Generalisable Long COVID Subtypes: Findings from the NIH N3C and RECOVER Programmes

Justin T. Reese,[1] Hannah Blau,[2] Elena Casiraghi,[1,3] Timothy Bergquist,[4] Johanna J. Loomba,[5] Tiffany J. Callahan,[6] Bryan Laraway,[7] Corneliu Antonescu,[8] Ben Coleman,[2] Michael Gargano,[2] Kenneth J. Wilkins,[9] Luca Cappelletti,[3] Tommaso Fontana,[3] Nariman Ammar,[10] Blessy Antony,[11] T. M. Murali,[11] J. Harry Caufield[1], Guy Karlebach,[2] Julie A McMurry,[7] Andrew Williams,[12,13,14] Richard Moffitt,[15] Jineta Banerjee,[4] Anthony E. Solomonides[16], Hannah Davis,[17] Kristin Kostka,[14] Giorgio Valentini,[3] David Sahner,[18] Christopher G. Chute,[19] Charisse Madlock-Brown,[10] Melissa A Haendel,[7] Peter N. Robinson[2,20,*] on behalf of the N3C consortium[†] and the RECOVER Consortium[††]

1. Environmental Genomics and Systems Biology Division, Lawrence Berkeley National Laboratory, Berkeley, CA, USA
2. The Jackson Laboratory for Genomic Medicine, 10 Discovery Drive, Farmington, CT, USA
3. AnacletoLab, Dipartimento di Informatica, Università degli Studi di Milano, Italy
4. Sage Bionetworks. Seattle, WA, USA
5. The Integrated Translational Health Research Institute of Virginia (iTHRIV), University of Virginia, Charlottesville, Virginia, USA.
6. Department of Biomedical Informatics, Columbia University Irving Medical Center, New York, NY, USA
7. University of Colorado Anschutz Medical Campus, Aurora, CO, USA
8. University of Arizona - Banner Health, Phoenix, AZ, USA
9. Biostatistics Program, National Institute of Diabetes and Digestive and Kidney Diseases, National Institutes of Health, Bethesda, MD, USA.
10. University of Tennessee Health Science Center, Memphis, TN, USA
11. Department of Computer Science, Virginia Tech, Blacksburg, VA, USA.
12. Tufts Medical Center Clinical and Translational Science Institute, Tufts Medical Center, Boston, MA, USA
13. Tufts University School of Medicine, Institute for Clinical Research and Health Policy Studies, Boston, MA, USA
14. Northeastern University, OHDSI Center at the Roux Institute, Boston, MA, USA
15. Stony Brook University Department of Biomedical Informatics and Stony Brook Cancer Center, Stony Brook, NY, USA
16. NorthShore University HealthSystem Research Institute, Evanston, IL
17. Patient-Led Research Collaborative, NY, USA
18. Axle Informatics, Rockville, MD, USA
19. Schools of Medicine, Public Health, and Nursing; Johns Hopkins University, Baltimore, MD, USA
20. Institute for Systems Genomics, University of Connecticut, Farmington, CT, USA
*Corresponding author: Peter Robinson email: Peter.Robinson@jax.org

ABSTRACT

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.

Funding

KEYWORDS

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.

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 two 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 6,469 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 have been developed, which includes 92,367 conditions, 8,615 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 4,767 OMOP concept IDs to 3,804 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 descendent terms (for instance, the HPO resource currently comprises 8,260 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 descendent 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 $IC(MICA(t_1, t_2))$. 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 $p_1, p_2, \ldots, p_n \in P$. The set of m HPO terms associated with patient i is represented as $t_1, \ldots, t_m \in p_i$. Then the similarity from patient p_i to p_j is calculated as

$$sim(p_i \rightarrow p_j) \;=\; \frac{1}{m}\left[\sum_{t_1 \in p_i} max_{t_2 \in p_j}\, IC(MICA(t_1, t_2))\right].$$

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

$$sim(p_i, p_j) \;=\; 0.5 \times sim(p_i \rightarrow p_j) \;+\; 0.5 \times sim(p_j \rightarrow p_i).$$

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). Using a previously described method, we chose c cluster centroids such that centroids were distant from one another (24). 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 (25).

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:

$$s(p, C) \;=\; mean_{q \in C}\,(sim(p, q))$$

Assuming we have k clusters from data partner 1, then a 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 between the test patient p and the cluster C_i. If the patient is equally similar to each of the k clusters, then $s_1 = s_2 =... = s_k = 1/k$. 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 $j \neq i$ will be larger. 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 1,000 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 $z = (x - μ)/σ$, 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-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 (Figure 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).

```
N3C patients
n = 13,998,246
        ↓
COVID-19+ patients
n = 5,434,528
        ↓
Patients with U09.9 (Long COVID) diagnosis
n = 20,532
```

Data partners with average long COVID HPO terms per patient > 5 AND > 750 U09.9 patients, patients >= 1 HPO term

```
Long COVID patients with
rich HPO data
n = 6,469
```

For clustering — *Data partner with most U09.9 patients*

Other data partners — For assessing generalizability

Not analyzed

```
Patients from data partner 1
n = 2,256
```
```
Patients from data partners 2-6
n = 4,213
```
```
Patients from other data partners
n = 12,609
```

Figure 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.

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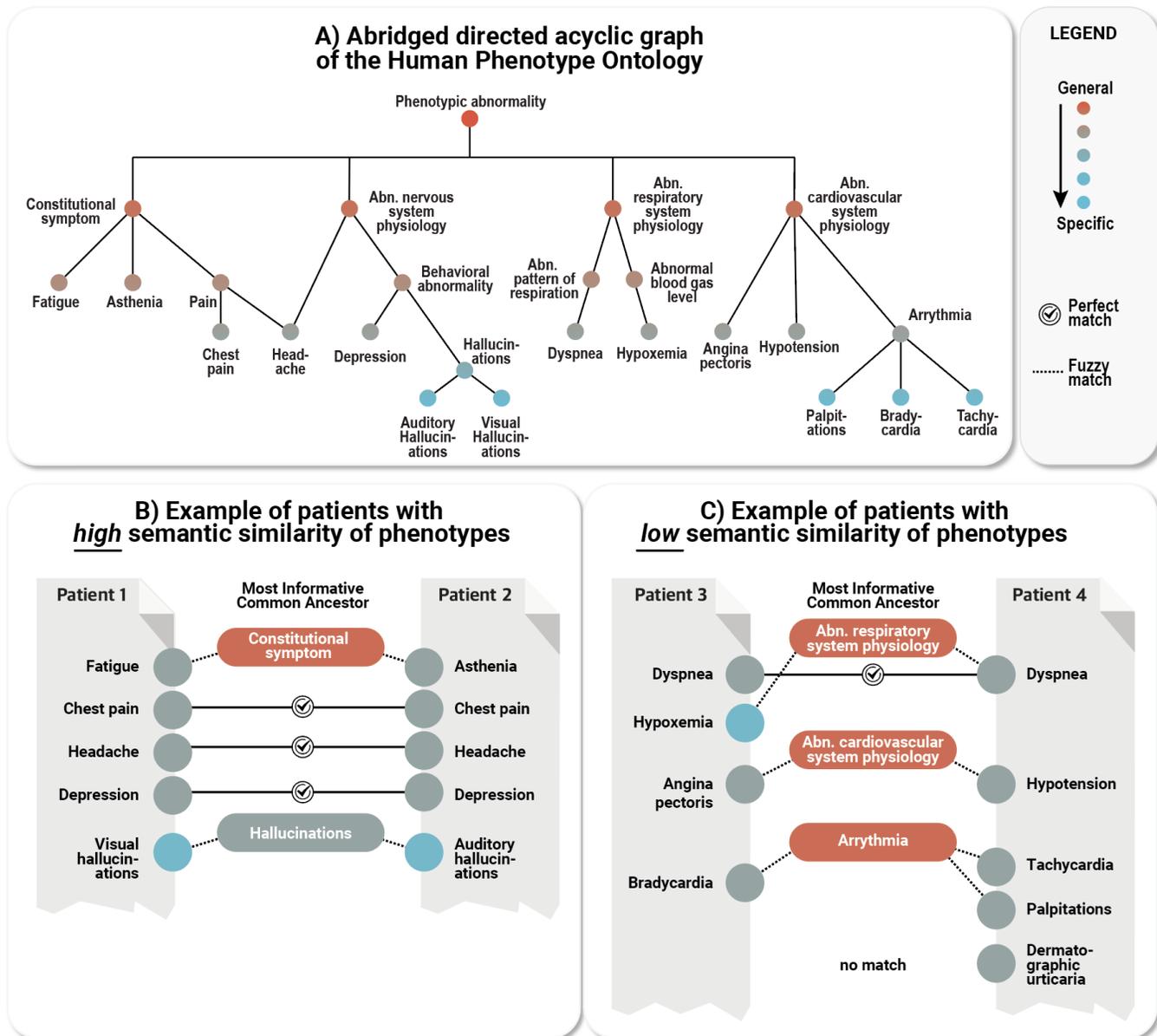

Figure 2 Calculating patient semantic similarity based on HPO phenotypes. **A)** HPO terms are arranged in a directed acyclic graph with specific terms such as *Bradycardia* (HP:0001662) being related to more general terms (here: *Arrhythmia*; HP:0011675) by subtype relations. An excerpt of the entire 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

8

similarity matrix that is used for clustering (Figure 3).

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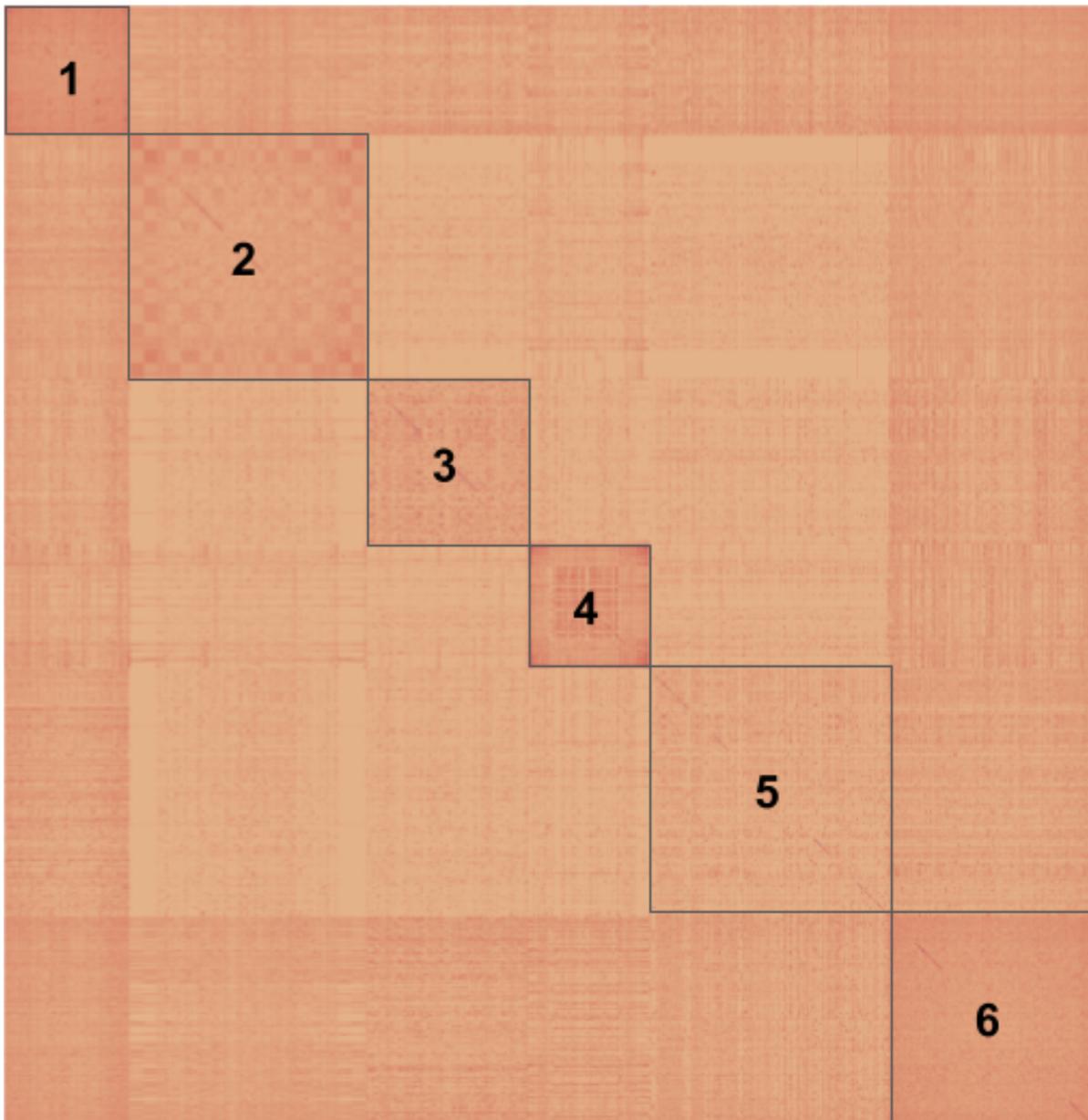
Figure 3. **Patient similarity matrix illustrating long COVID subtypes in data partner 1.** A heatmap representing the 6 clusters created by *k*-means clustering is shown. Cluster hierarchy was calculated using the nearest point algorithm and Euclidean distance.

In order to calculate pairwise phenotypic similarity of patients at data partner 1 for clustering, we leveraged the Phenomizer algorithm to calculate a 2256 x 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 (Figure 3; Supplemental Figure 1). 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 Figure S2-S3).

Feature	Overall	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5	Cluster 6
n	2256	262	491	334	250	500	419

10

	Overall	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5	Cluster 6
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%)

Table 1. **Characteristics of the study population in data partner 1**. 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 Figure S14). For instance, a cluster marked "a" is significantly different from a cluster marked "b" but not from another cluster marked "a" or "ab".

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 Supplementary 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 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 (Figure 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.

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 (Figure 4).
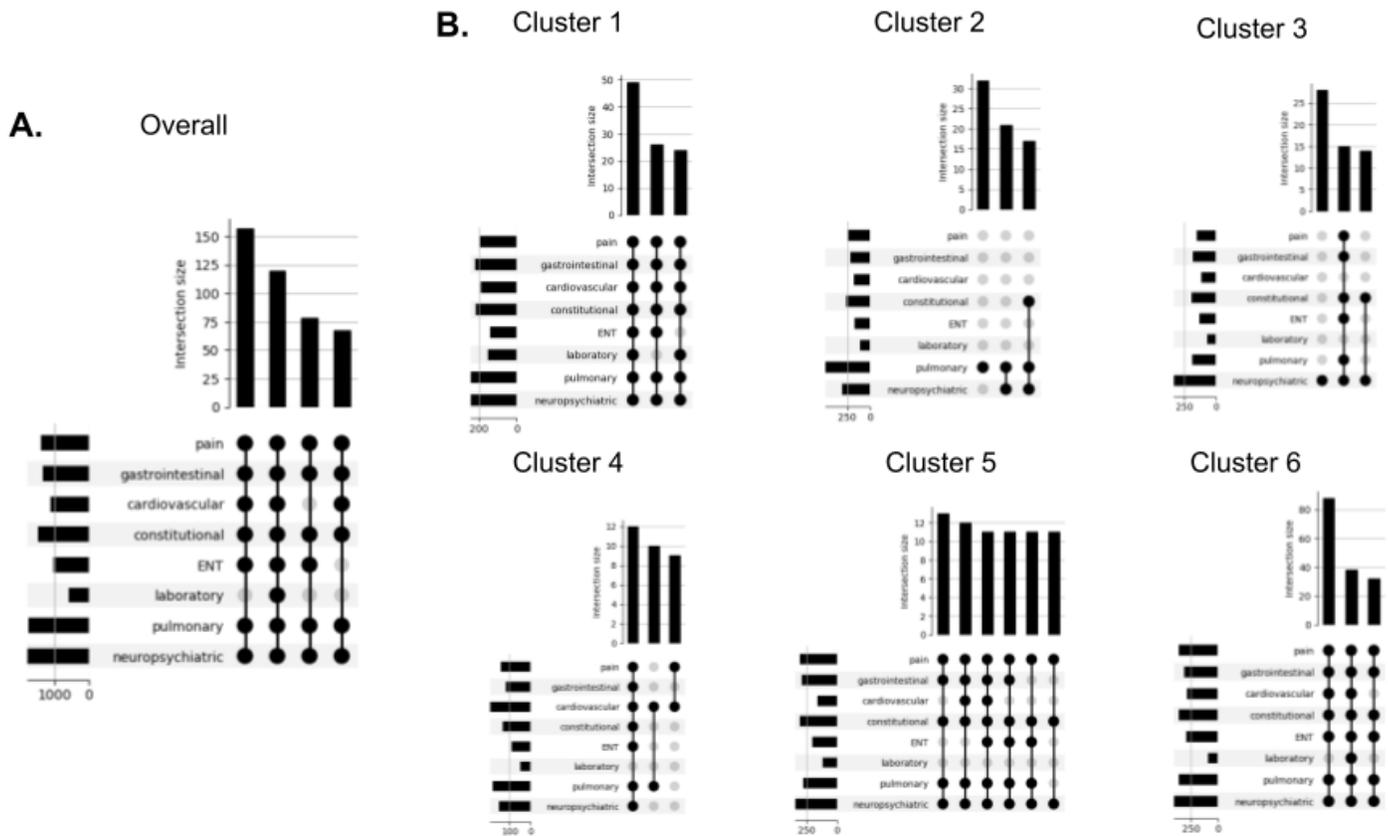
Figure 4. **Phenotypically characterising long COVID subtype clusters.** Shown are the most frequently co-occurring combinations of categories of HPO terms representing long COVID phenotypic features for patients in the overall cohort (A) and for each of the 6 clusters (B). Only those categories are shown that were found to be significantly correlated with cluster membership (chi-squared test). 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).

A)

HPO term	Cluster 1 multisystem+lab	Cluster 2 pulmonary	Cluster 3 neuropsychiatric	Cluster 4 cardiovascular	Cluster 5 pain/fatigue	Cluster 6 multisystem+pain
Hyperglycemia	78.6259542	13.64562118				
Hypocalcemia	52.29007634	1.83299389				
Lymphopenia	32.0610687	2.443991853				
Elevated circulating creatinine	31.67938931	2.036659878				
Thrombocytopenia	30.53435115	3.054989817				
Elevated circulating AST	30.15267176	0.6109979633				
Elevated circulating ALP	28.6259542	0.6109979633				
Headache	26.71755725	6.109979633				
Insomnia	24.42748092	5.49898167				
Depression	23.66412214	3.869653768				
Sleep apnea	20.99236641	5.702647658				
Abnormality of movement	16.03053435	2.851323829				
Paresthesia	10.30534351	1.221995927				
Hypoxemia	62.21374046	56.21181263				
Cough	43.89312977	57.02647658				
Pulmonary embolism	16.03053435	1.629327902				
Pain	66.79389313	15.88594705				
Chest pain	42.7480916	10.79429735				
Myalgia	19.08396947	1.425661914				
Fatigue	51.90839695	18.94093686				
Asthenia	37.78625954	7.535641548				
Fever	25.57251908	5.906313646				
Tachycardia	37.78625954	8.553971487				
Hypotension	23.66412214	2.036659878				
Palpitations	14.1221374	1.018329939				
Diarrhea	32.82442748	2.443991853				
Abdominal pain	25.19083969	0.6109979633				
Nausea	22.51908397	2.036659878				
Vertigo	17.55725191	3.258655804				

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%)

Figure 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 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 Figure 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.

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 Supplementary 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)

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 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.

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

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

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

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

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, Figure 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 to 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).

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

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

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 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 ($p = 5 \times 10^{-5}$) and in data partners 2-6 ($p = 5 \times 10^{-5}$) using a Fisher's exact test calculated by the Monte Carlo method with 100,000 permutations.

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	▏

Table 5. Recorded deaths according to cluster. 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.

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 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

18

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 Figure S2-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: JTR, PNR
Methodology: JTR, HB, TB, JJL, TC, BL, EC, BC, MG, KW, LC, TF, NA, BA, TMM, GK, JAM, GV, DS, CGC, CM-B, AW, RM, JB
Investigation: JTR, HB, CA, AES, HD, KK
Funding acquisition: MAH, PNR
Supervision: JTR, MAH, PNR
Writing – original draft: JTR, PNR
Writing – review & editing: JTR, DLA, PRB, JHC, JLS, EH
All authors read and approved the final version of the manuscript.
JTR and PNR verified the underlying data.

DECLARATION OF INTERESTS

Authors declare that they have no competing interests.

ACKNOWLEDGMENTS

California, San Francisco — UL1TR001872: UCSF Clinical and Translational Science Institute

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".

REFERENCES

1. Weekly operational update on COVID-19 - 30 March 2022 [Internet]. [cited 2022 Apr 20]. Available from: https://www.who.int/publications/m/item/weekly-operational-update-on-covid-19---30-march-2022

2. Raveendran AV, Jayadevan R, Sashidharan S. Long COVID: An overview. Diabetes Metab Syndr. 2021 May;15(3):869–75.

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 Sep;18(9):e1003773.

4. Michelen M, Manoharan L, Elkheir N, Cheng V, Dagens A, Hastie C, et al. Characterising long COVID: a living systematic review. BMJ Glob Health [Internet]. 2021 Sep;6(9). Available from: http://dx.doi.org/10.1136/bmjgh-2021-005427

5. Nalbandian A, Sehgal K, Gupta A, Madhavan MV, McGroder C, Stevens JS, et al. Post-acute COVID-19 syndrome. Nat Med. 2021 Apr;27(4):601–15.

6. Crook H, Raza S, Nowell J, Young M, Edison P. Long covid-mechanisms, risk factors, and management. BMJ. 2021 Jul 26;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 Aug 11;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 Apr;22(4):e102–7.

9. Kenny G, McCann K, O'Brien C, Savinelli S, Tinago W, Yousif O, et al. Identification of Distinct Long COVID Clinical Phenotypes Through Cluster Analysis of Self-Reported Symptoms. Open Forum Infect Dis. 2022 Apr;9(4):ofac060.

10. 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 Dec 1;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):610–5.

12. Köhler S, Gargano M, Matentzoglu N, Carmody LC, Lewis-Smith D, Vasilevsky NA, et al. The Human Phenotype Ontology in 2021. Nucleic Acids Res. 2021 Jan 8;49(D1):D1207–17.

13. Voss EA, Makadia R, Matcho A, Ma Q, Knoll C, Schuemie M, et al. Feasibility and utility of applications of the common data model to multiple, disparate observational health databases. J Am Med Inform Assoc. 2015 May;22(3):553–64.

14. Robinson PN, Webber C. Phenotype ontologies and cross-species analysis for translational research. PLoS Genet. 2014 Apr;10(4):e1004268.

15. Köhler S, Vasilevsky NA, Engelstad M, Foster E, McMurry J, Aymé S, et al. The Human Phenotype Ontology in 2017. Nucleic Acids Res. 2017 Jan 4;45(D1):D865–76.

16. Robinson PN, Mundlos S. The human phenotype ontology. Clin Genet. 2010 Jun;77(6):525–34.

17. Groza T, Köhler S, Moldenhauer D, Vasilevsky N, Baynam G, Zemojtel T, et al. The Human Phenotype Ontology: Semantic Unification of Common and Rare Disease. Am J Hum Genet. 2015 Jul;97(1):111–24.

18. Köhler S, Doelken SC, Mungall CJ, Bauer S, Firth HV, Bailleul-Forestier I, et al. The Human Phenotype Ontology project: linking molecular biology and disease through phenotype data. Nucleic Acids Res. 2014 Jan;42(Database issue):D966–74.

19. Zhang XA, Yates A, Vasilevsky N, Gourdine JP, Callahan TJ, Carmody LC, et al. Semantic integration of clinical laboratory tests from electronic health records for deep phenotyping and biomarker discovery. NPJ Digit Med [Internet]. 2019 May 2;2. Available from: http://dx.doi.org/10.1038/s41746-019-0110-4

20. Callahan TJ, Stefanski AL, Wyrwa JM, Zeng C, Ostropolets A, Banda JM, et al. Ontologizing Health Systems Data at Scale: Making Translational Discovery a Reality. 2022 Sep 10 [cited 2022 Sep 22]; Available from: http://dx.doi.org/10.48550/arXiv.2209.04732

21. Jackson R, Matentzoglu N, Overton JA, Vita R, Balhoff JP, Buttigieg PL, et al. OBO Foundry in 2021: operationalizing open data principles to evaluate ontologies. Database [Internet]. 2021 Oct 26;2021. Available from: http://dx.doi.org/10.1093/database/baab069

22. Köhler S, Schulz MH, Krawitz P, Bauer S, Dölken S, Ott CE, et al. Clinical diagnostics in human genetics with semantic similarity searches in ontologies. Am J Hum Genet. 2009 Oct;85(4):457–64.

23. Pesquita C, Faria D, Falcão AO, Lord P, Couto FM. Semantic similarity in biomedical ontologies. PLoS Comput Biol. 2009 Jul;5(7):e1000443.

24. Arthur D, Vassilvitskii S. k-means++: the advantages of careful seeding. SODA '07 [Internet]. 2007 [cited 2022 May 11]; Available from: https://www.semanticscholar.org/paper/5e0c61b7ee4a2de183a197f32c5013ad109531fa

25. Steinley D. K-means clustering: a half-century synthesis. Br J Math Stat Psychol. 2006 May;59(Pt 1):1–34.

26. Hennig C. Dissolution point and isolation robustness: Robustness criteria for general cluster analysis methods. J Multivar Anal. 2008 Jul 1;99(6):1154–76.

27. García-Escudero LÁ, Gordaliza A. Robustness Properties of k Means and Trimmed k Means. J Am Stat Assoc. 1999 Sep 1;94(447):956–69.

28. Barak S, Mokfi T. Evaluation and selection of clustering methods using a hybrid group MCDM. Expert Syst Appl. 2019 Dec 30;138:112817.

29. Dinno A. Nonparametric Pairwise Multiple Comparisons in Independent Groups using Dunn's Test. Stata J. 2015 Apr 1;15(1):292–300.

30. Piepho HP. An Algorithm for a Letter-Based Representation of All-Pairwise Comparisons. J Comput Graph Stat. 2004 Jun 1;13(2):456–66.

31. Crawford K, Xian J, Helbig KL, Galer PD, Parthasarathy S, Lewis-Smith D, et al. Computational analysis of 10,860 phenotypic annotations in individuals with SCN2A-related disorders. Genet Med. 2021 Jul;23(7):1263–72.

32. Robinson PN, Köhler S, Oellrich A, Sanger Mouse Genetics Project, Wang K, Mungall CJ, et al. Improved exome prioritization of disease genes through cross-species phenotype comparison. Genome Res. 2014 Feb;24(2):340–8.

33. Robinson PN, Ravanmehr V, Jacobsen JOB, Danis D, Zhang XA, Carmody LC, et al. Interpretable Clinical Genomics with a Likelihood Ratio Paradigm. Am J Hum Genet. 2020 Sep 3;107(3):403–17.

34. Yang H, Robinson PN, Wang K. Phenolyzer: phenotype-based prioritization of candidate genes for human diseases. Nat Methods. 2015 Sep;12(9):841–3.

35. Dunn OJ. Multiple Comparisons among Means [Internet]. Vol. 56, Journal of the American Statistical Association. 1961. p. 52–64. Available from: http://dx.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 Dec;20(12):1983–92.

37. Yong SJ. Long COVID or post-COVID-19 syndrome: putative pathophysiology, risk factors, and treatments. Infect Dis. 2021 Oct;53(10):737–54.

38. Pott-Junior H, Bittencourt NQP, Chacha SFG, Luporini RL, Cominetti MR, Anibal F de F. Elevations in Liver Transaminases in COVID-19: (How) Are They Related? Front Med. 2021 Jul 15;8:705247.

39. Gómez-Pastora J, Weigand M, Kim J, Wu X, Strayer J, Palmer AF, et al. Hyperferritinemia in critically ill COVID-19 patients - Is ferritin the product of inflammation or a pathogenic mediator? Clin Chim Acta. 2020 Oct;509:249–51.

40. Gan Q, Gong B, Sun M, Cao Z, Zheng Y, Zhang Y, et al. A High Percentage of Patients Recovered From COVID-19 but Discharged With Abnormal Liver Function Tests. Front Physiol. 2021 Mar 17;12:642922.

41. Martha JW, Wibowo A, Pranata R. Hypocalcemia is associated with severe COVID-19: A systematic review and meta-analysis. Diabetes Metab Syndr. 2021 Jan;15(1):337–42.

42. Litvinov RI, Evtugina NG, Peshkova AD, Safiullina SI, Andrianova IA, Khabirova AI, et al. Altered platelet and coagulation function in moderate-to-severe COVID-19. Sci Rep. 2021 Aug 11;11(1):16290.

43. Marshall M. The four most urgent questions about long COVID. Nature. 2021 Jun;594(7862):168–70.

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 Apr;76(4):405–7.

45. Kayaaslan B, Eser F, Kalem AK, Kaya G, Kaplan B, Kacar D, et al. Post-COVID syndrome: A single-center questionnaire study on 1007 participants recovered from COVID-19. J Med Virol. 2021 Dec;93(12):6566–74.

46. Sonnweber T, Tymoszuk P, Sahanic S, Boehm A, Pizzini A, Luger A, et al. Investigating phenotypes of pulmonary COVID-19 recovery: A longitudinal observational prospective multicenter trial. Elife [Internet]. 2022 Feb 8;11. Available from: http://dx.doi.org/10.7554/eLife.72500

47. Fernández-de-Las-Peñas C, Martín-Guerrero JD, Florencio LL, Navarro-Pardo E, Rodríguez-Jiménez J, Torres-Macho J, 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 [Internet]. 2022 Apr 22; Available from: http://dx.doi.org/10.1007/s15010-022-01822-x

48. Ziauddeen N, Gurdasani D, O'Hara ME, Hastie C, Roderick P, Yao G, et al. Characteristics and impact of Long Covid: Findings from an online survey. PLoS One. 2022 Mar 8;17(3):e0264331.

49. 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 Aug;38:101019.

50. Sudre CH, Murray B, Varsavsky T, Graham MS, Penfold RS, Bowyer RC, et al. Attributes and predictors of long COVID. Nat Med. 2021 Apr;27(4):626–31.

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 Dec;24(6):675–89.

52. Ganesan S, Galer PD, Helbig KL, McKeown SE, O'Brien M, Gonzalez AK, et al. A longitudinal footprint of genetic epilepsies using automated electronic medical record interpretation. Genet Med. 2020 Aug 10;22(12):2060–70.

53. 100,000 Genomes Project Pilot Investigators, Smedley D, Smith KR, Martin A, Thomas EA, McDonagh EM, et al. 100,000 Genomes Pilot on Rare-Disease Diagnosis in Health Care - Preliminary Report. N Engl J Med. 2021 Nov 11;385(20):1868–80.

54. Al-Aly Z, Xie Y, Bowe B. High-dimensional characterization of post-acute sequelae of COVID-19. Nature. 2021 Jun;594(7862):259–64.

55. Jacobsen JOB, Baudis M, Baynam GS, Beckmann JS, Beltran S, Buske OJ, et al. The GA4GH Phenopacket schema defines a computable representation of clinical data. Nat Biotechnol. 2022 Jun;40(6):817–20.

56. Duerlund LS, Shakar S, Nielsen H, Bodilsen J. Positive Predictive Value of the ICD-10 Diagnosis Code for Long-COVID. Clin Epidemiol. 2022 Feb 9;14:141–8.

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 Jul 7;11(1):14042.

58. 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 Mar 1;28(3):427–43.