the gastrointestinal tract in post-COVID conditions. *JGH Open* **6**, 667–676 (2022).
43. Gupta, A. *et al.* Extrapulmonary manifestations of COVID-19. *Nat. Med.* **26**, 1017–1032 (2020).
44. Sylvester, S. V. *et al.* Sex differences in sequelae from COVID-19 infection and in long COVID syndrome: a review. *Curr. Med. Res. Opin.* **38**, 1391–1399 (2022).

45. Seeley, M.-C. *et al.* High Incidence of Autonomic Dysfunction and Postural Orthostatic Tachycardia Syndrome in Patients with Long COVID: Implications for Management and Health Care Planning. *Am. J. Med.* (2023) doi:10.1016/j.amjmed.2023.06.010.
46. Fedorowski, A. & Sutton, R. Autonomic dysfunction and postural orthostatic tachycardia syndrome in post-acute COVID-19 syndrome. *Nat. Rev. Cardiol.* **20**, 281–282 (2023).
47. Bologva, E. V., Prokusheva, D. I., Krikunov, A. V., Zvartau, N. E. & Kovalchuk, S. V. Human-Computer Interaction in Electronic Medical Records: From the Perspectives of Physicians and Data Scientists. *Procedia Comput. Sci.* **100**, 915–920 (2016).
48. Crossley, S., Dascalu, M. & McNamara, D. How Important Is Size? An Investigation of Corpus Size and Meaning in Both Latent Semantic Analysis and Latent Dirichlet Allocation. in *The Thirtieth International Flairs Conference* (2017).
49. OMOP CDM v5.3. https://ohdsi.github.io/CommonDataModel/cdm53.html.
50. Mei, Q., Shen, X. & Zhai, C. Automatic labeling of multinomial topic models. in *Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining* 490–499 (Association for Computing Machinery, 2007).
51. R Core Team. R: A Language and Environment for Statistical Computing. https://www.R-project.org/ (2020).
52. Patefield, W. M. Algorithm AS 159: An Efficient Method of Generating Random R × C Tables with Given Row and Column Totals. *J. R. Stat. Soc. Ser. C Appl. Stat.* **30**, 91–97 (1981).
53. Højsgaard, S., Halekoh, U. & Yan, J. The R Package geepack for Generalized Estimating Equations. *J. Stat. Softw.* **15**, 1–11 (2006).
54. Lenth, R., Singmann, H., Love, J., Buerkner, P. & Herve, M. Emmeans: Estimated marginal means, aka least-squares means. *R package version* (2018).

Supplemental Figures

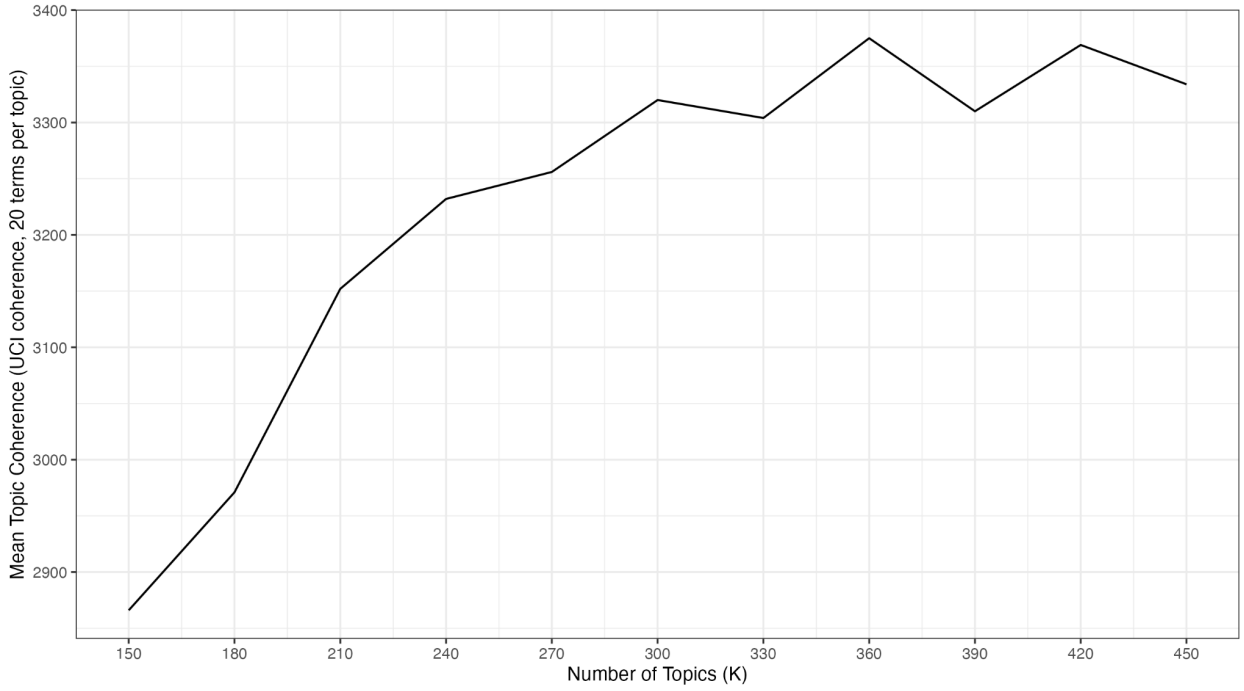
Suppl. Figure 1

Consort diagram illustrating stratification of patients into sets and cohorts, number of unique sites represented by those groups, and how each is used in analysis. The site quality filter removed sites with inpatient serum creatinine or white blood cell count results for fewer than 25% of patients, the site U09.9 filter removed patients from sites with no U09.9 diagnoses, and filter variables are as described for specific tests (see Suppl. Methods).



Suppl. Figure 2

Mean topic coherence scores for LDA models varying the number of topics generated (K). Topic coherences are computed as intrinsic UCI Coherence[26] using the top 20 terms per topic. UCI coherence evaluates, for all term pairs amongst these top 20, how frequently they occur together in patient histories compared to the expectation assuming terms occur independently, on the validation data set. K=300 was chosen as the final number of topics.



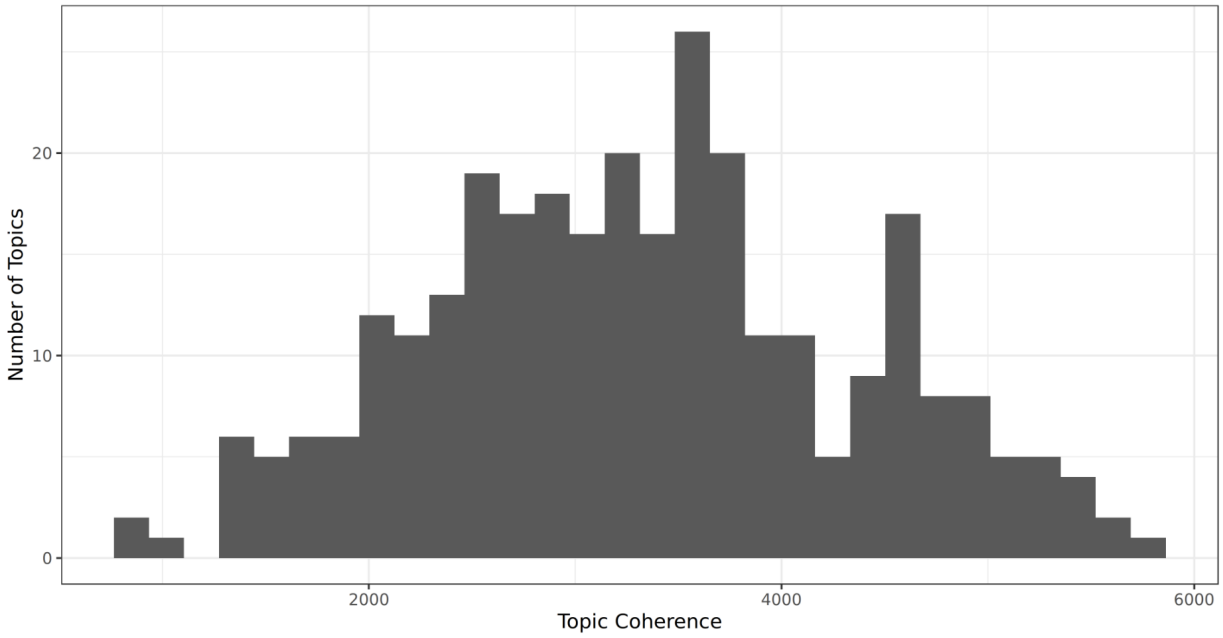
Suppl. Figure 3

Full topic clouds for all 300 topics generated and visualizations of corresponding contrasts.

Available at https://doi.org/10.5281/zenodo.7960028

Suppl. Figure 4

Histogram of topic coherence values.

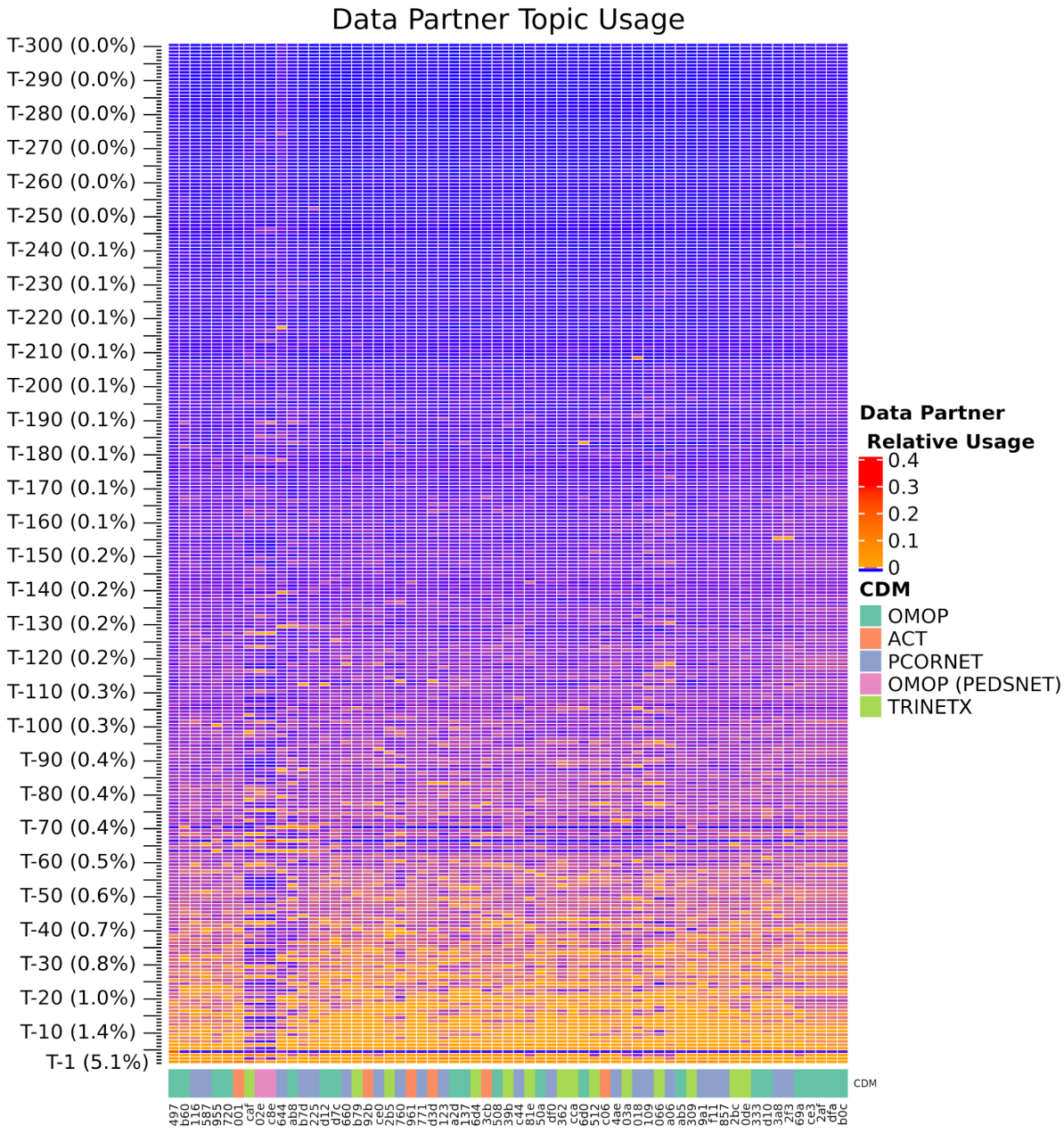


Suppl. Figure 5

Mean UCI coherence scores per topic and contributing data site. Site identifiers are masked, but labeled with the source common data model in use at the site.

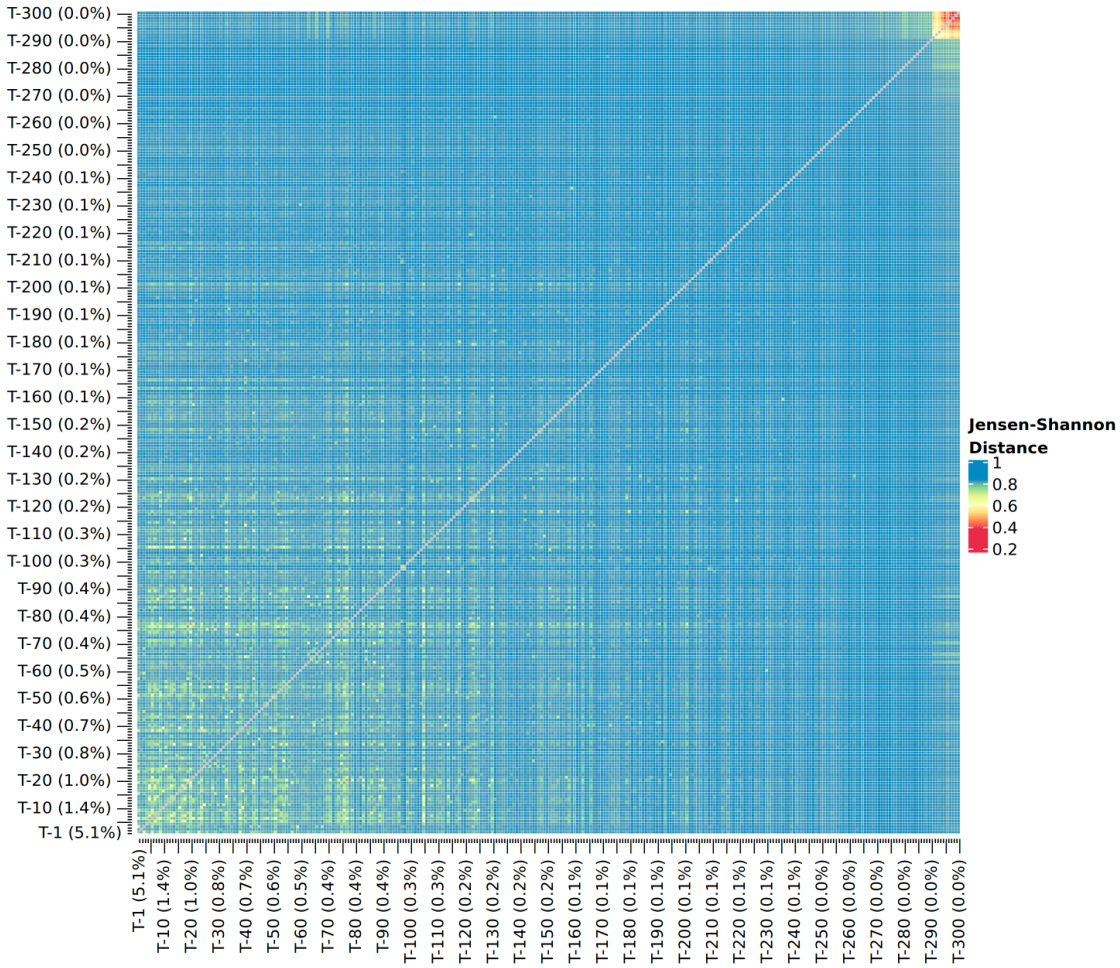
Suppl. Figure 6

Relative usage of topics per contributing site. For a given site and topic, relative usage is computed as the sum of assigned weights to that topic for patients from that site divided by the number of patients, representing a distribution over topics per site.



Suppl. Figure 7

Topic/topic dissimilarity as Jensen-Shannon Distance. Topic self-distances of 0 are not shown.



Supplemental Methods

Data Cleaning

EHR data from the National COVID Cohort Collaborative (N3C), released Aug. 2, 2022 represent records from 75 contributing sites. All analyses were restricted to data from 63 sites passing minimal quality checks. Sites were excluded if greater than 25% of inpatient visits were not accompanied by serum creatinine or white blood cell count measures (N=11), or if greater than 5% of COVID-19 confirmed patients were indicated as inpatient continuously for 200 or more days prior to and including their confirmed COVID-19 date (as potential long-term care facilities, N=1).

Records from the condition_era OMOP v5.3 table[49] were filtered to exclude suspicious entries beginning prior to Jan. 1, 2018 (the earliest date N3C requests records for) and those ending after each site's last reported contribution date. Data used for model training, evaluation, and testing excluded several OMOP concepts, including *COVID-19*, *Viral disease* and *Disease due to coronaviridae*, uninformative entries *No matching concept* and *Clinical finding*, and various *Findings of sexual activity* used primarily by a single site passing other quality checks. Individual concepts excluded are listed in Suppl. Table 6, and we also excluded all concepts not in the OMOP "Condition" domain.

Model Training

Model training utilized the online Latent Dirichlet Allocation (LDA) method of Hoffman et al.[24] as implemented in Apache Spark (pyspark.ml.clustering.LDA) version 3.2.1.[25] Parameters used include k (the number of topics, 300 in the final model), seed (42, a random seed to initialize the training), and maxIter (200, providing 10 passes over the training data in batches of 5% each). Condition counts passing data cleaning (above) were included in the training data. Determination of term-topic and topic-patient distributions were produced by the fitted LDA model.

Topic Annotations

As illustrated in Figure 2, each topic is annotated with three values: U, representing the relative usage of the topic by total weight assigned to patients (range 0-100%), H, a measure how uniformly the topic is used by N3C-contributing sites (range 0-1, with values closer to 0 being site-specific), and C, a measure of each topics' coherence compared to the mean over all topics (as a z-score). All three are computed over the training and validation sets.

U is computed as the sum over patients of the weight assigned to the topic, divided by the number of patients (which is also the total weight assigned over all topics).

H is computed as the information entropy of the relative usage of the topic across sites, normalized to a maximum value of 1.0 when the usage is uniformly distributed. Relative usage for a given site is computed as the total weight assigned to the topic for patients from the site, divided by the total number of patients from that site; these are also plotted in Suppl. Figure 11.

Finally, per-topic coherence C is calculated for each topic using the UCI Coherence metric (see Model Validation below). These values are not meant to be interpreted on an absolute scale, but since they are normally distributed amongst topics (Suppl. Figure 13) we adjust them to z-scores for comparative use. Suppl. Figure 10 illustrates non-adjusted scores broken out by site and topic.

Suppl. Figure 12 illustrates topic/topic similarity via pairwise Jensen-Shannon Distance. The Jensen-Shannon Distance between topics t_i and t_j is a true metric and is defined as the square root of the Jensen-Shannon divergence:

$$\text{JS Distance}(t_i, t_j) = \sqrt{\frac{\sum_{c \in terms} c_{t_i} \log\left(c_{t_i}/M\right)}{2} + \frac{\sum_{c \in terms} c_{t_i} \log\left(c_{t_i}/M\right)}{2}},$$

where $c_{t_x} = p(c|t_x)$ (the probability assigned to term c in topic t_x) and M is $(c_{t_i} + c_{t_j})/2$.

Topic Term Relevance

Term relevance provides a measure of term-topic-specificity, with values greater than zero indicating terms more likely for the topic than overall.[50] For term c_i and topic t_j, we define relevance as

$$\text{relevance}(c_i) = \ln \frac{p(c_i|t_j)}{p(c_i)} .$$

Model Validation

UCI coherence for a given topic t_i is computed over the top N terms by probability for the topic, where we used N = 20. Letting T_i be the set of top 20 terms for t_i, a sum score is computed for each distinct pair of terms a and b, where the score for a given pair is the log of the measured

probability of their occurring together in a patient compared to the joint probability assuming independence. To avoid undefined scores, 0 is used for pairs where the denominator is 0, and 1 is added to the joint probability.[26]

$$\text{Coherence}(t_i) = \sum_{b,c \in T_i,\, b<c} \begin{cases} \log_2\left(\frac{1+p(b\cap c|t_i)}{p(b|t_i)\cdot p(c|t_i)}\right) & \text{if } p(b \mid t_i)\cdot p(c \mid t_i) > 0 \\ 0 & \text{else} \end{cases}$$

Overall model quality was evaluated as the mean of coherence scores across topics, computed over the validation dataset only.

Per-Condition Tests

All tests were performed in R v3.6.[51] As described in the main text, patients in the test data set were included for evaluation of new-onset conditions if they satisfied requirements for being in the PASC, COVID, or Control cohorts. The top 20 conditions from each topic with relevance score > 0 were evaluated by considering only patients without the condition in the pre phase, comparing counts of PASC (and COVID) patients further indicated and not indicated for the post phase, to those same counts in the Control cohort. R's fisher.test() was used with simulate.p.value = TRUE to support tests where counts are large.[52] Reported p values were multiple-test corrected using Bonferroni's method.

BMI and Quan Comorbidity Scores

Patient BMI values used in modeling were the maximum over those reported after Jan. 1 2018, or the maximum of those computed as weight/(height^2) if no BMI measurement was directly available. Weight values outside 5kg– 300kg and height values outside 0.6m–2.43m were excluded from BMI calculations. Quan comorbidity scores[29] were computed from available source ICD code prefixes as shown in Suppl. Table 7.

Topic Regression Tests

Regression models were fitted using the geepack v1.3.2,[53] with contrasts computed using emmeans v1.6.0.[54] Individual patient histories defined by their pre- and post- phase data were assigned topic probability distributions by the fitted LDA model. For each topic, a logistic regression model was fitted predicting the model-assigned topic probability as the trial success rate with equal weight, from covariates phase (pre or post), cohort (PASC, COVID, or Control), patient life stage and wave of the index date (see main Methods), sex, race, Quan comorbidity

score, BMI, source CDM (PCORnet, ACT, OMOP, TrinetX, and OMOP (PedsNet)). To account for potential differential usage of PASC labels or topics, we also included percentage of patients at the given patients' site in the PASC cohort, and baseline usage of the topic by the patients' site relative to all sites (summing to 1.0 across sites). Interactions were included for terms of interest for contrast using the R/geepack formula `topic_probability ~ phase * cohort * (index_wave + sex + life_stage) + site_percent_pasc * phase * cohort + site_relative_topic_usage + race + quan_score + bmi + cdm`. Only patients from the assessment set with complete information for all variables were included. Because the assigned topic probability for a phase models the probability that a new condition will be drawn from the topic, the equal-weight approach considers the probability that, should a new condition be recorded, it is generated by the topic of interest.

Supplemental Tables

Suppl. Table 1

OMOP Concepts describing COVID-19 PCR or Antigen tests.

Concept Name	OMOP Concept Id
SARS-CoV-2 (COVID-19) N gene [Presence] in Respiratory specimen by Nucleic acid amplification using CDC primer-probe set N2	586525
SARS-CoV-2 (COVID-19) RdRp gene [Presence] in Saliva (oral fluid) by NAA with probe detection	36032174
SARS-related coronavirus RNA [Presence] in Specimen by NAA with probe detection	723472
SARS-CoV-2 (COVID-19) N gene [Cycle Threshold #] in Specimen by Nucleic acid amplification using CDC primer-probe set N2	706155
SARS-CoV-2 (COVID-19) S gene [Cycle Threshold #] in Specimen by NAA with probe detection	723468
SARS-CoV-2 (COVID-19) N gene [#/volume] (viral load) in Respiratory specimen by NAA with probe detection	36661370
SARS-CoV-2 (COVID-19) S gene [Cycle Threshold #] in Respiratory specimen by NAA with probe detection	723467
SARS-CoV-2 (COVID-19) N gene [Presence] in Serum or Plasma by NAA with probe detection	586520
SARS-CoV-2 (COVID-19) S gene [Presence] in Respiratory specimen by NAA with probe detection	723465
SARS-CoV-2 (COVID-19) [Presence] in Specimen by Organism specific culture	586516
SARS-CoV-2 (COVID-19) N gene [Cycle Threshold #] in Specimen by NAA with probe detection	706167
SARS-CoV-2 (COVID-19) Ag [Presence] in Respiratory specimen by Rapid immunoassay	723477
SARS-CoV-2 (COVID-19) RNA [Log #/volume] (viral load) in Specimen by NAA with probe detection	715262
SARS-related coronavirus N gene [Cycle Threshold #] in Specimen by Nucleic acid amplification using CDC primer-probe set N3	706172
SARS-CoV-2 (COVID-19) RNA [Presence] in Saliva (oral fluid) by NAA with probe	715260

detection	
SARS-CoV-2 (COVID-19) S gene [Presence] in Serum or Plasma by NAA with probe detection	586519
SARS-CoV-2 (COVID-19) ORF1ab region [Cycle Threshold #] in Respiratory specimen by NAA with probe detection	723469
SARS-CoV-2 (COVID-19) RNA [Cycle Threshold #] in Specimen by NAA with probe detection	586529
SARS-related coronavirus E gene [Presence] in Respiratory specimen by NAA with probe detection	586523
SARS-CoV-2 (COVID-19) ORF1ab region [Presence] in Saliva (oral fluid) by NAA with probe detection	36031506
SARS-CoV-2 (COVID-19) S gene [Presence] in Specimen by NAA with probe detection	723466
SARS-CoV-2 (COVID-19) RNA [Presence] in Nasopharynx by NAA with non-probe detection	723476
SARS-CoV-2 (COVID-19) N gene [Presence] in Saliva (oral fluid) by Nucleic acid amplification using CDC primer-probe set N1	36032258
SARS-CoV-2 (COVID-19) RNA [Presence] in Nasopharynx by NAA with probe detection	586526
SARS-related coronavirus E gene [Presence] in Serum or Plasma by NAA with probe detection	586518
SARS-CoV-2 (COVID-19) S gene [Presence] in Respiratory specimen by Sequencing	36031213
SARS-CoV-2 (COVID-19) RNA [Presence] in Nose by NAA with probe detection	757677
SARS-CoV-2 (COVID-19) N gene [Presence] in Specimen by Nucleic acid amplification using CDC primer-probe set N2	706154
SARS-CoV-2 (COVID-19) RNA panel - Respiratory specimen by NAA with probe detection	706158
SARS-CoV-2 (COVID-19) N gene [Presence] in Respiratory specimen by NAA with probe detection	706161
SARS-CoV-2 (COVID-19) RdRp gene [Cycle Threshold #] in Specimen by NAA with probe detection	723470
SARS-CoV-2 (COVID-19) RdRp gene [Presence] in Lower respiratory specimen by NAA with probe detection	36031652
SARS-CoV-2 (COVID-19) N gene [Presence] in Saliva (oral fluid) by NAA with probe detection	36661378

SARS-related coronavirus+MERS coronavirus RNA [Presence] in Respiratory specimen by NAA with probe detection	706159
SARS-related coronavirus E gene [Presence] in Specimen by NAA with probe detection	706174
SARS-CoV-2 (COVID-19) N gene [Presence] in Specimen by Nucleic acid amplification using CDC primer-probe set N1	706156
SARS-CoV-2 (COVID-19) RNA [Cycle Threshold #] in Respiratory specimen by NAA with probe detection	586528
Measurement of Severe acute respiratory syndrome coronavirus 2 antigen	37310257
SARS-related coronavirus E gene [Cycle Threshold #] in Specimen by NAA with probe detection	706166
SARS-CoV-2 (COVID-19) Ag [Presence] in Upper respiratory specimen by Immunoassay	36032419
SARS-CoV-2 (COVID-19) RNA panel - Specimen by NAA with probe detection	706169
SARS-CoV-2 (COVID-19) RNA [Presence] in Respiratory specimen by NAA with non-probe detection	36031238
SARS-CoV-2 (COVID-19) RdRp gene [Presence] in Respiratory specimen by NAA with probe detection	706160
SARS-CoV-2 (COVID-19) N gene [Presence] in Nasopharynx by NAA with probe detection	715272
SARS-CoV-2 (COVID-19) N gene [Presence] in Nose by NAA with probe detection	757678
SARS-CoV-2 (COVID-19) RNA [Presence] in Saliva (oral fluid) by Sequencing	715261
SARS-CoV-2 (COVID-19) RNA [Presence] in Specimen by NAA with probe detection	706170
SARS-CoV-2 (COVID-19) N gene [Cycle Threshold #] in Specimen by Nucleic acid amplification using CDC primer-probe set N1	706157
SARS-CoV-2 (COVID-19) ORF1ab region [Presence] in Respiratory specimen by NAA with probe detection	723478
SARS-related coronavirus N gene [Presence] in Specimen by Nucleic acid amplification using CDC primer-probe set N3	706171
SARS-CoV+SARS-CoV-2 (COVID-19) Ag [Presence] in Respiratory specimen by Rapid immunoassay	757685
SARS-CoV-2 (COVID-19) RNA [Presence] in Respiratory specimen by Sequencing	36661377
SARS-CoV-2 (COVID-19) N gene [Log #/volume] (viral load) in Respiratory specimen by NAA with probe detection	36661371

SARS-CoV-2 (COVID-19) RdRp gene [Cycle Threshold #] in Respiratory specimen by NAA with probe detection	723471
SARS-CoV-2 (COVID-19) RdRp gene [Presence] in Upper respiratory specimen by NAA with probe detection	36031453
SARS-CoV-2 (COVID-19) RdRp gene [Presence] in Specimen by NAA with probe detection	706173
SARS-CoV-2 (COVID-19) N gene [Presence] in Specimen by NAA with probe detection	706175
SARS-CoV-2 (COVID-19) ORF1ab region [Cycle Threshold #] in Specimen by NAA with probe detection	706168
SARS-CoV-2 (COVID-19) N gene [Presence] in Respiratory specimen by Nucleic acid amplification using CDC primer-probe set N1	586524
SARS-CoV-2 (COVID-19) ORF1ab region [Presence] in Specimen by NAA with probe detection	723464
SARS-related coronavirus RNA [Presence] in Respiratory specimen by NAA with probe detection	706165
SARS-CoV-2 (COVID-19) RNA panel - Saliva (oral fluid) by NAA with probe detection	36032061
SARS-CoV-2 (COVID-19) RNA [Presence] in Respiratory specimen by NAA with probe detection	706163
SARS-CoV-2 (COVID-19) specific TCRB gene rearrangements [Presence] in Blood by Sequencing	36031944
SARS-CoV-2 (COVID-19) RNA [Presence] in Serum or Plasma by NAA with probe detection	723463

Suppl. Table 2

All indicators of COVID-19 infection (except for PCR and Antigen tests, Suppl. Table 3).

Concept Name	Concept Id
SARS-CoV-2 (COVID-19) IgG Ab [Presence] in Serum, Plasma or Blood by Rapid immunoassay	706181
SARS-CoV-2 (COVID-19) IgA Ab [Units/volume] in Serum or Plasma by Immunoassay	723459
SARS-CoV-2 (COVID-19) IgM Ab [Presence] in Serum, Plasma or Blood by Rapid immunoassay	706180
SARS-CoV-2 (COVID-19) IgM Ab [Presence] in DBS by Immunoassay	36659631
SARS-CoV-2 (COVID-19) IgM Ab [Titer] in Serum or Plasma by Immunofluorescence	36661373
SARS-CoV-2 (COVID-19) neutralizing antibody [Presence] in Serum by pVNT	757680
SARS-CoV-2 (COVID-19) IgG+IgM Ab [Presence] in Serum or Plasma by Immunoassay	723479
SARS-CoV-2 (COVID-19) Ab panel - Serum, Plasma or Blood by Rapid immunoassay	706176
SARS-CoV-2 (COVID-19) IgG Ab [Titer] in Serum or Plasma by Immunofluorescence	36661374
SARS-CoV-2 (COVID-19) IgM Ab [Units/volume] in Serum or Plasma by Immunoassay	706178
SARS-CoV-2 (COVID-19) IgA Ab [Presence] in Serum or Plasma by Immunoassay	723473
SARS-CoV-2 (COVID-19) neutralizing antibody [Titer] in Serum by pVNT	757679
SARS-CoV-2 (COVID-19) Ab [Presence] in Serum or Plasma by Immunoassay	586515
SARS-CoV-2 (COVID-19) IgG Ab [Units/volume] in Serum or Plasma by Immunoassay	706177
SARS-CoV-2 (COVID-19) S protein RBD neutralizing antibody [Presence] in Serum or Plasma by sVNT	36031734
SARS-CoV-2 (COVID-19) IgA Ab [Titer] in Serum or Plasma by Immunofluorescence	36661372
SARS-CoV-2 (COVID-19) Ab [Units/volume] in Serum or Plasma by Immunoassay	586522
SARS-CoV-2 (COVID-19) IgA+IgM [Presence] in Serum or Plasma by Immunoassay	757686
Measurement of Severe acute respiratory syndrome coronavirus 2 antibody	37310258

SARS-CoV-2 (COVID-19) IgG Ab [Presence] in Serum or Plasma by Immunoassay	723474
SARS-CoV-2 (COVID-19) Ab panel - Serum or Plasma by Immunoassay	706179
SARS-CoV-2 stimulated gamma interferon [Presence] in Blood	36031969
SARS-CoV-2 stimulated gamma interferon release by T-cells [Units/volume] in Blood	36032309
SARS-CoV-2 (COVID-19) IgA Ab [Presence] in Serum, Plasma or Blood by Rapid immunoassay	586521
SARS-CoV-2 (COVID-19) Ab [Presence] in DBS by Immunoassay	36031197
SARS-CoV-2 (COVID-19) Ab [Presence] in Serum, Plasma or Blood by Rapid immunoassay	36661369
SARS-CoV-2 (COVID-19) IgM Ab [Presence] in Serum or Plasma by Immunoassay	723475
SARS-CoV-2 (COVID-19) Ab [Interpretation] in Serum or Plasma	723480
SARS-CoV-2 (COVID-19) IgG Ab [Presence] in DBS by Immunoassay	586527
SARS-CoV-2 stimulated gamma interferon release by T-cells [Units/volume] corrected for background in Blood	36031956

Suppl. Table 3

All significant single-condition tests. Listed estimates are odds ratios for the given cohort pre-to-post compared to Controls, and p-values are adjusted across all condition tests for both cohorts (Bonferroni, prior to filtering to significance). Available at https://doi.org/10.5281/zenodo.7960028

Suppl. Table 4

All topic-level logistic model tests. Estimates are odds ratios for the given cohort and demographic compared to Controls for the same demographic. Ratios where the demographic is listed as NA are for demographic contrasts independent of phase or cohort (model effectiveness checks, see main Methods). P-values are adjusted across all contrast tests (Holm). Available at https://doi.org/10.5281/zenodo.7960028

Suppl. Table 5

Summary statistics for pediatric and adolescent patients in the assessment set. Young PASC patients are labeled with many more post-infection conditions than similar COVID and Control patients, while being similar prior to infection. These patients also represent a diversity of sites. Note that the pre-infection phase covers 1 year of patient history, while the post-infection phase covers 6 months post-acute.

Cohort	Life Stage	Phase	Mean # Conditions	SD # Conditions	# Patients	# Sites
Control	adolescent	post	8.10	7.78	10789	32
Control	adolescent	pre	10.84	10.45	10789	32
Control	pediatric	post	7.045	6.66	16029	32
Control	pediatric	pre	10.88	10.98	16029	32
COVID	adolescent	post	8.19	8.95	3703	31
COVID	adolescent	pre	11.05	11.33	3703	31
COVID	pediatric	post	7.66	8.02	3724	29
COVID	pediatric	pre	11.35	11.39	3724	29
PASC	adolescent	post	17.75	16.33	66	20
PASC	adolescent	pre	12.87	12.65	66	20
PASC	pediatric	post	15.85	10.61	49	18
PASC	pediatric	pre	14.36	9.744	49	18

Suppl. Table 6

OMOP Concepts excluded from model training, evaluation, and testing.

Concept Name	OMOP Concept Id
No matching concept	0
Clinical finding	441840
COVID-19	37311061
Viral disease	440029
Disease due to coronaviridae	4100065
Sexually abstinent	764423
Single current sexual partner	4043045
New sexual partner	44813701
Sexually active with men	43021202
Single historical sexual partner	43021216
Number of current sexual partners - finding	4276728
Bigamy	4336540
Sexual activity - two to three times per month	4012347
Sexual activity - two to three times per week	4012202
Finding of number of historical sexual partners	43021214
No longer sexually active	4043041
Multiple current sexual partners	4038723
Sexually active with transgender person	43021204
Number of sexual partners - finding	4269990
Satisfactory sexual experience	44811373
Sexual activity - daily	4012377
Currently not sexually active	4012376
Never been sexually active	4145811
Fornication	4031991
Sexual activity - monthly	4012348
Sexual activity - weekly	4012203

Sexual contact with high risk partner	44789379
Finding of frequency of sexual activity	4188013
Engages in sexual activity outside marriage	43021163
Sexually active with women	43021203
Purposely unmarried and sexually abstinent	43021238
Sex within a relationship only	4021660
Sexually active in last month	37017764
Sexually active	4043042
Finding relating to sexual activity	4114865
Sexually active in last year	37017763
Engages in sexual activity before marriage	43021162
Sexually active in last six months	37017762
Multiple historical sexual partners	43021215

Suppl. Table 7

Source ICD code prefixes used to generate Quan-based comorbidity scores.

ICD Prefixes	Charleson Group	Quan Score
'I21','I22','I252'	1: Acute or historical MI	0
'I43','I50','I099','I110','I130','I132','I255','I420','I425','I426','I427','I428','I429','P290'	2: CHF	2
'I70','I71','I731','I738','I739','I771','I790','I792','K551','K558','K559','Z958','Z959'	3: Peripheral vascular disease	0
'G45','G46','I60','I61','I62','I63','I64','I65','I66','I67','I68','I69','H340'	4: Cerebrovascular disease	0
'F00','F01','F02','F03','G30','F051','G311'	5: Dementia	2
'J40','J41','J42','J43','J44','J45','J46','J47','J60','J61','J62','J63','J64','J65','J66','J67','I278','I279','J684','J701','J703'	6: COPD	1
'M32','M33','M34','M06','M05','M315','M351','M353','M360'	7: Rheumatic disease	1
'K25','K26','K27','K28'	8: Peptic ulcer	0
'B18','K73','K74','K700','K701','K702','K703','K709','K717','K713','K714','K715','K760','K762','K763','K764','K768','K769','Z944'	9: Mild liver disease	2
'E100','E101','E106','E108','E109','E110','E111','E116','E118','E119','E120','E121','E126','E128','E129','E130','E131','E136','E138','E139','E140','E141','E146','E148','E149'	10: Diabetes	0
'E102','E103','E104','E105','E107','E112','E113','E114','E115','E117','E122','E123','E124','E125','E127','E132','E133','E134','E135','E137','E142','E143','E144','E145','E147'	11: Diabetes with chronic complications	1
'G81','G82','G041','G114','G801','G802','G830','G831','G832','G833','G834','G839'	12: Paralysis	2
'N18','N19','N052','N053','N054','N055','N056','N057','N250','I120','I131','N032','N033','N034','N035','N036','N037','Z490','Z491','Z492','Z940','Z992'	13: Renal disease	1

'C00','C01','C02','C03','C04','C05','C06','C07','C08','C09','C10','C11', 'C12','C13','C14','C15','C16','C17','C18','C19','C20','C21','C22','C23', 'C24','C25','C26','C30','C31','C32','C33','C34','C37','C38','C39','C40', 'C41','C43','C45','C46','C47','C48','C49','C50','C51','C52','C53','C54', 'C55','C56','C57','C58','C60','C61','C62','C63','C64','C65','C66','C67', 'C68','C69','C70','C71','C72','C73','C74','C75','C76','C81','C82','C83', 'C84','C85','C88','C90','C91','C92','C93','C94','C95','C96','C97'	14: Localized cancer/leukemia/lymphoma	2
'K704','K711','K721','K729','K765','K766','K767','I850','I859','I864','I982'	15: Moderate/severe liver disease	4
'C77','C78','C79','C80'	16: Metastatic cancer	6
'B20','B21','B22','B24'	17: HIV/AIDS	4