Generalisable long COVID subtypes: Findings from the NIH N3C and RECOVER programmes

Justin T. Reese,[a] Hannah Blau,[b] Elena Casiraghi,[a,c] Timothy Bergquist,[d] Johanna J. Loomba,[e] Tiffany J. Callahan,[f] Bryan Laraway,[g] Corneliu Antonescu,[h] Ben Coleman,[b] Michael Gargano,[b] Kenneth J. Wilkins,[i] Luca Cappelletti,[c] Tommaso Fontana,[c] Nariman Ammar,[j] Blessy Antony,[k] T. M. Murali,[k] J. Harry Caufield,[a] Guy Karlebach,[b] Julie A. McMurry,[g] Andrew Williams,[l,m,n] Richard Moffitt,[o] Jineta Banerjee,[d] Anthony E. Solomonides,[p] Hannah Davis,[q] Kristin Kostka,[n] Giorgio Valentini,[c] David Sahner,[r] Christopher G. Chute,[s] Charisse Madlock-Brown,[j] Melissa A. Haendel,[g] and Peter N. Robinson,[b,t,*] on behalf of the N3C Consortium[u] and the RECOVER Consortium[v]

[a]Environmental Genomics and Systems Biology Division, Lawrence Berkeley National Laboratory, Berkeley, CA, USA
[b]The Jackson Laboratory for Genomic Medicine, 10 Discovery Drive, Farmington, CT, USA
[c]AnacletoLab, Dipartimento di Informatica, Università Degli Studi di Milano, Milan, Italy
[d]Sage Bionetworks, Seattle, WA, USA
[e]The Integrated Translational Health Research Institute of Virginia (iTHRIV), University of Virginia, Charlottesville, VA, USA
[f]Department of Biomedical Informatics, Columbia University Irving Medical Center, New York, NY, USA
[g]Departments of Biomedical Informatics and Pediatrics, University of Colorado Anschutz Medical Campus, Aurora, CO, USA
[h]University of Arizona - Banner Health, Phoenix, AZ, USA
[i]Biostatistics Program, National Institute of Diabetes and Digestive and Kidney Diseases, National Institutes of Health, Bethesda, MD, USA
[j]Health Science Center, University of Tennessee, Memphis, TN, USA
[k]Department of Computer Science, Virginia Tech, Blacksburg, VA, USA
[l]Tufts Medical Center Clinical and Translational Science Institute, Tufts Medical Center, Boston, MA, USA
[m]Tufts University School of Medicine, Institute for Clinical Research and Health Policy Studies, Boston, MA, USA
[n]Northeastern University, OHDSI Center at the Roux Institute, Boston, MA, USA
[o]Department of Biomedical Informatics and Stony Brook Cancer Center, Stony Brook University, Stony Brook, NY, USA
[p]HealthSystem Research Institute, NorthShore University, Evanston, IL, USA
[q]Patient-Led Research Collaborative, NY, USA
[r]Axle Informatics, Rockville, MD, USA
[s]Schools of Medicine, Public Health and Nursing, Johns Hopkins University, Baltimore, MD, USA
[t]Institute for Systems Genomics, University of Connecticut, Farmington, CT, USA

eBioMedicine
2023;87: 104413
Published Online
https://doi.org/10.1016/j.ebiom.2022.104413

Summary

Background Stratification of patients with post-acute sequelae of SARS-CoV-2 infection (PASC, or long COVID) would allow precision clinical management strategies. However, long COVID is incompletely understood and characterised by a wide range of manifestations that are difficult to analyse computationally. Additionally, the generalisability of machine learning classification of COVID-19 clinical outcomes has rarely been tested.

Methods We present a method for computationally modelling PASC phenotype data based on electronic healthcare records (EHRs) and for assessing pairwise phenotypic similarity between patients using semantic similarity. Our approach defines a nonlinear similarity function that maps from a feature space of phenotypic abnormalities to a matrix of pairwise patient similarity that can be clustered using unsupervised machine learning.

Findings We found six clusters of PASC patients, each with distinct profiles of phenotypic abnormalities, including clusters with distinct pulmonary, neuropsychiatric, and cardiovascular abnormalities, and a cluster associated with broad, severe manifestations and increased mortality. There was significant association of cluster membership with a range of pre-existing conditions and measures of severity during acute COVID-19. We assigned new patients from other healthcare centres to clusters by maximum semantic similarity to the original patients, and showed that the clusters were generalisable across different hospital systems. The increased mortality rate originally identified in one cluster was consistently observed in patients assigned to that cluster in other hospital systems.

Interpretation Semantic phenotypic clustering provides a foundation for assigning patients to stratified subgroups for natural history or therapy studies on PASC.

*Corresponding author.
E-mail address: Peter.Robinson@jax.org (P.N. Robinson).
[u]National COVID Cohort Collaborative.
[v]Researching COVID to Enhance Recovery.

Funding NIH (TR002306/OT2HL161847-01/OD011883/HG010860), U.S.D.O.E. (DE-AC02-05CH11231), Donald A. Roux Family Fund at Jackson Laboratory, Marsico Family at CU Anschutz.

Keywords: Long COVID; COVID-19; Semantic similarity; Machine learning; Precision medicine; Human Phenotype Ontology

Research in context

Evidence before this study

Previous studies demonstrated that a substantial fraction of those infected with SARS-CoV-2 go on to develop long COVID. Current evidence is insufficient to determine whether distinct subtypes of long COVID exist. We searched PubMed for studies on data-driven clustering of individuals with post-acute sequelae SARS-CoV-2 infection (PASC). One prospective study on 233 individuals (PMID: 35265728) employed multiple correspondence analysis (MCA) on commonly reported symptoms and identified three clusters characterised by the predominance of symptoms such as pain or cardiovascular manifestations, or by a paucity of symptoms. We did not identify previous studies that used cluster analysis of comprehensive phenotypic manifestations recorded in EHR data from individuals with long COVID, nor did we find computational methods to assess the generalisability of the resulting clusters across different patient cohorts.

Added value of this study

We describe an unsupervised machine learning method that uses semantic similarity of phenotype data to stratify long COVID patients into clusters. These clusters correlate with pre-existing comorbidities, markers of clinical severity of COVID-19, and mortality in the post-acute long COVID-19 period.

Implications of all the available evidence

This study demonstrates the existence of subtypes of long COVID that differ with respect to clinical outcome and pre-existing clinical features. This demonstrates the feasibility of stratifying long COVID patients and provides a foundation for characterising the natural history of long COVID and developing precision clinical management strategies.

Introduction

Hundreds of millions of cases of acute Coronavirus disease 2019 (COVID-19) have been recorded since the beginning of the pandemic, and more than six million deaths had been reported by the World Health Organisation (WHO) by the end of March, 2022.[1] The clinical presentation of COVID-19 ranges from asymptomatic infection to fatal disease, with many patients continuing to have heterogeneous, long-term, multi-system symptoms including fatigue, post-exertional malaise, dyspnea, cough, chest pain, palpitations, headache, arthralgia, weakness (asthenia), paresthesias, diarrhoea, alopecia, rash, impaired balance, and memory or cognitive dysfunction.[2,3] Although there is still no detailed and widely accepted case definition, post-acute sequelae of SARS-CoV-2 infection (PASC, long-haul COVID or long COVID) generally refers to a range of persistent or new symptoms beyond three or four weeks of the initial infection.[4–7] The NIH REsearching COVID to Enhance Recovery (RECOVER) Initiative program defines PASC as 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). The WHO has developed a case definition of "post COVID-19 condition" suggesting that the syndrome is usually diagnosed several months after the onset of acute symptoms of COVID-19 based on new-onset or lingering symptoms (e.g., fatigue, dyspnea, cognitive dysfunction) which cannot be explained by an alternative aetiology and which continue for at least 2 months.[8] In this work, we will use the term long COVID to refer to patients given a diagnosis using the newly introduced ICD-10 U09.9 code ("Post COVID-19 condition"). Although presumably only a small subset of all individuals with PASC are identified by this code, we chose to focus on it since it marks patients diagnosed with PASC by a physician.

Our understanding of the natural history of long COVID is still incomplete. Limited emerging evidence suggests the existence of clinical subtypes or clusters characterised by the predominance of symptoms such as pain or cardiovascular manifestations, or by a paucity of symptoms.[9] However, computational methods to characterise long COVID subtypes based on comprehensive phenotypic analysis are lacking, as are approaches to assess the generalisability of the resulting clusters across different patient cohorts. In this study, we constructed a cohort of 6469 patients diagnosed with

long COVID using the U09.9 code from multicentre electronic health record (EHR) data available through the National COVID Cohort Collaborative (N3C), a harmonised EHR repository with 5,434,528 COVID-19 positive patients as of August 10, 2022. Previous work mapped 287 unique clinical findings previously reported in studies of long COVID[10] to the Human Phenotype Ontology (HPO), which is widely used to support differential diagnosis and translational research in human genetics.[11,12] Here, we introduce an approach that calculates the semantic similarity between patients by transforming EHR data to phenotypic profiles using the HPO. The method identifies distinct clusters of long COVID patients that show highly significant correlations with pre-existing conditions and generalise across different hospital systems.

Methods

Ethics

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.

Setting

We obtained patient data from the National COVID Cohort Collaborative (N3C; covid.cd2h.org). N3C aggregates and integrates EHR data across multiple clinical organisations in the United States, including the Clinical and Translational Science Awards (CTSA) Program hubs. N3C harmonises EHR data across four clinical data models and provides a unified analytical platform in which data are encoded using version 5.3.1 of the Observational Medical Outcomes Partnership (OMOP) common data model.[13]

Cohort

The Centers for Disease Control (CDC) announced an International Classification of Diseases, version 10 (ICD-10) code (U09.9) for emergency/provisional use in the United States of America on June 30, 2021. The code represents Post COVID-19 condition, unspecified. Use of the code was approved for implementation effective October 1, 2021. The code should be used for patients with a history of probable or confirmed SARS CoV-2 infection who are identified with a post-COVID condition. The data freeze date was August 10, 2022 (v87 release). Only patients with an initial COVID-19 diagnosis within the Enclave were included in the cohort. At the time of the data freeze for this analysis, 38 participating data partners were using the code, and a total of 20,532 patients were coded in this way.

Human Phenotype Ontology (HPO)

The HPO is a rich representation of the diversity of phenotypic features associated with human disease and is the de facto standard for the computational analysis and exchange of phenotype data in human genetics.[11,14–18] The HPO comprises over 16,000 terms that denote specific phenotypic abnormalities at increasingly specific granularity, for example, *Atrial septal defect* (HP:0001631) and *Interrupted inferior vena cava with azygous continuation* (HP:0011671). We recently identified 287 unique clinical findings reported in cohorts of patients with long COVID and mapped them to existing HPO terms and in some cases created new HPO terms to cover COVID-specific features such as *Pseudo-chilblains on toes* (HP:0034036).[10] The 2022-08-11 release of the HPO was used in our study.

Mapping OMOP codes to HPO terms

To obtain mappings between standard OMOP condition concept identifiers and HPO concepts, we used OMOP2OBO (https://github.com/callahantiff/OMOP2OBO) and LOINC2HPO.[19,20] The OMOP2OBO algorithm was developed to generate mappings between clinical vocabularies in the OMOP common data model and eight Open Biomedical Foundry ontologies[21] spanning diseases, phenotypes, anatomical entities, organisms, chemicals, vaccines, and proteins. Using this algorithm, a large-scale set of mappings was developed, which includes 92,367 conditions, 8615 drug ingredients, and 10,673 measurement results.[20] For this project, we filtered the v1.0.0 release of mappings to only include exact 1:1 mappings at the concept level. This mapping set aligned 4767 OMOP concept IDs to 3804 unique HPO concepts (1.25 OMOP concept IDs/HPO concept). To apply LOINC2HPO mappings from OMOP to HPO concepts, we reimplemented the LOINC to HPO mappings in the N3C Enclave. For any HPO term that was among the 287 HPO terms associated with long COVID, we determined for each patient in our study group the LOINC codes present in the measurement OMOP table determined to be 'low', 'high', or 'positive' compared to the reference range for the test in question, and assigned the HPO term to the patient if the test occurred during the long COVID period for that patient (starting 28 days after diagnosis of acute COVID-19 for outpatients, and 28 days after hospitalisation for inpatients).

Specificity-weighted fuzzy phenotype matching

We previously developed a method called Phenomizer for clinical diagnostics that uses the semantic structure of the HPO to weight clinical features on the basis of specificity and to identify those clinical features that best distinguish among the top candidate differential diagnoses.[22] The algorithm represents the clinical

specificity of a finding as the information content (IC) of a term. Given a set of diseases of interest in the differential diagnosis process, the frequency of each HPO term is defined as the proportion of diseases in a database that are annotated by the term or any of its descendant terms (for instance, the HPO resource currently comprises 8260 Mendelian diseases).[12] The IC is then defined as the negative natural logarithm of the term frequency.[23] The annotation propagation rule applies to all terms in the HPO. That is, if a disease is annotated to the term t, it is implicitly annotated to all ancestors of t recursively. For instance, Marfan syndrome is annotated to *Aortic root aneurysm* (HP:0002616), and it is therefore implicitly annotated to the parent term *Thoracic aortic aneurysm* (HP:0012727) and its parent term *Aortic aneurysm* (HP:0004942), and so on. Thus, the IC of terms increases as we move from the root term of the HPO ontology to the more specific descendant terms.

To define the similarity between any two HPO terms t_1 and t_2, we find the most specific common ancestor of t_1 and t_2 in the HPO hierarchy, which we call the Most Informative Common Ancestor of t_1 and t_2, $MICA(t_1, t_2)$. We calculate its IC as $\mathrm{IC}(MICA(t_1, t_2))$ In essence, this procedure leverages the ontological structure of the HPO to perform specificity-weighted fuzzy matching.

In the Phenomizer algorithm, the similarity between a set of query terms (symptoms, signs, etc.) entered by a physician for an individual case is used to calculate a similarity score for each of the diseases in the HPO database as an aid in differential diagnosis. In the current work, we adapt this algorithm to implement semantic phenotypic-based clustering by using the Phenomizer framework to calculate a matrix of pairwise phenotypic similarities between all patients in the long COVID cohort. In the following, we represent the set of n long COVID patients as $p_1, p_2, ..., p_n \in P$. The set of m HPO terms associated with patient i is represented as $t_1, t_2, ..., t_m \in p_i$. Then the similarity from patient p_i to p_j is calculated as

$$\mathrm{sim}\left(p_i \to p_j\right) = \frac{1}{m} \sum_{t_1 \in p_i} max_{t_2 \in p_j} \mathrm{IC}(MICA(t_1, t_2))$$

This equation is not symmetric, so the final similarity score is calculated as

$$sim\left(p_i, p_j\right) = 0.5 \times sim\left(p_i \to p_j\right) + 0.5 \times sim\left(p_j \to p_i\right)$$

k-means clustering

For n patients, we calculated a similarity matrix $X^{n \times n}$ using the Phenomizer algorithm. We then applied k-means clustering to partition the patients into c clusters, denoted $C_1, C_2, ..., C_c$, where C_i is the set of n_i objects in cluster i and c is the number of clusters (a user-chosen hyperparameter). We randomly initialized the c cluster centroids so that the centroids were maximally distant from one another.[24,25] Clusters were then formed iteratively such that the Euclidean distance between the vector that represents any object and the centroid vector of its cluster was at least as small as that between the object and any of the other clusters. In each iteration, objects were moved to the cluster with the closest centroid, following which the centroids were recalculated until no further improvement was obtained or the maximum number of 100 iterations was reached.[26]

We used the elbow method to choose a suitable number of clusters, as k-means clustering does not provide this value. The elbow method computes the total within-cluster sum of squares error (SSE) for each candidate number of clusters. The SSE is plotted against the number of clusters and an 'elbow' in the curve is used to determine the number of clusters.

Statistics

Assessing cluster reproducibility between data partners

We first performed clustering on patients from the data partner with the greatest number of U09.9 long COVID patients. To maintain data privacy, we refer to this as data partner 1. We then assessed reproducibility of clustering results in data partners 2–6 (hereafter referred to as test data partners) as explained below. This approach was chosen given the inherent challenge owing to the lack of a generally applicable method for assessing any given clustering approach.[26–28] The HPO terms for patients from data partner 1 and their assignment to k-means clusters were recorded. We reasoned that if the clustering results in data partner 1 are generalizable, then patients of the test data partners will tend to display more similarity to one or other cluster of data partner 1 than one would expect by chance. To this end we introduce a similarity measure s between a patient p and cluster C of patients that assesses the average similarity of patient p to all patients in cluster C:

$$\mathrm{s}(\mathrm{p}, \mathrm{C}) = \mathrm{mean}_{q \in \mathrm{C}} \mathrm{sim}(\mathrm{p}, \mathrm{q})$$

Assuming we have k clusters from data partner 1, then a normalized weighted similarity vector can be calculated for each patient p from a test data partner as $[s_1, s_2, ..., s_k] / \sum_i s_i$ where the index refers to the cluster. In other words, $s_i = s(p, C_i)$ is the similarity between the test patient p and the cluster C_i. If the patient is equally similar to each of the k clusters, then $s_1 = s_2 = ... = s_k = \frac{1}{k}$. If, on the other hand, the patient is much more similar to one of the clusters, say cluster i, then we expect $s_i \gg s_j$, for $j \neq i$. We therefore define the test statistic $s_{max} = \max_i s_i$ for patient p. To assess generalizability, we calculate s_{max} for each patient p in the test

data partner and take the mean value of s_{max} between a random 10% sample of patients in the test data partner and a random 10% sample of patients from data partner 1 as our test statistic $\overline{s_{max}}$. To generate a null distribution of this statistic, we create 1000 permuted cluster assignments by assigning a random 10% sample of patients from data partner 1 uniformly at random to one of the k clusters. We compute the test statistic for each of these random cluster assignments as described above and record the mean, μ, and standard deviation, σ, of these values. We present the results as a z score calculated as $z = (x - \mu)/\sigma$, where x = $\overline{s_{max}}$. Note that this procedure does not cluster patients from the test data partners. Instead, it calculates similarities of patients from the test data partners to the clusters defined in data partner 1.

Assessing covariate distribution

The HPO terms assessed in the above procedures were derived from clinical data at least 28 days after the initial bout of COVID-19. We analysed additional clinical covariates covering items such as comorbidities and medications prior to and during acute COVID-19 (Supplemental Tables S12 and S13). Categorical variables were assessed with a chi-squared test if at least five counts were present for each cell of the contingency table and numerical variables were assessed with one-way ANOVA. Analysis was done using R version 3.5.1.

Post hoc testing

To improve the informativeness of the cluster descriptions, post hoc tests were conducted to detect differences in the distribution of covariates deemed as significantly different between clusters by the chi-squared test or one-way ANOVA. Pairwise chi-squared tests with Bonferroni correction were performed for categorical covariates (to assess which specific category distributions are significantly different from the others), while non-parametric Dunn's test with Bonferroni correction was used for numeric covariates to assess which means are significantly different from the others.[29] To summarise the results, the Compact Letter Display (CLD) method[30] was used.

Role of the funding source

The funders had no role in study design, data collection, analysis, interpretation, writing of the report, or in the decision to submit for publication.

Results

A cohort of patients diagnosed with PASC

As of August 10, 2022, the N3C platform ("Enclave") contained data for 5,434,528 patients diagnosed with acute COVID-19, and 38 data partners had begun to use the newly introduced ICD-10 diagnosis code U09.9 for use in US hospitals to denote Post COVID-19 condition. These 38 data partners provided data for 20,532 patients with this diagnosis (Fig. 1). Phenotypic features observed in the post-acute COVID-19 period were mapped from OMOP codes to HPO terms. The post-acute COVID-19 period was defined as starting 28 days after the earliest COVID-19 index date for outpatients, and 28 days after the end of hospitalisation for inpatients. The COVID-19 index date for each patient was defined as the earliest date of any positive PCR or antigen SARS-CoV-2 test or diagnosis with ICD-10 U07.1 (acute COVID-19).

Phenotypic clustering of patients with long COVID

We hypothesised that consistent subgroups of patients with long COVID can be defined based on the spectrum of phenotypic features in the patients' electronic health records (EHR). Our previous analysis identified 287 clinical findings previously reported in studies on long COVID and coded these findings using terms of the Human Phenotype Ontology (HPO).[10,12] Numerous algorithms have been developed that define a fuzzy, specificity-weighted similarity metric between a patient and a computational disease model or between pairs of patients.[31–34] Here, we adapted an algorithm called Phenomizer that calculates semantic similarity between a pair of patients based on their phenotypic features (Methods).[22] Common clustering methods define feature vectors with one field for each measured quantity. In principle, one could define a feature vector with 287 dimensions, one for each of the clinical findings related to long COVID, and for each clinical finding identified in a patient, a "1" would be placed in the corresponding field of the vector, otherwise a "0". Patient similarity could then be measured by calculating the cosine between any two such vectors, which essentially counts the number of exact matches normalised by the total number of features in each vector. This procedure would not capture the fact that some features are similar. For instance, although dyspnea and hypoxemia are both abnormalities of respiratory physiology, they are represented by different fields in the feature vector and thus if one patient was recorded to have dyspnea and another hypoxemia, this would not contribute to the similarity score. Another drawback to a simple 0/1 feature vector for the 287 clinical findings would be that matches between more or less specific findings would be weighted equally. The Phenomizer algorithm uses the structure of the ontological hierarchy to identify partial matches between related clinical findings, and it leverages the information content of each term, which is a measure of specificity, to weight the matches. The Phenomizer is thus a nonlinear mapping from the original feature space of clinical findings to a pairwise similarity matrix that implements a fuzzy, specificity-weighted matching strategy. The resulting similarity

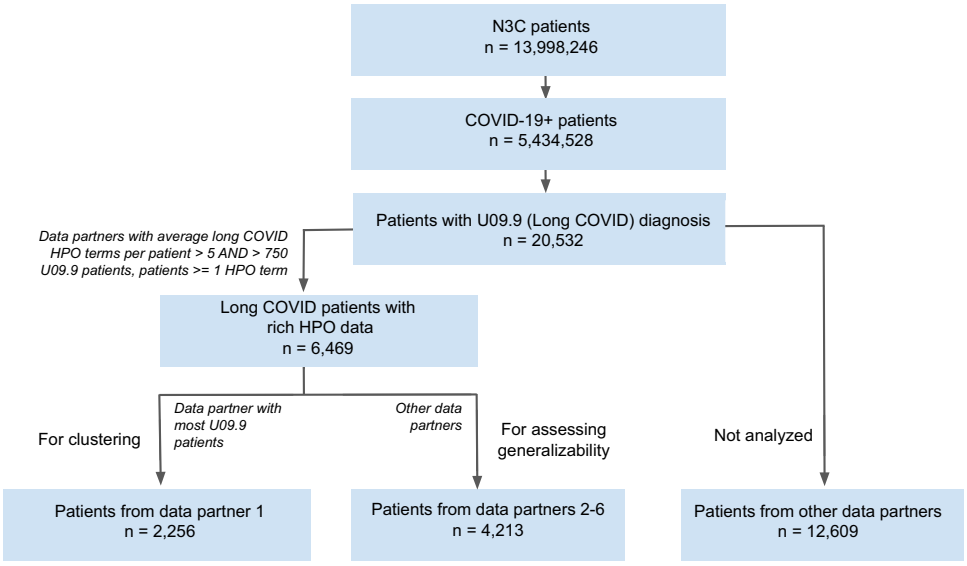


***Fig. 1:* Cohort construction.** Patients with long COVID (U09.9 diagnosis) were extracted from the much larger dataset of the N3C. Long COVID patients were selected from the six data partners that provided data for at least 300 U09.9 patients and had an average of at least 7 long COVID HPO terms per patient. The data partner with the most U09.9 patients (data partner 1) was chosen for clustering, and additional U09.9 patients from five other data partners (data partners 2–6) were chosen to assess generalizability.

matrix can be used as input to a number of clustering algorithms (Fig. 2).

To leverage this procedure for analysis of N3C data, we mapped the 287 long COVID-associated HPO terms[10] to corresponding Observational Medical Outcomes Partnership (OMOP) codes[13] (see Methods). Of these, 118 terms were identified in the data (Supplemental Tables S1–S11). The terms not found in the data largely were clinical or patient-reported features that are not commonly represented in EHR data, such as *Centrilobular ground-glass opacification on pulmonary HRCT* (HP:0025180) or *Ocular pruritus* (HP:0033841), and were not included in further analyses.

We selected data partners that provided at least 750 U09.9 patients and an average of at least five HPO terms per patient (Fig. 1). This threshold was chosen to include data partners with a sufficient number of patients with a sufficient depth of phenotypic information available in EHR data to assess patient similarity. For clustering, we selected U09.9 patients from the data partner (referred to here as data partner 1, as data regulations disallow use of real data partner names or IDs) that supplied data for the greatest number of U09.9 patients (2256 patients with at least one long COVID HPO phenotypic feature). For assessment of the generalizability of the clusters to other data partners, we selected the remaining U09.9 patients who had at least one long HPO phenotypic feature from the remaining data partners (referred to here as data partners 2–6, again due to data regulations) (4213 patients). We calculated the frequency with which each term was used in the total group of 2256 patients from data partner 1 and used this value to determine the information content (a measure of specificity; see Methods) for each term.

In order to calculate pairwise phenotypic similarity of patients at data partner 1 for clustering, we leveraged the Phenomizer algorithm to calculate a 2256 × 2256 similarity matrix for the 2256 patients with at least one HPO term at data partner 1. *K*-means clustering was applied to the data and the number of clusters was determined to be 6 based on visual inspection of the 'elbow' curve (Fig. 3; Supplemental Figure S1). We note that although the determination of cluster number by this method is subjective, the major findings were similar with 4 or 5 clusters (Supplemental Figures S2 and S3).

Characterization of PASC clusters

We characterised the features of each of the six clusters with respect to age, gender, and race/ethnicity (Table 1). The six clusters contained between 250 and 500 patients, and differed significantly with respect to rate of hospitalisation, age, gender, and ethnicity. Results of post-hoc analysis (see Table 1 and additional details in Supplemental Tables S14–S17) found statistically significant differences suggesting that Cluster 1 contains a larger proportion of patients with acute infection, Cluster 6 contains a larger proportion of females, Cluster 1 and Cluster 2 contain older patients, Cluster 3 contains a higher proportion of White non-Hispanic people, while Cluster 5 contains a lower proportion of

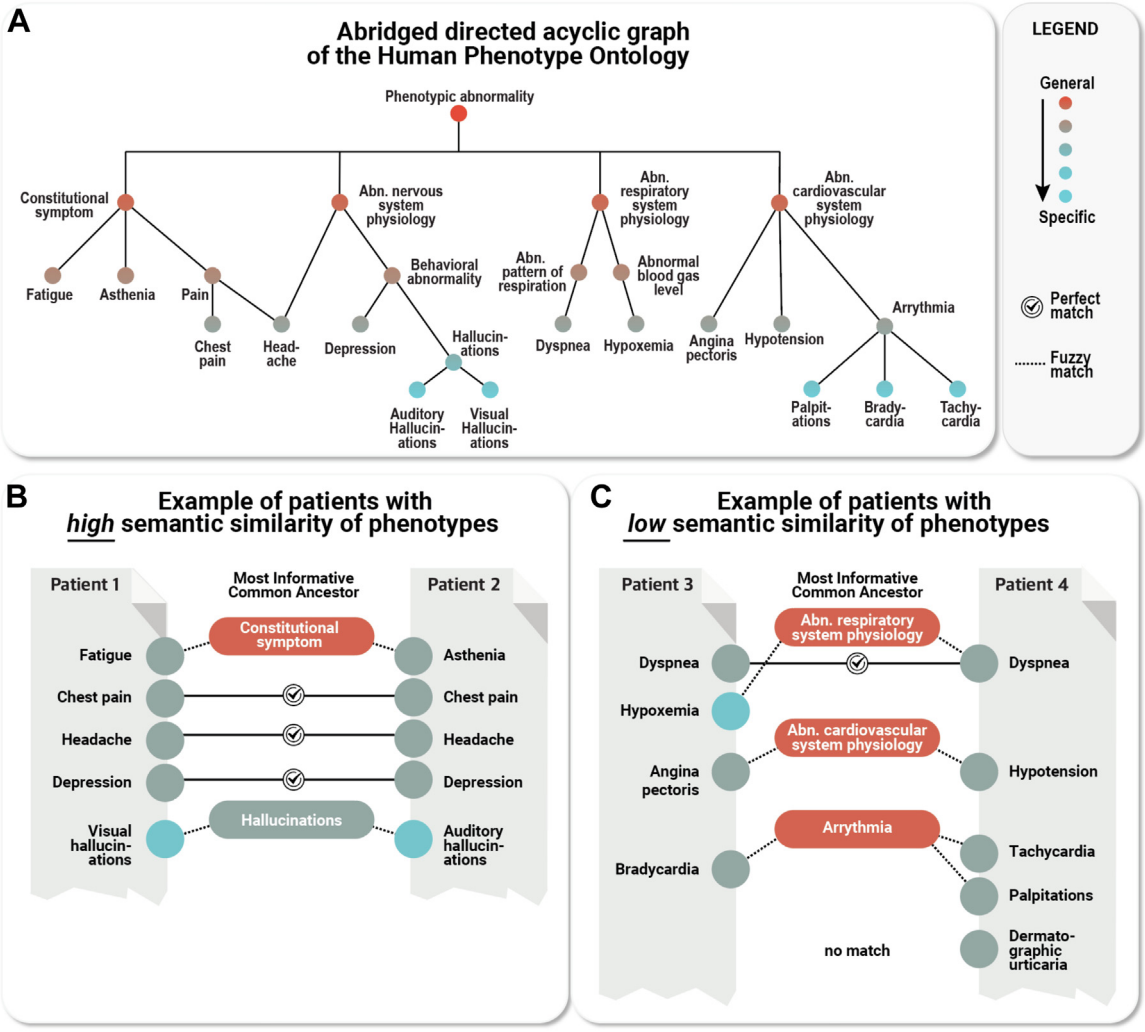


***Fig. 2:* Calculating patient semantic similarity based on HPO phenotypes. A)** HPO terms are arranged in a directed acyclic graph with specific terms such as *Bradycardia* (HP:0001662) being related to more general terms (here: *Arrhythmia*; HP:0011675) by subtype relations. An excerpt of the entire ontology (15,247 terms) is shown. **B)** Example showing a pair of patients with relatively high phenotypic similarity; for each of the HPO terms in patient 1, the best match is sought in patient 2. If an exact match is not found, the algorithm searches for the most informative common ancestor (MICA) in the ontology; the information content (a measure of specificity) of the exact matching term or most specific ancestor term is calculated to determine the specificity. For instance, *Visual hallucinations* (HP:0002367) and *Auditory hallucinations* (HP:0008765) are not an exact match, so the information content of their MICA *Hallucinations* (HP:0000738) is chosen. *Hallucinations* (HP:0002367) is still relatively specific (and shown in grey), while the MICA of *Angina pectoris* (HP:0001681) and *Hypotension* (HP:0002615) is more general (shown in red) and contributes less to the matching score. **C)** Example of a pair of patients with a relatively lower similarity due to (specific) fewer exact matches and one unmatched term. The pairwise similarity is calculated in this way for all pairs of patients to construct the similarity matrix that is used for clustering (Fig. 3).

White non-Hispanic people (significant differences are shown in Table 1 using CLD notation).

To further characterise each of the six clusters, we identified HPO terms that tended to occur among patients in certain clusters (Fig. 4). Of the 287 HPO terms we identified as being used in published cohort studies on long COVID,[10] only 118 were identified in our data. The presence or absence of each of the 118 HPO terms used for clustering was treated as a categorical variable whose distribution among the six clusters was assessed using a chi-squared test. Of the 118 HPO terms, 63 were significantly correlated with cluster membership following Bonferroni correction. Of these, 29 terms had a corrected p-value of less than 10^{-5} and were present in at least 20% of patients in one or more clusters. These terms were therefore considered to be the features that best defined the clustering.

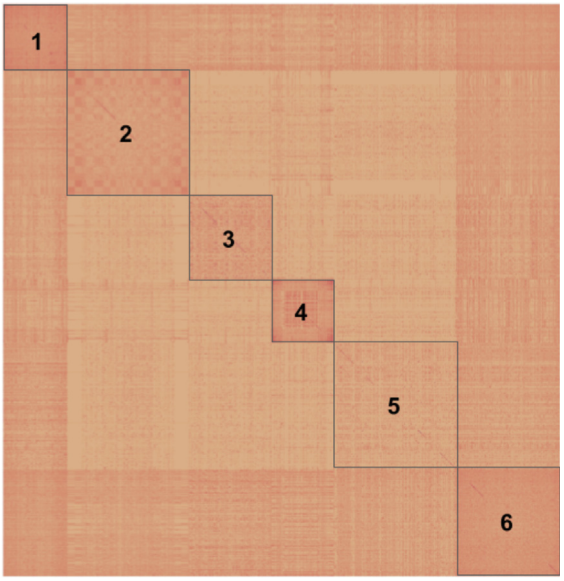


***Fig. 3*: Patient similarity matrix illustrating long COVID subtypes in data partner 1.** A heatmap representing the 6 clusters created by *k*-means clustering is shown. Cluster hierarchy was calculated using the nearest point algorithm and Euclidean distance.

HPO terms were classified into these categories: cardiovascular, constitutional, endocrine, ear nose and throat (ENT), eye, gastrointestinal, immunology, laboratory, neuropsychiatric, pulmonary, and skin. The constitutional category encompasses symptoms and findings such as *Fatigue* (HP:0012378), *Night sweats* (HP:0030166), and *Xerostomia* (HP:0000217) that cannot be unambiguously assigned to a single organ system. UpSet plots[36] were used to visualise the salient characteristics of each cluster according to these categories. UpSet visualisations show not only the most common categories, but also the most common combinations of categories. For instance, in cluster 1, patients most commonly had HPO terms from the categories pulmonary, neuropsychiatric, laboratory, constitutional, gastrointestinal, cardiovascular, and ear nose throat (ENT), and pulmonary. Although there was some overlap in the distribution of features, the profiles of terms and categories were distinct for the six clusters (Fig. 4).

The six PASC clusters differ with respect to frequencies of clinical manifestations

For ease of exposition, we refer to the six clusters according to the category or categories of the HPO terms showing the highest degree of enrichment. We refer to cluster 1 as multisystem + lab, because patients in this cluster had a high frequency of terms in the multiple categories: neuropsychiatric, pulmonary, constitutional, cardiovascular, gastrointestinal and ENT (vertigo) as well as multiple laboratory abnormalities. Patients in cluster 2, which we refer to as the pulmonary cluster, had high frequencies of *Hypoxemia* and *Cough*. We refer to cluster 3 as neuropsychiatric because of the relatively high frequencies of the terms *Headache*, *Insomnia*, *Depression*, *Sleep apnea*, *Abnormality of movement*, and *Paresthesia*. We refer to cluster 4 as cardiovascular because of the high frequency of *Tachycardia*, *Palpitations*, *Hypoxemia* (and also *Pulmonary embolism*, which because of the ontological structure of the HPO is a subclass of both the pulmonary and the cardiovascular subhierarchies). Cluster 5 is referred to as the pain/fatigue cluster because of the relatively high frequencies of *Pain*, *Chest pain*, and *Fatigue*. Finally, cluster 6 had a similar distribution of terms as cluster 1, but substantially lower frequencies of the laboratory abnormalities. Cluster 6 had the highest frequency of *Pain* of any cluster. Therefore, we refer to cluster 6 as the multisystem-pain cluster. Details are shown in Fig. 5A, and the results of post hoc testing are shown in 5B. For the latter, the proportion of patients in each cluster who had one or more of the manifestations in each category was compared by pairwise chi-squared testing.

Feature	Overall	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5	Cluster 6
n	2256	262	491	334	250	500	419
Inpatient**	440 (19.5%)	89 (34.0%) a	103 (21.0%) b	55 (16.5%) b	38 (15.2%) b	89 (17.8%) b	66 (15.8%) b
Age - mean ± SD**	53.0 ± 16.7	57.7 ± 15.0 a	55.1 ± 17.1 a	51.9 ± 16.9 b	51.7 ± 16.1 b	52.1 ± 17.8 b	50.2 ± 15.2 b
Female**	1403 (62.2%)	141 (53.8%) a	268 (54.6%) a	211 (63.2%) a	165 (66.0%) ab	306 (61.2%) a	312 (74.5%) b
White Non-Hispanic*	1787 (79.2%)	207 (79.0%) ab	398 (81.1%) ab	281 (84.1%) a	204 (81.6%) ab	377 (75.4%) b	320 (76.4%) ab
Black or African American Non-Hispanic	109 (4.8%)	<20	20 (4.1%)	<20	<20	22 (4.4%)	25 (6.0%)
Other/Unknown	360 (16%)	43 (16.4%)	73 (14.9%)	40 (12%)	29 (11.6%)	35 (12.8%)	74 (17.7%)

For the overall study population and for each cluster, age, gender, and race/ethnicity are shown. Data for characteristics for which there were fewer than 20 patients, and data about race/ethnicities for which there were fewer than 20 patients overall (Other Non-Hispanic, Native Hawaiian or Other Pacific Islander Non-Hispanic, Asian Non-Hispanic) are not shown to reduce the risk of patient re-identification. **p < 0.001 by one-way ANOVA (age) or chi squared test (all others). *p < 0.05 by chi squared test. We applied post-hoc tests on categorical and numeric variables that were significant by omnibus tests. For categorical variables, we computed pairwise chi-square tests while we used Dunn's test[35] for numerical variables. Bonferroni correction was performed in both cases. The results of adjusted pairwise comparisons are summarised using Compact Letter Display (CLD). The CLD method uses letters to mark groups for which the differences were not statistically significant (details in Supplemental Table S14). For instance, a cluster marked "a" is significantly different from a cluster marked "b" but not from another cluster marked "a" or "ab".

***Table 1*: Characteristics of the study population in data partner 1.**

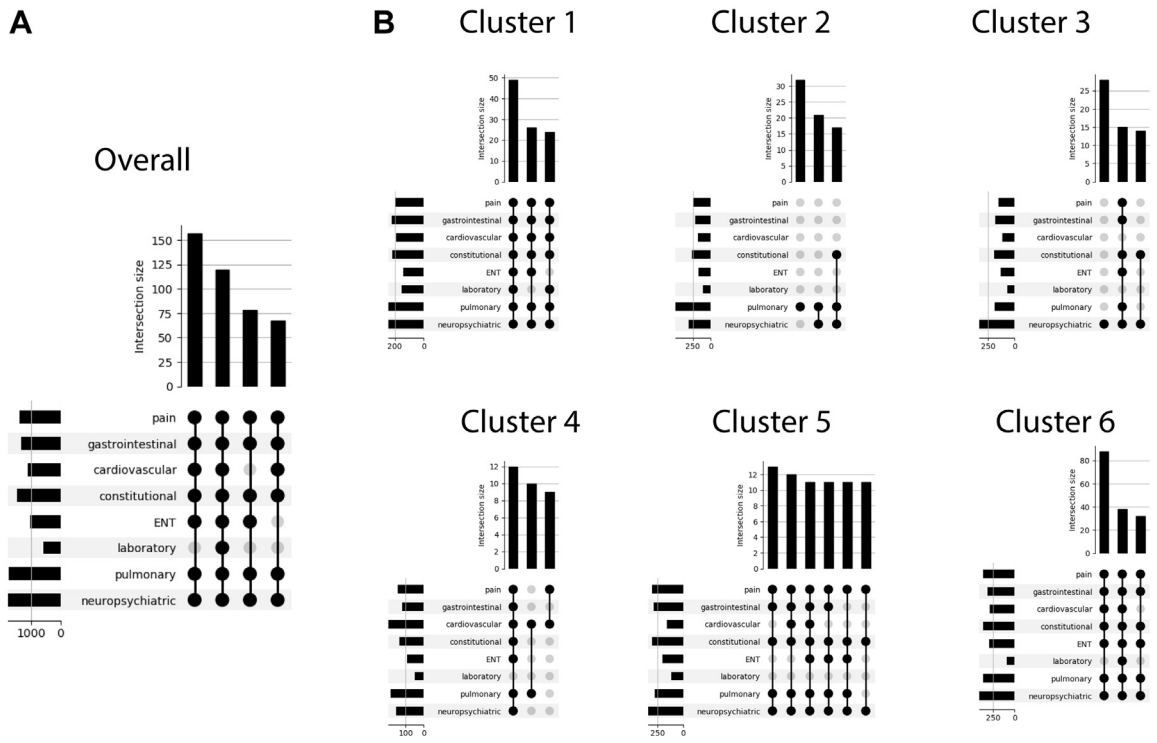


Fig. 4: **Phenotypically characterising long COVID subtype clusters.** Shown are the most frequently co-occurring combinations of categories of HPO terms representing long COVID phenotypic features for patients in the overall cohort **(A)** and for each of the 6 clusters **(B)**. Only those categories are shown that were found to be significantly correlated with cluster membership (chi-squared test, $p < 0.00001$). For the overall population of patients in data partner 1 and for each cluster, the frequency of each category of long COVID HPO terms (left) and the frequency of the three most common combinations of HPO categories (top) are shown (Six combinations are shown for cluster 5 because of a tie.) Notably, most clusters contain some widely shared features, but also distinguishing features such as symptoms in the pulmonary, neuropsychiatric, and cardiovascular systems. Data are shown as UpSet plots, which visualise set intersections in a matrix layout and show the counts of patients with the combination indicated by the black dots as bars above the matrix.[36]

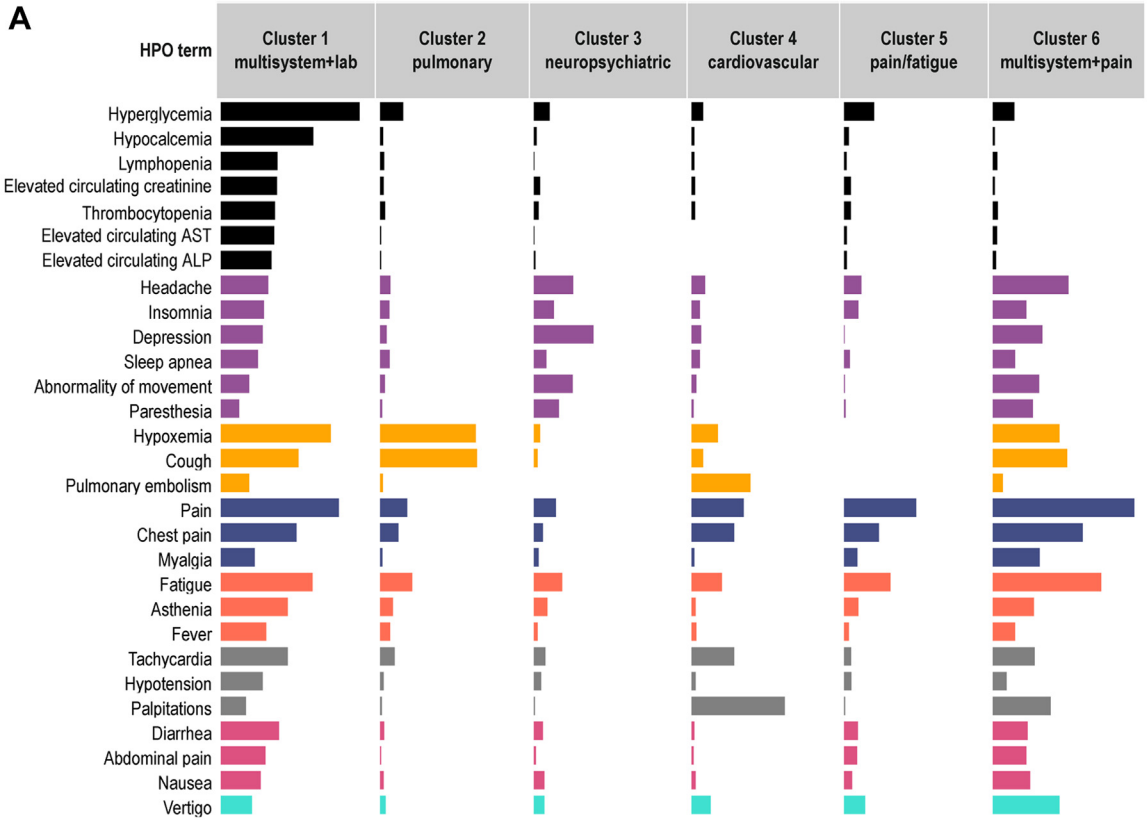
The multisystem-lab cluster 1 is characterised by manifestations suggesting increased clinical severity

The clustering described above relied solely on HPO terms that represent phenotypic abnormalities identified 4 weeks or more following COVID-19 diagnosis. We analysed the clusters for differences in the distribution of other variates. As shown by post-hoc tests the multisystem-lab cluster contained a higher proportion of inpatients (34.0%) compared to any other cluster and the mean age of 57.7 years was higher (see Table 1 and Supplemental Tables S14–S17).

The multisystem-lab cluster showed a high frequency of post-acute COVID-19 laboratory abnormalities that have been associated with severe course of acute COVID-19, namely, *Lymphopenia* (HP:0001888), *Elevated circulating alanine aminotransferase concentration* (HP:0031964), *Increased circulating ferritin concentration* (HP:0003281), *Elevated circulating alkaline phosphatase concentration* (HP:0003155), *Hypocalcemia* (HP:0002901), and *Thrombocytopenia* (HP:0001873).[37–42] Further, post-hoc tests suggested that this cluster contains higher proportions of patients with (either pre-existing and/or contextual to COVID-19) acute kidney injury (AKI, see Tables 2 and 3) and steroid usage (Table 3). This suggests that this cluster may represent patients with residual manifestations of more severe COVID-19 and/or long COVID manifestations, although severity cannot unambiguously be inferred from EHR data. Patients in cluster 1 showed a higher mortality, a finding that was generalizable to other data centres (see below).

In the entire cohort, 61.2% of patients were female. In the multisystem-lab cluster characterised by a severe clinical course, only 53.8% of patients were female. This was significantly different from the multisystem-pain cluster in which 74.5% of patients were female (See Table 1). Evidence available prior to our study suggests that sex differences exist that influence the clinical course of COVID-19. For instance, although males are more likely to be hospitalised or die with acute COVID-19, females are more likely to develop long COVID.[43]

A



B

		Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5	Cluster 6
laboratory	31%	A (100%)	BCD (20%)	BC (17%)	B (12%)	D (28%)	CD (26%)
neuropsychiatric	58%	A (82%)	B (30%)	C (99%)	B (32%)	B (31%)	D (92%)
pulmonary	53%	A (85%)	B (100%)	C (8%)	D (48%)	C (5%)	E (72%)
pain	35%	A (56%)	B (14%)	B (11%)	C (29%)	C (36%)	D (70%)
constitutional	39%	A (67%)	B (25%)	B (22%)	B (20%)	C (34%)	A (71%)
cardiovascular	29%	A (55%)	B (13%)	B (10%)	C (68%)	B (8%)	A (49%)
gastrointestinal	29%	A (65%)	B (12%)	BC (18%)	B (10%)	C (26%)	D (52%)
ENT	28%	A (39%)	BC (14%)	B (13%)	BC (20%)	C (21%)	D (60%)

Fig. 5: **Summary of phenotypic feature distribution in the six clusters. A)** The HPO terms corresponding to different phenotypic features are grouped in HPO categories shown on the left. Categories are colour-coded and are in the same order as shown in panel **B**. Laboratory abnormalities are grouped together because of their association with severe COVID-19 (see text). HPO terms are shown if at least 20% of patients in at least one cluster had the corresponding phenotypic feature and if Pearson's chi-squared test found a significant difference ($p < 0.00001$) in the phenotypic feature distribution. **B)** Post hoc analysis of categories of long COVID HPO phenotypic features by cluster. For each category of Long COVID HPO phenotypic feature, we performed a post hoc analysis (pairwise chi-squared test with Bonferroni correction) to assess differences between clusters. For each category, the percent of patients from each cluster that have at least one HPO term in the given category are shown, and red and blue cells mark the CLD group having the highest and lowest proportion, respectively. Letters a–e indicate CLD groups between which differences for the given category are statistically significant according to post hoc analysis (Methods).

The six PASC clusters differ with respect to pre-existing comorbidities

To investigate how clinical features before or during COVID-19 infection correlated with cluster membership, we assessed the distribution across the six clusters of 44 clinical features determined prior to acute COVID-19 or during acute COVID-19. Of these, 13 displayed a statistically significant difference between clusters and are shown in Tables 2 and 3. Among parameters that were present before acute COVID-19 (Table 2), 10 differed significantly between clusters, mainly showing a higher frequency in the multisystem-lab cluster (as per post-hoc

Pre-existing clinical feature	Cluster 1. Multisystem + lab	Cluster 2. Pulmonary	Cluster 3. Neuropsychiatric	Cluster 4. Cardiovascular	Cluster 5. Pain/fatigue	Cluster 6. Multisystem-pain
Acute kidney injury	24.8% a	8.1% b	11.1% b	9.2% b	10.4% b	7.9% b
Chronic lung disease	44.7% a	28.7% bc	29.6% bc	30.8% bc	23.8% b	33.9% ac
Depression	32.1% a	21.2% b	37.7% a	16.4% b	20.8% b	37.0% a
Diabetes (complicated)	28.6% a	12.0% b	8.7% b	7.2% b	10.2% b	11.0% b
Diabetes (uncomplicated)	38.9% a	20.4% b	16.2% b	17.2% b	19.2% b	21.0% b
Hypertension	60.7% a	44.2% b	40.4% b	36.0% b	36.2% b	42.7% b
Immunocompromised (other)	14.9% a	7.9% ab	7.5% ab	4.0% bc	2.4% c	7.4% b
Kidney disease	29.0% a	12.0% b	14.1% b	9.6% b	9.8% b	12.6% b
Mild liver disease	17.9% a	5.5% b	8.1% b	6.4% b	6.8% b	9.5% b
Peripheral vascular disease	14.1% a	5.5% b	5.1% b	2.0% b	5.0% b	4.1% b

The 10 of 45 clinical features present before COVID-19 infection (Supplemental Table S12) that were significantly overrepresented in clusters (chi squared $p < 0.001$ after Bonferonni correction) and the percent of patients in each cluster with each clinical feature are shown. Letters a-c indicate CLD groups between which differences for the given pre-clinical feature are statistically significant according to post hoc analysis (Methods).

Table 2: **Clinical features of patients before acute COVID-19 infection by cluster.**

analysis). The risk of long COVID has been shown to be associated with the number of comorbidities.[44] These observations are consistent with the notion that the multisystem-lab cluster is composed of patients with more severe clinical manifestations, and that there may be different risk factors for clusters 2–6.

Post-hoc analysis also confirmed that covariates during acute COVID-19 whose frequencies were higher in the multisystem-lab cluster included acute kidney injury (AKI) and corticosteroid medications that also may be proxies for a severe clinical course (Table 3). Severity of acute COVID has been associated with risk of persistent symptoms in some studies.[45] Although the frequency of depression as a pre-existing comorbidity was highest in the neuropsychiatric cluster (Table 2), post hoc tests failed to find statistically significant differences when comparing it to the proportions of pre-existing depression in the multisystem-lab and the multisystem pain clusters.

Generalisability of clusters to new data partners

The results presented in the previous sections were generated with data from data partner 1. We assessed the generalizability of the clustering results for four additional data partners (data partners 2–6, Fig. 1) by comparing each patient from these data partners to the patients in each cluster from data partner 1 and also to randomly permuted clusters (Methods). If the clusters in data partner 1 did not generalise at all to other data partners, we would expect that patients from other data partners would be equally similar to the patients of any of the clusters in data partner 1.

We observed that patients from data partners 2–6 were much more similar to clusters from data partner 1 compared to randomly permuted clusters. The mean similarity ranged from 0.202 to 0.211 for test data partners 2–6 for the randomly permuted clusters, but the observed mean similarities to the original clusters at data partner 1 ranged from 0.283 to 0.319, corresponding to z-scores of 28.6–65. The mean similarity score for the randomly permuted clusters was never as high as the observed score over 1000 permutations, corresponding to an empirical p-value of less than 0.001 for each of the data partners 2–6. This strongly suggests that clusters identified in data partner 1 generalise to patients from other data partners (Table 4).

The multisystem-lab clusters is characterised by higher mortality reproducibly across data partners 1–6

Because of the indications that the multisystem-lab cluster may be characterised by greater clinical

Clinical feature during COVID-19	Cluster 1. Multisystem + lab	Cluster 2. Pulmonary	Cluster 3. Neuropsychiatric	Cluster 4. Cardiopulmonary	Cluster 5. Pain/fatigue	Cluster 6. Multisystem-pain
Acute kidney injury	14.5% a	6.3% b	4.2% b	4.4% b	4.8% b	4.1% b
Corticosteroid regimen	30.2% a	19.8% b	14.1% b	14.4% b	15.4% b	13.8% b
COVID diagnosis during hospitalisation	34.0% a	21.0% b	16.2% b	15.2% b	17.6% b	15.8% b

The 3 of 43 clinical features present during COVID-19 infection (Supplemental Table S13) that were significantly overrepresented in clusters (chi squared $p < 0.001$ after Bonferonni correction) and the percent of patients in each cluster with each clinical feature are shown. Letters a and b indicate CLD groups between which differences for the given clinical feature are statistically significant according to post hoc analysis (Methods).

Table 3: **Clinical features of patients during acute COVID-19 infection by cluster.**

Test data partner	Similarity to permuted clusters	Observed mean similarity	Z-score	Empirical p-value
2	0.211 ± 0.0032	0.302	28.6	<0.001
3	0.208 ± 0.0017	0.318	65.0	<0.001
4	0.21 ± 0.0026	0.319	42.6	<0.001
5	0.202 ± 0.0022	0.283	36.3	<0.001
6	0.204 ± 0.0022	0.294	40.5	<0.001

The similarity of patients from test data partners 2–6 to patients in clusters made from data partner 1 clusters and to patients from randomly permuted clusters was measured as in Fig. 2. For each test data partner, a random 10% sample of patients from test data partner and data partner 1 was selected. The average similarity of its patients to the best matching randomly permuted cluster and to the best matching cluster from data partner 1 are shown along with the Z-score and p-value. Results are representative of five duplicate experiments with different random samples. The empirical p-value reflects the number of times that the similarity of a permuted dataset was higher than that of the observed clusters (this never occurred).

Table 4: **Generalisability of clusters in patients from new data partners.**

severity, we assessed recorded mortality in the time period subsequent to acute COVID-19. We assigned patients from data partners 2–6 to the original six clusters according to the maximum mean similarity of patients in those clusters (Methods). In these patients, the majority of cases of recorded mortality occurred in patients assigned to clusters 1 (counts less than 20 are masked for data privacy reasons). We performed a chi-squared test of the null hypothesis that the proportion of mortalities in the clusters was uniform. The observed correlation between mortality and cluster membership was statistically significant for the analysis of clustered patients in data partner 1 ($p = 5\times10^{-5}$) and in data partners 2–6 ($p = 5\times10^{-5}$) using a Fisher's exact test calculated by the Monte Carlo method with 100,000 permutations (Table 5).

Discussion

According to the WHO, approximately 10–20% of patients with COVID-19 may experience new-onset, lingering or recurrent clinical symptoms after acute infection. This has been termed 'post-acute sequelae of SARS-CoV-2 infection' (PASC) or long COVID. Definitions of long COVID in the literature vary, and the frequencies and time course of phenotypic manifestations following acute COVID-19 are highly heterogeneous.[10] This observation raises the question of whether long COVID can be stratified into well delineated and reproducible subtypes, or whether the degree of heterogeneity is so high that stratification is impossible. This is critically relevant for defining sub-cohorts in clinical research studies such as the NIH program "Researching COVID to Enhance Recovery (RECOVER)," and for identifying candidate therapeutics. ML clustering methods offer a data-driven approach to stratification of patients that can reveal such subtypes in the face of this new and heterogeneous disease.

Evidence available prior to our study suggests that important clinical differences do exist that influence the susceptibility to subsequent complications of COVID-19. For instance, although males are more likely to be hospitalised or die with acute COVID-19, females are more likely to develop long COVID.[43] It is possible that the pathophysiology of long COVID may be multifactorial in origin. Conceivably, the biological underpinnings of long COVID may vary among individuals as a function of

Cluster	Data Partner 1			Data Partners 2-6		
	deaths	total	%	deaths	total	%
1 - multisystem+lab	33	262		92	1490	
2 - pulmonary	<20	491		<20	435	
3 - neuropsychiatric	<20	334		<20	322	
4 - cardiovascular	<20	250		<20	539	
5 - pain/fatigue	<20	500		0	<20	
6 - multisystem-pain	<20	419		<20	1312	

Data partner 1 was the source of data for generating the six clusters. Patients from data partners 2–6 were assigned to these clusters (Methods). Number of recorded deaths, total number of patients, and percentage of patients with recorded death in each cluster are shown.

Table 5: **Recorded deaths according to cluster.**

baseline risk factors, resulting in different general phenotypes of long COVID, the treatment or prevention of which may need to be specifically tailored using precision medicine in order to achieve optimal outcomes. As a first step, we sought to use unsupervised learning to delineate potential subtypes of patients with long COVID with differing clinical characteristics. We identified six published studies that present clusters from either patient-reported data (in four studies) or manually recorded clinical data (two studies) with cohorts of between 145 and 3762 patients. The studies report two or three clusters based on different types of input data, making study comparison challenging. None of the studies were based on EHR data and no assessment of generalisability to other data partners was presented.[9,46–50]

Here we have presented a method for semantic clustering of long COVID patients based on HPO-encoded EHR data. We further present a method for assessing generalisability of the identified subtypes or clusters across different data contributing sites. Ontology-based algorithms differ from machine learning and other algorithms in many ways. Coding numerical data with HPO implies that parameters are simplified into categories. Although this loss of numerical data reduces precision in data granularity, simplification allows powerful simultaneous analysis of all phenotypic observations using semantic similarity that can take the relatedness of concepts into account.

Our method for assessing patient–patient similarity using the Phenomizer algorithm generates an essentially continuous similarity value from arbitrary sets of HPO terms that characterise any two patients. An alternative method would be to encode the 287 HPO terms as a 287-dimensional feature vector and to measure similarity for example using dot product (cosine) of these vectors. The Phenomizer algorithm has several advantages over the feature vector method: it does not suffer from sparse count issues that may make clustering less robust,[51] and it takes advantage of the similarity between individual items using the structure of the HPO in a way that a feature vector cannot.[22] This approach has proven powerful both in the support of differential diagnosis of rare disease and in efforts to enable longitudinal analysis of EHR data as a means of identifying gene–phenotype associations with Mendelian forms of epilepsy,[52,53] but has never before been applied in the context of infectious disease EHR data and methods for assessing generalisability have not previously been presented.

We have shown that unsupervised learning based on semantic clustering identifies phenotypic profiles that are reproducible across data partners with a high degree of statistical significance. The six clusters that emerged demonstrated non-uniform frequencies of symptoms and clinical findings across an array of features, spanning constitutional/systemic symptoms and pain, cardiac, respiratory, gastrointestinal, and neurologic symptom domains, with some degree of overlap but clear distinctions between various groups. We interpret our multisystem-lab cluster as comprising patients with a severe course of acute COVID-19 because of the higher hospitalisation rates (Table 1) and mortality (Table 3). It is possible that this cluster represents a subtype of long COVID that results from severe acute COVID-19. Our findings confirm and extend previous findings of a steeper risk gradient for long COVID manifestations that increases according to the severity of the acute COVID-19 infection.[54]

We suggest that analogous algorithms could be used to evaluate data gathered from prospective studies of long COVID patients to extend and deepen our characterization of phenotypic clusters by including data that are currently difficult to ascertain reliably from EHR data, including symptoms such as *Asthenia* (HP:0025406) or *Exertional dyspnea* (HP:0002875) and radiology findings (which are typically not represented using structured fields in EHR data and are underrepresented in OMOP datasets). The recently released Phenopacket Schema of the Global Alliance for Genomics and Health (GA4GH) provides a standardised way to record clinical findings including phenotypic features, measurements, biospecimens, and medical actions over the time course of a disease as a computational case report.[55] Recording clinical data with the Phenopacket Schema would promote data sharing and comparability of results from different studies.

Study limitations

While our study provides insight into the variability and natural history of long COVID, there are limitations that should be considered. While the U09.9 code provides a simple inclusion criterion, its application in health systems across the country is not uniform and may differ across data partners. Also, since the use of the code began only recently, patients with long COVID that were diagnosed prior to the introduction of the code are not included, limiting our ability to compare the current clinical manifestations with those observed earlier in the pandemic before widespread vaccination and with different distributions of SARS-CoV2 strains and variants. However, in a pilot study in Denmark, coding with U09.9 was found to have a positive predictive value of 94% for long COVID.[56]

Our ability to capture clinical manifestations of long COVID is limited by the accessibility of clinical data in EHR systems. Of the 287 HPO terms we identified as being mentioned in published cohort studies on long COVID,[10] only 118 are present in our data. The reasons for this presumably include unstructured data such as symptoms and radiological findings that are not well represented in the OMOP data that is the source of our data. Examples include *Gaze-evoked nystagmus* (HP:0000640), *Pericardial effusion* (HP:0001698), and

Exercise intolerance (HP:0003546) that are typically diagnosed using specialist examinations or medical history that may not be easily coded in structured EHR fields. Additionally, several common manifestations of long COVID, including dysautonomia,[57] are less documented in EHR data in part due to the difficulties in recognizing these illnesses clinically and the fact that relevant findings may not be well represented in structured fields including the OMOP data available in N3C.

Our study uses the newly minted ICD code U09.9 to identify patients with PASC/long COVID. At the time of this writing, a relatively small number of labelled patients was available for analysis. Furthermore, the population defined by these patients is not fully representative of the American population; for instance, the proportion of African Americans in our study (~5%) is lower than the proportion of African Americans among the entire population. As more data accrues, future work will be required to characterise the role of social determinants of health that are confounded with race in our society in determining long COVID subtypes. It is likely that many additional long COVID patients are present in the N3C dataset who have not received the U09.9 diagnosis code, and it is possible that this fact could introduce a bias into the data analysed in this study. Additionally, the group of patients who present for medical care for long COVID symptoms and receive a U09.9 diagnostic code may not be representative of the entire population of patients with long COVID manifestations.

Our exploration of k-means clustering results with different values of k from 2 to 8 showed that increasing the number of clusters tended to subdivide existing clusters hierarchically. Although numerous methods for determining the 'best' number of clusters are available, there is no objective definition of optimum that applies to all applications, and the choice of k is perforce subjective in nature. Our main findings of generalisable phenotypic clusters pertain also for values of k of 4 and 5 (Supplemental Figures S2 and S3).

Conclusions

We have presented a novel algorithm for semantic clustering that identifies patient similarity by transforming EHR data to phenotypic profiles using the HPO, and identified long COVID subtypes that show a statistically significant degree of generalizability of clusters across different medical centres. There was a significant association of cluster membership with a range of pre-existing conditions and with measures of severity during acute COVID-19. One of the clusters (multisystem-lab) was associated with severe manifestations and displayed increased mortality, and other clusters showed enrichment for pulmonary, neuropsychiatric, cardiovascular, pain/fatigue, and a multisystem/pain profile not associated with significantly increased mortality. Additionally, we show that the identified clusters were generalizable across different hospital systems and that the increased mortality rate was consistently observed in the multisystem-lab cluster. Semantic phenotypic clustering could provide a basis for assigning patients to stratified subgroups for natural history or therapy studies.

Contributors

Conceptualization: J.T.R., P.N.R. Methodology: J.T.R., H.B., T.B., J.J.L., T.J.C., B.L., E.C., B.C., M.G., K.J.W., L.C., T.F., N.A., B.A., T.M.M., G.K., J.A.M., G.V., D.S., C.G.C., C.M.-B., A.W., R.M., J.B. Investigation: J.T.R., H.B., C.A., A.E.S., H.D., K.K. Funding acquisition: M.A.H., P.N.R. Supervision: J.T.R., M.A.H., P.N.R. Writing – original draft: J.T.R., P.N.R. Writing – review & editing: J.T.R., D.L.A., P.R.B., J.H.C., J.L.S., E.H. All authors read and approved the final version of the manuscript. J.T.R. and P.N.R. verified the underlying data.

Data sharing statement

The analyses described in this publication were conducted with data accessed through the NCATS N3C Data Enclave covid.cd2h.org/enclave. Researchers can apply for access to the data as described in https://ncats.nih.gov/n3c/.

The code for performing the semantic analysis, clustering, and generalizability assessment is freely available at https://github.com/National-COVID-Cohort-Collaborative/semanticsimilarity and https://github.com/National-COVID-Cohort-Collaborative/kernelkm. Additional code for defining the cohort and transforming raw OMOP data for this analysis is available through the NCATS N3C Data Enclave covid.cd2h.org/enclave with access procedures as described above. The project is available under the Data Use Request RP-5677B5 "Characterization of long-COVID: definition, stratification, and multi-modal analysis".

Declaration of interests

T. Bergquist received other support from Bill and Melinda Gates Foundation, H. Davis received support from Balvi Foundation and is a cofounder of Patient Led Research Collaborative. The other authors declare that they have no other competing interests.

Acknowledgments

The authors acknowledge the following funding sources: National Institutes of Health grant CD2H NCATS U24 TR002306 (J.T.R., C.C., H.B., N.A., B.L., K.K., M.A.H., P.N.R.). National Institutes of Health grant NHLBI RECOVER Agreement OT2HL161847-01 (J.T.R., K.K., B.L., M.A.H., P.N.R.). National Institutes of Health grant Office of the Director Monarch Initiative R24 OD011883 (M.A.H., P.N.R.). National Institutes of Health grant NHGRI Center of Excellence in Genome Sciences RM1 HG010860 (M.A.H., P.N.R.). National Institutes of Health grant NCATS UL1TR003015 (B.A., T.M.M.). National Institutes of Health grant NCATS KL2TR003016 (B.A., T.M.M.). Director, Office of Science, Office of Basic Energy Sciences of the U.S. Department of Energy Contract No. DE-AC02-05CH11231 (J.T.R.). Donald A. Roux Family Fund at the Jackson Laboratory (P.N.R.). Marsico Family at the University of Colorado Anschutz (M.A.H.). K. Wilkins is an employee of NIH. D. Sahner is a contractor to NIH through Axle Informatics. 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; and was conducted under the N3C DUR RP-5677B5. 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. Medical Authorship was determined using ICMJE recommendations. 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 organisations (covid.cd2h.org/dtas) and the organisations and scientists (covid.cd2h.org/duas) who have contributed to the on-going development of this community resource.[58] 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. We gratefully acknowledge the following core contributors to N3C: Anita Walden, Leonie Misquitta, Joni L. Rutter, Kenneth R. Gersing, Penny Wung Burgoon, Samuel Bozzette, Mariam Deacy, Christopher Dillon, Rebecca Erwin-Cohen, Nicole Garbarini, Valery Gordon, Michael G. Kurilla, Emily Carlson Marti, Sam G. Michael, Lili Portilla, Clare Schmitt, Meredith Temple-O'Connor, David A. Eichmann, 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, Davera Gabriel, Stephanie S. Hong, Kristin Kostka, Harold P. Lehmann, Michele Morris, Matvey B. Palchuk, Xiaohan Tanner Zhang, Richard L. Zhu, Benjamin Amor, Mark M. Bissell, Marshall Clark, Andrew T. Girvin, Stephanie S. Hong, Kristin Kostka, Adam M. Lee, Robert T. Miller, Michele Morris, Matvey B. Palchuk, Kellie M. Walters, Will Cooper, Patricia A. Francis, Rafael Fuentes, Alexis Graves, Julie A. McMurry, Shawn T. O'Neil, Usman Sheikh, Elizabeth Zampino, Katie Rebecca Bradwell, Andrew T. Girvin, Amin Manna, Nabeel Qureshi, Christine Suver, Julie A. McMurry, Carolyn Bramante, Jeremy Richard Harper, Wenndy Hernandez, Farukh M Koraishy, Amit Saha, Satyanarayana Vedula, Johanna Loomba, Andrea Zhou, Steve Johnson, Evan French, Alfred (Jerrod) Anzalone, Umit Topaloglu, Amy Olex, Hythem Sidkey. Details of contributions available at covid.cd2h.org/acknowledgements.

We acknowledge support from many grants; the content is solely the responsibility of the authors and does not necessarily represent the official views of the N3C Program, the NIH or other funders. In addition, access to N3C Data Enclave resources does not imply endorsement of the research project and/or results by NIH or NCATS.

We acknowledge 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.

Appendix A. Supplementary data

Supplementary data related to this article can be found at https://doi.org/10.1016/j.ebiom.2022.104413.

References

1 Weekly operational update on COVID-19-30 March 2022 [Internet]. Available from: https://www.who.int/publications/m/item/weekly-operational-update-on-covid-19—30-march-2022. Accessed April 20, 2022.
2 Raveendran AV, Jayadevan R, Sashidharan S. Long COVID: an overview. *Diabetes Metabol Syndr*. 2021;15(3):869–875.
3 Taquet M, Dercon Q, Luciano S, Geddes JR, Husain M, Harrison PJ. Incidence, co-occurrence, and evolution of long-COVID features: a 6-month retrospective cohort study of 273,618 survivors of COVID-19. *PLoS Med*. 2021;18(9):e1003773.
4 Michelen M, Manoharan L, Elkheir N, et al. Characterising long COVID: a living systematic review. *BMJ Glob Health*. 2021;6(9):e005427. https://doi.org/10.1136/bmjgh-2021-005427.
5 Nalbandian A, Sehgal K, Gupta A, et al. Post-acute COVID-19 syndrome. *Nat Med*. 2021;27(4):601–615.
6 Crook H, Raza S, Nowell J, Young M, Edison P. Long covid-mechanisms, risk factors, and management. *BMJ*. 2021;374:n1648.
7 Greenhalgh T, Knight M, A'Court C, Buxton M, Husain L. Management of post-acute covid-19 in primary care. *BMJ*. 2020;370:m3026.
8 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*. 2022;22(4):e102–e107.
9 Kenny G, McCann K, O'Brien C, et al. Identification of distinct long COVID clinical phenotypes through cluster analysis of self-reported symptoms. *Open Forum Infect Dis*. 2022;9(4):ofac060.
10 Deer RR, Rock MA, Vasilevsky N, et al. Characterizing long COVID: deep phenotype of a complex condition. *eBioMedicine*. 2021;74:103722.
11 Robinson PN, Köhler S, Bauer S, Seelow D, Horn D, Mundlos S. The Human Phenotype Ontology: a tool for annotating and analyzing human hereditary disease. *Am J Hum Genet*. 2008;83(5):610–615.
12 Köhler S, Gargano M, Matentzoglu N, et al. The human phenotype ontology in 2021. *Nucleic Acids Res*. 2021;49(D1):D1207–D1217.
13 Voss EA, Makadia R, Matcho A, et al. Feasibility and utility of applications of the common data model to multiple, disparate observational health databases. *J Am Med Inform Assoc*. 2015;22(3):553–564.
14 Robinson PN, Webber C. Phenotype ontologies and cross-species analysis for translational research. *PLoS Genet*. 2014;10(4):e1004268.
15 Köhler S, Vasilevsky NA, Engelstad M, et al. The Human Phenotype Ontology in 2017. *Nucleic Acids Res*. 2017;45(D1):D865–D876.
16 Robinson PN, Mundlos S. The Human Phenotype Ontology. *Clin Genet*. 2010;77(6):525–534.
17 Groza T, Köhler S, Moldenhauer D, et al. The human phenotype ontology: semantic unification of common and rare disease. *Am J Hum Genet*. 2015;97(1):111–124.
18 Köhler S, Doelken SC, Mungall CJ, et al. The Human Phenotype Ontology project: linking molecular biology and disease through phenotype data. *Nucleic Acids Res*. 2014;42(Database issue):D966–D974.
19 Zhang XA, Yates A, Vasilevsky N, et al. Semantic integration of clinical laboratory tests from electronic health records for deep phenotyping and biomarker discovery. *NPJ Digit Med*. 2019;2:32. https://doi.org/10.1038/s41746-019-0110-4.
20 Callahan TJ, Stefanski AL, Wyrwa JM, et al. Ontologizing health systems data at scale: making translational discovery a reality. *arXiv*. 2022. https://doi.org/10.48550/arXiv.2209.04732.
21 Jackson R, Matentzoglu N, Overton JA, et al. OBO Foundry in 2021: operationalizing open data principles to evaluate ontologies. *Database (Oxford)*. 2021;2021:baab069. https://doi.org/10.1093/database/baab069.
22 Köhler S, Schulz MH, Krawitz P, et al. Clinical diagnostics in human genetics with semantic similarity searches in ontologies. *Am J Hum Genet*. 2009;85(4):457–464.
23 Pesquita C, Faria D, Falcão AO, Lord P, Couto FM. Semantic similarity in biomedical ontologies. *PLoS Comput Biol*. 2009;5(7):e1000443.
24 Arthur D, Vassilvitskii S. k-means++: the advantages of careful seeding. SODA '07 [Internet]. Available from: https://www.semanticscholar.org/paper/5e0c61b7ee4a2de183a197f32c5013ad109531fa; 2007. Accessed May 11, 2022.
25 Steinley D. K-means clustering: a half-century synthesis. *Br J Math Stat Psychol*. 2006;59(Pt 1):1–34.
26 Hennig C. Dissolution point and isolation robustness: robustness criteria for general cluster analysis methods. *J Multivar Anal*. 2008;99(6):1154–1176.
27 García-Escudero LÁ, Gordaliza A. Robustness properties of k means and trimmed k means. *J Am Stat Assoc*. 1999;94(447):956–969.
28 Barak S, Mokfi T. Evaluation and selection of clustering methods using a hybrid group MCDM. *Expert Syst Appl*. 2019;138:112817.

29 Dinno A. Nonparametric pairwise multiple comparisons in independent groups using Dunn's test. *STATA J*. 2015;15(1):292–300.
30 Piepho HP. An algorithm for a letter-based representation of all-pairwise comparisons. *J Comput Graph Stat*. 2004;13(2):456–466.
31 Crawford K, Xian J, Helbig KL, et al. Computational analysis of 10,860 phenotypic annotations in individuals with SCN2A-related disorders. *Genet Med*. 2021;23(7):1263–1272.
32 Robinson PN, Köhler S, Oellrich A, et al. Improved exome prioritization of disease genes through cross-species phenotype comparison. *Genome Res*. 2014;24(2):340–348.
33 Robinson PN, Ravanmehr V, Jacobsen JOB, et al. Interpretable clinical genomics with a likelihood ratio paradigm. *Am J Hum Genet*. 2020;107(3):403–417.
34 Yang H, Robinson PN, Wang K. Phenolyzer: phenotype-based prioritization of candidate genes for human diseases. *Nat Methods*. 2015;12(9):841–843.
35 Dunn OJ. Multiple comparisons among means. *J Am Stat Assoc*. 1961;56:52–64. https://doi.org/10.1080/01621459.1961.10482090.
36 Lex A, Gehlenborg N, Strobelt H, Vuillemot R, Pfister H. UpSet: visualization of intersecting sets. *IEEE Trans Vis Comput Graph*. 2014;20(12):1983–1992.
37 Yong SJ. Long COVID or post-COVID-19 syndrome: putative pathophysiology, risk factors, and treatments. *Infect Dis*. 2021;53(10):737–754.
38 Pott-Junior H, Bittencourt NQP, Chacha SFG, Luporini RL, Cominetti MR, Anibal FDF. Elevations in liver transaminases in COVID-19: (how) are they related? *Front Med (Lausanne)*. 2021;8:705247.
39 Gómez-Pastora J, Weigand M, Kim J, et al. Hyperferritinemia in critically ill COVID-19 patients - is ferritin the product of inflammation or a pathogenic mediator? *Clin Chim Acta*. 2020;509:249–251.
40 Gan Q, Gong B, Sun M, et al. A high percentage of patients recovered from COVID-19 but discharged with abnormal Liver function tests. *Front Physiol*. 2021;12:642922.
41 Martha JW, Wibowo A, Pranata R. Hypocalcemia is associated with severe COVID-19: a systematic review and meta-analysis. *Diabetes Metabol Syndr*. 2021;15(1):337–342.
42 Litvinov RI, Evtugina NG, Peshkova AD, et al. Altered platelet and coagulation function in moderate-to-severe COVID-19. *Sci Rep*. 2021;11(1):16290.
43 Marshall M. The four most urgent questions about long COVID. *Nature*. 2021;594(7862):168–170.
44 Stavem K, Ghanima W, Olsen MK, Gilboe HM, Einvik G. Persistent symptoms 1.5-6 months after COVID-19 in non-hospitalised subjects: a population-based cohort study. *Thorax*. 2021;76(4):405–407.
45 Kayaaslan B, Eser F, Kalem AK, et al. Post-COVID syndrome: a single-center questionnaire study on 1007 participants recovered from COVID-19. *J Med Virol*. 2021;93(12):6566–6574.
46 Sonnweber T, Tymoszuk P, Sahanic S, et al. Investigating phenotypes of pulmonary COVID-19 recovery: a longitudinal observational prospective multicenter trial. *Elife*. 2022;11:e72500. https://doi.org/10.7554/eLife.72500.
47 Fernández-de-Las-Peñas C, Martín-Guerrero JD, Florencio LL, et al. Clustering analysis reveals different profiles associating long-term post-COVID symptoms, COVID-19 symptoms at hospital admission and previous medical co-morbidities in previously hospitalized COVID-19 survivors. *Infection*. 2022. https://doi.org/10.1007/s15010-022-01822-x.
48 Ziauddeen N, Gurdasani D, O'Hara ME, et al. Characteristics and impact of long covid: findings from an online survey. *PLoS One*. 2022;17(3):e0264331.
49 Davis HE, Assaf GS, McCorkell L, et al. Characterizing long COVID in an international cohort: 7 months of symptoms and their impact. *EClinicalMedicine*. 2021;38:101019.
50 Sudre CH, Murray B, Varsavsky T, et al. Attributes and predictors of long COVID. *Nat Med*. 2021;27(4):626–631.
51 Gates KM, Fisher ZF, Arizmendi C, Henry TR, Duffy KA, Mucha PJ. Assessing the robustness of cluster solutions obtained from sparse count matrices. *Psychol Methods*. 2019;24(6):675–689.
52 Ganesan S, Galer PD, Helbig KL, et al. A longitudinal footprint of genetic epilepsies using automated electronic medical record interpretation. *Genet Med*. 2020;22(12):2060–2070.
53 100,000 Genomes Project Pilot Investigators, Smedley D, Smith KR, et al. 100,000 genomes pilot on rare-disease diagnosis in health care - preliminary report. *N Engl J Med*. 2021;385(20):1868–1880.
54 Al-Aly Z, Xie Y, Bowe B. High-dimensional characterization of post-acute sequelae of COVID-19. *Nature*. 2021;594(7862):259–264.
55 Jacobsen JOB, Baudis M, Baynam GS, et al. The GA4GH phenopacket schema defines a computable representation of clinical data. *Nat Biotechnol*. 2022;40(6):817–820.
56 Duerlund LS, Shakar S, Nielsen H, Bodilsen J. Positive predictive value of the ICD-10 diagnosis code for long-COVID. *Clin Epidemiol*. 2022;14:141–148.
57 Barizien N, Le Guen M, Russel S, Touche P, Huang F, Vallée A. Clinical characterization of dysautonomia in long COVID-19 patients. *Sci Rep*. 2021;11(1):14042.
58 Haendel MA, Chute CG, Bennett TD, et al. The National COVID Cohort Collaborative (N3C): rationale, design, infrastructure, and deployment. *J Am Med Inform Assoc*. 2021;28(3):427–443.