Multi-method phenotyping of Long COVID patients using high-dimensional symptom data

Tessa D. Green*

tessa.d.green@proton.me

Patient-Led Research Collaborative, Washington, DC, US; Harvard Medical School, Department of Systems Biology, Boston, MA, US https://orcid.org/0000-0002-6075-2058

Christopher McWilliams*

Patient-Led Research Collaborative, Washington, DC, US; University of Bristol, Department of Engineering Mathematics, Bristol, UK

Leonardo de Figueiredo*

Patient-Led Research Collaborative, Washington, DC, US

Letícia Soares

Patient-Led Research Collaborative, Washington, DC, US https://orcid.org/0000-0002-6933-8048

Beth Pollack

Patient-Led Research Collaborative, Washington, DC, US; Massachusetts Institute of Technology, Department of Biological Engineering, Cambridge, MA, US

Allison K. Cohen

Patient-Led Research Collaborative, Washington, DC, US; University of California San Francisco, Department of Epidemiology & Biostatistics and Philip R. Lee Institute for Health Policy Studies, San Francisco, US

Tan Zhi-Xuan

Massachusetts Institute of Technology, Department of Electrical Engineering and Computer Science and Department of Brain and Cognitive Science, Cambridge, MA, US

Tess Falor

Patient-Led Research Collaborative, Washington, DC, US; Renegade Research, Denver, CO, US https://orcid.org/0000-0003-0111-3114

Hannah E. Davis

Patient-Led Research Collaborative, Washington, DC, US

Article

Keywords: Long COVID, symptom phenotyping, COVID-19, unsupervised machine learning, clustering

Posted Date: October 2nd, 2024

DOI: https://doi.org/10.21203/rs.3.rs-4901463/v1

Additional Declarations: Yes there is potential Competing Interest. At the time of the research, Tessa Green was affiliated with Patient-Led Research Collaborative and Harvard Medical School; Tessa Green is currently an employee of KiraGen Bio.

Abstract

Background

Long COVID, characterized by symptoms that remain or emerge in the months after infection with COVID-19, has complex and highly variable patient presentations, with myriad seemingly disconnected symptoms.

Methods

We apply three different machine learning techniques to identify groups of patients with similar symptoms in a large patient-reported symptom dataset with the aim of identifying robust Long COVID phenotypes.

Results

All three methods produced clinically plausible symptom clusters which are technically valid partitions of the high-dimensional symptom space. However, concordance across methods was low. Some features did recur, such as low-symptom count clusters having the highest average age and lowest proportion of women, and specific recurrent clusters or subclusters across pairs of methods.

Conclusions

The high sensitivity of observed patient clusters to algorithm choice has implications for other studies reporting Long COVID phenotype clustering, as it suggests that a single method may provide an incomplete or unstable partition of the cohort, particularly in studies with fewer symptoms observed. With the 162 reported symptoms considered here, patient presentations vary smoothly and segmentation, while internally consistent, was not reproducible across methods; this suggests that the complexity of LC symptom presentation may easily be missed by clustering approaches that use insufficient data or overly-simplistic clustering methods. Future work would likely benefit from semi-supervised approaches matching patients to pre-defined phenotypes or diagnoses, or from the inclusion of additional patient data. Overall, our multi-method analysis highlights the importance of assessing clustering robustness and considering the full scope of patient symptoms when evaluating treatments.

***Tessa D. Green, Chris McWilliams, and Leonardo de Figueiredo share first authorship.**

Introduction

Long COVID is a debilitating condition that affects multiple tissues and organ systems and is likely to have a multifactorial and overlapping symptom etiology (Davis et al., 2023; Turner et al., 2023). Accumulating empirical evidence supports multiple but compatible pathogenesis of Long COVID including, but not limited to, vascular endothelial dysfunction (Fogarty et al., 2021; Prasannan et al.,

2022), fibrinaloid microclots (Pretorius et al., 2022), persisting reservoirs of SARS-CoV-2 (Stein et al., 2022; Swank et al., 2023; Zollner et al., 2022), autoimmunity (Lim et al., 2023), T-cell dysregulation and exhaustion (Klein et al., 2023; Peluso et al., 2021; Yin et al., 2024), reactivation of latent viral infections (Klein et al., 2023), organ damage (Dennis et al., 2021), gut dysbiosis (Ancona et al., 2023; D. Zhang et al., 2023), and neuroinflammation (Soung et al., 2022). This multisystemic and complex pathophysiology gives rise to a wide range of symptoms, which may be associated with phenotypic variation in disease presentation. A patient-led survey of Long COVID patients (H. E. Davis et al., 2021b) identified 203 symptoms in 10 organ systems, and a Human Phenotype Ontology curation (Deer et al., 2021) identified 287 phenotypic abnormalities associated with Long COVID. Patients frequently experience post-exertional malaise (PEM), a physiological state characterized by worsened symptoms after disproportionately minor physical or cognitive exertion.

Improving understanding of Long COVID natural history can support clinical decision-making and research, including clinical trial design. Disease phenotypes combine observable traits to classify patients into clinical subtypes, which can potentially facilitate pathophysiological and therapeutic research, prognosis prediction, and clinical care. Machine learning approaches can be used to identify complex disease phenotypes; for Long COVID, machine learning approaches have been applied to high-dimensional data to characterize sex-disaggregated subphenotypes (H. Zhang et al., 2023), to discriminate clinical presentations in a cohort of pediatric patients (Lorman et al., 2023), and to describe temporal changes in symptom signals (Dagliati et al., 2023). However, current Long COVID phenotype clustering studies have limitations, including not surveying enough symptoms and/or core symptoms (e.g., post-exertional malaise, neurocognitive symptoms), non-representative sampling of patients, being restricted to intrinsically biased EHR data, sampling only hospitalized patients, or not involving patient lived experience, thus resulting in non-representative clusters with limited generalizability.

Here, we use data from a patient-led survey (H. E. Davis et al., 2021b) to group people with Long COVID into phenotypic clusters, and compare how the choice of clustering algorithm affects dimensionality reduction of Long COVID phenotypic variation. We characterize the number of phenotypes, the traits that define them, and describe the characteristics of patients grouped within each phenotype.

Ultimately, we assess and compare these findings using three clustering methods: an autoencoder, ensemble clustering, and latent class analysis, henceforth method A, B and C, respectively. Our goal in comparing the phenotypes obtained through different clustering methods stems from the inconsistent phenotypes reported in the literature obtained through dimensionality reduction of Long COVID traits. We hypothesized that any robust structure in the dataset should be detectable and replicable across different methodologies, whereas structures that are not reproducible may be artifacts of a particular method. Hence, our multi-method approach aimed to address a fundamental challenge inherent to all unsupervised machine learning, namely the lack of ground-truth labels. While we identified some recurring features, we found relatively low replicability of identified clusters across algorithms. Our findings underscore the challenges in clustering symptom data and highlight the pivotal role of algorithm selection in shaping outcomes.

Methods

Sample

[[Study details masked for blind review]] Data analyzed here is an updated version of data collected in an online cross-sectional survey using Qualtrics (v. Sept 2020) that began recruitment in September 2020; data included in this paper are through August 2023. Survey design and recruitment processes for the source data are described in detail in dataset originating publications (H. E. Davis et al., 2021b; Re'em et al., 2023) and are summarized below, along with data cleaning and processing that aligns with prior publications on earlier stopping points of this dataset. The survey questions are available at (H. E. Davis et al., 2021a). The survey study was approved by the [[masked for blind review]] Research Ethics Committee, and all participants provided written informed consent with no financial incentives or compensation.

Inclusion criteria for the survey were adults with confirmed or suspected COVID-19 at least 1 week past symptom onset date. Since access to and documentation of COVID tests varies widely (J. T. Davis et al., 2021; "Estimated COVID–19 Burden | CDC," 2023; Kucirka et al., 2020; Pecoraro et al., 2022), the World Health Organization consensus criteria of probable or confirmed COVID-19 infection was used. Among people who meet the WHO criteria for probable or confirmed COVID-19 infection, prior research indicates clinical manifestation of Long COVID does not meaningfully differ between those who had a positive COVID test and those who had no test (H. E. Davis et al., 2021b). At least 90 days of illness were required (Soriano et al., 2022) to define Long COVID.

Out of the 14,169 responses received, the following responses were removed from the dataset: did not start survey (only completed consent form) or did not complete symptom questions (n = 5,557), flagged as spam by Qualtrics (n = 8), symptom onset date before December 2019 (n = 71), 0 days of symptoms (n = 7), test entry (n = 1), duplicated participants (n = 247), responses with ongoing symptoms who had not reached 3 months (n = 2,091), and those who had recovered (n = 157). The final analytic sample was 6,031 individual survey responses. Patients had been ill for a mean of 243.7 days, median of 197 days at the time of survey (min 90, max 1209, IQR 173.5–254 days). Data were cleaned in Python version 3.7.1.

Measures

A total of 162 symptoms were originally sampled in 10 recollected time points, weekly for the first four weeks after COVID onset and monthly for six consecutive months up to month seven of illness. Symptom questions were binary reports of whether the respondent had ever experienced the symptom, and these binary responses were the only features used to perform our clustering. We also collected two severity scales for physical and cognitive symptoms of Post-Exertional Malaise (PEM). PEM was defined to participants in the survey as worsening or relapse of symptoms after physical and/or mental activity. PEM severity as reported in Tables 1–3 was rescaled to range from 0–1, where 1 indicates a survey

response of 10 (strong PEM) and 0 indicates a response of no PEM. Where PEM severity averages are reported, averages are taken only over patients who reported a numerica PEM score.

Demographic questions included age, sex, gender, and whether participants had a menstrual cycle. Age was reported as a categorical range (18–29, 30–39, 40–49, 50–59, 60–69, 70–79, 80+). We used the range midpoint to report average ages, and all participants who reported being 80 + were assumed to be age 80. Gender was categorized as Female/Woman, Male/Man, Non-binary/Genderqueer/Gender non-conforming, Prefer not to say, or Other, and sex was operationalized in the survey as a followup question asking if participants gender matched their gender assigned as birth, with yes or no options.

Analysis approach

We used three different analytic techniques to identify Long COVID disease phenotypes that are robust across modeling approaches. Specifically, we used a genetically optimized autoencoder (Method A), an ensemble clustering method (Method B), and latent class analysis (Method C). All modeling and analyses were conducted in Python and the code is available at https://anonymous.4open.science/r/PLR/README.md Clusterings were performed on the symptom data only; metadata consisting of age, gender, and reported cognitive and physical PEM severity, was only used for analysis of symptom-based clusterings. Symptom groupings used for visualization were determined by conceptually grouping symptom clusters into categories (Supp Data 1) .

Analysis metrics

Throughout the analyses, we used the Adjusted Mutual Information (AMI) metric to compare the similarity of pairs of clusterings while accounting for chance agreement (Vinh et al., 2010). AMI is considered the best metric for comparing clusterings with different cluster numbers or sizes (Romano et al., 2016). To compute the AMI where one clustering only has a partial label assignment, it is necessary to produce a complete assignment by either predicting or imputing the missing labels. This situation was handled as follows for two different sources of partial label assignment:

1. When subsampling of the dataset to evaluate robustness of a given method: patients not included in the sample were not assigned a cluster label. Model-based clustering method C predicted missing labels based on the model. Ensemble-based method B applied k-nearest neighbor imputation to predict the missing labels.
2. When HDBSCAN assigned cluster label − 1: such patients are part of the noise cluster and were considered part of the same group for the purpose of AMI calculation. In practice, we found that retaining the − 1 label made little difference as the noise cluster was small (< 4% of the dataset for method A).

Method A: Autoencoder - HDBSCAN

We developed a neural network based approach to cluster the high-dimensional symptom data. We first refined the original feature set by dropping some symptom features, which was achieved through setting

two thresholds: one based on dropping features skewed heavily towards either 0 or 1 and the other based on symptom correlation (Supp Table A1), effectively allowing us to reduce the dimensionality of the dataset. This approach was inspired by the questionnaire being launched in the early stage of the pandemic where relevant symptom features were still not clear.

Since clustering techniques perform better in lower dimensions (Domeniconi et al., 2007), we subsequently employed an autoencoder neural network to learn a 2-dimensional embedding of the dataset. HDBSCAN (Campello et al., 2013) clustering was then applied to the resulting dataset. Various hyperparameters, including the skewness and correlation thresholds, the number of layers and neurons per layer, and finally, the HDBSCAN parameters were all optimized using a genetic algorithm that utilized the silhouette score (Rousseeuw, 1987) as a measure of individual fitness.

We refined our approach by using the genetic algorithm's outcomes to instead define a narrower space for a grid search. The outcomes of this secondary optimisation had greater clinical interpretability. For robustness, an autoencoder pipeline, initialized with Glorot initialization (Glorot and Bengio, 2010), was run five times with different initial weights and the AMI was calculated across clusters. Supp Table A2 details the parameters used in the pipeline.

In the analysis of our clusters, we used symptom prevalence as a key measure. This is defined as the mean of the binary symptom value over a group of patients and is equivalent to the frequency of occurrence of the symptom within that group. To understand which symptoms were characteristic of each cluster we defined the metric δ as the difference between the symptom prevalence at the cohort level and the prevalence within the cluster. Symptoms with $\delta > 0$ and $\delta < 0$ are referred to as enriched and dis-enriched respectively. Those symptoms with $|\delta|>0.1$ are considered to be strongly (dis-)enriched, whereas $|\delta|<0.1$ is referred to as mild. These enriched and dis-enriched symptoms were used to define cluster characteristics for ease of interpretation. The listed characteristics sometimes combine multiple symptoms. They were chosen for interpretability and should not be taken as fully descriptive. Complete symptom prevalences and enrichments are available in Supp Data 2.

Method B: Ensemble clustering

We developed an ensemble clustering method based on a pipeline consisting of a dimensionality reduction algorithm and a shallow clustering algorithm. Following some initial experimentation (see Supp Section SB.1), UMAP (McInnes et al., 2020) plus k-means (Lloyd, 1982) were used for this pipeline. We generated a library of 500 diverse clustering solutions with random parameterizations of the pipeline (the search space is defined in Supp Table B1), and then employed an ensemble selection method (Fern and Lin, 2008) that greedily adds clustering solutions in order to jointly optimize the 'quality' and 'diversity' of the ensemble (see Supp Section SB.1 for full details). We then applied a standard consensus function (Boongoen and Iam-On, 2018) to aggregate the ensemble solutions into a single clustering. This consensus function involves building a co-association matrix A that counts how many times each pair of patients occur in the same cluster across the ensemble, normalizing this matrix by the size of the ensemble such that each element $a_{ij} \in (0,1)$, and then running this matrix through a

similarity-based clustering algorithm. In our case, spectral clustering was found to produce the most stable solutions. The full ensemble method was repeated 10 times with different random seeds to assess the stability of the solution. The optimal number of clusters was chosen using a combination the eigengap heuristic (von Luxburg, 2007) and the stability of the clusters. The full ensemble method was repeated using three subsampling regimes to determine robustness to data removal. The three regimes used random samples of 1) 80% of symptoms; 2) 80% of patients; and 3) 80% of both patients and symptoms. Cluster characteristics and symptom prevalences were analyzed as described for Method A.

Method C: Latent Class Analysis

Latent class analysis (LCA) was performed using StepMix (Morin et al., 2024) v. 2.1.3 in Python 3.12. Default parameter settings were used when not otherwise specified. Symptoms present in more than 95% or fewer than 5% of patients were removed. Grid search was performed over 2–25 clusters, with the Bayesian information criterion (BIC) computed for each cluster count. The optimal BIC was 13 clusters for the single run and an average of 13 clusters when comparing across ten random seeds, so n_components was set to 13. For the consensus clustering, which defines clusters using commonalities between ten clusterings with different random seeds, consensus was determined using a co-association matrix as described for method B above. Hierarchical clustering was performed on the vector of LCA clusters for each patient using scipy v. 1.11.3 fcluster with criterion distance and threshold 0.026, with the threshold chosen such that the number of clusters was maintained at 13. tSNE plots of the LCA clustering were created using scikit-learn v. 1.3.2 function TSNE with default parameters. The input data used for tSNE were the 13 cluster membership probabilities for each patient as outputted from single-run LCA.

For robustness analysis, the model was trained using 80% sub-samplings of patients ten times. Clusters were then assigned to all points using the StepMix function predict. The AMI of the full dataset was compared across runs. For further evaluation and exploration of performance in settings with fewer symptom counts, symptoms were subsampled to 0.1–0.9 of reported symptoms and a grid search was performed for 1–20 clusters for random seeds 1–10.

In contrast to methods A and B, to identify symptoms characterizing each cluster, we examined the parameters of the mixture model corresponding to each symptom as produced by the StepMix function get_mm_df. The value for each parameter corresponds to the probability that a patient in that cluster reported the relevant symptom. We define enriched and dis-enriched symptoms in the LCA model by subtracting the parameter value in the cluster of interest from the average parameter value for that symptom across all other clusters. Clusters had highly variable numbers of (dis-)enriched symptoms, so a per-cluster threshold was defined manually for describing cluster characteristics.

Results

We applied three different clustering methodologies, each using different machine learning modalities to partition the same high-dimensional dataset. Across all three clustering approaches, we find that the

overall structure of the data is relatively continuous (Fig. 1a). Similar symptoms tend to occur together (e.g. patients with one symptom reflecting impaired speech likely have others from the same category) (Supp Fig G2). Well-defined subgroups of patients that reflect distinct and diagnosable phenotypes were not observed. Although groupings are internally consistent within each of the three methods, most groupings are not reproducible across methods. Below, we individually present the clusterings A, B and C produced by the three methods, followed by a comparative analysis.

Method A: Genetically Optimised Autoencoder

The method produced six distinct clusters, with additional outliers identified by the HDBSCAN algorithm labeled as ‘AX’. The inclusion of this outlier cluster is how HDBSCAN explicitly handles noise in the dataset. The optimal clustering solution was determined by selecting for solutions with the highest silhouette score that had fewer than 13 clusters. Relaxing this second constraint was found to significantly hinder clinical interpretability of the solution. Our optimal clustering solution exhibits a silhouette score of 0.6133 in the representation space learned by the autoencoder (Fig. 2).

To enhance the interpretability of Fig. 2, the embedded meaning of both of the dimensions was estimated by taking as input a set of clusters $C = \{1, 2, ..., N\}$, a set of cluster pairs $P = \{(i, j) \mid i, j \in C\}$, and a set of features $F = \{1, 2, ..., M\}$. For each cluster pair $(i, j) \in P$, the algorithm calculates the absolute difference $d_{ijk} = |x_{ik} - x_{jk}|$ for each feature $k \in F$, where x_{ik} and x_{jk} are the values of feature k in clusters i and j, respectively. It then selects the top 3 features $T_{ij} \subseteq F$ with the greatest distances d_{ijk} and stores the results as tuples (i, j, k, d_{ijk}) in a set R. Finally, the algorithm sorts the elements of R based on the cluster pairs (i, j) and distances d_{ijk} in descending order and outputs the top 3 feature differences for each cluster pair. The results indicate that Dimension 1 most likely pertains primarily to sleep disturbances but also to motor related symptoms (Supp Table A3) and that Dimension 2 relates to temperature regulation symptoms and to a lesser extent gastrointestinal and musculoskeletal related symptoms (Supp Table A4). A Ward method dendrogram was created to further demonstrate similarity between clusters (Supp Fig A2).

Table 1

Summary of clustering method A. For each numeric value the cluster with the highest value is shown in bold. Descriptive names have been assigned to each cluster based on characteristic symptoms, however more details are given in the text. Age corresponds to the midpoint average of the age class of patients within a cluster. Women lists the fraction of patients whose reported gender was woman. Cognitive and physical PEM are averaged reported cognitive and physical symptom severity during PEM respectively, calculated as the average of patients who reported experiencing the reported form of PEM.

Cluster	Name	Size	Symptom count	Women	Age	Physical PEM Severity	Cognitive PEM Severity
AX	Outliers	222	16	0.667	47.6	0.619	0.271
A0	Temperature dysregulation (chills and flushing sweats). Cardio-respiratory (shortness of breath, heart palpitations and dizziness). Cognition (short term memory and attention/concentration).	1125	44	0.795	47.3	0.773	0.549
A1	Largest cluster. Highest proportion of women. Highest symptom burden. Severe insomnia and sleep interruption. Severe physical fatigue and PEM.	**2126**	**60**	**0.830**	46.7	**0.811**	**0.636**
A2	Elevated temperature. Generally elevated cardio-respiratory symptoms.	321	35	0.791	45.9	0.747	0.485
A3	Very generalized cluster. Symptom severity scattered over: cardio,cognition, short term memory, sleep disturbances, temperature dysregulation and headaches	498	46	0.791	46.3	0.773	0.594
A4	Lowest symptom burden of non outlier clusters.	850	28	0.714	**47.7**	0.719	0.480

Cluster	Name	Size	Symptom count	Women	Age	Physical PEM Severity	Cognitive PEM Severity
	Lowest severity of cognitive PEM. Notably low sleep disturbance.						
A5	Elevated insomnia, short term memory	889	37	0.721	47.1	0.735	0.530
Cohort		6031	42	0.78	47.0	0.778	0.637

Although method A reveals substantial overlap (Supp Table A5) between the clusters and their average symptom scores, there are some notable takeaways.The outliers (AX) are more likely to be male and have the lowest symptom count as well as PEM severity, with the standout feature being low Cognitive PEM Severity. In contrast, cluster A1, the cluster where patients were most likely to be women, exhibits the highest symptom count as well as highest Physical and Cognitive PEM Severity. We found a positive correlation between the fraction of women in a cluster and symptom count: (r = 0.883, p = 0.008, df = 6), Physical (r = 0.973, p = 0.0002, df = 6) and Cognitive PEM severity (r = 0.883, p = 0.008, df = 6), see Supp Fig A3. The patients in this cluster also experience high sleep and temperature disturbance related symptoms with a δ of 0.275 and 0.288 respectively (Z-scores of 1.391 and 1.282) respectively.

Cluster A4 is characterized by a low symptom count and severity, with notably low temperature related symptoms, with a mean severity of 0.0522 and a Z-Score of -1.076 or -1.237 if including or excluding outliers respectively. Further analysis was conducted to determine what made cluster A4 a significantly low burden cluster. Cluster A5 has an overall symptom burden close to the mean symptom burden of all clusters, being dis-enriched by -0.009 yet it is the nearest neighbor to Cluster A4. The symptom group that differs most in terms of severity between these two neighbors is Sleep, with a difference of 0.313; suggesting that improved sleep or the lack of sleep related symptoms improved outcomes.

Method B: Ensemble clustering

An ensemble of 50 clusterings was selected using the joint criterion (Fern and Lin, 2008) from a library of 500 base k-means clusterings, and this ensemble was used to construct a co-association matrix A. Details of the selected ensemble are provided in section SB.2. The eigenvalues of the normalized graph Laplcian L_{norm} invoked by A were computed (Zelnik-Manor and Perona, 2004), and the largest eigengaps were used to select candidates for the optimal number of clusters (Supp Fig B1, left panel). The stability of the resulting structures was then assessed by repeat clustering using different random seeds and the most stable was found to be that with 8 clusters with a mean AMI of 0.88 across ten repeats (Supp Fig B1, right panel). It should be noted that 6 clusters was also a strong candidate, with an AMI of 0.85. However, the 8-cluster structure was preferred due to its slightly better stability and the increased clinical discrimination enabled by more groups. This structure also showed good robustness

to the 3 subsampling regimes detailed in methods section B above, with mean AMIs of 0.75, 0.82 and 0.74 (Supp Fig F2) with the final clustering B.

Table 2
Similar to Table 1, but for clustering B.

Cluster	Name	Size	Symptom count	Women	Age	Physical PEM Severity	Cognitive PEM Severity
B0	Speech, memory, cognition, sleep and eye/vision (neurological).	615	53	0.833	45.1	0.806	0.739
B1	Cognitive PEM much less severe than physical PEM. Shortness of breath, chest burning pain, short term memory, heart palpitations.	872	30	0.748	47.6	0.740	0.478
B2	Highest symptom burden. Severe physical and cognitive PEM. Highest proportion of women. Sensory sensitivity, difficulty communicating verbally and processing information, neuropathy, itchy eyes and blurred vision.	506	**89**	**0.866**	45.5	**0.876**	**0.786**
B3	Largest cluster. Cognitive dysfunction symptoms most enriched. Cognition, long term memory, speech, PEM, and fatigue.	**1105**	34	0.693	47.3	0.749	0.654
B4	Lowest symptom burden. Highest average age. Lowest PEM severity. Altered smell and taste including loss, respiratory.	887	21	0.692	**50.3**	0.676	0.479
B5	Nausea, temperature dysregulation, sleep disturbance, paresthesia and	727	54	0.834	45.0	0.794	0.611

Cluster	Name	Size	Symptom count	Women	Age	Physical PEM Severity	Cognitive PEM Severity
	vibrations, palpitations, dizziness and vertigo.						
B6	Paresthesia, speech, vision problems, neuropathy, tremors, temperature dysregulation.	611	69	0.856	45.7	0.848	0.736
B7	Temperature dysregulation, dry cough, loss of smell, taste and appetite, shortness of breath.	708	40	0.836	47.2	0.763	0.554
Cohort		6031	42	0.78	47.0	0.778	0.637

A summary of the final 8-cluster structure is provided in Table 2, and also visualized in the middle panels of Fig. 1. The cluster sizes were relatively constant, with the smallest cluster B2 being just under half (0.46) the size of the largest cluster B3. This clustering produced a higher maximum silhouette score (0.24) than clusterings A and C (-0.03 and 0.01 respectively) across the set of tSNE embeddings produced (see section SG.1). This suggests, as demonstrated in Fig. 1, that it is easier for clustering B to produce a 2D embedding of the full symptom data in which these clusters are well separated than it is for clusterings A and C. This localization of the clusters in the embedding space means that they can be interpreted by visualizing the variation in symptom groupings across the space, as in Supp Fig G2. In general the symptom count increases along the long diagonal axis of the dataset (final panel, Supp Fig G2) and most symptom groups vary accordingly with their highest prevalences visible at the top left of the embedding (all panels, Supp Fig G2). This region corresponds to cluster B2, which had strong enrichment for almost all symptoms (Supp Table B2).

In terms of symptom prevalence, the differences between clusters were subtle. Symptoms which were common across the cohort tended to be common across the 8 clusters with small but clinically meaningful variation. The three most prevalent symptoms at the cohort level were fatigue (0.98), short term memory loss (0.92), and PEM (0.88). Only cluster B4 showed strong disenrichment for any of these symptoms, with a prevalence of 0.58 for PEM. Cluster B4 had the lowest symptom burden of all the clusters with a mean of 21 symptoms, highest average age (50.3), lowest cognitive PEM severity (mean 0.479), lowest proportion of women (0.69), and was characterized by four mildly enriched symptoms: loss of smell, altered sense of smell, loss of taste and respiratory (other) (Supp Table B3). These correspond to an increased prevalence of olfaction- and temperature-related symptoms that is visible toward the bottom right of the corresponding panels in Supp Fig G2, in the locality of cluster B4.

Cluster B1 is notable because of the difference between the mean severity of physical and cognitive PEM, which were 0.74 and 0.48 respectively. By comparison, cluster B3 has a similar symptom burden

(34 versus 30) and a similar severity of physical PEM (0.75) but a much higher severity of cognitive PEM (0.65). This difference is reflected in the enriched symptoms which for B3 are primarily speech and cognition-related symptoms, none of which are enriched in B1. In general, there appears to be a differentiation on either side of the long diagonal axis in the embedding space (Supp Fig G2), with patients to the right of this axis displaying more speech-, cognitive- and memory related symptoms. The other two clusters that display a reduced cognitive versus physical PEM severity are B4 and B7, which are to the left of the diagonal axis and in a region of markedly reduced cognitive-symptom prevalence.

B7 is similar to B5 as evidenced by their proximity and overlap in the embedding space. Both clusters share strongly enriched temperature dysregulation (elevated temperature, chills, flushing, sweats) and respiratory symptoms (*s*hortness of breath, tightness of chest). However, B5 includes strongly enriched cognitive and sleep disturbance components which are not present in B7. Conversely, three symptoms are strongly enriched in B7 but dis-enriched in B5: loss of smell (0.51 versus 0.31), loss of taste (0.48 versus 0.28) and fever (0.55 versus 0.37).

Method C: Latent class analysis

Latent class analysis is a probabilistic model-based clustering method that identifies groups of related cases within a heterogeneous population (Sinha et al., 2021). Patients are assumed to be sampled from a finite mixture of *k* latent classes, with each class characterized by the conditional probabilities that patients in that class experience each recorded symptom. The model learns likely classes and the parameters of those classes (symptom probabilities), which can then be used to probabilistically predict the class membership of patients. By modeling class membership probabilities, patients could be assigned to the most likely latent class. We assigned each patient to the cluster which had the highest probability, resulting in somewhat variable cluster sizes (Fig. 3a). The output of the individual LCA run are presented as the main output of this method and referred to as clustering C throughout. The model infers coefficients describing the likelihood of observing each symptom in a patient from each cluster (Supp Data 4). Most patients are well-assigned to a single cluster (Fig. 3b), and the groups are well-distinguished; there aren't any clusters where patients have a frequent contribution to another group. Cluster C1, which is the smallest, is also the most different from the other groups. Only 1.7% of patients had ambiguous assignments, defined as cases where no single cluster had probability above 0.5. Because these patients are a small fraction, and not overrepresented in any particular cluster, they were assigned to their highest probability cluster for further analysis.

The observed clusters have a high degree of coherence among both unusual and prevalent symptoms, with the most distinctive symptoms generally coming from a semantically similar group (e.g. the top distinguishing symptoms for cluster C2 were all from the speech symptom subgroup). Notably, C9 is composed of patients who seldom report sleep-related symptoms, including dramatically reduced rates of difficulty falling asleep (0.00 vs 0.44 of the full dataset), waking up during the night (0.00 vs 0.53), and insomnia (0.03 vs 0.76). Even though these patients do not report experiencing disturbed sleep, their physical and cognitive PEM severity is in line with the remainder of the dataset. Cluster C4 has the

lowest average symptom count (11, compared to 45 for the full dataset). A majority of patients in C4 still reported short-term memory impairment (0.80) and PEM (0.57). This cluster also includes all 7 participants who reported only a single symptom (short-term memory impairment). Other symptoms reported in this cluster varied by patient, but included significant symptoms such as coughing up blood, chest pain and tightness, and difficulty swallowing. The survey does not include information as to symptom severity, intensity, or duration, so these patients having a lower symptom count does not necessarily indicate a mild disease course. PEM itself often also involves multiple symptoms (Hartle et al., 2021), so a patient reporting PEM is likely experiencing a variety of additional symptoms, even if they did not report them in the survey.

Table 3
Similar to Table 1, but for clustering C.

cluster	Name	size	Symptom count	Women	Age	Physical PEM Severity	Cognitive PEM Severity
C0	Temperature dysregulation, sleep disturbance, cognition (attention/concentration) and shortness of breath	**656**	40	0.794	45.2	0.752	0.595
C1	High symptom burden with cognitive dysfunction, difficulty reading and communicating, and sensory and allergic symptoms; Elevated musculoskeletal symptom count; high physical and cognitive PEM severity	178	**101**	0.848	44.0	**0.897**	**0.818**
C2	Difficulty reading and communicating, high prevalence of cognitive symptoms	399	56	0.769	45.7	0.804	0.772
C3	Low symptom burden, cognitive symptoms reduced with short term memory difficulty still reported	549	27	0.745	47.9	0.701	0.416
C4	Lowest symptom burden, Lowest cognitive PEM severity, reduced frequency of PEM, cognitive symptoms reduced	389	11	0.640	47.6	0.637	0.394
C5	Chest tightness, dizziness/vertigo, shortness of breath, gasping for air while oxygen normal	352	44	0.835	46.8	0.783	0.593
C6	Impaired reading and communication, elevated temperature and chills/sweats, sleep difficulty and elevated average musculoskeletal symptom count	449	58	**0.902**	44.2	0.821	0.703
C7	High symptom burden with impaired speech,	544	78	0.881	45.8	0.854	0.757

cluster	Name	size	Symptom count	Women	Age	Physical PEM Severity	Cognitive PEM Severity
	auditory and visual comprehension, blurred vision, pain: muscle aches, muscle spasms, neuralgia and skin burning without associated rash.						
C8	Insomnia, difficulty falling asleep and waking up in night, lower symptom burden	490	23	0.673	**49.4**	0.732	0.539
C9	Consistent sleepers with PEM and cognitive impairment, lower symptom burden	534	24	0.738	47.3	0.696	0.574
C10	Insomnia, tremors, sensory features: tingling/prickling, skin burning without rash, vibrations.	509	45	0.800	46.9	0.751	0.555
C11	Speech and cognition with moderate symptom burden	608	38	0.745	46.3	0.767	0.682
C12	High symptom burden with reduced cognitive/speech, elevated musculoskeletal symptom count	374	63	0.813	45.3	0.820	0.675
All		6031	42	0.78	47.0	0.778	0.637

The exact clusters produced by this method varied moderately with random seed selection, with an average AMI of 0.62 (Supp Fig C1a). As such, we also produced a combined clustering where ten different maximum-likelihood estimate assignments were combined as for Clustering B, with the resulting co-association matrix hierarchically clustered to produce 13 groupings for comparison. The consensus clusters can be visualized on the same tSNE plot produced using the single-run probabilistic assignment vectors (Fig. 3c). We observe considerable overlap with the initial clustering (AMI 0.70), with the membership of 7 clusters of the original run maintaining at least 75% identity in the second (Fig. 3c). For example, 98% of the patients assigned to C1 are assigned to consensus cluster 11. In contrast, 75% of patients in C0 are in the closest corresponding consensus cluster 4; however this subgroup of patients makes up 98% of consensus cluster 4, and are distinguished from the remainder of C0 by substantially decreased likelihood of reporting headaches after mental exertion (22% of C0 patients in consensus cluster 4 vs 46% of C0 patients not in consensus cluster 4), with pressure sensation (23% vs 45%), stiff neck (38% vs 61%), diffuse sensation (32% vs 54%), or migraine (22% vs 38%); as well as

decreased reports of blurred vision (18% vs 40%) and light sensitivity (19% vs 39%). The categorical similarity of these symptoms suggests that the consensus clustering has identified a true subgroup associated with migraine and headache symptoms, perhaps indicating a shared causal origin for these varied types of headaches that is connected with vision disturbances, and sometimes occurs alongside temperature dysregulation, sleep disturbance, and cognitive impairments. Notably, patients from some clusters which appear semantically similar (C2 and C11 are both characterized by speech and cognitive impairment) remain largely separate in the aggregated clustering, likely reflecting increased prevalence of sensory and vision-related symptoms in C2. Similarly, C8 and C10 are both characterized by high rates of insomnia, but also remain separate, with patients in C10 reporting more sensory symptoms, tremors, neuralgia and nausea. This suggests a possible pathological difference between generalized insomnia and insomnia accompanied by sensory abnormalities, in which such difference may be indicative of PEM phenotypes (Stussman et al., 2020).

To evaluate robustness to held-out patients, we repeated this procedure on ten random subsets of 80% of patients. The average pairwise AMI between these runs was 0.61, comparable to the difference between runs on the full dataset with different random seeds (Supp Fig C1b). We also explored model changes with respect to held out symptoms, focusing on how the optimal number of clusters changes when fewer symptoms are reported. The optimal number of clusters decreased as symptom numbers were reduced. While the full 13 clusters required the full dataset, 70% of symptoms (approximately 100) were required for complexity of 12 clusters, and 10 clusters remained optimal with as few as 30% of the reported symptoms (approximately 44) (Supp Fig C1c). This demonstrates that, unsurprisingly, in studies with significantly fewer symptoms assessed, clusters may be missed.

Comparative analysis

The three methodologies produced very different patient subgroups, indicating poor reproducibility of these groups overall, despite some recurring features discussed below. The AMI between the pairs of clusterings ranged between 0.13 (A,B) and 0.40 (B,C). This reflects a lack of clear structure in the dataset due to its approximately homogeneous density across the symptom space, such that the phenotypes detected by one methodology are not reproducible by another, in general. Exemplifying this, the highest symptom burden clusters (A1, B2, and C1) differ substantially in membership. While most patients in clusters B2 and C1 are also in cluster A1, the majority of patients in A1 belong to different clusters in Clusterings B and C. This suggests that there is no clear separation between the high-symptom-burden patients and the rest of the dataset, despite clusters B2 and C1 having similar enriched symptoms. This poor separability appears to be a general feature of the dataset, with manifold learning struggling to produce structured 2D embeddings (Supp Section SG.1). The feature engineering of method A was able to produce a structured 2D embedding with a high degree of separation between clusters (Fig. 2), but with cluster assignments that were less similar, in terms of AMI to clusterings B (0.13) and C (0.18). Similarly, probabilistic outputs of Method C could be used to produce a highly separated visualization (Fig. 3b), but when clusterings A and B are visualized on the same axes, clusters are not well separated (Supp Fig C2).

Despite the lack of reproducibility across methods, we do see some recurrent features, and some clusters with high shared membership. For example, cluster B2 is composed primarily of patients in cluster C7 and cluster A1. While members of A1 are spread throughout the other two clustering methodologies, both B2 and C7 demonstrate significantly impaired speech (Fig. 1, Supp Tables B3 and C1). The reproducibility of this cluster suggests that speech-impaired patients may be more likely to be a clinically relevant subgroup, as has been observed in prior studies of long COVID patients (Cummings, 2024, 2023). We also see that B4 and C4 are composed predominantly of each other. These two groupings are the lowest symptom burden groupings for both of those methodologies. Interestingly, the patients who did not have a strong cluster fit in methodology A (Cluster AX) predominantly sorted into both of these groups. This is likely due to the pre-autoencoder dimensionality reduction, in which symptom features were dropped using skewness and correlation measures. A common feature was that these low-symptom count clusters have the highest average age and the lowest proportion of women. Conversely, the highest symptom count clusters have a high proportion of women and low average age. Highest symptom count clusters also had the highest physical PEM severity and highest cognitive PEM severity in all three methods.

Cluster C9, which includes patients reporting significant PEM but not sleep disturbance, is of potential clinical interest in the context of PEM phenotypes. A majority (57%) of patients in this group also appear in cluster A4, which was characterized by having the lowest symptom burden of non-noise clusters by methodology A. PEM is a hallmark symptom of ME/CFS, which includes fatigue as well as symptoms related to pain, immune dysfunction and sleep disturbance (Chu et al., 2018; Holtzman et al., 2019), including sleep onset latency, frequent awakenings, and apnea (Chang et al., 2021; Jackson and Bruck, 2012; Le Bon et al., 2000). Additionally, unrefreshing sleep is part of the diagnostic criteria for the illness. Unrefreshing sleep and sleep issues that usually require a medical diagnosis (e.g. restless leg syndrome and sleep apnea) were not among the symptoms surveyed in our dataset. Thus, we cannot infer with certainty that C9 patients experience PEM without any sleep disturbance, although this finding suggests that there is phenotypic variation in PEM-related symptoms. However, this reproducible cluster suggests the existence of a patient population with exertion intolerance that may not meet ME/CFS diagnostic criteria; this merits further investigation..

Across all methodologies, cognitive and physical PEM severity tend to move in tandem and both are correlated with symptom count in all three clusterings. Physical PEM is more severe than cognitive PEM for 89% of patients; while the ratio of mean cognitive PEM severity (0.78) to physical PEM severity (0.64) is 0.82 for the cohort. All three methodologies identified clusters where the ratio of cognitive to physical PEM severity is markedly reduced, namely A2 (0.65), A4 (0.65), B1 (0.65), B4 (0.70), C3 (0.59) and C4 (0.59). These are all low symptom count clusters, which is consistent with the observation that patients with fewer symptoms reported less cognitive impairment. The clusters with the highest fraction of patients reporting more severe cognitive than physical PEM were B0 (20%) and B3 (18%), which are both intermediate symptom count clusters with enrichment of symptoms associated with cognitive impairment, suggesting a phenotype with more cognitive challenges.

The splitting of cluster C0 into two clinically distinct subgroups under latent class consensus clustering is also reflected in clustering B, where clusters B7 and B5, which are the main constituents of C0 (Fig. 4), display similarly decreased likelihoods of reporting headaches and migraines. This lends weight to the suggestion that there is a distinct subgroup that is associated with migraine and headache symptoms, perhaps indicative of a common pathogenesis among patients in these subgroups. While both subgroups share temperature dysregulation, the headaches appear to be connected with vision disturbances, sleep disturbance, and cognitive impairments, which are enriched in B5.

Discussion

We identified three clinically plausible symptom-based patient clusterings that could be used to describe phenotypic subgroupings of patients. However, the specific groupings observed depended heavily on the analytic approach used, with low similarity scores between the structures produced using each method. Despite this lack of reproducibility, each clustering represents a technically valid partition of the high-dimensional symptom space and can enhance our understanding of how symptom presentation in Long COVID varies across regions of that space. Some features were common across methodologies. For example, across all three methodologies, clusters with the fewest symptoms also had the fewest women. This aligns with prior studies showing that female patients with Long COVID had higher symptom frequencies than male patients, and were more likely to have complex multisystemic symptoms (Silva et al., 2024; Sylvester et al., 2022). This adds a new angle to how Long COVID disproportionately affects women; it has also been established that premenopausal women and trans people have an elevated risk of developing Long COVID, thought to be related to the role of sex hormones in immune dysfunction and immune response to infection (Pollack et al., 2023; Silva et al., 2024). Individual clustering methodologies were also able to point to interpretable patient groups; clustering B contained adjacent cluster pairs (B5/B7, B1/B3) that were differentiated primarily by the relative prevalence of cognitive symptoms. This difference related to coherent regions of enriched cognitive, memory, and speech related symptoms and was observable at a high-level as differences in physical and cognitive PEM severity between clusters. Similarly, clusters with significant overlap such as A4 and C9 both are characterized by PEM without sleep disturbance. This combination is unusual, as PEM is generally considered specific to ME/CFS, which often occurs with sleep disturbances; that it was observed in two different clusterings suggests that the grouping may be robust.

Our results emphasize the difficulty of clustering with high-dimensional symptom data, and the strong dependence of the detected structure on the choice of algorithm. We hypothesized that if a robust clustering structure was present in the symptom data, it should be discoverable by multiple methods. We instead found that methodology choice profoundly altered the location of cluster boundaries, such that it was not possible to assign patients to a robust phenotype based on their symptom data. This appeared to reflect a lack of inherent structure in the dataset, which exhibits an approximately homogeneous density in the low-dimensional embeddings, even when the UMAP embedding parameters were specifically selected to emphasize clusters (Fig. 1 and Supplemental Fig G3-4). While method A was able to produce clear structural breaks via an autoencoder with significant feature preprocessing (Fig. 2),

these boundaries were not detected by the other two methods. Even in clusters which seem similar when assessing aggregated symptoms, such as the recurrent high symptom burden cluster with enriched musculoskeletal symptoms (B2, C1), the membership of that grouping varies considerably with methodological choice. As noted above, more members of B2 come from cluster C7 than from C1, which also has a high symptom burden but with somewhat distinctive features from C1 (more speech impairment rather than sensory deficits).

Source of Noise

Our lack of detection of robust symptom phenotypes of Long COVID could be an inherent feature of Long COVID, but it could also be due to a methodological issue, noise in the dataset, or other data quality issues.There are certainly several sources of noise in this data, and that noise may be masking some underlying structure. Our analysis collapses time information and considers only whether patients have experienced a symptom at any point post-COVID, which may be a contributing source of noise, as symptoms vary over time and change with illness duration (H. E. Davis et al., 2021b; Fjelltveit et al., 2023) and application of self-management techniques such as pacing (Parker et al., 2023). The survey method may also be a source of noise owing to the inherent variability in symptom self-reporting, as well as the fact that patients were asked to recall symptoms in the past rather than tracking them as they occurred. Validated survey tools and patient-reported outcomes developed for Long COVID and other infection-associated chronic illnesses, such as the DePaul Questionnaire for PEM (Cotler et al., 2018) and the symptom burden questionnaire for Long COVID (Hughes et al., 2022) could reduce noise while ensuring standardization and methodological replicability across studies in the future. It may also be beneficial to collect data on medications, self-management and lifestyle factors, which could be analyzed as confounding variables. Methodologically, the clustering algorithms used here are each designed to handle noise. HDBSCAN/Method A by design aims to ‘not be wrong’ and achieves this principle by including a noise cluster consisting of all samples that could not be assigned to one of the main clusters with high certainty. As an ensemble approach, method B handles noise by aggregating the results over a large library of base clusterings thereby improving the stability of the output. Similarly, Method C includes modeling noise as part of its data generation model. Therefore, although there is clearly noise present in this dataset, the three methods should be able to cope with this and pick up any structure that is present.

Methodological Considerations

By comparing and combining three separate methods, we aimed to improve cluster robustness. However, each of the clustering methods used here come with their own set of drawbacks. The effectiveness of methodology A hinges greatly on data quality, as the dropping of features based on correlation may overlook important aspects in noisy datasets. Features that might appear redundant or less correlated in a noisy dataset might still hold significant predictive power once the noise is removed (Wagner et al., 1993). Through testing we found that the embeddings generated by the autoencoder were quite dependent on the initialized weights, potentially explaining this methodology producing the lowest

AMI across runs. Future work applying this methodology could consider more refined feature selection techniques as well as methods to increase the stability of embeddings across autoencoder initialisation states. Method B uses k-means to generate the base clusterings for the ensemble. Although the k-means algorithm in isolation is a weak clustering method, ensembles of k-means clusterings have been shown to be able to detect non-linearly separable and high-dimensional clusters (Wu et al., 2022). Nevertheless, it may be possible to improve on clustering B by using multiple algorithms to generate the base clusterings and experimenting with different consensus functions (Golalipour et al., 2021), and perhaps incorporating feature engineering similar to method A into the pipeline. Method C will perform best in cases where a mixture model is a good representation of the data (McLachlan et al., 2019); if, for example, patients are better modeled as arising from overlapping phenotypes, where each patient could have more than one label, the single-cluster assumption may not hold. The mixture model used assumes a finite number of classes, that all observed patients fall into one of those classes, and that the classes themselves are homogenous (Sinha et al., 2021). If any of these assumptions do not hold, Method C may provide an inaccurate clustering.

Limitations and Data Quality

A key difference between our study and previous attempts to cluster Long COVID patients is that those prior studies used significantly fewer symptoms (12–40) (Fernández-de-las-Peñas et al., 2023; Kenny et al., 2022; Ziauddeen et al., 2022). We found that the observed clusters depended heavily on the number of symptoms surveyed, as when we reduced the total number of symptoms used in the analyses, optimal cluster counts decreased significantly under latent class analysis (Supp Fig C1c) and the other two methods were sensitive to the removal of features also. This suggests that more studies using fewer symptoms do not capture the full heterogeneous symptomatology of Long COVID. Additionally, these prior studies of Long COVID phenotypes did not perform cross-methodological comparison of machine learning algorithms for clustering, so phenotype consistency cannot be directly assessed or compared. It is possible that clustering using electronic health record (EHR) data could provide a similarly high-dimensional insight into patient experiences. However, EHR studies in this domain are extremely noisy due to issues such as practitioner bias (Yan et al., 2023), insurance billing practices (Verheij et al., 2018), and low rates of diagnosis due to delay or misdiagnosis of known Long COVID-associated conditions (Au et al., 2022; Committee on the Diagnostic Criteria for Myalgic Encephalomyelitis/ChronicFatigue Syndrome et al., 2015; Shaw et al., 2019; Solomon and Reeves, 2004). Moreover, not all relevant symptoms have ICD codes: for example, there is not an ICD code for PEM, which was reported by 88% of participants to our survey. Most EHR queries on Long COVID also require a medical record of a positive test of COVID-19, which undermines issues around testing accessibility, false negatives (Davis et al., 2023), and reduced testing over the course of the pandemic. The survey data reported here provides a uniquely high-dimensional view of symptoms, with the advantage of using patient-reports where all symptoms surveyed were asked of all patients.

Additional limitations of this study is that it does not include children and adolescents, does not account for asymptomatic organ damage, and participants had experienced a median of 198 days of symptoms.

We also did not collect symptom severity; symptom severity measured using the Likert scale and information about symptom timing could both be beneficial in future clustering studies. Additionally, some of our observed symptom frequencies indicate that our cohort may be non-representative, or a different representation of Long COVID than observed in prior studies. Most notably, published studies of Long COVID have identified PEM in approximately half of participants, whereas 88.4% of our participants reported experiencing PEM. Prior work has found that most Long COVID patients who do not meet ME/CFS diagnostic criteria nonetheless exhibit exertion intolerance with shorter-duration PEM < 14 hours (Kedor et al., 2022); some of the difference between our survey and other results may be patients with this form of exertion intolerance. Slight differences in wording of questions have been shown to significantly influence rates of PEM reported (Jason et al., 2015). It is also possible that patients with Long COVID that maps onto known diagnoses, such as those who develop post-COVID diabetes (Harding et al., 2023), are less likely to participate in Long COVID research studies, as their illness is better understood biomedically.

Implications

Further innovations for clustering noisy data could improve these results; however, some of this noise is likely a fundamental feature of Long COVID, as it is associated with several co-occuring conditions with overlapping symptoms (Davis et al., 2023). For example, tachycardia is associated with both dysautonomia and mast cell activation syndrome, two clinically associated disorders (Kohno et al., 2021) often diagnosed in Long COVID patients. We argue that this extensive symptom overlap between individuals with differing disease mechanisms is an inherent complexity in Long COVID that hampers attempts at symptom-based clustering. Importantly, our results do not preclude the implication of distinct and characterizable endotypes in Long COVID, but do suggest that these endotypes may not map onto resolvable phenotypes at the symptom level. Moving forward, Long COVID clusters will have improved clinical utility when endotype (Turner et al., 2023) and phenotype data can be analyzed in tandem, which will be enabled by deployment of biomarkers for the several pathological mechanisms described to date (Klein et al., 2023).

Taken together, our findings suggest that clustering studies of Long COVID patients need to be skeptical of cluster outputs, and particularly cluster reproducibility. Machine learning algorithms excel at finding statistically meaningful data subdivisions, but in the context of a homogeneously varying dataset, these partitions are fundamentally arbitrary. There were some reproducible features across methods, though the adjusted mutual information between the methods was low. Some of these features, such as lower symptom burden being associated with higher proportion of men and higher age, should be further researched to be of potential use in clinical practice. Other reproducible findings, such as those regarding differences between cognitive and physical PEM, indicate a need for future research to not treat PEM as a monolith, but rather probe further to explore phenotypes of PEM, such as differences in triggers, symptoms, and severity. While studies with fewer reported symptoms may achieve more consistent results, these outcomes may be artifacts of the limited number of symptoms under

consideration. More research is needed to determine not only what symptoms, but what conditions and biological markers can help consistently reproduce Long COVID phenotypes.

Declarations

CRediT author statement:

Conceptualization: Hannah Davis

Methodology: Tessa Green, Chris McWilliams, Leonardo de Figueiredo, Hannah Davis, Tan Zhi-Xuan

Formal analysis: Tessa Green, Chris McWilliams, Leonardo de Figueiredo

Data curation: Tessa Green, Chris McWilliams, Leonardo de Figueiredo, Hannah Davis

Writing- original draft: Tessa Green, Chris McWilliams, Leonardo de Figueiredo, Hannah Davis Beth Pollack, Alison Cohen, Letícia Soares

Writing- review and editing: Tessa Green, Chris McWilliams, Leonardo de Figueiredo, Hannah Davis, Beth Pollack, Alison Cohen, Letícia Soares, Tess Falor, Tan Zhi-Xuan

Data visualization: Tessa Green, Chris McWilliams, Leonardo de Figueiredo

Funding acquisition: Hannah Davis

Acknowledgments: This research was supported by the Kanro Foundation.

References

Ancona, G., Alagna, L., Alteri, C., Palomba, E., Tonizzo, A., Pastena, A., Muscatello, A., Gori, A., Bandera, A., 2023. Gut and airway microbiota dysbiosis and their role in COVID-19 and long-COVID. Front. Immunol. 14. https://doi.org/10.3389/fimmu.2023.1080043

Au, L., Capotescu, C., Eyal, G., Finestone, G., 2022. Long covid and medical gaslighting: Dismissal, delayed diagnosis, and deferred treatment. SSM - Qual. Res. Health 2, 100167. https://doi.org/10.1016/j.ssmqr.2022.100167

Boongoen, T., Iam-On, N., 2018. Cluster ensembles: A survey of approaches with recent extensions and applications. Comput. Sci. Rev. 28, 1–25. https://doi.org/10.1016/j.cosrev.2018.01.003

Campello, R.J.G.B., Moulavi, D., Sander, J., 2013. Density-Based Clustering Based on Hierarchical Density Estimates, in: Pei, J., Tseng, V.S., Cao, L., Motoda, H., Xu, G. (Eds.), Advances in Knowledge Discovery and Data Mining. Springer, Berlin, Heidelberg, pp. 160–172. https://doi.org/10.1007/978-3-642-37456-2_14

Chang, C.-J., Hung, L.-Y., Kogelnik, A.M., Kaufman, D., Aiyar, R.S., Chu, A.M., Wilhelmy, J., Li, P., Tannenbaum, L., Xiao, W., Davis, R.W., 2021. A Comprehensive Examination of Severely Ill ME/CFS Patients. Healthcare 9, 1290. https://doi.org/10.3390/healthcare9101290

Chu, L., Valencia, I.J., Garvert, D.W., Montoya, J.G., 2018. Deconstructing post-exertional malaise in myalgic encephalomyelitis/ chronic fatigue syndrome: A patient-centered, cross-sectional survey. PLOS ONE 13, e0197811. https://doi.org/10.1371/journal.pone.0197811

Committee on the Diagnostic Criteria for Myalgic Encephalomyelitis/ChronicFatigue Syndrome, Board on the Health of Select Populations, Institute of Medicine, 2015. Beyond Myalgic Encephalomyelitis/Chronic Fatigue Syndrome: Redefining an Illness, The National Academies Collection: Reports funded by National Institutes of Health. National Academies Press (US), Washington (DC).

Cotler, J., Holtzman, C., Dudun, C., Jason, L.A., 2018. A Brief Questionnaire to Assess Post-Exertional Malaise. Diagnostics 8, 66. https://doi.org/10.3390/diagnostics8030066

Cummings, L., 2024. Cognitive-linguistic difficulties in adults with Long COVID: A follow-up study. Lang. Health 2, 1–21. https://doi.org/10.1016/j.laheal.2023.09.001

Cummings, L., 2023. Long COVID: The impact on language and cognition. Lang. Health. https://doi.org/10.1016/j.laheal.2023.05.001

Dagliati, A., Strasser, Z.H., Abad, Z.S.H., Klann, et al, 2023. Characterization of long COVID temporal sub-phenotypes by distributed representation learning from electronic health record data: a cohort study. eClinicalMedicine 64. https://doi.org/10.1016/j.eclinm.2023.102210

Davis, H.E., Assaf, G., McCorkell, L., Wei, H., Re'em, Y., Akrami, A., 2021a. Questionnaire to Characterize Long COVID: 200+ symptoms over 7 months. https://doi.org/10.6084/m9.figshare.13642553.v2

Davis, H.E., Assaf, G.S., McCorkell, L., Wei, H., Low, R.J., Re'em, Y., Redfield, S., Austin, J.P., Akrami, A., 2021b. Characterizing long COVID in an international cohort: 7 months of symptoms and their impact. eClinicalMedicine 38. https://doi.org/10.1016/j.eclinm.2021.101019

Davis, H.E., McCorkell, L., Vogel, J.M., Topol, E.J., 2023. Long COVID: major findings, mechanisms and recommendations. Nat. Rev. Microbiol. 21, 133–146. https://doi.org/10.1038/s41579-022-00846-2

Davis, J.T., Chinazzi, M., Perra, N., Mu, K., Pastore y Piontti, A., Ajelli, M., Dean, N.E., Gioannini, C., Litvinova, M., Merler, S., Rossi, L., Sun, K., Xiong, X., Longini, I.M., Halloran, M.E., Viboud, C., Vespignani, A., 2021. Cryptic transmission of SARS-CoV-2 and the first COVID-19 wave. Nature 600, 127–132. https://doi.org/10.1038/s41586-021-04130-w

Deer, R.R., Rock, M.A., Vasilevsky, N. et al, 2021. Characterizing Long COVID: Deep Phenotype of a Complex Condition. eBioMedicine 74. https://doi.org/10.1016/j.ebiom.2021.103722

Dennis, A., Wamil, M., Alberts, J., Oben, J., Cuthbertson, D.J., Wootton, D., Crooks, M., Gabbay, M., Brady, M., Hishmeh, L., Attree, E., Heightman, M., Banerjee, R., Banerjee, A., 2021. Multiorgan impairment in low-risk individuals with post-COVID-19 syndrome: a prospective, community-based study. BMJ Open 11, e048391. https://doi.org/10.1136/bmjopen-2020-048391

Domeniconi, C., Gunopulos, D., Ma, S., Yan, B., Al-Razgan, M., Papadopoulos, D., 2007. Locally adaptive metrics for clustering high dimensional data. Data Min. Knowl. Discov. 14, 63–97. https://doi.org/10.1007/s10618-006-0060-8

Estimated COVID-19 Burden | CDC [WWW Document], 2023. URL https://web.archive.org/web/20230920190907/https://www.cdc.gov/coronavirus/2019-ncov/cases-updates/burden.html (accessed 8.5.24).

Fern, X.Z., Lin, W., 2008. Cluster Ensemble Selection. Stat. Anal. Data Min. ASA Data Sci. J. 1, 128–141. https://doi.org/10.1002/sam.10008

Fernández-de-las-Peñas, C., Martín-Guerrero, J.D., Florencio, L.L., Navarro-Pardo, E., Rodríguez-Jiménez, J., Torres-Macho, J., Pellicer-Valero, O.J., 2023. 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 51, 61–69. https://doi.org/10.1007/s15010-022-01822-x

Fjelltveit, E.B., Blomberg, B., Kuwelker, K., Zhou, F., Onyango, T.B., Brokstad, K.A., Elyanow, R., Kaplan, I.M., Tøndel, C., Mohn, K.G.I., Özgümüş, T., Cox, R.J., Langeland, N., Bergen COVID-19 Research Group, 2023. Symptom Burden and Immune Dynamics 6 to 18 Months Following Mild Severe Acute Respiratory Syndrome Coronavirus 2 Infection (SARS-CoV-2): A Case-control Study. Clin. Infect. Dis. 76, e60–e70. https://doi.org/10.1093/cid/ciac655

Fogarty, H., Townsend, L., Morrin, H., Ahmad, A., Comerford, C., Karampini, E., Englert, H., Byrne, M., Bergin, C., O'Sullivan, J.M., Martin-Loeches, I., Nadarajan, P., Bannan, C., Mallon, P.W., Curley, G.F., Preston, R.J.S., Rehill, A.M., McGonagle, D., Cheallaigh, C.N., Baker, R.I., Renné, T., Ward, S.E., O'Donnell, J.S., O'Connell, N., Ryan, K., Kenny, D., Fazavana, J., 2021. Persistent endotheliopathy in the pathogenesis of long COVID syndrome. J. Thromb. Haemost. 19, 2546–2553. https://doi.org/10.1111/jth.15490

Glorot, X., Bengio, Y., 2010. Understanding the difficulty of training deep feedforward neural networks, in: Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics. Presented at the Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics, JMLR Workshop and Conference Proceedings, pp. 249–256.

Golalipour, K., Akbari, E., Hamidi, S.S., Lee, M., Enayatifar, R., 2021. From clustering to clustering ensemble selection: A review. Eng. Appl. Artif. Intell. 104, 104388. https://doi.org/10.1016/j.engappai.2021.104388

Harding, J.L., Oviedo, S.A., Ali, M.K., Ofotokun, I., Gander, J.C., Patel, S.A., Magliano, D.J., Patzer, R.E., 2023. The bidirectional association between diabetes and long-COVID-19 – A systematic review. Diabetes Res. Clin. Pract. 195. https://doi.org/10.1016/j.diabres.2022.110202

Hartle, M., Bateman, L., Vernon, S.D., 2021. Dissecting the nature of post-exertional malaise. Fatigue Biomed. Health Behav. 9, 33–44. https://doi.org/10.1080/21641846.2021.1905415

Holtzman, C.S., Bhatia, S., Cotler, J., Jason, L.A., 2019. Assessment of Post-Exertional Malaise (PEM) in Patients with Myalgic Encephalomyelitis (ME) and Chronic Fatigue Syndrome (CFS): A Patient-Driven Survey. Diagnostics 9, 26. https://doi.org/10.3390/diagnostics9010026

Hughes, S.E., Haroon, S., Subramanian, A., McMullan, C., Aiyegbusi, O.L., Turner, G.M., Jackson, L., Davies, E.H., Frost, C., McNamara, G., Price, G., Matthews, K., Camaradou, J., Ormerod, J., Walker, A., Calvert, M.J., 2022. Development and validation of the symptom burden questionnaire for long covid (SBQ-LC): Rasch analysis. BMJ 377, e070230. https://doi.org/10.1136/bmj-2022-070230

Jackson, M.L., Bruck, D., 2012. Sleep Abnormalities in Chronic Fatigue Syndrome/Myalgic Encephalomyelitis: A Review. J. Clin. Sleep Med. JCSM Off. Publ. Am. Acad. Sleep Med. 8, 719–728. https://doi.org/10.5664/jcsm.2276

Jason, L.A., Evans, M., So, S., Scott, J., Brown, A., 2015. Problems in Defining Post-Exertional Malaise. J. Prev. Interv. Community 43, 20–31. https://doi.org/10.1080/10852352.2014.973239

Kedor, C., Freitag, H., Meyer-Arndt, L., Wittke, K., Hanitsch, L.G., Zoller, T., Steinbeis, F., Haffke, M., Rudolf, G., Heidecker, B., Bobbert, T., Spranger, J., Volk, H.-D., Skurk, C., Konietschke, F., Paul, F., Behrends, U., Bellmann-Strobl, J., Scheibenbogen, C., 2022. A prospective observational study of post-COVID-19 chronic fatigue syndrome following the first pandemic wave in Germany and biomarkers associated with symptom severity. Nat. Commun. 13, 5104. https://doi.org/10.1038/s41467-022-32507-6

Kenny, G., McCann, K., O'Brien, C., Savinelli, S., Tinago, W., Yousif, O., Lambert, J.S., O'Broin, C., Feeney, E.R., De Barra, E., Doran, P., Mallon, P.W.G., All-Ireland Infectious Diseases (AIID) Cohort Study Group, 2022. Identification of Distinct Long COVID Clinical Phenotypes Through Cluster Analysis of Self-Reported Symptoms. Open Forum Infect. Dis. 9, ofac060. https://doi.org/10.1093/ofid/ofac060

Klein, J., Wood, J., Jaycox, J.R., Dhodapkar, R.M., Lu, P., Gehlhausen, J.R., Tabachnikova, A., Greene, K., Tabacof, L., Malik, A.A., Silva Monteiro, V., Silva, J., Kamath, K., Zhang, M., Dhal, A., Ott, I.M., Valle, G., Peña-Hernández, M., Mao, T., Bhattacharjee, B., Takahashi, T., Lucas, C., Song, E., McCarthy, D., Breyman, E., Tosto-Mancuso, J., Dai, Y., Perotti, E., Akduman, K., Tzeng, T.J., Xu, L., Geraghty, A.C., Monje, M., Yildirim, I., Shon, J., Medzhitov, R., Lutchmansingh, D., Possick, J.D., Kaminski, N., Omer, S.B., Krumholz, H.M., Guan, L., Dela Cruz, C.S., van Dijk, D., Ring, A.M., Putrino, D., Iwasaki, A., 2023. Distinguishing features of long COVID identified through immune profiling. Nature 623, 139–148. https://doi.org/10.1038/s41586-023-06651-y

Kohno, R., Cannom, D.S., Olshansky, B., Xi, S.C., Krishnappa, D., Adkisson, W.O., Norby, F.L., Fedorowski, A., Benditt, D.G., 2021. Mast Cell Activation Disorder and Postural Orthostatic Tachycardia Syndrome: A Clinical Association. J. Am. Heart Assoc. 10, e021002. https://doi.org/10.1161/JAHA.121.021002

Kucirka, L.M., Lauer, S.A., Laeyendecker, O., Boon, D., Lessler, J., 2020. Variation in False-Negative Rate of Reverse Transcriptase Polymerase Chain Reaction–Based SARS-CoV-2 Tests by Time Since Exposure. Ann. Intern. Med. 173, 262–267. https://doi.org/10.7326/M20-1495

Le Bon, O., Fischler, B., Hoffmann, G., Murphy, J.R., De Meirleir, K., Cluydts, R., Pelc, I., 2000. How significant are primary sleep disorders and sleepiness in the chronic fatigue syndrome? Sleep Res. Online SRO 3, 43–48.

Lim, S.H., Ju, H.J., Han, J.H., Lee, J.H., Lee, W.-S., Bae, J.M., Lee, S., 2023. Autoimmune and Autoinflammatory Connective Tissue Disorders Following COVID-19. JAMA Netw. Open 6, e2336120. https://doi.org/10.1001/jamanetworkopen.2023.36120

Lloyd, S., 1982. Least squares quantization in PCM. IEEE Trans. Inf. Theory 28, 129–137. https://doi.org/10.1109/TIT.1982.1056489

Lorman, V., Song, X., Rao, S., Allen, A.J., Utidjian, L., Charles Bailey, L., 2023. 1362. Pediatric long COVID subphenotypes: an EHR-based study from the RECOVER program. Open Forum Infect. Dis. 10, ofad500.1199. https://doi.org/10.1093/ofid/ofad500.1199

McInnes, L., Healy, J., Melville, J., 2020. UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction. https://doi.org/10.48550/arXiv.1802.03426

McLachlan, G.J., Lee, S.X., Rathnayake, S.I., 2019. Finite Mixture Models. Annu. Rev. Stat. Its Appl. 6, 355–378. https://doi.org/10.1146/annurev-statistics-031017-100325

Morin, S., Legault, R., Laliberté, F., Bakk, Z., Giguère, C.-É., de la Sablonnière, R., Lacourse, É., 2024. StepMix: A Python Package for Pseudo-Likelihood Estimation of Generalized Mixture Models with External Variables. https://doi.org/10.48550/arXiv.2304.03853

Parker, M., Sawant, H.B., Flannery, T., Tarrant, R., Shardha, J., Bannister, R., Ross, D., Halpin, S., Greenwood, D.C., Sivan, M., 2023. Effect of using a structured pacing protocol on post-exertional symptom exacerbation and health status in a longitudinal cohort with the post-COVID-19 syndrome. J. Med. Virol. 95, e28373. https://doi.org/10.1002/jmv.28373

Pecoraro, V., Negro, A., Pirotti, T., Trenti, T., 2022. Estimate false-negative RT-PCR rates for SARS-CoV-2. A systematic review and meta-analysis. Eur. J. Clin. Invest. 52, e13706. https://doi.org/10.1111/eci.13706

Peluso, M.J., Deitchman, A.N., Torres, L., Iyer, N.S., Munter, S.E., Nixon, C.C., Donatelli, J., Thanh, C., Takahashi, S., Hakim, J., Turcios, K., Janson, O., Hoh, R., Tai, V., Hernandez, Y., Fehrman, E.A., Spinelli, M.A., Gandhi, M., Trinh, L., Wrin, T., Petropoulos, C.J., Aweeka, F.T., Rodriguez-Barraquer, I., Kelly, J.D.,

Martin, J.N., Deeks, S.G., Greenhouse, B., Rutishauser, R.L., Henrich, T.J., 2021. Long-term SARS-CoV-2-specific immune and inflammatory responses in individuals recovering from COVID-19 with and without post-acute symptoms. Cell Rep. 36. https://doi.org/10.1016/j.celrep.2021.109518

Pollack, B., von Saltza, E., McCorkell, L., Santos, L., Hultman, A., Cohen, A.K., Soares, L., 2023. Female reproductive health impacts of Long COVID and associated illnesses including ME/CFS, POTS, and connective tissue disorders: a literature review. Front. Rehabil. Sci. 4, 1122673. https://doi.org/10.3389/fresc.2023.1122673

Prasannan, N., Heightman, M., Hillman, T., Wall, E., Bell, R., Kessler, A., Neave, L., Doyle, A., Devaraj, A., Singh, D., Dehbi, H.-M., Scully, M., 2022. Impaired exercise capacity in post–COVID-19 syndrome: the role of VWF-ADAMTS13 axis. Blood Adv. 6, 4041–4048. https://doi.org/10.1182/bloodadvances.2021006944

Pretorius, E., Venter, C., Laubscher, G.J., Kotze, M.J., Oladejo, S.O., Watson, L.R., Rajaratnam, K., Watson, B.W., Kell, D.B., 2022. Prevalence of symptoms, comorbidities, fibrin amyloid microclots and platelet pathology in individuals with Long COVID/Post-Acute Sequelae of COVID-19 (PASC). Cardiovasc. Diabetol. 21, 148. https://doi.org/10.1186/s12933-022-01579-5

Re'em, Y., Stelson, E.A., Davis, H.E., McCorkell, L., Wei, H., Assaf, G., Akrami, A., 2023. Factors associated with psychiatric outcomes and coping in Long COVID. Nat. Ment. Health 1, 361–372. https://doi.org/10.1038/s44220-023-00064-6

Romano, S., Vinh, N.X., Bailey, J., Verspoor, K., 2016. Adjusting for chance clustering comparison measures. J Mach Learn Res 17, 4635–4666.

Rousseeuw, P.J., 1987. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. J. Comput. Appl. Math. 20, 53–65. https://doi.org/10.1016/0377-0427(87)90125-7

Shaw, B.H., Stiles, L.E., Bourne, K., Green, E.A., Shibao, C.A., Okamoto, L.E., Garland, E.M., Gamboa, A., Diedrich, A., Raj, V., Sheldon, R.S., Biaggioni, I., Robertson, D., Raj, S.R., 2019. The face of postural tachycardia syndrome – insights from a large cross-sectional online community-based survey. J. Intern. Med. 286, 438–448. https://doi.org/10.1111/joim.12895

Silva, J., Takahashi, T., Wood, J., Lu, P., Tabachnikova, A., Gehlhausen, J.R., Greene, K., Bhattacharjee, B., Monteiro, V.S., Lucas, C., Dhodapkar, R.M., Tabacof, L., Peña-Hernandez, M., Kamath, K., Mao, T., Mccarthy, D., Medzhitov, R., Dijk, D. van, Krumholz, H.M., Guan, L., Putrino, D., Iwasaki, A., 2024. Sex differences in symptomatology and immune profiles of Long COVID. https://doi.org/10.1101/2024.02.29.24303568

Sinha, P., Calfee, C.S., Delucchi, K.L., 2021. Practitioner's Guide to Latent Class Analysis: Methodological Considerations and Common Pitfalls. Crit. Care Med. 49, e63–e79. https://doi.org/10.1097/CCM.0000000000004710

Solomon, L., Reeves, W.C., 2004. Factors Influencing the Diagnosis of Chronic Fatigue Syndrome. Arch. Intern. Med. 164, 2241–2245. https://doi.org/10.1001/archinte.164.20.2241

Soriano, J.B., Murthy, S., Marshall, J.C., Relan, P., Diaz, J.V., 2022. A clinical case definition of post-COVID-19 condition by a Delphi consensus. Lancet Infect. Dis. 22, e102–e107. https://doi.org/10.1016/S1473-3099(21)00703-9

Soung, A.L., Vanderheiden, A., Nordvig, A.S., Sissoko, C.A., Canoll, P., Mariani, M.B., Jiang, X., Bricker, T., Rosoklija, G.B., Arango, V., Underwood, M., Mann, J.J., Dwork, A.J., Goldman, J.E., Boon, A.C.M., Boldrini, M., Klein, R.S., 2022. COVID-19 induces CNS cytokine expression and loss of hippocampal neurogenesis. Brain 145, 4193–4201. https://doi.org/10.1093/brain/awac270

Stein, S.R., Ramelli, S.C., Grazioli, A., Chung, J.-Y., Singh, M., Yinda, C.K., Winkler, C.W., Sun, J., Dickey, J.M., Ylaya, K., Ko, S.H., Platt, A.P., Burbelo, P.D., Quezado, M., Pittaluga, S., Purcell, M., Munster, V.J., Belinky, F., Ramos-Benitez, M.J., Boritz, E.A., Lach, I.A., Herr, D.L., Rabin, J., Saharia, K.K., Madathil, R.J., Tabatabai, A., Soherwardi, S., McCurdy, M.T., Peterson, K.E., Cohen, J.I., de Wit, E., Vannella, K.M., Hewitt, S.M., Kleiner, D.E., Chertow, D.S., 2022. SARS-CoV-2 infection and persistence in the human body and brain at autopsy. Nature 612, 758–763. https://doi.org/10.1038/s41586-022-05542-y

Stussman, B., Williams, A., Snow, J., Gavin, A., Scott, R., Nath, A., Walitt, B., 2020. Characterization of Post–exertional Malaise in Patients With Myalgic Encephalomyelitis/Chronic Fatigue Syndrome. Front. Neurol. 11. https://doi.org/10.3389/fneur.2020.01025

Swank, Z., Senussi, Y., Manickas-Hill, Z., Yu, X.G., Li, J.Z., Alter, G., Walt, D.R., 2023. Persistent Circulating Severe Acute Respiratory Syndrome Coronavirus 2 Spike Is Associated With Post-acute Coronavirus Disease 2019 Sequelae. Clin. Infect. Dis. 76, e487–e490. https://doi.org/10.1093/cid/ciac722

Sylvester, S.V., Rusu, R., Chan, B., Bellows, M., O'Keefe, C., Nicholson, S., 2022. Sex differences in sequelae from COVID-19 infection and in long COVID syndrome: a review. Curr. Med. Res. Opin. 38, 1391–1399. https://doi.org/10.1080/03007995.2022.2081454

Turner, S., Khan, M.A., Putrino, D., Woodcock, A., Kell, D.B., Pretorius, E., 2023. Long COVID: pathophysiological factors and abnormalities of coagulation. Trends Endocrinol. Metab. 34, 321–344. https://doi.org/10.1016/j.tem.2023.03.002

Verheij, R.A., Curcin, V., Delaney, B.C., McGilchrist, M.M., 2018. Possible Sources of Bias in Primary Care Electronic Health Record Data Use and Reuse. J. Med. Internet Res. 20, e185. https://doi.org/10.2196/jmir.9134

Vinh, N.X., Epps, J., Bailey, J., 2010. Information Theoretic Measures for Clusterings Comparison: Variants, Properties, Normalization and Correction for Chance. J Mach Learn Res 11, 2837–2854.

von Luxburg, U., 2007. A tutorial on spectral clustering. Stat. Comput. 17, 395–416. https://doi.org/10.1007/s11222-007-9033-z

Wu, H.-F., Yu, W., Saito-Diaz, K., Huang, C.-W., Carey, J., Lefcort, F., Hart, G.W., Liu, H.-X., Zeltner, N., 2022. Norepinephrine transporter defects lead to sympathetic hyperactivity in Familial Dysautonomia models. Nat. Commun. 13, 7032. https://doi.org/10.1038/s41467-022-34811-7

Yan, C., Zhang, X., Yang, Y., Kang, K., Were, M.C., Embí, P., Patel, M.B., Malin, B.A., Kho, A.N., Chen, Y., 2023. Differences in Health Professionals' Engagement With Electronic Health Records Based on Inpatient Race and Ethnicity. JAMA Netw. Open 6, e2336383. https://doi.org/10.1001/jamanetworkopen.2023.36383

Yin, K., Peluso, M.J., Luo, X., Thomas, R., Shin, M.-G., Neidleman, J., Andrew, A., Young, K.C., Ma, T., Hoh, R., Anglin, K., Huang, B., Argueta, U., Lopez, M., Valdivieso, D., Asare, K., Deveau, T.-M., Munter, S.E., Ibrahim, R., Ständker, L., Lu, S., Goldberg, S.A., Lee, S.A., Lynch, K.L., Kelly, J.D., Martin, J.N., Münch, J., Deeks, S.G., Henrich, T.J., Roan, N.R., 2024. Long COVID manifests with T cell dysregulation, inflammation and an uncoordinated adaptive immune response to SARS-CoV-2. Nat. Immunol. 25, 218–225. https://doi.org/10.1038/s41590-023-01724-6

Zelnik-Manor, L., Perona, P., 2004. Self-tuning spectral clustering, in: Proceedings of the 17th International Conference on Neural Information Processing Systems, NIPS'04. MIT Press, Cambridge, MA, USA, pp. 1601–1608.

Zhang, D., Zhou, Y., Ma, Y., Chen, P., Tang, J., Yang, B., Li, H., Liang, M., Xue, Y., Liu, Y., Zhang, J., Wang, X., 2023. Gut Microbiota Dysbiosis Correlates With Long COVID-19 at One-Year After Discharge. J. Korean Med. Sci. 38, e120. https://doi.org/10.3346/jkms.2023.38.e120

Zhang, H., Zang, C., Xu, Z., Zhang, Yongkang, Xu, J., Bian, J., Morozyuk, D., Khullar, D., Zhang, Yiye, Nordvig, A.S., Schenck, E.J., Shenkman, E.A., Rothman, R.L., Block, J.P., Lyman, K., Weiner, M.G., Carton, T.W., Wang, F., Kaushal, R., 2023. Data-driven identification of post-acute SARS-CoV-2 infection subphenotypes. Nat. Med. 29, 226–235. https://doi.org/10.1038/s41591-022-02116-3

Ziauddeen, N., Gurdasani, D., O'Hara, M.E., Hastie, C., Roderick, P., Yao, G., Alwan, N.A., 2022. Characteristics and impact of Long Covid: Findings from an online survey. PLOS ONE 17, e0264331. https://doi.org/10.1371/journal.pone.0264331

Zollner, A., Koch, R., Jukic, A., Pfister, A., Meyer, M., Rössler, A., Kimpel, J., Adolph, T.E., Tilg, H., 2022. Postacute COVID-19 is Characterized by Gut Viral Antigen Persistence in Inflammatory Bowel Diseases. Gastroenterology 163, 495-506.e8. https://doi.org/10.1053/j.gastro.2022.04.037

Figures

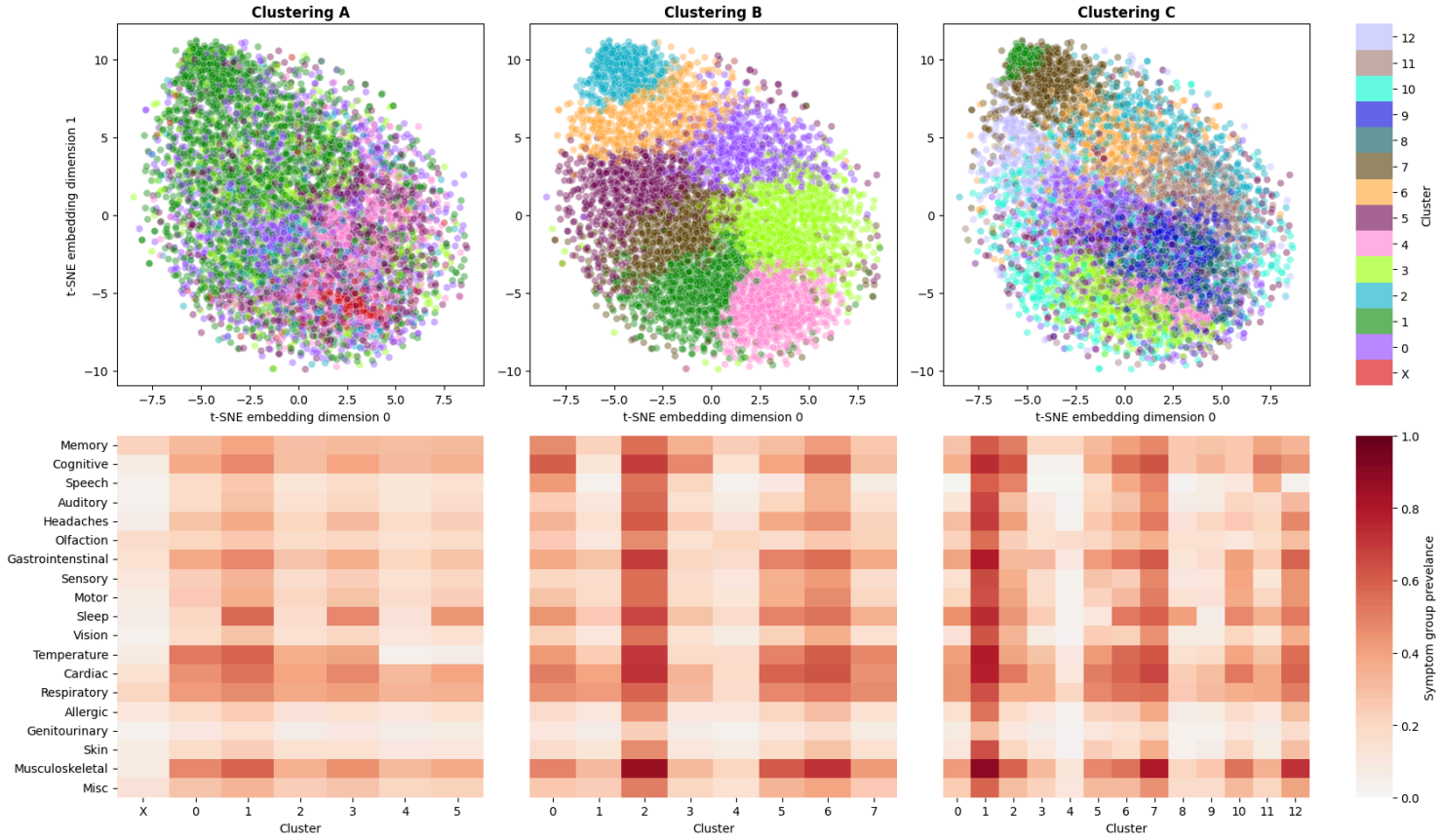

Figure 1

Left: method A (genetically optimized autoencoder) with 6 clusters and one noise cluster; *Middle*: method B (ensemble) with 8 clusters; *Right*: method C (LCA) with 13 clusters. *Top row:* The 3 clusterings shown in the same 2D t-SNE embedding. Each point is a patient and each color represents a cluster. Patients in a given cluster are not shared across clusterings, and sharing a cluster number does not imply that two clusters from different methods are similar. *Bottom row:* Average fraction of symptoms reported in each symptom grouping for each cluster. Cluster AX denotes patients who were labeled as noise (no cluster) by HDBSCAN. Symptom groups are defined in Supp Data 1.

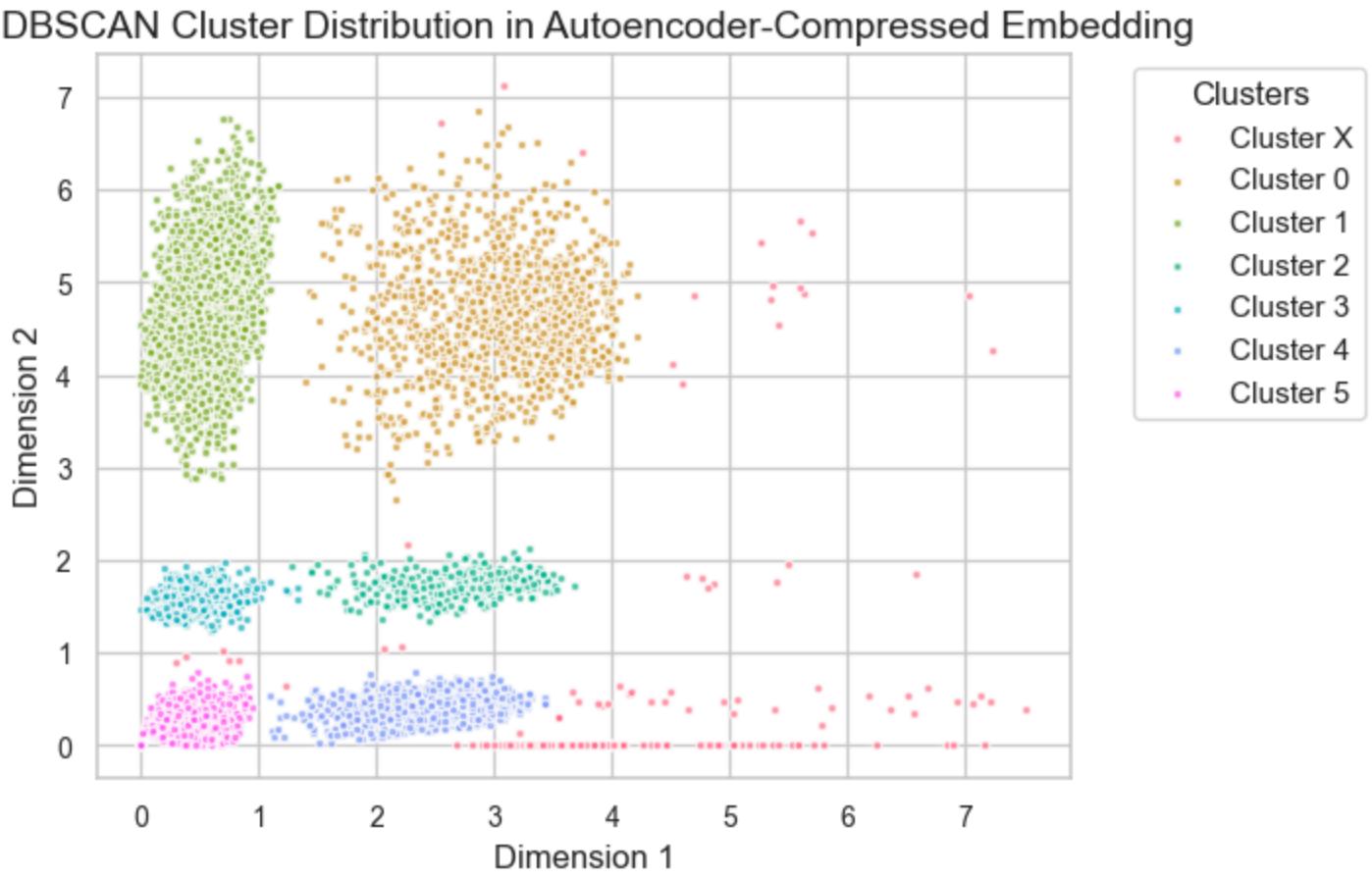

Figure 2

Representation space learned by the method A autoencoder clustered using HDBSCAN. Each point represents a patient, and colors represent the six clusters. The ‘X’ cluster are points labeled as noise by HDBSCAN. The plot dimensions are the compressed representations generated by the autoencoder, capturing essential patterns and relationships within the data in a reduced-dimensional space for clustering. The parameters used for the clustering technique above are detailed in STA2.

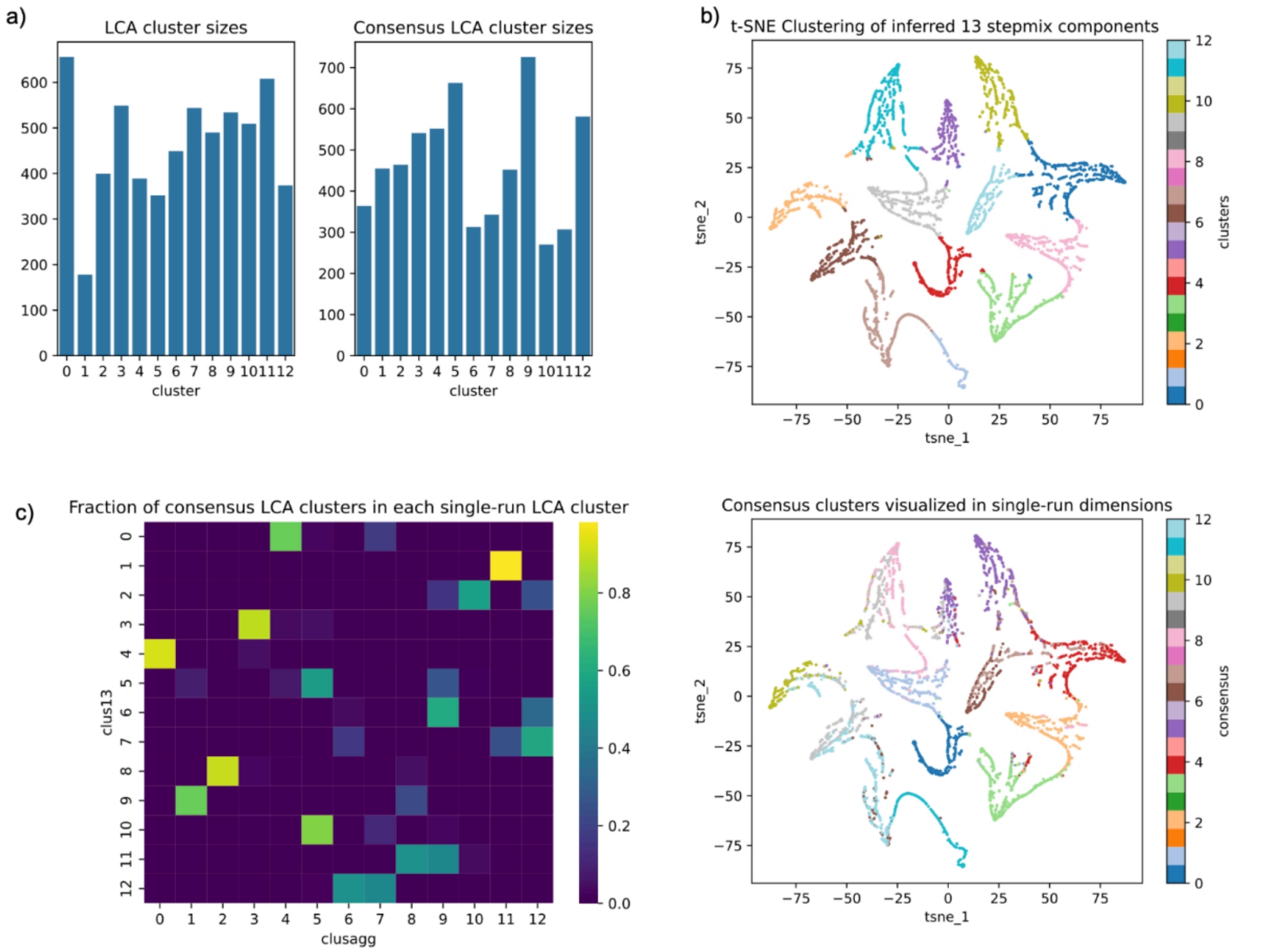

Figure 3

a) Cluster sizes for an individual LCA run presented as clustering C (left) and a consensus of ten LCA clusterings with different random seeds (right). Cluster numbers are arbitrary and do not indicate overlapping patient assignments. b) tSNE of probabilistic cluster assignments for all patients. Patients are colored by their most probable cluster. This visualization shows that clusters are well-defined, with only a few ambiguous assignments. c) Consensus clustering overlap (left) and on the same tSNE plot (right) visually clarify the substantial consensus across LCA clusterings.

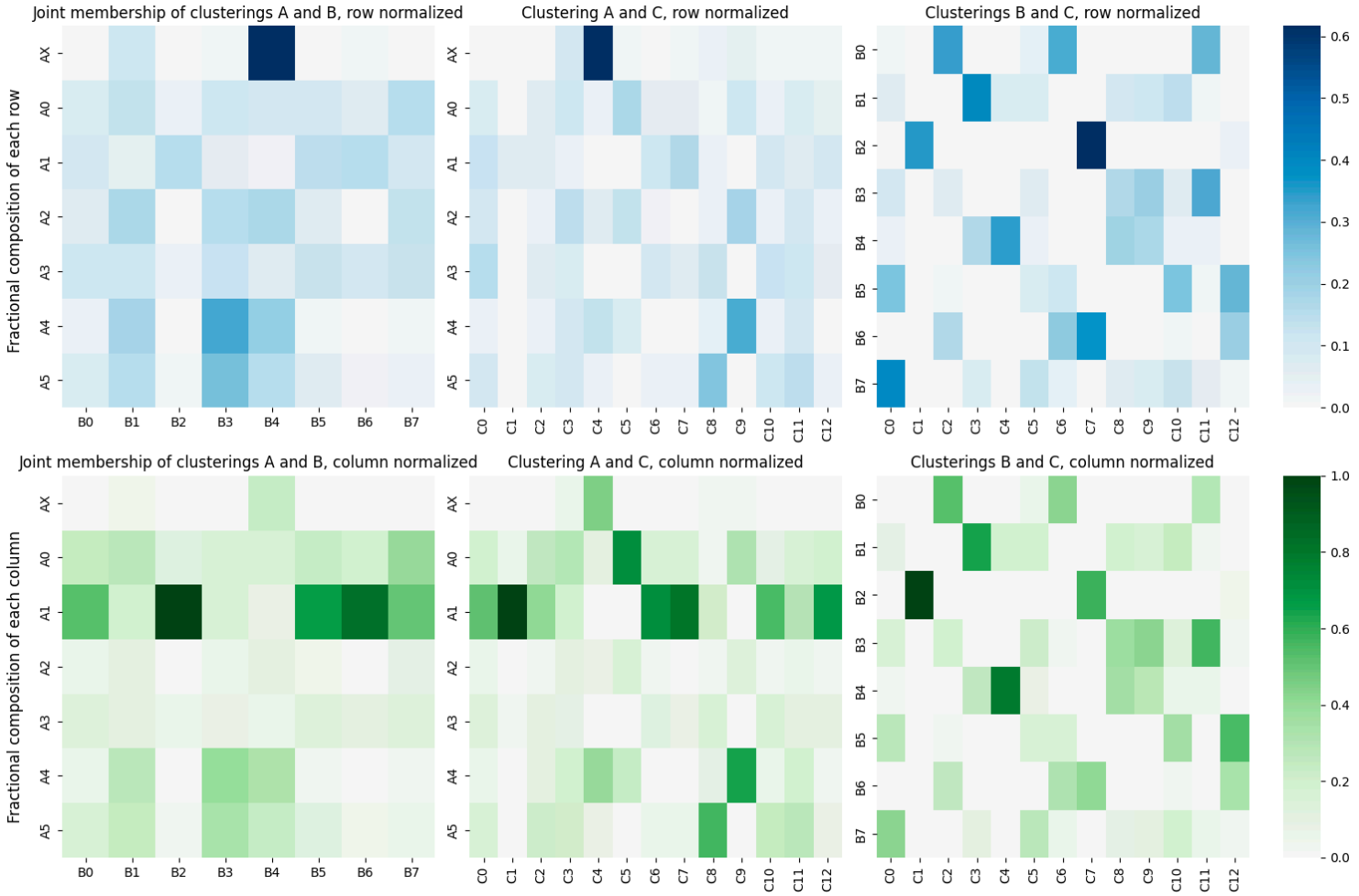

Figure 4

Cluster overlaps between the three methodologies. Top row depicts row-normalized data, in which each square is colored by the fraction of each row which is made up by patients from each column. A dark box indicates that patients from the group corresponding the row are predominantly in the corresponding column. The bottom row of plots is column-normalized, depicting the fraction of each column made up by each row. A dark box indicates that patients from the column cluster are predominantly also assigned to the corresponding row cluster. For example, comparing the joint membership of clusters A and B, cluster B2 is mostly composed of members of cluster A1 (bottom left graph), but cluster A1 is mostly composed of members of B4 (top right graph).

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- SupplementalData1.csv
- SupplementalData2.xlsx
- SupplementalData3.xlsx

- SupplementalData4.csv
- SupplementalData5.xlsx
- SupplementaryMaterial.docx