Title

COVID Symptoms, Symptom Clusters, and Predictors for Becoming a Long-Hauler Looking for Clarity in the Haze of the Pandemic.

Permalink
https://escholarship.org/uc/item/26t106w9

Journal
Clinical nursing research, 31(8)

ISSN
1054-7738

Authors
Huang, Yong
Pinto, Melissa D
Borelli, Jessica L
et al.

Publication Date
2022-11-01

DOI
10.1177/10547738221125632

Clinical Nursing Research
1–9
© The Author(s) 2022
Article reuse guidelines:
sagepub.com/journals-permissions
DOI: 10.1177/10547738221125632
journals.sagepub.com/home/cnr
SAGE

COVID Symptoms, Symptom Clusters, and Predictors for Becoming a Long-Hauler Looking for Clarity in the Haze of the Pandemic

Yong Huang[1], Melissa D. Pinto[2], Jessica L. Borelli[3], Milad Asgari Mehrabadi[1], Heather L. Abrahim[2], Nikil Dutt[1], Natalie Lambert[4], Erika L. Nurmi[5], Rana Chakraborty[6], Amir M. Rahmani[1,2], and Charles A. Downs[7]

Abstract

Post-acute sequelae of SARS-CoV-2 (PASC) is defined as persistent symptoms after apparent recovery from acute COVID-19 infection, also known as COVID-19 long-haul. We performed a retrospective review of electronic health records (EHR) from the University of California COvid Research Data Set (UC CORDS), a de-identified EHR of PCR-confirmed SARS-CoV-2-positive patients in California. The purposes were to (1) describe the prevalence of PASC, (2) describe COVID-19 symptoms and symptom clusters, and (3) identify risk factors for PASC. Data were subjected to non-negative matrix factorization to identify symptom clusters, and a predictive model of PASC was developed. PASC prevalence was 11% (277/2,153), and of these patients, 66% (183/277) were considered asymptomatic at days 0–30. Five PASC symptom clusters emerged and specific symptoms at days 0–30 were associated with PASC. Women were more likely than men to develop PASC, with all age groups and ethnicities represented. PASC is a public health priority.

Keywords

COVID-19, long-COVID, electronic health record, machine learning

Background and Significance

In the United States, over 83 million people have been infected with SARS-CoV-2, the virus responsible for COVID-19 (Johns Hopkins University, 2021), and the cumulative hospitalization rate has exceeded 1,300 persons per 100,000 since early 2020 (Centers for Disease Control and Prevention, 2021). Hospitalized patients account for 1% of COVID-19 patients, yet most of the COVID-19 research focuses on in-patients with severe disease (Bergquist et al., 2020). Among non-hospitalized patients, little is known about the medium- and long-term consequences of COVID-19. The most recent statistics show that 10–30% non-hospitalized patients, those with mild to moderate COVID-19 cases, will not recover quickly, within the expected time-frame for symptom resolution (Carfi et al., 2020; Greenhalgh et al., 2020; Lambert & Survivor Corps, 2020; Rubin, 2020). These individuals, termed "long-haulers," or persons with post-acute sequelae of SARS-CoV-2 (PASC) infection, as recently termed by the United States National Institutes of Health, struggle with debilitating, persistent, and ever-evolving symptoms that last for weeks and can exceed 1 year after SARS-CoV-2 infection. In short, COVID-19 has resulted in a cohort of millions of long-haulers worldwide, and we know little about the diagnosis and treatment. There is no cure.

It is not uncommon for infectious diseases to have late sequelae; however, the reason is unclear. For PASC, some scientists believe that late sequelae reflects primary organ involvement during acute infection; others believe that long-term signs and symptoms are promoted by aberrant inflammatory immune responses. PASC is yet to be

[1]Donald Bren School of Information and Computer Science, Irvine, CA, USA
[2]Sue & Bill Gross School of Nursing, Irvine, CA, USA
[3]University of California Irvine, USA
[4]Indiana University, Indianapolis, IN, USA
[5]UCLA Semel Institute for Neuroscience and Human Behavior, Los Angeles, CA, USA
[6]Mayo Clinic, Rochester, MN, USA
[7]University of Miami, Coral Gables, FL, USA

Corresponding Author:
Charles A. Downs, School of Nursing & Health Studies, University of Miami, 5030 Brunson Drive, Suite 334, Coral Gables, FL 33146, USA.
Email: cxd826@miami.edu

clinically and biologically characterized. We know little about PASC diagnosis, treatment, and what medical, community, and societal resources are needed to optimally support survivors in their recovery. Little is also known about the prevalence of PASC, risk factors to develop PASC, individual symptoms and symptom clusters, and symptoms that evolve over time among non-hospitalized patients. To date, PASC research has been limited by small sample sizes and omission of non-hospitalized survivors (classified as "mild" to "moderate" infection), thereby limiting population-level data that are essential for development of evidence-based management guidelines for PASC.

SARS-CoV-2 was thought to target the respiratory systems; however, we have since learned that multiple other organs are involved during infection; there is a wide range of symptom presentations. Clinical observations suggest the co-occurrence of hallmark COVID-19 symptoms. For example, dyspnea and chest pain tend to co-occur as do loss of sense of taste with loss of sense of smell. The underlying pathophysiology of these phenomena is unknown. To this end, it is important to investigate symptoms, symptom clusters, and their associations.

In this study, medically documented symptoms via electronic health records (EHRs) of non-hospitalized patients ($N=2,153$) with confirmed SARS-CoV-2 infection (via PCR) were examined to identify symptoms and symptom clusters. Specifically, we evaluated symptoms at presentation (days 0–30 following a positive PCR test for SARS-CoV-2) and at 180+ days. Symptoms present within the year prior to SARS-CoV-2 infection were excluded to mitigate potential overlap with pre-existing comorbidities. There is not yet consensus on the definition of PASC. For the purpose of this paper, PASC is defined as persistent symptoms at 180+ days, a time in which symptoms would be expected to have abated. We examined how early symptoms and pre-infection non-modifiable factors (age, ethnicity) could predict the likelihood of persistent symptoms at 180+ days (e.g., long-hauler) and/or assignment within any given symptom cluster.

Objectives

The objectives of this study were to (1) determine the prevalence of PASC, (2) to describe medically documented symptoms (EHR) of non-hospitalized patients with PCR-confirmed SARS-CoV-2-positive ($n=2,153$) tests at days 0–30 and 180+ days, and (3) to identify the factors that increase risk to develop symptoms at 180+ days.

Materials and Methods

Sample Size and Inclusion Criteria

University of California COvid Research Data Set (UC CORDS as of 03/17/2021) is a de-identified EHR for 52,083 patients treated in facilities throughout California and is available to University of California researchers. Because this is one of the first studies to characterize long-haul symptoms, it was important to ensure accurate temporal ordering of symptoms that were not subject to limitations of patient recall and used the gold standard of recording medical data, the EHR. Symptoms recorded in the EHR are documented by a medical provider, who receives information from the patient, and conducts a medical assessment. Using medically documented symptoms through the EHR increased our confidence in the accurate reporting of symptoms and their temporal order. This approach worked to overcome the limitations of prior work that relied on patient retrospective recall of symptoms by patients months following SARS-CoV-2 infection and lacked symptom history data and was plagued by the inability to determine the temporal order of symptoms or if symptoms were new or pre-existing before COVID-19. To validly capture the symptoms that could be attributed to COVID-19, we employed strict inclusion/exclusion criteria that yielded a sample ($N=2,153$) of never-hospitalized SARS-CoV-2-infected individuals with COVID-19 symptoms. Patients hospitalized for COVID-19 were excluded as were out-patients with false-positive PCR results—having a positive and negative test within the same visit—or documented reinfection with SARS-CoV-2. In addition, patients needed to have records within the system before SARS-CoV-2 infection for at least a year and symptoms reported before infection were excluded in the analysis. The data captured in the EHR reflect a time before vaccines were publicly available. Figure 1 provides a schematic of how we achieved the sample size of 2,153.

Non-Negative Matrix Factorization for Subgroup Identification

Non-negative matrix factorization (NMF) was employed to identify and quantitatively derive the subgroups of patients based on initial symptom presentation data. NMF is an unsupervised learning algorithm that can extract meaningful features from high-dimensional data sets (Lee & Seung, 1999). It is widely used to extract interpretable components from data in various domains such as image processing, data mining, and genomics. In principle, NMF could be used in any application for meaningful components extraction where the elements in input data are non-negative. We formulate our input data as a matrix A of m rows and n columns with each element $A_{ij} \geq 0$, where m is the number of subjects and n is the total number of symptoms of interest, each element denotes the count of a symptom documented within a specific time range for a given patient. NMF searches for matrices W of size $m*k$ and H of size $k*n$, such that $A \approx WH$. The matrices W and H are derived by minimizing the squared Frobenius norm $||A-WH||2$ with some suitable optimization algorithms such as coordinate descent. Here k is a hyperparameter and its value is determined by the user. In the context

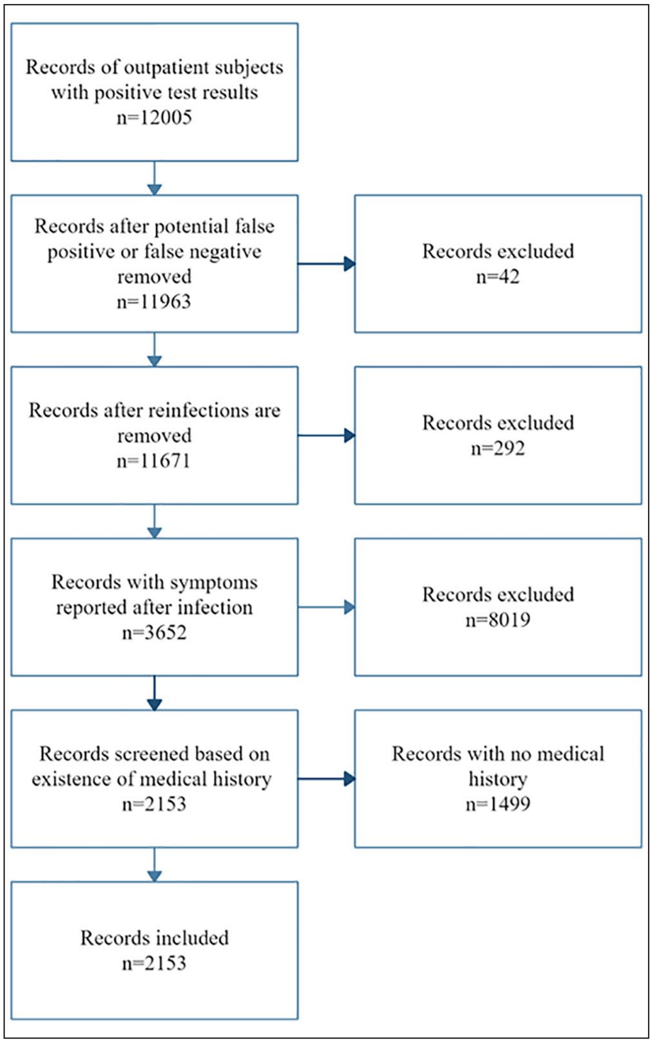
Records of outpatient subjects with positive test results n=12005
Records after potential false positive or false negative removed n=11963
Records excluded n=42
Records after reinfections are removed n=11671
Records excluded n=292
Records with symptoms reported after infection n=3652
Records excluded n=8019
Records screened based on existence of medical history n=2153
Records with no medical history n=1499
Records included n=2153

Figure 1. Flowchart depicting how records were screened for inclusion in the study.

of symptom cluster discovery, k would be the number of symptom clusters; its value should be smaller than the smallest of m and n. The matrix W is also known as the basis matrix, and H is called coefficient matrix. Note that the value of an element inside matrix A can be approximated in terms of the dot product of a basis vector (rows of W) and the corresponding coefficients (columns of H). Plenty of questions can be explained by the W matrix and H matrix. First of all, we can recognize which subgroup a patient resides in by inspecting the basis vector where the ith element of the basis vector can be interpreted as how likely the patient belongs to the ith subgroup, and given that the patient belongs to the ith subgroup, the chances of developing each symptom are represented by the coefficient vector in matrix H. Second, we can identify the most representative symptoms in each subgroup by aggregating and sorting the coefficient vectors.

Model Selection of NMF

We used Akaike information criterion (AIC) (Akaike, 1974), a method originated from information theory and widely applied in statistics and machine learning for model selection to determine the optimal k, the formula of AIC is given as follows:

$$\text{AIC} = -2\log L + K(N + M)$$

where L is the likelihood of the model, and $K(N+M)$ is the number of parameters in the model. When there are several candidate models, the best choice is identified as the one corresponding to the minimum AIC, and the intuition is that when the likelihood is identical for two models, the better model is the one with less parameters. We further normalized likelihood term and model complexity term in the original AIC formula because without proper normalization, the second term would usually dominate the equation and thus suggest $K=1$ all the time. We found out that AIC works the best and the optimal Ks derived are 5 for both early-stage symptoms and long-haul symptoms. More details are found in our supplemental materials.

Network Analysis to Determine Symptom Associations

Association study methods are typically based on correlations and to derive the associations, a common way is to set an arbitrary threshold (e.g., significance of $\alpha = .05$) via null-hypothesis testing, one often has to execute a considerable number of significance tests and deal with it through Bonferroni corrections, which will lead to a loss of power. To circumvent the above problems, we applied Graphical lasso (Glasso) which uses model selection to find the simplest model (sparse network) with spasticity achieved by imposing L1 penalty on model coefficients. Glasso operates on symptom occurrence counts data and yields pairwise associations of symptoms which are represented by an undirected network where each vertex represents a symptom. The network structure can be estimated by applying a lasso regression using each variable in the graph as the response and the others as predictors. After network structure estimation using Glasso, we employed the walktrap algorithm which is a standard community detection algorithm on graphs to further cluster nodes in the network (Friedman et al., 2008; Pons & Latapy, 2005).

Analysis to Identify Predictors of PASC

To identify factors leading to the development of persistent symptoms, we developed a predictive model that inputs multiple potential key factors to predict whether a subject with SARS-CoV-2 infection will become a long-hauler. By inspecting the model's coefficient strength, we can identify the key predictors. We formulated socio-demographics, symptoms the individuals experienced in the first 30 days, and whether a patient was asymptomatic in the first 30 days as input features and then applied logistic regression as our predictive model. To alleviate the effect of multicollinearity, which may cause misinterpretation of predictor importance, we performed hierarchical clustering on the Spearman rank-order correlations of input features.

One of the key properties of supervised machine learning techniques (e.g., logistic regression) is their predictive nature. In these techniques also known as classification, by analyzing a set of features (e.g., symptoms from the first 30 days) from an existing set of labeled samples (e.g., COVID-19 long-hauler vs. not COVID-19 long-hauler), a model can be learned to predict the probability of those classes/labels (e.g., becoming a long-hauler or not) for a new sample. Such predictive models do not require a control group for predictions since the labels (e.g., positive/negative) intrinsically provide information regarding both classes (in binary classification), assuming that enough samples were observed by the model to learn from. Once can then analyze the learned models to calculate the contribution/correlation of each of the features (e.g., symptoms) to the final classification.

Results

Features and Symptoms Among Community-Dwelling Individuals with COVID-19: Days 0–30 and 180+ Days

Table 1 shows the sample distribution related to age, sex, and ethnicity at days 0–30 and at 180+ days. Comorbidities included asthma (~13%), hypertension (~43%), diabetes mellitus (~24%), chronic obstructive pulmonary disease (~1%), hyperlipidemia (~38%), major depressive disorder (~15%), congestive heart failure (~7%), anemia (~21%), cerebrovascular disease (~1%), and hypothyroidism (~13%). Approximately 22% had 1 comorbidity, 16% had 2 comorbidities, 32% had three or more, and 30% of the sample did not have any of these comorbidities. Figure 2 shows the prevalence of symptoms reported at days 0–30 and 180+ days. Prevalent symptoms at days 0–30 include (in descending order) cough, dyspnea, fever, chest pain, headache, among others. Symptoms reported among those with PASC (180+ days) include (in descending order) dyspnea, chest pain, abdominal pain, headache, low back pain, among other symptoms.

Using NMF five symptom clusters with the co-occurrence of symptoms were identified at days 0–30 (Figure 3a) and 180+ days (Figure 3b). At days 0–30 symptom clusters included: dyspnea—diarrhea (N=283), cough–diarrhea (N=378), chest pain–heart palpitations (N=173), fever–nausea (N=270), and tachycardia–heart palpitations (N=337). Symptom network analysis was used to identify the strength of association between symptoms, wherein the darker the line connecting nodes indicates a stronger relationship. Symptom clusters observed among PASC (180+ days) included the following: dyspnea–cough (N=41), chest pain–cough (N=40), heart palpitations–anxiety (N=29), headache–low back pain (N=53), and nausea–fatigue (N=64). Symptom network analysis was again used to identify the strength of association between symptoms, wherein the darker the line connecting nodes indicates a stronger relationship.

Table 1. Demographics of Patients Seen for SARS-CoV-2 Infection at days 0–30 and 180+ days.

	0–30 Days	180+ Days
	n=2,153	n=277 (11%)
Age (years)	Number (%)	Number (%)
<18	46 (2%)	3 (1%)
18–29	261 (12%)	32 (14%)
30–39	336 (16%)	38 (17%)
40–49	367 (17%)	38 (17%)
50–59	433 (20%)	44 (19%)
60–69	357 (17%)	28 (12%)
70–79	225 (10%)	22 (10%)
>80	126 (6%)	22 (10%)
Gender		
Female	1,218 (57%)	129 (57%)
Male	935 (43%)	98 (43%)
Race/ethnicity		
Asian	121 (6%)	14 (6%)
Black	81 (4%)	7 (3%)
Hispanic	1,004 (47%)	115 (51%)
White	707 (33%)	75 (33%)
Other	240 (11%)	16 (7%)

Symptoms at Days 0–30 and Their Predictive Value for PASC and Specific Symptom Cluster

Table 1 provides the distribution of age, sex, and ethnicity among patients with PASC, those reporting symptoms at 180+ days. This group was distributed across all age groups, including among those <18, ethnicities, and included more women than men. In Figure 4, features associated with developing PASC were identified and included an initial asymptomatic presentation at the time of testing, as well as reporting alopecia, chronic rhinitis, joint or throat pain, and tinnitus, among many others. Conversely, an initial presentation of dysgeusia, or loss of taste, was negatively associated with developing PASC.

Features present at days 0–30 that most likely result in grouping within one of five symptom clusters identified among PASC survivors are reported in Figure 5. The greatest

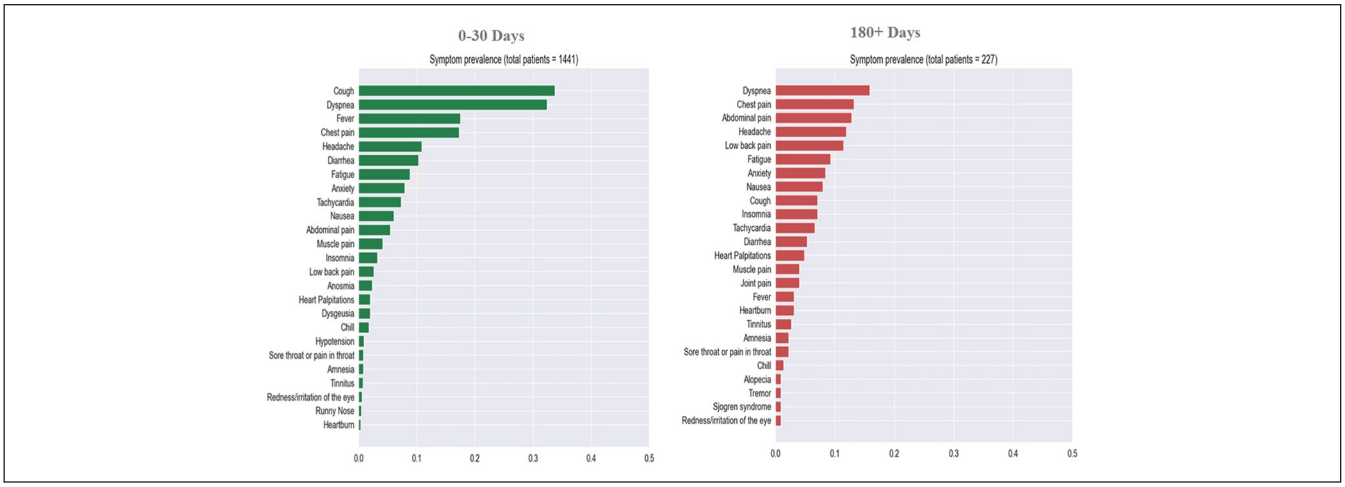

Figure 2. Symptoms prevalence among SARS-CoV-2 infected community dwellers at days 0–30 and 180+ days.
Bar graphs showing prevalence of symptoms reported at days 0–30 and 180+ days. Symptoms with very low prevalence are omitted in this graph.

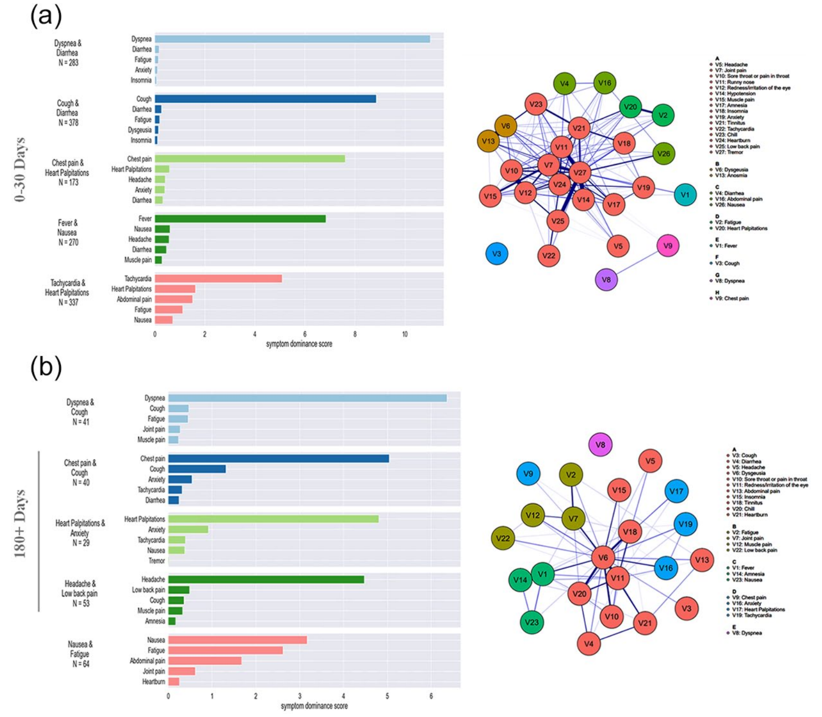

Figure 3. Symptom clusters among SARS-CoV-2-infected community dwellers at days (a) 0–30 and (b) 180+ days.
NMF determined symptom clusters depicted in bar graphs with symptom ranking within each cluster, graph demonstrating optimal *k* means clustering, and graph demonstrating symptom network analysis showing relationship between each reported symptom. Each symptom is denoted as a node, with darker lines connecting symptoms indicating stronger relationships.
NMF = non-negative matrix factorization.

magnitude between a feature and membership within a cluster included initial presentation with fatigue with the dyspnea–cough cluster, asymptomatic with heart palpitations–anxiety cluster, and muscle pain with nausea–fatigue cluster. Of the non-modifiable factors such as age, sex, and ethnicity, age 18–29 was associated with the nausea–fatigue symptom cluster and White race was associated with headache–low back pain cluster.

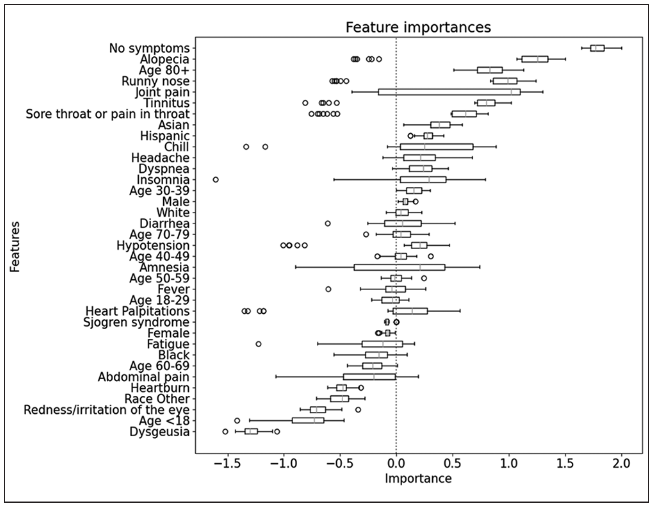

Figure 4. Key features during days 0–30 and their potential as indicators for developing prolonged COVID-19 symptoms or being a long-hauler.

Bar graph showing factors that positively or negatively affect the probability of developing persistent symptoms among COVID+ community dwellers.

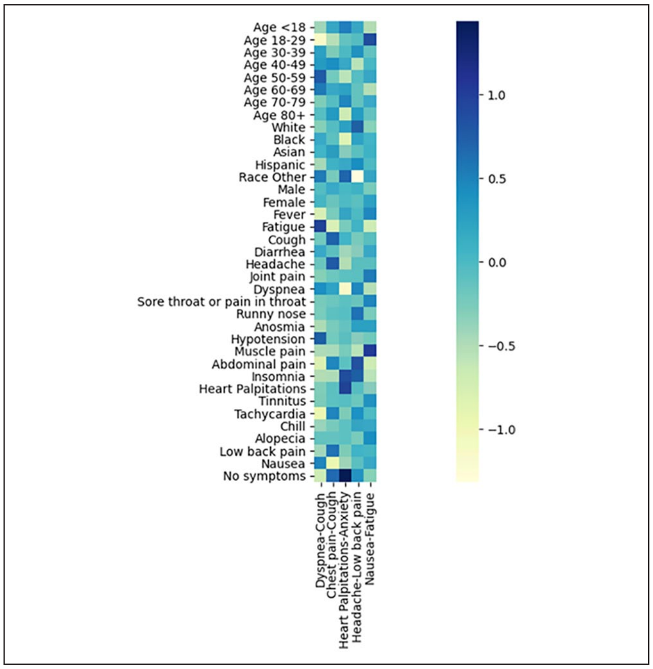

Figure 5. Presence of key indicators at days 0–30 predict inclusion into specific symptom clusters reported at 180+ days.

Heat map demonstrating magnitude of association between key predictors reported at days 0–30 and assignment to a cluster; darker coloring indicates greater positive magnitude of association.

Discussion

The current study provides much needed insight into early factors predisposing individuals for developing PASC. These novel findings warrant additional investigations, discussion, and context within current knowledge about long-haulers/PASC. In reviewing our findings, we believe there are several key take-home points from our analyses.

First, we found the PASC prevalence to be 11%. While our method captured symptoms attributable to COVID-19, it is likely the symptoms reported in the EHR are not comprehensive. Patient visits are typically short, and patients often report the most bothersome symptoms, the symptoms they see as most related to the suspected illness, and refrain from discussing symptoms that may be stigmatized or cause embarrassment. The understanding of COVID-19 symptoms and openness of providers in validating patient accounts have evolved over time. While this study estimated PASC prevalence at 11%, it is possible that there are more patients with symptoms that did not disclose for one the aforementioned reasons. Therefore, our estimate is conservative and is likely at the lower bounds of prevalence. Of note, of those that developed PASC, 66% were asymptomatic days 0–30. The reason for testing is unclear, but the implication that many asymptomatic, SARS-CoV-2-infected persons suggest PASC may be difficult to diagnose because of a lack of COVID-19 symptoms or COVID-19 disease among non-hospitalized people. Moreover, this could represent a significant number of people that may be omitted from any supportive services that would assist in recovery where documentation of a PCR-positive test is required.

Second, the UC CORDS data set provides both patient-reported and clinician-documented symptoms from SARS-CoV-2-infected patients. These symptoms are reported and recorded in real time, which minimizes retrospective recall that has been used in the limited studies to date and includes any symptoms uncovered by the medical provider (Carfi et al., 2020; Carvalho-Schneider et al., 2020). A few other important strengths from using the UC CORDS data set and the analytic techniques are exclusion of symptoms attributed to chronic diseases (e.g., asthma, heart failure, etc.). By excluding symptoms reported prior to SARS-CoV-2 infection, we increase confidence in the identified symptoms being attributable to becoming a long-hauler. However, it is important to note that while exclusion of symptomatic comorbidities may raise confidence in the symptoms attributable to PASC, a potential limitation with our model is that it did not account for asymptomatic or well-controlled comorbid conditions, nor does it account for the potential impact of comorbidities on PASC and symptoms/symptom clusters. However, the use of the data set allowed for a broad swath of symptoms, rather than being limited to a narrowly focused checklist of symptoms, which allows for a more sophisticated understanding of symptoms among long-haulers. Symptoms were obtained through a clinical encounter and are documented by the provider; however, this does not mean that all symptoms experienced were captured. It is likely that symptoms that were causing the patient the most problems were reported, and only symptoms that patients believe were associated with the primary problem. Thus, it is possible that dominant symptoms of each symptom cluster were potentially the most bothersome to patients. Also, it is worth considering that a feature of PASC may be the acceleration of symptoms that could be attributed to worsening or progression of underlying comorbidities, a nuance not currently being considered in attempting to understand and define PASC.

Third, our observations suggest a developing picture of long-haulers potentially reflecting that middle age and female sex as common features specific to a subset of long-haulers. Although similar descriptions have been provided in other investigations (Huang et al., 2021) and the lay media, further corroboration is warranted. We observed a near normal distribution of age among long-haulers (Table 1), including those under the age of 18—with the mean age at 10.92 years. Although our study supported a potential association with male sex and higher likelihood of becoming a long-hauler, race also appeared to be predictive for both Caucasian and Hispanic ethnicity. Blacks were at decreased risk (Figure 4), and a larger sample is needed to validate this finding. Age distribution of all SARS-CoV-2-infected individuals at days 0–30 very closely mimicked that of the long-haulers, suggesting the latter group are distributed across all age groups with persons ages 50–59 range (±20 years) representing more than 75% of the long-hauler population.

Next, the symptom experience among those who become long-haulers changes over time and may be influenced by the circulating viral strain. In this study, founder and alpha strains were the dominant strains, and symptoms associated with SARS-CoV-2 infection have evolved as newer strain emerge. However, data from multiple studies converge to illustrate that many hospitalized and non-hospitalized survivors of COVID-19 experience persistent symptoms (Chopra et al., 2020; Davis et al., 2020; Goërtz et al., 2020; Halpin et al., 2021; Huang et al., 2021; Mandal et al., 2020; Meys et al., 2020). The reported incidence of persistent symptoms varies; however, in the current study, we report that 11% of people-reported symptoms after 180+ days. Some of the variability in symptom reporting and symptom association with long-haulers in other studies may be due to limitations inherent in rapid screening questionnaires in as much as these questionnaires inquire about symptoms that predominantly impact those with severe disease. Also, questionnaires may fail to inquire about emerging symptoms such as cognitive dysfunction (including "brain fog"), limiting the ability to accurately document such symptoms. Asymptomatic individuals may be less often intensely monitored due to an inherent notion of low risk for severe acute disease; however, this is problematic as asymptomatic individuals account for 66% of the long-haulers observed in this study. Since the start of the pandemic, we have learned that COVID-19 affects nearly every body system, and there is a wide variety of symptoms. It is possible that patients who reported symptoms did not mention during their office visit with the provider symptoms that they did not connect as related to their primary complaint. Because long-haul among non-hospitalized patients has not been well characterized, we took a conservative approach to symptom identification with increased confidence that symptoms documented were not pre-existing and because information was collected in real time during office visits, it was not subject to recall. We see these symptom clusters as a foundation to build from in future studies and enrich and continue to validate with other data sources. The symptom clusters observed among long-haulers vary compared to those at initial presentation. The evolution of these clusters may provide insight into the etiology of long-haulers in which elucidating sites of evolving tissue damage, and alterations in innate and adaptive immune inflammatory pathways might provide clarity in understanding the underlying pathophysiology.

In October 2020, the Tony Blair Institute for Global Change identified key characteristics among long-haulers, specifically that women appear to be at greater risk and those who are of working age (mean of age 45) (Sleat et al., 2020). Our data align with these observations. Therefore, to our third key point, we observed that all ethnicities and races were affected as well as individuals who were initially asymptomatic. However, our use of ethnicity and race is limited to broad groups and lacks needed specificity, a limitation imposed by how data are recorded in the EHR.

Larger population-based studies will be needed to confirm and expand upon these observations, particularly studies using large databases that can incorporate the methods used here would be insightful in determining emerging variants and their symptom patterns with the development of PASC. Undertaking detailed immune profiling through emerging technologies such as the -omics platforms may identify key host phenotypes associated with the symptom clusters that we have described. We hope this article will prompt the development and implementation of longitudinal prospective studies that garner patient-generated reports of symptoms, rather than patient responses to questions generated by researchers—this latter approach inherently constrained the answers we obtained. With such a new phenomenon, an ethnographic approach that focuses on understanding patients' experiences would add an important lens to our analyses.

Conclusion

This study utilized machine learning techniques to develop and test an algorithm designed to predict individuals at risk for the development of PASC. In addition, we identified symptoms and symptom clusters among individuals with persistent symptoms at 180+ days post-SARS-CoV-2 infection. Indeed, data suggest that infection with SARS-CoV-2 leads to prolonged and persistent symptoms in a subset of persons known as long-haulers. However, the long-term consequences of becoming a long-hauler are unclear, and further research is urgently needed to corroborate our findings.

These findings include identifying a cohort of long-haulers with non-modifiable risk factors, which may have predicted the likelihood of persistent symptoms and/or assignment within given symptom clusters. Further research is needed to understand the underlying pathophysiology including host phenotypes associated with aberrant innate and adaptive immune responses following SARS-CoV-2 infection.

Declaration of Conflicting Interests

The authors declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The authors disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: The data set used in this study, University of California COvid Research Data Set (UC CORDS), is made available by the University of California Office of the President and the University of California Biomedical, Research, Acceleration, Integration, and Development. Biomedical computing facilities are supported by the National Center for Research Resources and the National Center for Advancing Translational Sciences, National Institutes of Health, through Grant UL1 TR001414. The content is solely the responsibility of the authors and does not necessarily represent the official views of the NIH.

ORCID iDs

Melissa D. Pinto https://orcid.org/0000-0002-1003-180X
Charles A. Downs https://orcid.org/0000-0003-4859-7929

Supplemental Material

Supplemental material for this article is available online.

References

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, *19*(6), 716–723. https://doi.org/10.1109/TAC.1974.1100705

Bergquist, S. H., Partin, C., Roberts, D. L., O'Keefe, J. B., Tong, E. J., Zreloff, J., Jarrett, T. L., & Moore, M. A. (2020). Non-hospitalized adults with COVID-19 differ noticeably from hospitalized adults in their demographic, clinical, and social characteristics. *SN Comprehensive Clinical Medicine*, *2*(9), 1349–1357. https://doi.org/10.1007/s42399-020-00453-3

Carfi, A., Bernabei, R., & Landi, F. Gemelli Against COVID-19 Post-Acute Care Study Group. (2020). Persistent symptoms in patients after acute COVID-19. *Jama*, *324*(6), 603–605. https://doi.org/10.1001/jama.2020.12603

Carvalho-Schneider, C., Laurent, E., Lemaignen, A., Beaufils, E., Bourbao-Tournois, C., Laribi, S., Flament, T., Ferreira-Maldent, N., Bruyère, F., Stefic, K., Gaudy-Graffin, C., Grammatico-Guillon, L., & Bernard, L. (2020). Follow-up of adults with noncritical COVID-19 two months after symptom onset. *Clinical Microbiology and Infection : The Official Publication of the European Society of Clinical Microbiology and Infectious Diseases*, *27*(2), 258–263. https://doi.org/10.1016/j.cmi.2020.09.052 [pii]

Centers for Disease Control. (2021). COVID-19 laboratory confirmed hospitalizations. https://gis.cdc.gov/grasp/covidnet/COVID19_5.html

Chopra, V., Flanders, S. A., O'Malley, M., Malani, A. N., & Prescott, H. C. (2020). Sixty-day outcomes among patients hospitalized with COVID-19. *Annals of Internal Medicine*, *174*(4), 576–578. https://doi.org/10.7326/M20-5661

Davis, H. E., Assaf, G. S., McCorkell, L., Wei, H., Low, R. J., Re'em, Y., Redfield, S., Austin, J. P., & Akrami, A. (2020). Characterizing long COVID in an international cohort: 7 months of symptoms and their impact. *MedRxiv*, *38*, 101019 . https://doi.org/10.1101/2020.12.24.20248802

Friedman, J., Hastie, T., & Tibshirani, R. (2008). Sparse inverse covariance estimation with the graphical lasso. *Biostatistics (Oxford, England)*, *9*(3), 432–441. https://doi.org/10.1093/biostatistics/kxm045

Goërtz, Y. M. J., Van Herck, M., Delbressine, J. M., Vaes, A. W., Meys, R., Machado, F. V. C., Houben-Wilke, S., Burtin, C., Posthuma, R., Franssen, F. M. E., van Loon, N., Hajian, B., Spies, Y., Vijlbrief, H., van 't Hul, A. J., Janssen, D. J. A., & Spruit, M. A. (2020). Persistent symptoms 3 months after a SARS-CoV-2 infection: The post-COVID-19 syndrome? *ERJ Open Research*, *6*(4), 00542–2020. https://doi.org/10.1183/23120541.00542-2020

Greenhalgh, T., Knight, M., A'Court, C., Buxton, M., & Husain, L. (2020). Management of post-acute covid-19 in primary care. *BMJ*, *370*, m3026. https://doi.org/10.1136/bmj.m3026

Halpin, S. J., McIvor, C., Whyatt, G., Adams, A., Harvey, O., McLean, L., Walshaw, C., Kemp, S., Corrado, J., Singh, R., Collins, T., O'Connor, R. J., & Sivan, M. (2021). Postdischarge symptoms and rehabilitation needs in survivors of COVID-19 infection: A cross-sectional evaluation. *Journal of Medical Virology*, *93*(2), 1013–1022. https://doi.org/10.1002/jmv.26368

Huang, C., Huang, L., Wang, Y., Li, X., Ren, L., Gu, X., Kang, L., Guo, L., Liu, M., Zhou, X., Luo, J., Huang, Z., Tu, S., Zhao, Y., Chen, L., Xu, D., Li, Y., Li, C., Peng, L., . . . Cao, B. (2021). 6-month consequences of COVID-19 in patients discharged from hospital: A cohort study. *The Lancet (London, England)*, *397*(10270), 220–232. https://doi.org/S0140-6736(20)32656-8

Johns Hopkins University. (2021) *COVID-19 dashboard by the center for systems science and engineering*. https://gisanddata.maps.arcgis.com/apps/opsdashboard/index.html#/bda7594740fd40299423467b48e9ecf6

Lambert, N. J., & Survivor Corps. (2020). *COVID-19 "long-hauler" symptoms survey report* (Survey report). Indiana University School of Medicine.

Lee, D. D., & Seung, H. S. (1999). Learning the parts of objects by non-negative matrix factorization. *Nature*, *401*(6755), 788–791. https://doi.org/10.1038/44565

Mandal, S., Barnett, J., Brill, S. E., Brown, J. S., Denneny, E. K., Hare, S. S., Heightman, M., Hillman, T. E., Jacob, J., Jarvis, H. C., Lipman, M. C. I., Naidu, S. B., Nair, A., Porter, J. C., Tomlinson, G. S., Hurst, T.R., & ARC Study Group. (2020). 'Long-COVID': A cross-sectional study of persisting symptoms, biomarker and imaging abnormalities following hospitalisation for COVID-19. *Thorax*, *76*(4), 396–398. https://doi.org/10.1136/thoraxjnl-2020-215818

Meys, R., Delbressine, J. M., Goërtz, Y. M. J., Vaes, A. W., Machado, F. V. C., Van Herck, M., Burtin, C., Posthuma, R., Spaetgens, B., Franssen, F. M. E., Spies, Y., Vijlbrief, H., van't Hul, A. J., Janssen, D. J. A., Spruit, M. A., & Houben-Wilke, S. (2020). Generic and respiratory-specific quality of life in non-hospitalized patients with COVID-19. *Journal of Clinical Medicine*, *9*(12), 3993. https://doi.org/10.3390/jcm9123993

Pons, P., & Latapy, M. (2005). Computing communities in large networks using random walks. In: Yolum, p., Güngör, T., Gürgen, F., & Özturan, C. (eds.), *Computer and information sciences - ISCIS 2005. ISCIS 2005. Lecture notes in computer science* (long version, vol. 3733, pp. 284–293). Berlin, Heidelberg: Springer.

Rubin, R. (2020). As their numbers grow, COVID-19 "Long haulers" stump experts. *JAMA*, *324*(14), 1381–1383. https://doi.org/10.1001/jama.2020.17709

Sleat, D., Wain, R., & Miller, B. (2020). *Long Covid: Reviewing the science and assessing the risk*. https://institute.global/sites/default/files/articles/Long-Covid-Reviewing-the-Science-and-Assessing-the-Risk.pdf

Author Biographies

Yong Huang is a PhD student in the Department of Computer Science, Donald Bren School of Information and Computer Science, University of California, Irvine.

Melissa D. Pinto, PhD,RN, FSHAM, FAAN is an associate professor in the Sue and Bill Gross School of Nursing, University of California, Irvine. Dr. Pinto's research focuses on mHealth and mental health among adolescents, including depression and cognition, as well as long-COVID and long-COVID symptomatology.

Jessica L. Borelli, PhD is an associate professor in the School of Social Ecology, University of California, Irvine. She is a clinical psychologist specializing in the field of developmental psychopathology with a research focus on links between close relationships, emotions, health, and development, with a particular focus on risk for anxiety and depression.

Milad Asgari Mehrabadi is a PhD candidate in the Department of Computer Science, Donald Bren School of Information and Computer Science, University of California, Irvine.

Heather L. Abrahim is a PhD student in the Sue and Bill Gross School of Nursing, University of California, Irvine.

Nikil Dutt, PhD is Chancellor's Professor in Computer Science, Electrical Engineering, and Computer Science , and Cognitive Sciences at the University of California, Irvine. Dr. Dutt's research interests are in embedded systems, electronic automation, computer architecture, optimizing compliers, systems specification techniques, distributed systems, and formal methods.

Natalie Lambert, PhD is an associate professor in the Department of Biostatistics and Health Data Sciences at Indiana University School of Medicine. Dr. Lambert's research focuses on understanding the patients' experiences with chronic diseases, including social, attitudinal, and environmental factors that infliuence health.

Erika L. Nurmi, MD, PhD is the medical director of the UCLA Obsessive-Compulsive Disorder Intensive Outpatient Program, the associate director of the psychiatry residency research track, and a member of the Child and Adolescent Psychiatry Division faculty in the Department of Psychiatry and Biobehavioral Sciences at the UCLA Semel Institute for Neuroscience and Human Behavior. Dr. Nurmi's research focuses on the genetic basis of childhood OCD and tic disorders.

Rana Chakraborty, MD, D.Phil is a professor and pediatric infectious disease specialist in the Department fo Pediatrics and Adolescent Medicine at the Mayo Clinic. Dr. Chakraborty's research interests involve placental immunology in the context of HIV, Cytomegalovirus, and Zika virus.

Amir M. Rahmani, PhD is an associate professor of Nursing and Computer Science at the University of California, Irvine. Dr. Rahmani is the Associate Director of the UCI Institute for Future Health and leads the multi-disciplinary HealthScieTech Group at UCI. Dr. Rahamani's research includes medical cyber-physical systems, Internet-of-Things, and e-health.

Charles A. Downs PhD, ACNP-BC, FAAN is associate professor and Director of the Biobehavioral Laboratory in the School of Nursing & Health Studies at the University of Miami. Dr. Downs's research interest include use of high-dimensional and molecualr and biochemcial methods to understand the underpinnings of biological responses to injury and their application to understanding the development and evolution of symptoms, particulary Long-COVID.