Coding Long COVID: Characterizing a new disease through an ICD-10 lens

Emily R Pfaff (co-first), University of North Carolina at Chapel Hill
Charisse Madlock-Brown (co-first), University of Tennessee Health Science Center
John M. Baratta, University of North Carolina at Chapel Hill
Abhishek Bhatia, University of North Carolina at Chapel Hill
Hannah Davis, Patient-Led Research Collaborative
Andrew Girvin, Palantir Technologies
Elaine Hill, University of Rochester
Liz Kelly, University of North Carolina at Chapel Hill
Kristin Kostka, Northeastern University
Johanna Loomba, University of Virginia
Julie A. McMurry, University of Colorado Anschutz Medical Campus
Rachel Wong, Stony Brook University
Tellen D Bennett, University of Colorado Anschutz Medical Campus
Richard Moffitt, Stony Brook University
Christopher G Chute, Johns Hopkins University
Melissa Haendel, University of Colorado Anschutz Medical Campus
The N3C Consortium
The RECOVER Consortium

Abstract

Naming a newly discovered disease is always challenging; in the context of the COVID-19 pandemic and the existence of post-acute sequelae of SARS-CoV-2 infection (PASC), which includes Long COVID, it has proven especially challenging. Disease definitions and assignment of a diagnosis code are often asynchronous and iterative. The clinical definition and our understanding of the underlying mechanisms of Long COVID are still in flux. The deployment of an ICD-10-CM code for Long COVID in the US took nearly two years after patients had begun to describe their condition. Here we leverage the largest publicly available HIPAA-limited dataset about patients with COVID-19 in the US to examine the heterogeneity of adoption and use of U09.9, the ICD-10-CM code for “Post COVID-19 condition, unspecified.”

Our results include a characterization of common diagnostics, treatment-oriented procedures, and medications associated with U09.9-coded patients, which give us insight into current practice patterns around Long COVID. We also established the diagnoses most commonly co-occurring with U09.9, and algorithmically clustered them into three major categories: cardiopulmonary, neurological, and metabolic. We aim to apply the patterns gleaned from this analysis to flag probable Long COVID cases occurring prior to the existence of U09.9, thus establishing a mechanism to ensure patients with earlier cases of Long-COVID are no less ascertainable for current and future research and treatment opportunities.

Introduction

Naming diseases is an ever present challenge, and there is no shortage of efforts that aim to better standardize, disambiguate, and keep track of disease nomenclature and definitions[1–4]. Disease naming has always been controversial–for example, there are more than 400 names for syphilis dating back to the 15th century[5]. Naming a disease requires defining it, and assigning a standard code to the disease facilitates research, care, and patient engagement due to ease of patient classification and knowledge exchange. However, naming and coding a disease does not mean the disease did not exist prior to its naming or coding. For instance, although “SARS-CoV-2” and “COVID-19” were both coined February 11, 2020, by the International Committee on the Taxonomy of Viruses and the WHO, respectively[6,7], we know that cases of COVID-19 began to surface in Wuhan, China in late December 2019[8]. In the US, most diagnostic coding

uses the ICD-10-CM terminology; however the ICD-10-CM code for COVID-19, U07.1, was not made available for use until April 1, 2020. The implications of this naming delay are wide-ranging. To this day, US COVID-19 cases prior to April 1, 2020 are difficult to retrospectively ascertain. Even after that date, use of U07.1 for COVID-19 phenotyping came with caveats–use of the new code was inconsistent and of variable sensitivity and specificity, and studies have shown both underuse and overuse of U07.1 in different contexts and health systems[9–11].

Long COVID, which is included in the more general term of post-acute sequelae of SARS CoV-2 infection (PASC), is now also subject to the effects of delayed naming. By Spring of 2020, patients suffering from Long COVID had coined various terms to describe the condition, including the COVID-19 long tail, long-haul COVID, and Long COVID[12–14]. Long COVID is defined by ongoing, relapsing, or new symptoms or other health effects occurring after the acute phase of SARS-CoV-2 infection (i.e., present four or more weeks after the acute infection). Heterogeneous symptoms may include, but are not limited to, fatigue, difficulty breathing, brain fog, insomnia, joint pain, and cardiac issues[15–17]. As the impact of Long COVID on health and quality of life became increasingly clear at a population level, patients worldwide came together to urge healthcare systems and policymakers to acknowledge this condition[18,19].

Despite the relatively early recognition of this condition, an ICD-10-CM code (U09.9, “Post COVID-19 condition, unspecified”) was not made available for use in the clinical setting until October of 2021. Moreover, this single code may prove insufficient: considering the phenotypic and severity variation seen in Long COVID patients, it is likely that subtypes of Long COVID exist, and such subtypes may correlate with specific underlying mechanisms that should be targeted by different interventions. There is thus more naming to be done, and a particular need to define and refine computable phenotypes for Long COVID and its subtypes. In doing so, we can appropriately define cohorts for clinical studies and provide more precise treatment and clinical decision support. This is a key priority for the parent program for this work, the NIH Researching COVID to Enhance Recovery (RECOVER) Initiative,[20] which seeks to understand, treat, and prevent PASC through a wide variety of research modalities, including electronic health record (EHR) and real-world data.

In response to the COVID-19 pandemic, the informatics and clinical community harmonized an enormous amount of EHR data to reveal candidate risk factors and therapies associated with COVID-19. The NIH’s National COVID Cohort Collaborative (N3C) is now the largest publicly available HIPAA limited dataset in U.S. history, with over 13 million patients, and is a testament to the partnership of over 290 organizations. Due to the scale and demographic and geographic diversity of data within the N3C, it is uniquely well-suited to characterize the early use of the new Long COVID ICD-10-CM code.

In prior work, we proposed a machine learning-based computable phenotype definition for Long COVID using the N3C data[21]. Now that U09.9 is available, the presence of the code will be a valuable addition to that existing Long COVID model, especially since ascertainment of presumptive cases based on EHR data in the absence of a U09.9 diagnosis code is limited by the non-specificity of the clinical manifestations of the disease, the frequency with which these symptoms are seen in the general population, and the observation that the diagnosis of Long COVID is one of exclusion. However, due to the caveats noted above regarding newly introduced codes, we first sought to characterize the early clinical use patterns of U09.9 before accepting it into our model and cohort definition at face value. This characterization revealed interesting patterns that may enable us to glean a better understanding of both rough subtypes of Long COVID and current clinical practices for diagnosis and treatment of Long COVID. Ultimately, identifying patients with Long COVID based upon multiple means of inquiry (including U09.9) is critically important to recruit participants for research studies, assess the public health burden, and support nimble analytics across our heterogenous health care systems.

Methods

To characterize the use of the U09.9 code, we used EHR data integrated and harmonized inside the NIH-hosted N3C Secure Data Enclave to identify clinical features co-occurring around the time of patients’ U09.9 index date. The methods for patient identification, data acquisition, ingestion, and harmonization into the N3C Enclave have been described previously[22–24]. Briefly, N3C contains EHR data for patients who (1)

tested positive for SARS-CoV-2 infection, (2) whose symptoms are consistent with a COVID-19 diagnosis, or (3) are demographically matched controls who have tested negative for SARS-CoV-2 infection (and have never tested positive) to support comparative studies. Lookback data are available from January 2018 forward for each patient.

In this analysis, we defined our initial population (*n* = 9,571, sourced from 28 different health care systems) as any non-deceased patient with one or more U09.9 diagnosis codes recorded on or after October 1, 2021. U09.9 codes appearing prior to this date were likely retroactively applied to these patients' records (e.g., as "onset dates" in an EHR Problem List), therefore making it difficult to determine an index date that reflects the actual date of diagnosis. We excluded patients (*n* = 1,497) whose U09.9 index occured during an inpatient hospitalization, due to the difficulty of distinguishing co-occurring clinical features related to Long COVID versus the primary reason for their hospitalization. After these exclusions, a base population of 8,074 remained (see **Supplemental Figure 1**). Note that we did not require patients in our cohort to have a COVID-19 diagnosis code (U07.1) or positive SARS-CoV-2 test on record, as many patients with Long COVID do not have this documentation[19].

Data from 28 of the 71 N3C sites were used for this analysis. The remaining sites either (1) did not use the U09.9 code in their N3C data or had not refreshed data since November 1, 2021, meaning the U09.9 code would not be present even if used at the site (*n* = 30 sites), or (2) did not meet the minimum criteria we set for site data for all RECOVER-related analyses (*n* = 13 sites): (a) >=25% of inpatients with at least one white blood cell count and at least one serum creatinine (to ensure lab measurement completeness); (b) 75% of inpatient visits have valid end dates; and (c) dates must not be shifted by the site more than 30 days. Additional N3C data quality criteria have been described previously, and also apply to this work.[23] The 28 sites used here are diverse in geographic location and institution size, but cannot be specifically named due to N3C governance policies.

We calculated person-level demographics and a number of social determinants of health variables at the area level. These variables are sourced from the Sharecare-Boston University School of Public Health Social Determinants of Health Index[23], and were linked to patients based on the preferred county (majority residence) associated with the patient's 5-digit ZIP code. We then characterized this cohort by examining diagnoses, procedures, and medications that occurred between each patient's U09.9 index date and 60 days after index (hereafter referred to as our "analysis window").

Diagnosis Analysis

Our objective in characterizing diagnoses around the U09.9 index date was not only to catalog conditions and symptoms that tend to co-occur with the U09.9 diagnosis, but also to determine which of those conditions and symptoms tend to co-occur with each other. In doing so, we begin to see clusters of conditions that are more likely to occur together within a single patient's record. First, we extracted all conditions in each patient's record within the analysis window, and identified the most frequently occurring conditions in the study population. We then constructed an adjacency matrix for the top 30 conditions, with values indicating the frequency of co-occurrence between two conditions in the study population. From this matrix, we constructed a weighted network with nodes representing individual diagnoses, edges between nodes representing co-occurrence, and edge weights corresponding to the count of patients with both conditions. In order to detect conditions that are more likely to co-occur in our study population than at random, we tested the Louvain [25], Walktrap,[26] and Girvan-Newman[27] algorithms for community detection. We selected the Louvain algorithm in our final model, as it maximized modularity while retaining a reasonable resolution of detection. For further subgroup analyses, we present clusters detected within age-stratified condition co-occurrence networks. Additional details on community detection, network stability and subgroup analyses are available in **Supplemental Methods**.

Procedure Analysis

Characterizing common procedures around the time of U09.9 allowed us to assess current practice patterns (i.e., diagnostics and treatments) for patients receiving the code. We defined a "procedure" as any medical diagnostics or treatments rendered by a healthcare provider. We excluded non-informative records that simply

reflect that an encounter took place (e.g., CPT 99212, “Office or other outpatient visit”), despite their technical classification as “procedure codes.” We then aggregated remaining procedures into high-level categories (e.g., “radiography,” “physical therapy”) in order to discern the diagnostics and treatments that occured within each patient’s analysis window.

Medication Analysis

As with diagnoses and procedures, we extracted all medication records occurring within each patient’s analysis window, in order to characterize newly prescribed medications that may be used to treat symptoms of Long COVID. In order to focus on newly prescribed medications and not long-standing prescriptions, we excluded medications for each patient for which there were records prior to the patient’s U09.9 index. Medications were categorized using the third level of the Anatomical Therapeutic Chemical (ATC) classification system[28].

Results

Each of the patients in our base population came from one of 28 N3C data-submitting health care organizations. **Table 1** shows the breakdown of the study cohort by person-level demographics and area-level social determinants of health. It should be noted that greater severity of acute SARS-CoV-2 infection does not appear to have outsize influence in determining which patients end up with a U09.9 code; 1,722 of the U09.9 patients (21.3%) were hospitalized during their acute SARS-CoV-2 infection.

	Age <21	21-45	46-65	66+	All ages
	n = 454	*n* = 2,816	*n* = 3,422	*n* = 1,382	*n* = 8,074
Person-level variables					
Sex (%)					
female	276 (60.8)	2001 (71.1)	2163 +/-5 (63.0)	803 (58.1)	**5243 (64.9)**
male	178 (39.2)	815 (28.9)	1257 +/-5 (36.7)	579 (41.9)	**2829 (35)**
unknown	0 (0.0)	0 (0.0)	<20	0 (0.0)	**<20**
Race (%)					
Black	55 (12.1)	402 (14.3)	448 (13.1)	133 (9.6)	**1038 (12.9)**
White	327 (72.0)	1987 (70.6)	2646 (77.3)	1140 (82.5)	**6100 (75.6)**
Other	20 (4.4)	101 (3.6)	75 (2.2)	31 (2.2)	**227 (2.8)**
Unknown	52 (11.5)	326 (11.6)	253 (7.4)	78 (5.6)	**709 (8.8)**
Ethnicity (%)					
Hispanic/Latino	40 (8.8)	324 (11.5)	258 (7.5)	75 (5.4)	**697 (8.6)**
Not Hispanic/Latino	370 (81.5)	2370 (84.2)	3001 (87.7)	1254 (90.7)	**6995 (86.6)**
Unknown	44 (9.7)	122 (4.3)	163 (4.8)	53 (3.8)	**382 (4.7)**
Area-level social determinants of health (county level)					
Households with Income below poverty (%)					
High (>15%)	148 (32.6)	931 (33.1)	1216 (35.5)	467 (33.8)	**2762 (34.2)**
Medium (11-15%)	109 (24.0)	684 (24.3)	786 (23.0)	364 (26.3)	**1943 (24.1)**
Low (<11%)	122 (26.9)	786 (27.9)	895 (26.2)	344 (24.9)	**2147 (26.6)**
Missing	75 (16.5)	415 (14.7)	525 (15.3)	207 (15.0)	**1222 (15.1)**
Residents with college degree (%)					
High (>25%)	89 (19.6)	580 (20.6)	622 (18.2)	265 (19.2)	**1556 (19.3)**
Medium (19-25%)	136 (30.0)	1044 (37.1)	1173 (34.3)	453 (32.8)	**2806 (34.8)**
Low (<19%)	154 (33.9)	777 (27.6)	1102 (32.2)	457 (33.1)	**2490 (30.8)**
Missing	75 (16.5)	415 (14.7)	525 (15.3)	207 (15.0)	**1222 (15.1)**
Residents 19-64 with public health insurance (%)					

High (>18%)	96 (21.1)	421 (15.0)	623 (18.2)	256 (18.5)	**1396 (17.3)**
Medium (13-18%)	142 (31.3)	1095 (38.9)	1240 (36.2)	528 (38.2)	**3005 (37.2)**
Low (<13%)	141 (31.1)	885 (31.4)	1034 (30.2)	391 (28.3)	**2451 (30.4)**
Missing	75 (16.5)	415 (14.7)	525 (15.3)	207 (15.0)	**1222 (15.1)**
MDs per 1000 residents (%)					
High (>3.61)	133 (29.3)	1031 (36.6)	1064 (31.1)	377 (27.3)	**2605 (32.3)**
Medium (1.91-3.61)	98 (21.6)	598 (21.2)	742 (21.7)	324 (23.4)	**1762 (21.8)**
Low (<1.91)	148 (32.6)	772 (27.4)	1091 (31.9)	474 (34.3)	**2485 (30.8)**

Table 1. Demographic breakdown of patients in N3C with a U09.9 diagnosis code. In addition to person-level demographics, we have included a number of social determinants of health variables at the *area* level (see Methods). In accordance with the N3C download policy, for demographics where small cell sizes (<20 patients) could be derived from context, we have shifted the counts +/- by a random number between 1 and 5. The accompanying percentages reflect the shifted number. All shifted counts are labeled as such, e.g. +/- 5.

In addition to demographics, the N3C data also enables us to examine medication use and procedures that occur in each patient's analysis window, as shown in **Figures 1 and 2**, respectively.

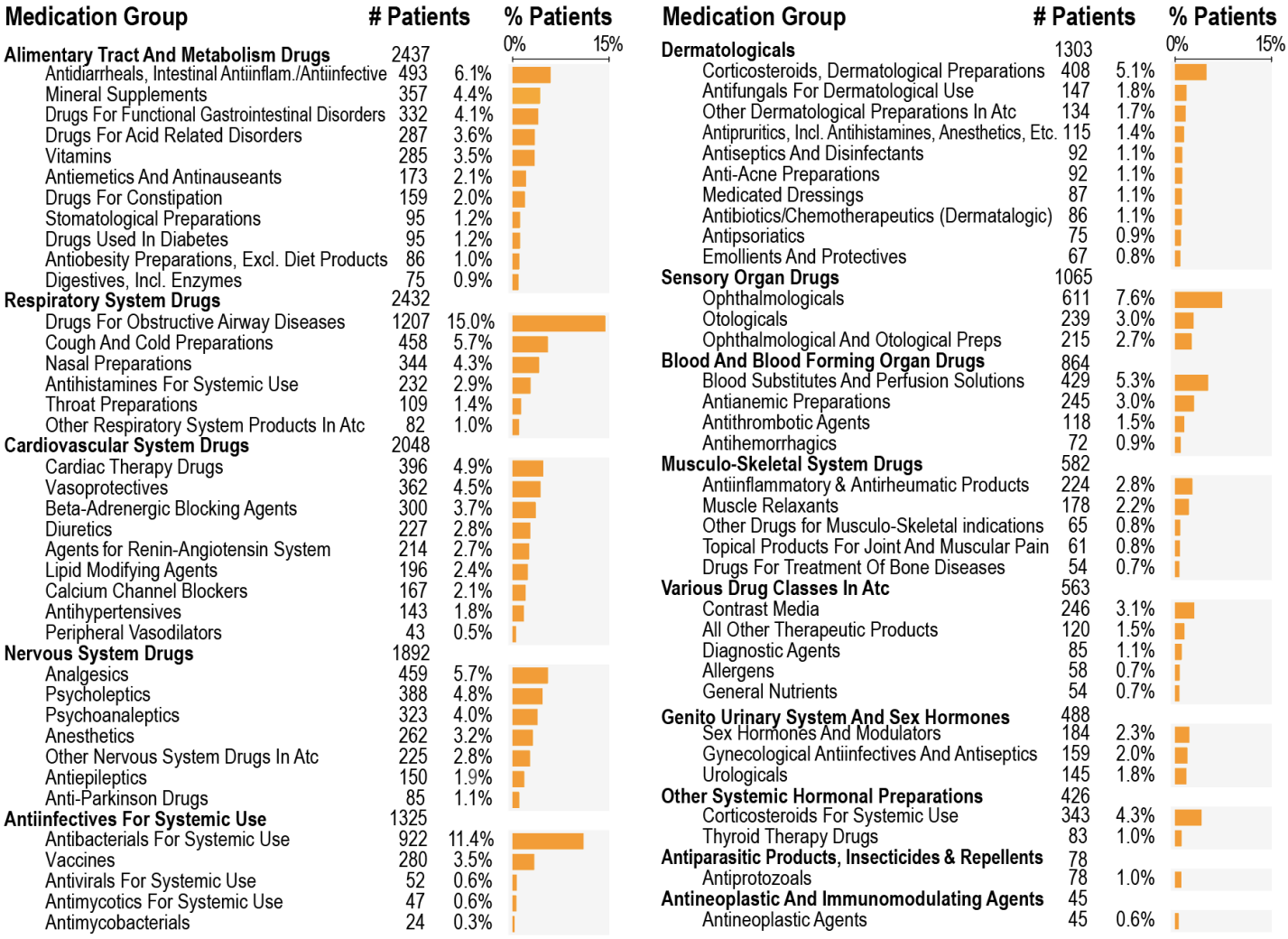


Figure 1. Common medications for 4,004 patients with a U09.9 code. Medications shown occur within 60 days after a patient's U09.9 diagnosis, and do *not* occur prior to the U09.9 (i.e., new medications). Medications are coded using the ATC terminology. Because a single drug can have multiple ATC codes, some medications are counted in more than one category. Category totals represent unique patient - drug pairs, not necessarily unique individuals. Medication classes

associated with fewer than 20 patients are not shown, per N3C download policy. An additional 4,070 patients had no new recorded medications in the analysis window; percentages are shown relative to all patients in the final study population (8,074).

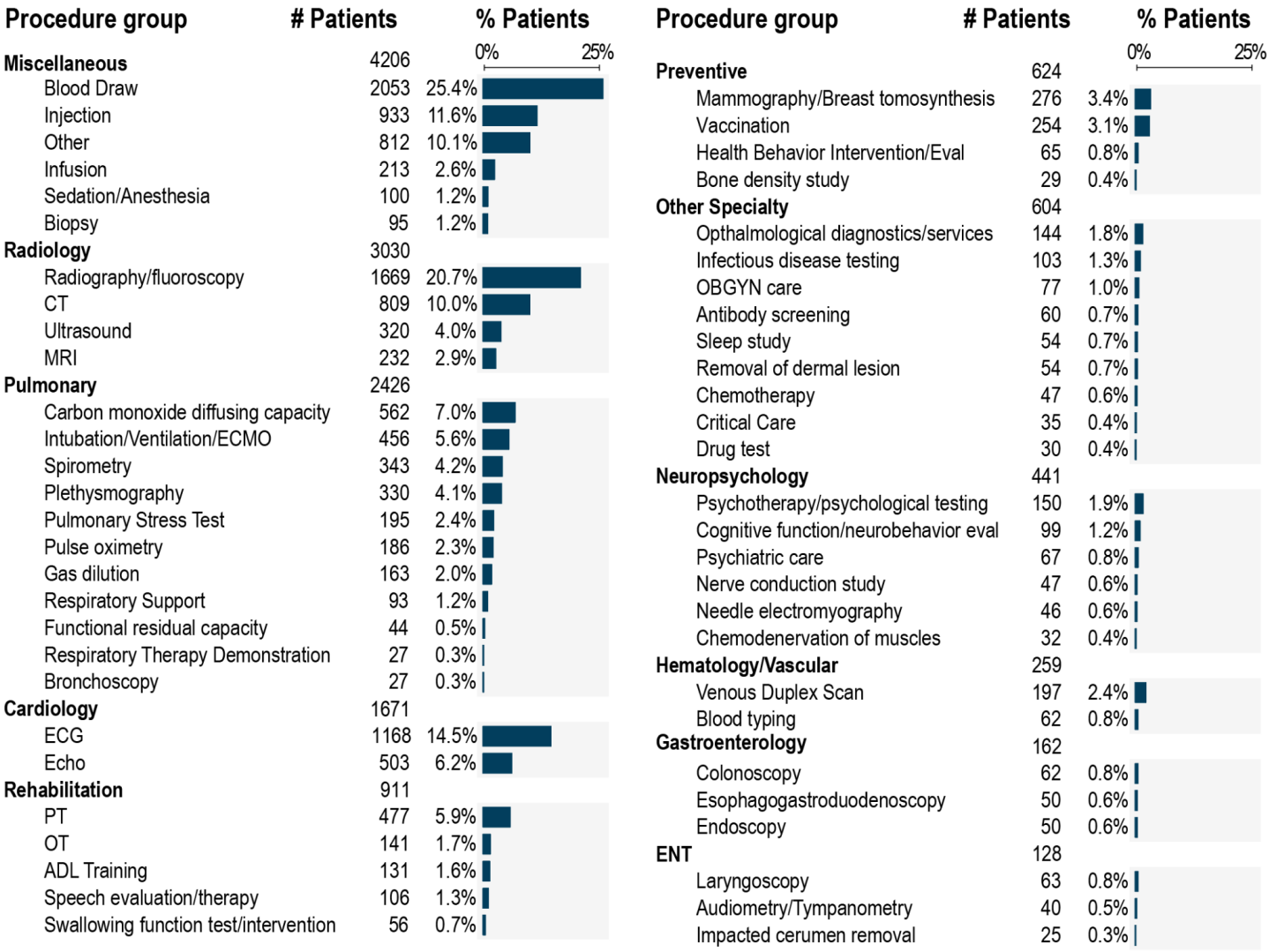


Figure 2. Common procedures for 5,111 patients with a U09.9 code. Procedures shown occur within 60 days after a patient's U09.9 diagnosis. Procedure records that simply reflect that an encounter took place (e.g., CPT 99212, "Office or other outpatient visit") are excluded. Category totals represent unique patient - procedure pairs, not necessarily unique individuals. Procedure classes associated with fewer than 20 patients are not shown, per N3C download policy. An additional 2,963 patients with a U09.9 code had no recorded procedures in the analysis window; percentages are shown relative to all patients in the final study population (8,074).

We also analyzed uptake of the code itself, among sites using the code. There is a rapid increase in use of U09.9 by sites following the code's release (**Figure 3**). Usage of U09.9 post-release is compared with usage of B94.8 ("Sequelae of other specified infectious and parasitic diseases") among COVID positive patients; some sites may have used B94.8 at the CDC's initial recommendation[29] as a placeholder code prior to U09.9's release. Once U09.9 became available, use of B94.8 at the same sites levels off but does not decrease. This suggests that both codes are still being used; indeed, we see both codes used in the records of 1,614 (20%) of N3C patients in our included U09.9 population.

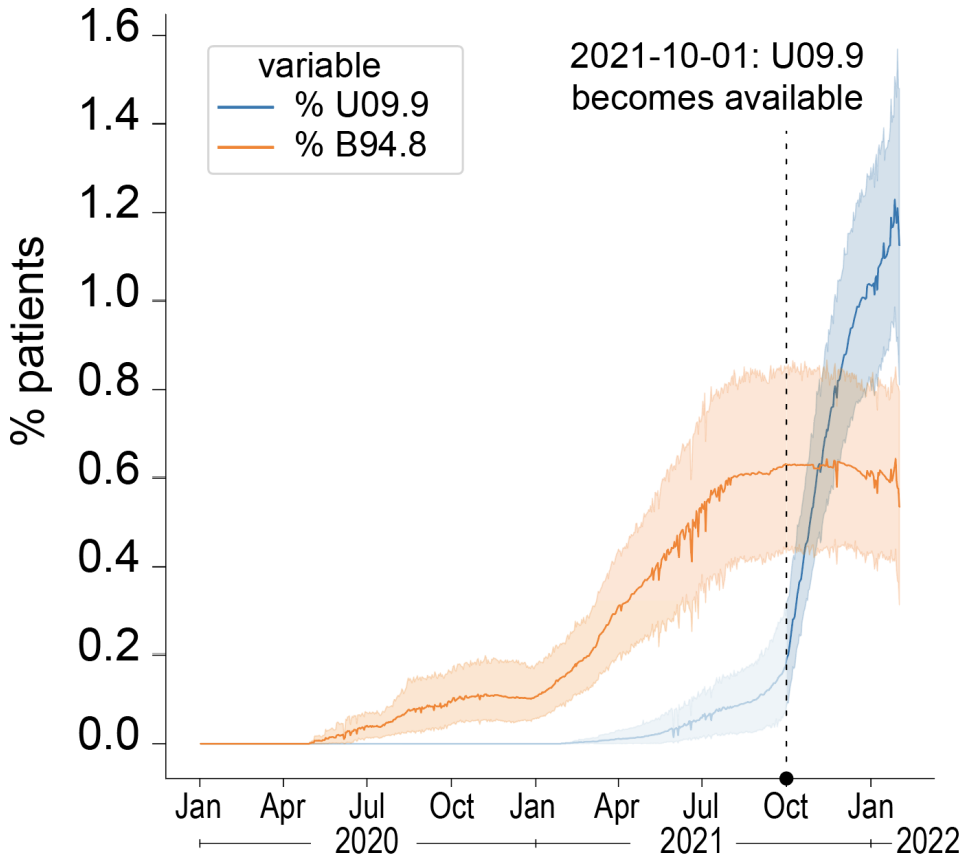
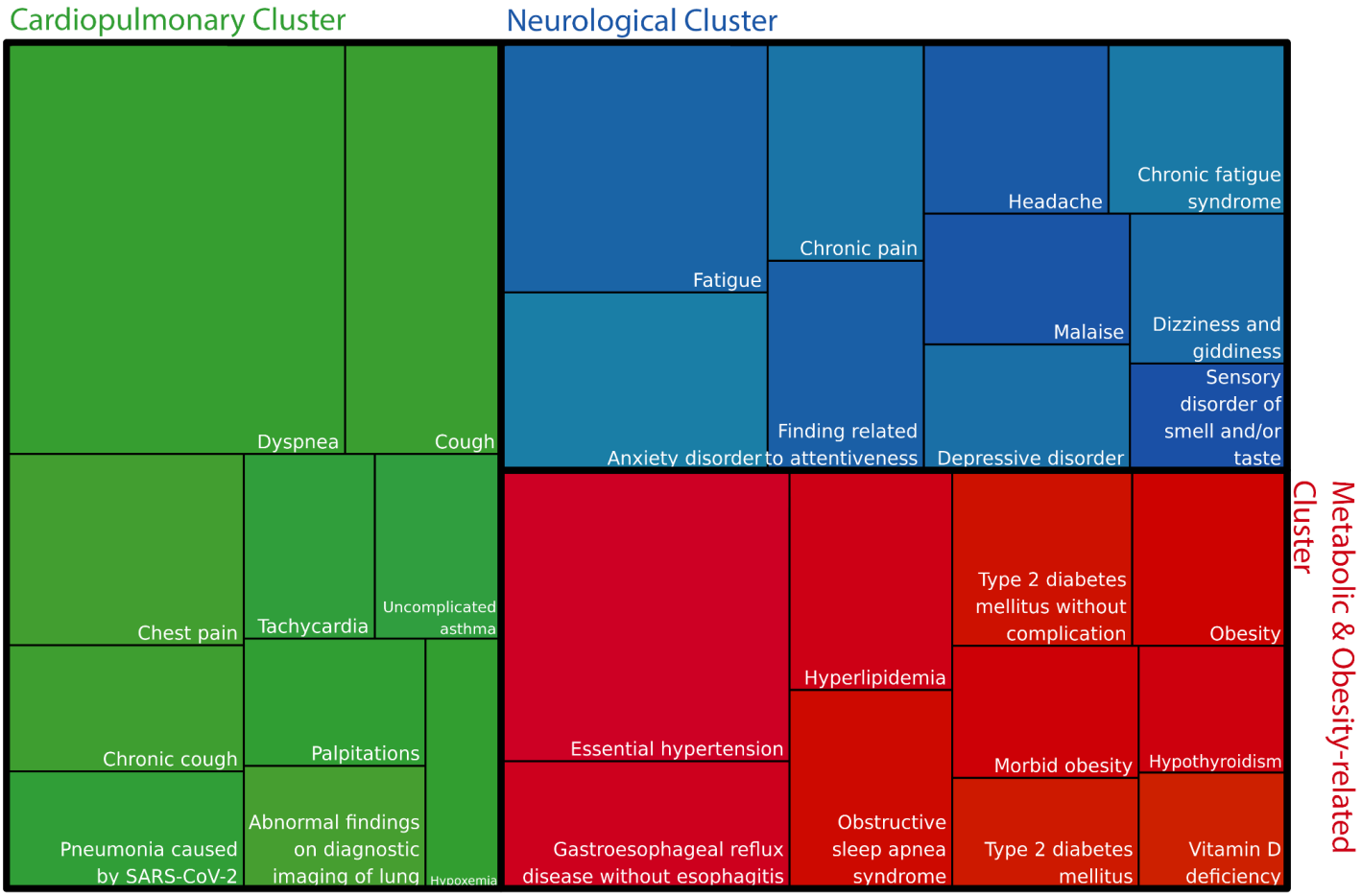


Figure 3. Clinical use of B94.8 levels off as U09.9 becomes available. Prior to U09.9's release, the CDC recommended use of B94.8 ("Sequelae of other specified infectious and parasitic diseases") as a placeholder code to signify Long COVID. Among the 28 sites using U09.9, we plotted the use of B94.8 (orange line) as a percentage of patients who had an acute COVID index (to exclude instances of B94.8, a general purpose code, used for non-COVID-related purposes). Compare this trajectory with U09.9's (blue line), which quickly ramps up in use after October 1, 2021. (U09.9 codes shown prior to that date have been retroactively applied to patients' records.)

The definition of Long COVID[30] includes a wide-ranging list of symptoms and clinical features. Many of those features appear below in **Figure 4**, a visualization of diagnoses that commonly co-occur with U09.9, and each other. The mix of co-occurring diagnoses as well as the clusters produced by the Louvain algorithm change when the cohort is subset into age groups. These age-based clusters are included as **Supplemental Figures 2a-d**. A full accounting of diagnoses co-occurring with U09.9 (i.e., within the analysis window) in at least 20 patients from our cohort is included as **Supplemental Table 1**.

Figure 4. **Clusters of co-occurring diagnoses among patients with a U09.9 code.** When the Louvain algorithm is applied to the top 30 most frequent pairs of co-occurring diagnoses for U09.9 patients (i.e., diagnoses co-occurring in the same patient 0 through 60 days from U09.9 diagnosis date), three distinct clusters emerge (cardiopulmonary, neurological, metabolic). These clusters may represent rough subtypes of Long COVID presentations. The size of each box within a cluster reflects the frequency of that diagnosis relative to others in the diagram. Condition names are derived from the SNOMED CT terminology, mapped from their ICD-10-CM equivalents.



Our findings suggest that Long COVID symptoms and associated functional disability may present differently depending on the patient, but commonly fall into one of these three identified clusters (cardiopulmonary, neurological, metabolic). When stratified by age, the diagnoses within each cluster change somewhat, though the themes remain constant (**Supplemental Figure 2**). For the youngest group (<21 years of age; **Supplemental Figure 2a**), note the appearance of multisystem inflammatory syndrome[31] within the respiratory cluster. Patients aged 65+ (**Supplemental Figure 2d**) were the most distinct, presenting with more chronic diseases associated with aging (e.g. congestive heart failure, atherosclerosis, atrial fibrillation).

Discussion

Diagnosis codes are frequently used as criteria to define patient populations. While diagnosis codes alone may not define a cohort with perfect accuracy, they are a useful mechanism to narrow a population from “everyone in the EHR” to a cohort highly enriched with the condition of interest. Our analysis of U09.9 shows that this code may serve in a similar capacity to identify Long COVID patients. However, temporality and rate of uptake by providers are critical issues that must be considered. U09.9 was released for use nearly two years into the COVID-19 pandemic, resulting in potentially millions of patients with Long COVID who “missed out” on being assigned the code. Moreover, nearly six months after the code was introduced, only about half of N3C sites

have utilized U09.9. Our findings must thus be interpreted through this lens of partial and incremental adoption. More work is needed to understand clinical variability and barriers to uptake by providers.

We investigated whether the use of non-specific coding such as B94.8 ("Sequelae of other specified infectious and parasitic diseases") could be used as a proxy for early case identification. Our findings show B94.8 use increasing among COVID patients from April 2021 to October 2021, indicating a potential shift in clinical practice patterns to code for Long COVID presentation as guided by the Centers for Disease Control[29]. However, B94.8 is used to code for any sequelae of any infectious disease, and is thus not likely specific enough to rely on for Long COVID case ascertainment in the EHR.

The common procedures and medications around the time of U09.9 index provide insight into diagnostics and treatments currently used by providers for patients presenting with Long COVID, for which treatment guidelines remain under development[32–35]. For new diseases where consensus is lacking, care is often ad hoc and informed by both the symptoms that patients present with and the available diagnostics and treatments that providers can offer. The identification and characterization of care patterns is an important step in designing future research to assess the efficacy and outcomes of these interventions. In our analysis of procedure and medication codes, the frequent use of respiratory medications and tests is unsurprising given that pulmonary manifestations of Long COVID are a predominant subtype of symptoms[19]. Interestingly, antibacterials were also used frequently; it is unclear whether patients with Long COVID are more susceptible to bacterial infections, or if there may be overuse of antibiotics in the setting of fluctuating respiratory Long COVID symptoms or viral infections [36,37]. Both systemic and topical corticosteroids were also commonly used, presumably to treat persistent inflammation as a possible mechanism mediating Long COVID symptoms. Other frequently prescribed medication categories, such as cardiac, neuropsychiatric, gastrointestinal, and dermatologic medications, reflect the potential multi-system organ involvement and symptom clusters in Long COVID that we see in the analysis of conditions. Also of interest is the fact that some patients are receiving a number of rehabilitation services in the 60 days after diagnosis, such as physical and occupational therapy, which lends insight into the burden of functional disability for patients with Long COVID.

Our diagnosis clusters suggest that Long COVID may not be a single phenotype, but rather a collection of sub-phenotypes that may benefit from different diagnostics and treatments. This may explain the hesitancy behind uptake of U09.9, as clinical presentation is not universal. Each of these clusters (cardiopulmonary, neurological, and metabolic) contains conditions and symptoms reported in existing Long COVID literature[38], and clearly suggests that the definition of Long COVID is more expansive than lingering respiratory symptoms[39]. Of particular note is the appearance of myalgic encephalomyelitis –a disease which parallels Long COVID in many ways[40–42]–in the neurological cluster, suggesting not only frequent co-occurrence with a U09.9 diagnosis, but also co-occurrence with other neurological symptoms. The metabolic cluster is also hypothesis-generating, and follows prior research on the complex relationship between type II diabetes and COVID-19[43,44]. The cluster differences we see among age groups (**Supplemental Figures 2a-d**) make a strong case for age stratification when studying U09.9, and Long COVID in general. Regardless, given Long COVID's heterogeneity in presentation, course, and outcome, the clustering of symptoms may prove informative for future development of classification and diagnostic criteria.[45]

We also investigated how demographics and social determinants of health may contribute to variation in use of U09.9. We found more women than men presenting with Long COVID across all age groups, consistent with literature and anecdotes from Long COVID clinic providers.[46] When evaluating the U09.9 cohort across age groups and socioeconomic status, Long COVID presentation was more heterogeneous. While our findings do not present a clear socio-demographic trend (see Table 1), the role of access to providers and the economic means to afford Long COVID care should continue to be studied for their role as confounders.

Limitations

All EHR data is limited in that patients with lower access or barriers to care are less likely to be represented. EHR heterogeneity across sites may mean that a U09.9 code at one site does not quite equate to a U09.9 code at another. Moreover, we are not able to know what type of provider issued the U09.9 diagnosis (i.e., specialty), and different clinical organizations have different coding practices.

As the U09.9 code is still quite new and our sample size is limited, we cannot yet confidently label these clusters as clear “Long COVID subtypes.” Rather, these clusters are intended to be hypothesis generating, with additional work underway by the RECOVER consortium to further develop and validate these clusters. It should also be noted that many symptoms are not coded in the EHR (and may, for example, be more likely to appear in free-text notes rather than diagnosis code lists). Future work will incorporate these non-structured sources of symptoms for use in our clustering methodology.

Given the variable uptake of the U09.9 code, it is challenging to accurately identify comparator groups for this population–i.e., the absence of a U09.9 code cannot, at this time, be interpreted as the absence of Long COVID. This will continue to be an issue in future research, especially when evaluating the effect of PASC on patient morbidity and utilization of diagnostic testing and treatments.

Conclusion

The recent release of ICD-10-CM code U09.9 to codify Long COVID will undoubtedly assist with future case ascertainment and computable phenotyping. However, a large number of patients who developed Long COVID prior to October 1, 2021 continue to be burdened with symptoms, and must also be included in data-driven cohort identification efforts for trial recruitment and retrospective analyses. Considering the caveats around rate of uptake among clinicians and late timing of the code’s release, we recommend that when characterizing Long COVID using EHRs, U09.9 should not be used alone, but rather in combination with other strategies such as more complex computable phenotypes[21]. Our findings from the characterization of patients with the U09.9 diagnosis may be of use in refining phenotypes to identify pre-U09.9 patients that might have Long COVID. There is clear utility to the characterization of early use of U09.9, as it represents the first “hook” in EHR data that can be used to identify and assess current diagnostic and treatment patterns at scale. Moreover, given the heterogeneous presentation of Long COVID, clustering of co-existing conditions and potential symptoms may be valuable in informing future development of more detailed criteria for diagnosis of Long COVID and its subtypes.

Acknowledgements

This research was funded by the National Institutes of Health (NIH) Agreement OT2HL161847-01. The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the NIH.

The analyses described in this publication were conducted with data or tools accessed through the NCATS N3C Data Enclave covid.cd2h.org/enclave and supported by NCATS U24 TR002306. This research was possible because of the patients whose information is included within the data from participating organizations (covid.cd2h.org/dtas) and the organizations and scientists (covid.cd2h.org/duas) who have contributed to the on-going development of this community resource[22].

The N3C data transfer to NCATS is performed under a Johns Hopkins University Reliance Protocol # IRB00249128 or individual site agreements with NIH. The N3C Data Enclave is managed under the authority of the NIH; information can be found at https://ncats.nih.gov/n3c/resources. The work was performed under DUR RP-5677B5.

Authorship was determined using ICMJE recommendations.

We gratefully acknowledge contributions from the following N3C core teams:
(Asterisks indicate leads)

- Workstream, subgroup and administrative leaders: David A. Eichmann, Justin Guinney, Warren A. Kibbe, Hongfang Liu, Philip R.O. Payne, Emily R. Pfaff, Peter N. Robinson, Joel H. Saltz, Heidi Spratt, Justin Starren, Christine Suver, Adam B. Wilcox, Andrew E. Williams, Chunlei Wu

- Individuals at the sites who are responsible for creating the datasets and submitting data to N3C Data Ingest and Harmonization Team: Davera Gabriel, Stephanie S. Hong, Emily Pfaff, Kristin Kostka, Harold P. Lehmann, Michele Morris, Matvey B. Palchuk, Xiaohan Tanner Zhang, Richard L. Zhu, Sofia Dard, Tim Schwab
- Phenotype Team (Individuals who create the scripts that the sites use to submit their data, based on the COVID and Long COVID definitions): Emily Pfaff, Marshall Clark, Kristin Kostka, Adam M. Lee, Robert T. Miller, Michele Morris, Matvey B. Palchuk, Kellie M. Walters
- Project Management and Operations Team: Anita Walden*, Will Cooper, Patricia A. Francis, Rafael Fuentes, Alexis Graves, Julie A. McMurry, Shawn T. O'Neil, Usman Sheikh, Elizabeth Zampino
- Partners from NIH and other federal agencies: Joni L. Rutter*, Kenneth R. Gersing*, Samuel Bozzette, Mariam Deacy, Nicole Garbarini, Michael G. Kurilla, Sam G. Michael, Joni L. Rutter, Meredith Temple-O'Connor
- Analytics Team (Individuals who build the Enclave infrastructure, help create codesets, variables, and help Domain Teams and project teams with their datasets): Benjamin Amor*, Katie Rebecca Bradwell, Andrew T. Girvin, Amin Manna, Nabeel Qureshi
- Publication Committee Management Team: Mary Morrison Saltz*, Christine Suver*
- Logic Liaison Core Workgroup: Johanna Loomba, Andrea Zhou, Steve Johnson, Evan French, Alfred (Jerrod) Anzalone, Umit Topaloglu, Amy Olex, Hythem Sidky
- Publication Committee Review Team: Carolyn Bramante, Jeremy Richard Harper, Wenndy Hernandez, Farrukh M Koraishy, Federico Mariona, Amit Saha, Satyanarayana Vedula

Data Partners with Released Data

Stony Brook University — U24TR002306 • University of Oklahoma Health Sciences Center — U54GM104938: Oklahoma Clinical and Translational Science Institute (OCTSI) • West Virginia University — U54GM104942: West Virginia Clinical and Translational Science Institute (WVCTSI) • 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 • Maine Medical Center — U54GM115516: Northern New England Clinical & Translational Research (NNE-CTR) Network • Wake Forest University Health Sciences — UL1TR001420: Wake Forest Clinical and Translational Science Institute • Northwestern University at Chicago — UL1TR001422: Northwestern University Clinical and Translational Science Institute (NUCATS) • University of Cincinnati — UL1TR001425: Center for Clinical and Translational Science and Training • The University of Texas Medical Branch at Galveston — UL1TR001439: The Institute for Translational Sciences • Medical University of South Carolina — UL1TR001450: South Carolina Clinical & Translational Research Institute (SCTR) • University of Massachusetts Medical School Worcester — UL1TR001453: The UMass Center for Clinical and Translational Science (UMCCTS) • University of Southern California — UL1TR001855: The Southern California Clinical and Translational Science Institute (SC CTSI) • Columbia University Irving Medical Center — UL1TR001873: Irving Institute for Clinical and Translational Research • George Washington Children's Research Institute — UL1TR001876: Clinical and Translational Science Institute at Children's National (CTSA-CN) • University of Kentucky — UL1TR001998: UK Center for Clinical and Translational Science • University of Rochester — UL1TR002001: UR Clinical & Translational Science Institute • University of Illinois at Chicago — UL1TR002003: UIC Center for Clinical and Translational Science • Penn State Health Milton S. Hershey Medical Center — UL1TR002014: Penn State Clinical and Translational Science Institute • The University of Michigan at Ann Arbor — UL1TR002240: Michigan Institute for Clinical and Health Research • Vanderbilt University Medical Center — UL1TR002243: Vanderbilt Institute for Clinical and Translational Research • University of Washington — UL1TR002319: Institute of Translational Health Sciences • Washington University in St. Louis — UL1TR002345: Institute of Clinical and Translational Sciences • Oregon Health & Science University — UL1TR002369: Oregon Clinical and Translational Research Institute • University of Wisconsin-Madison — UL1TR002373: UW Institute for Clinical and Translational Research • Rush University Medical Center — UL1TR002389: The Institute for Translational Medicine (ITM) • The University of Chicago — UL1TR002389: The Institute for Translational Medicine (ITM) • University of North Carolina at Chapel Hill — UL1TR002489: North Carolina Translational and Clinical Science Institute • University of Minnesota — UL1TR002494: Clinical and Translational Science Institute • Children's Hospital Colorado — UL1TR002535: Colorado Clinical and Translational Sciences Institute • The University of Iowa — UL1TR002537: Institute for Clinical and Translational Science • The

University of Utah — UL1TR002538: Uhealth Center for Clinical and Translational Science • Tufts Medical Center — UL1TR002544: Tufts Clinical and Translational Science Institute • Duke University — UL1TR002553: Duke Clinical and Translational Science Institute • Virginia Commonwealth University — UL1TR002649: C. Kenneth and Dianne Wright Center for Clinical and Translational Research • The Ohio State University — UL1TR002733: Center for Clinical and Translational Science • The University of Miami Leonard M. Miller School of Medicine — UL1TR002736: University of Miami Clinical and Translational Science Institute • University of Virginia — UL1TR003015: iTHRIV Integrated Translational health Research Institute of Virginia • Carilion Clinic — UL1TR003015: iTHRIV Integrated Translational health Research Institute of Virginia • University of Alabama at Birmingham — UL1TR003096: Center for Clinical and Translational Science • Johns Hopkins University — UL1TR003098: Johns Hopkins Institute for Clinical and Translational Research • University of Arkansas for Medical Sciences — UL1TR003107: UAMS Translational Research Institute • Nemours — U54GM104941: Delaware CTR ACCEL Program • University Medical Center New Orleans — U54GM104940: Louisiana Clinical and Translational Science (LA CaTS) Center • University of Colorado Denver, Anschutz Medical Campus — UL1TR002535: Colorado Clinical and Translational Sciences Institute • Mayo Clinic Rochester — UL1TR002377: Mayo Clinic Center for Clinical and Translational Science (CCaTS) • Tulane University — UL1TR003096: Center for Clinical and Translational Science • Loyola University Medical Center — UL1TR002389: The Institute for Translational Medicine (ITM) • Advocate Health Care Network — UL1TR002389: The Institute for Translational Medicine (ITM) • OCHIN — INV-018455: Bill and Melinda Gates Foundation grant to Sage Bionetworks • The University of Texas Health Science Center at Houston — UL1TR003167: Center for Clinical and Translational Sciences (CCTS) • NorthShore University HealthSystem — UL1TR002389: The Institute for Translational Medicine (ITM) • Weill Medical College of Cornell University — UL1TR002384: Weill Cornell Medicine Clinical and Translational Science Center • Montefiore Medical Center — UL1TR002556: Institute for Clinical and Translational Research at Einstein and Montefiore • Medical College of Wisconsin — UL1TR001436: Clinical and Translational Science Institute of Southeast Wisconsin • George Washington University — UL1TR001876: Clinical and Translational Science Institute at Children's National (CTSA-CN) • Regenstrief Institute — UL1TR002529: Indiana Clinical and Translational Science Institute • Cincinnati Children's Hospital Medical Center — UL1TR001425: Center for Clinical and Translational Science and Training • Boston University Medical Campus — UL1TR001430: Boston University Clinical and Translational Science Institute • The State University of New York at Buffalo — UL1TR001412: Clinical and Translational Science Institute • Brown University — U54GM115677: Advance Clinical Translational Research (Advance-CTR) • Rutgers, The State University of New Jersey — UL1TR003017: New Jersey Alliance for Clinical and Translational Science • Loyola University Chicago — UL1TR002389: The Institute for Translational Medicine (ITM) • New York University — UL1TR001445: Langone Health's Clinical and Translational Science Institute • University of Kansas Medical Center — UL1TR002366: Frontiers: University of Kansas Clinical and Translational Science Institute • Massachusetts General Brigham — UL1TR002541: Harvard Catalyst • University of California, Irvine — UL1TR001414: The UC Irvine Institute for Clinical and Translational Science (ICTS) • University of California, San Diego — UL1TR001442: Altman Clinical and Translational Research Institute • University of California, Davis — UL1TR001860: UCDavis Health Clinical and Translational Science Center • University of California, San Francisco — UL1TR001872: UCSF Clinical and Translational Science Institute • University of California, Los Angeles — UL1TR001881: UCLA Clinical Translational Science Institute • University of Vermont — U54GM115516: Northern New England Clinical & Translational Research (NNE-CTR) Network

Additional Data Partners Who Have Signed a DTA and Whose Data Release is Pending

The Scripps Research Institute — UL1TR002550: Scripps Research Translational Institute • 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 • Emory University — UL1TR002378: Georgia Clinical and Translational Science Alliance • University of New Mexico Health Sciences Center — UL1TR001449: University of New Mexico Clinical and Translational Science Center • Stanford University — UL1TR003142: Spectrum: The Stanford Center for Clinical and Translational Research and Education • Aurora Health Care — UL1TR002373: Wisconsin Network For Health Research • Children's Hospital of Philadelphia — UL1TR001878: Institute for Translational Medicine and Therapeutics • Icahn School of Medicine at Mount Sinai — UL1TR001433: ConduITS Institute for Translational Sciences • Ochsner Medical

Center — U54GM104940: Louisiana Clinical and Translational Science (LA CaTS) Center • HonorHealth — None (Voluntary) • Arkansas Children's Hospital — UL1TR003107: UAMS Translational Research Institute

References

1. Rasmussen SA, Hamosh A. What's in a name? Issues to consider when naming Mendelian disorders. Genet Med. 2020;22: 1573–1575.

2. Biesecker LG, Adam MP, Alkuraya FS, Amemiya AR, Bamshad MJ, Beck AE, et al. A dyadic approach to the delineation of diagnostic entities in clinical genomics. Am J Hum Genet. 2021;108: 8–15.

3. Haendel MA, McMurry JA, Relevo R, Mungall CJ, Robinson PN, Chute CG. A Census of Disease Ontologies. Annual Review of Biomedical Data Science. 2018. pp. 305–331. doi:10.1146/annurev-biodatasci-080917-013459

4. Harrison JE, Weber S, Jakob R, Chute CG. ICD-11: an international classification of diseases for the twenty-first century. BMC Med Inform Decis Mak. 2021;21: 206.

5. Piro A, Distante AE, Tagarelli A. On Allusive Names for the Syphilitic Patient From the 16th to the 19th Century: The Role of Dermatopathology. Am J Dermatopathol. 2017;39: 949–950.

6. Coronaviridae Study Group of the International Committee on Taxonomy of Viruses. The species Severe acute respiratory syndrome-related coronavirus: classifying 2019-nCoV and naming it SARS-CoV-2. Nat Microbiol. 2020;5: 536–544.

7. World Health Organization. Novel Coronavirus(2019-nCoV) Situation Report - 22. 2020 Feb.

8. Worobey M. Dissecting the early COVID-19 cases in Wuhan. Science. 2021;374: 1202–1204.

9. Bhatt AS, McElrath EE, Claggett BL, Bhatt DL, Adler DS, Solomon SD, et al. Accuracy of ICD-10 Diagnostic Codes to Identify COVID-19 Among Hospitalized Patients. J Gen Intern Med. 2021;36: 2532–2535.

10. Bodilsen J, Leth S, Nielsen SL, Holler JG, Benfield T, Omland LH. Positive Predictive Value of ICD-10 Diagnosis Codes for COVID-19. Clin Epidemiol. 2021;13: 367–372.

11. Lynch KE, Viernes B, Gatsby E, DuVall SL, Jones BE, Box TL, et al. Positive predictive value of COVID-19 ICD-10 diagnosis codes across calendar time and clinical setting. Clin Epidemiol. 2021;13: 1011–1018.

12. Doykov I, Hällqvist J, Gilmour KC, Grandjean L, Mills K, Heywood WE. "The long tail of Covid-19" - The detection of a prolonged inflammatory response after a SARS-CoV-2 infection in asymptomatic and mildly affected patients. F1000Res. 2020;9. doi:10.12688/f1000research.27287.2

13. Marshall M. The lasting misery of coronavirus long-haulers. Nature. 2020;585: 339–341.

14. Nabavi N. Long covid: How to define it and how to manage it. BMJ. 2020;370. doi:10.1136/bmj.m3489

15. Han Q, Zheng B, Daines L, Sheikh A. Long-Term Sequelae of COVID-19: A Systematic Review and Meta-Analysis of One-Year Follow-Up Studies on Post-COVID Symptoms. Pathogens. 2022;11. doi:10.3390/pathogens11020269

16. Soriano JB, Murthy S, Marshall JC, Relan P, Diaz JV, WHO Clinical Case Definition Working Group on Post-COVID-19 Condition. A clinical case definition of post-COVID-19 condition by a Delphi consensus. Lancet Infect Dis. 2021. doi:10.1016/S1473-3099(21)00703-9

17. Nasserie T, Hittle M, Goodman SN. Assessment of the Frequency and Variety of Persistent Symptoms Among Patients With COVID-19: A Systematic Review. JAMA Netw Open. 2021;4: e2111417.

18. McCorkell L, Assaf GS, Davis HE, Wei H, Akrami A. Patient-Led Research Collaborative: embedding patients in the Long COVID narrative. PAIN Reports. 2021. p. e913. doi:10.1097/pr9.0000000000000913

19. Davis HE, Assaf GS, McCorkell L, Wei H, Low RJ, Re'em Y, et al. Characterizing long COVID in an international cohort: 7 months of symptoms and their impact. EClinicalMedicine. 2021;38: 101019.

20. RECOVER: Researching COVID to Enhance Recovery. In: RECOVER: Researching COVID to Enhance Recovery [Internet]. [cited 15 Apr 2022]. Available: https://recovercovid.org

21. Pfaff ER, Girvin AT, Bennett TD, Bhatia A, Brooks IM, Deer RR, et al. Who has long-COVID? A big data approach. doi:10.1101/2021.10.18.21265168

22. Haendel MA, Chute CG, Bennett TD, Eichmann DA, Guinney J, Kibbe WA, et al. The National COVID Cohort Collaborative (N3C): Rationale, design, infrastructure, and deployment. J Am Med Inform Assoc. 2021;28: 427–443.

23. Pfaff ER, Girvin AT, Gabriel DL, Kostka K, Morris M, Palchuk MB, et al. Synergies between centralized and federated approaches to data quality: a report from the national COVID cohort collaborative. J Am Med Inform Assoc. 2022;29: 609–618.

24. Phenotype_Data_Acquisition Wiki. Github; Available: https://github.com/National-COVID-Cohort-Collaborative/Phenotype_Data_Acquisition

25. Blondel VD, Guillaume J-L, Lambiotte R, Lefebvre E. Fast unfolding of communities in large networks. Journal of Statistical Mechanics: Theory and Experiment. 2008. p. P10008. doi:10.1088/1742-5468/2008/10/p10008

26. Pons P, Latapy M. Computing Communities in Large Networks Using Random Walks. Journal of Graph Algorithms and Applications. 2006. pp. 191–218. doi:10.7155/jgaa.00124

27. Girvan M, Newman MEJ. Community structure in social and biological networks. Proceedings of the National Academy of Sciences. 2002. pp. 7821–7826. doi:10.1073/pnas.122653799

28. WHOCC. Structure and principles. [cited 11 Mar 2022]. Available: https://www.whocc.no/atc/structure_and_principles/

29. CDC. Public health recommendations. In: Centers for Disease Control and Prevention [Internet]. 10 Aug 2021 [cited 5 Apr 2022]. Available: https://www.cdc.gov/coronavirus/2019-ncov/hcp/clinical-care/post-covid-public-health-recs.html

30. Coronavirus disease (COVID-19): Post COVID-19 condition. [cited 28 Mar 2022]. Available: https://www.who.int/news-room/questions-and-answers/item/coronavirus-disease-(covid-19)-post-covid-19-condition

31. CDC. Information for healthcare providers about multisystem inflammatory syndrome in children (MIS-C). In: Centers for Disease Control and Prevention [Internet]. 29 Dec 2021 [cited 5 Apr 2022]. Available: https://www.cdc.gov/mis/mis-c/hcp/index.html?CDC_AA_refVal=https%3A%2F%2Fwww.cdc.gov%2Fmis%2Fhcp%2Findex.html

32. Mikkelsen ME, Abramoff B, Elmore JG, Kunins L. COVID-19: evaluation and management of adults following acute viral illness. UpToDate Waltham, MA: UpToDate Inc (Accessed on October 30, 2021 [Google Scholar]. 2021. Available: https://www.medilib.ir/uptodate/show/129312

33. Herrera JE, Niehaus WN, Whiteson J, Azola A, Baratta JM, Fleming TK, et al. Multidisciplinary collaborative consensus guidance statement on the assessment and treatment of fatigue in postacute sequelae of SARS-CoV-2 infection (PASC) patients. PM R. 2021;13: 1027–1043.

34. Fine JS, Ambrose AF, Didehbani N, Fleming TK, Glashan L, Longo M, et al. Multi-disciplinary collaborative consensus guidance statement on the assessment and treatment of cognitive symptoms in patients with post-acute sequelae of SARS-CoV-2 infection (PASC). PM and R. 2022;14: 96–111.

35. Maley JH, Alba GA, Barry JT, Bartels MN, Fleming TK, Oleson CV, et al. Multi-disciplinary collaborative consensus guidance statement on the assessment and treatment of breathing discomfort and respiratory sequelae in patients with post-acute sequelae of SARS-CoV-2 infection (PASC). PM and R. 2022;14: 77–95.

36. Harris AM, Hicks LA, Qaseem A, High Value Care Task Force of the American College of Physicians and for the Centers for Disease Control and Prevention. Appropriate antibiotic use for acute respiratory tract infection in adults: Advice for high-value care from the American college of physicians and the centers for disease control and prevention. Ann Intern Med. 2016;164: 425–434.

37. Havers FP, Hicks LA, Chung JR, Gaglani M, Murthy K, Zimmerman RK, et al. Outpatient Antibiotic Prescribing for Acute Respiratory Infections During Influenza Seasons. JAMA Netw Open. 2018;1: e180243.

38. Deer RR, Rock MA, Vasilevsky N, Carmody L, Rando H, Anzalone AJ, et al. Characterizing Long COVID: Deep Phenotype of a Complex Condition. EBioMedicine. 2021;74: 103722.

39. Clinical Services, Systems. A clinical case definition of post COVID-19 condition by a Delphi consensus, 6 October 2021. World Health Organization; 6 Oct 2021 [cited 1 Apr 2022]. Available: https://www.who.int/publications/i/item/WHO-2019-nCoV-Post_COVID-19_condition-Clinical_case_definition-2021.1

40. Komaroff AL, Lipkin WI. Insights from myalgic encephalomyelitis/chronic fatigue syndrome may help unravel the pathogenesis of postacute COVID-19 syndrome. Trends Mol Med. 2021;27: 895–906.

41. Proal AD, VanElzakker MB. Long COVID or Post-acute Sequelae of COVID-19 (PASC): An Overview of Biological Factors That May Contribute to Persistent Symptoms. Front Microbiol. 2021;12: 698169.

42. Komaroff AL, Bateman L. Will COVID-19 Lead to Myalgic Encephalomyelitis/Chronic Fatigue Syndrome? Front Med. 2020;7: 606824.

43. Scherer PE, Kirwan JP, Rosen CJ. Post-acute sequelae of COVID-19: A metabolic perspective. Elife. 2022;11. doi:10.7554/eLife.78200

44. Fernández-de-Las-Peñas C, Guijarro C, Torres-Macho J, Velasco-Arribas M, Plaza-Canteli S, Hernández-Barrera V, et al. Diabetes and the Risk of Long-term Post-COVID Symptoms. Diabetes. 2021;70: 2917–2921.

45. Aggarwal R, Ringold S, Khanna D, Neogi T, Johnson SR, Miller A, et al. Distinctions between diagnostic and classification criteria? Arthritis Care Res . 2015;67: 891–897.

46. Durstenfeld MS, Hsue PY, Peluso MJ, Deeks SG. Findings From Mayo Clinic's Post-COVID Clinic: PASC Phenotypes Vary by Sex and Degree of IL-6 Elevation. Mayo Clin Proc. 2022;97: 430.