Title

Precision phenotyping for curating research cohorts of patients with unexplained post-acute sequelae of COVID-19

Permalink
https://escholarship.org/uc/item/89g2c1s6

Journal
Med, 6(3)

ISSN
2666-6359

Authors
Azhir, Alaleh
Hügel, Jonas
Tian, Jiazi
et al.

Publication Date
2025-03-01

DOI
10.1016/j.medj.2024.10.009

Peer reviewed

Published in final edited form as:
Med. 2025 March 14; 6(3): 100532. doi:10.1016/j.medj.2024.10.009.

Precision phenotyping for curating research cohorts of patients with unexplained post-acute sequelae of COVID-19

Alaleh Azhir[1,2,12], **Jonas Hügel**[1,3,12], **Jiazi Tian**[1], **Jingya Cheng**[1], **Ingrid V. Bassett**[4], **Douglas S. Bell**[5], **Elmer V. Bernstam**[6], **Maha R. Farhat**[7], **Darren W. Henderson**[8], **Emily S. Lau**[4], **Michele Morris**[9], **Yevgeniy R. Semenov**[10], **Virginia A. Triant**[4], **Shyam Visweswaran**[9], **Zachary H. Strasser**[4], **Jeffrey G. Klann**[4], **Shawn N. Murphy**[11], **Hossein Estiri**[1,4,13,*]

[1]Clinical Augmented Intelligence Group, Massachusetts General Hospital, Boston, MA, USA

[2]Department of Medicine, Brigham and Women's Hospital, Boston, MA, USA

[3]Department of Medical Informatics, University Medical Center Göttingen, Göttingen, Germany

[4]Department of Medicine, Massachusetts General Hospital, Boston, MA, USA

[5]Department of Medicine, University of California, Los Angeles, Los Angeles, CA, USA

[6]McWilliams School of Biomedical Informatics, The University of Texas Health Science Center at Houston, Houston, TX, USA

[7]Department of Biomedical Informatics, Harvard Medical School, Boston, MA, USA

[8]Center for Clinical and Translational Science, University of Kentucky, Lexington, KY, USA

[9]Department of Biomedical Informatics, University of Pittsburgh, Pittsburgh, PA, USA

[10]Department of Dermatology, Massachusetts General Hospital, Boston, MA, USA

[*]Correspondence: hestiri@mgh.harvard.edu.

AUTHOR CONTRIBUTIONS

Conceptualization, H.E., J.H., and A.A.; methodology: J.H.; formal analysis, H.E., J.T., A.A., and J.C.; data curation, H.E., J.G.K., and J.T.; unrestricted access to all the data, H.E., J.T., A.A., and J.C.; writing – original draft, A.A., H.E., J.T., and J.C.; writing – review & editing, H.E., A.A., J.H., J.T., J.C., I.V.B., D.S.B., E.V.B., M.R.F., D.W.H., E.S.L., M.M., Y.R.S., V.A.T., S.V., Z.H.S., J.G.K., and S.N.M.; funding acquisition, H.E. and S.N.M. All authors agreed to submit the manuscript, read and approved the final article, and take full responsibility of its content, including the accuracy of the data and its statistical analysis.

DECLARATION OF INTERESTS

The authors declare no competing interests.

SUPPLEMENTAL INFORMATION

Supplemental information can be found online at https://doi.org/10.1016/j.medj.2024.10.009.

RESOURCE AVAILABILITY

Lead contact

Requests for further information and resources should be directed to and will be fulfilled by the lead contact, Hossein Estiri (hestiri@mgh.harvard.edu).

Materials availability

This study did not generate new unique reagents.

Data and code availability

Due to privacy regulations and per institutional and institutional review board approvals for this study, the patient-level data cannot be shared. All original code has been deposited at https://github.com/clai-group/long_covid_ai_scripts (or via docker at https://github.com/clai-group/long_covid_ai_scripts/pkgs/container/post_covid_ai_scripts) and is publicly available at https://doi.org/10.5281/zenodo.13835060 as of the date of publication. A short presentation of the methodologies can be found on GitHub: https://clai-group.github.io/long_covid_ai_implementation_guide (https://doi.org/10.5281/zenodo.13835071). Any additional information required to reanalyze the data reported in this paper is available from the lead contact upon request.

[11]Department of Neurology, Massachusetts General Hospital, Boston, MA, USA

[12]These authors contributed equally

[13]Lead contact

SUMMARY

Background: Scalable identification of patients with post-acute sequelae of COVID-19 (PASC) is challenging due to a lack of reproducible precision phenotyping algorithms, which has led to suboptimal accuracy, demographic biases, and underestimation of the PASC.

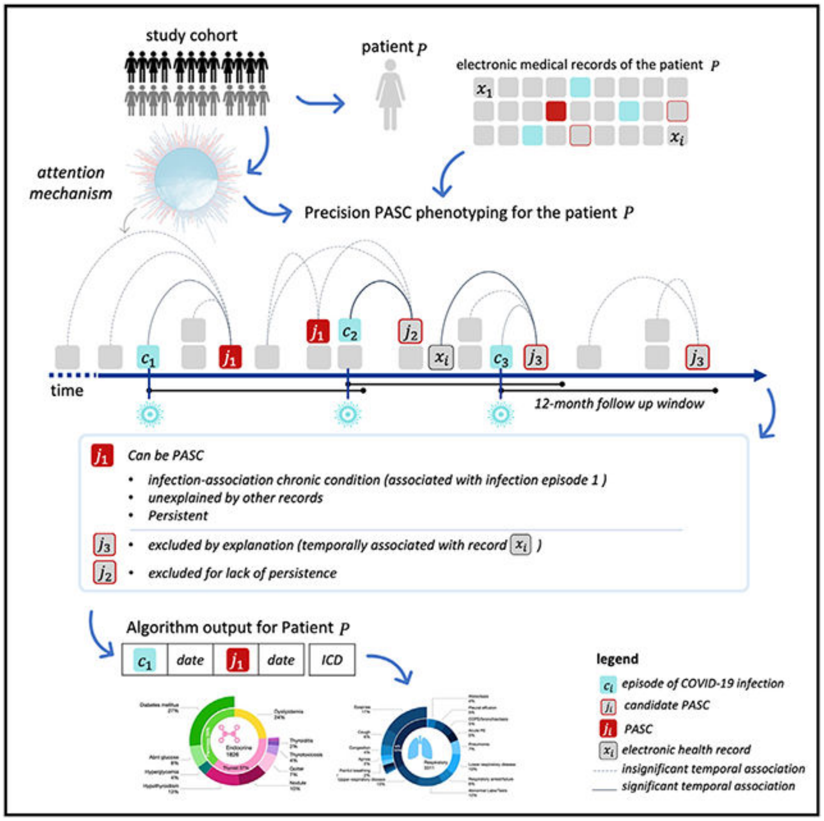
Methods: In a retrospective case-control study, we developed a precision phenotyping algorithm for identifying cohorts of patients with PASC. We used longitudinal electronic health records data from over 295,000 patients from 14 hospitals and 20 community health centers in Massachusetts. The algorithm employs an attention mechanism to simultaneously exclude sequelae that prior conditions can explain and include infection-associated chronic conditions. We performed independent chart reviews to tune and validate the algorithm.

Findings: The PASC phenotyping algorithm improves precision and prevalence estimation and reduces bias in identifying PASC cohorts compared to the ICD-10-CM code U09.9. The algorithm identified a cohort of over 24,000 patients with 79.9% precision. Our estimated prevalence of PASC was 22.8%, which is close to the national estimates for the region. We also provide in-depth analyses, encompassing identified lingering effects by organ, comorbidity profiles, and temporal differences in the risk of PASC.

Conclusions: PASC precision phenotyping boasts superior precision and prevalence estimation while exhibiting less bias in identifying patients with PASC. The cohort derived from this algorithm will serve as a springboard for delving into the genetic, metabolomic, and clinical intricacies of PASC, surmounting the constraints of prior PASC cohort studies.

Funding: This research was funded by the US National Institute of Allergy and Infectious Diseases (NIAID).

Graphical Abstract


study cohort
patient p
electronic medical records of the patient P
x_1
x_i
attention mechanism
Precision PASC phenotyping for the patient P
time
c_1
j_1
c_2
j_2
x_i
c_3
j_3
12-month follow up window
j_1 Can be PASC
• infection-association chronic condition (associated with infection episode 1)
• unexplained by other records
• Persistent
j_3 • excluded by explanation (temporally associated with record x_i)
j_2 • excluded for lack of persistence
Algorithm output for Patient P
c_1 date j_1 date ICD
legend
c_i episode of COVID-19 infection
j_i candidate PASC
j_i PASC
x_i electronic health record
insignificant temporal association
significant temporal association

In brief

Azhir et al. developed a novel precision phenotyping algorithm using electronic health records to precisely identify cases of post-acute sequelae of COVID-19 (PASC). This method enhances diagnostic precision and reduces demographic biases, offering a valuable tool for future research into the genetic, metabolomic, and clinical aspects of long COVID.

INTRODUCTION

The COVID-19 pandemic exerted widespread, long-lasting impacts on the full spectrum of human health. As we move into the endemic phase of SARS-CoV-2, a considerable proportion of the population continues to struggle with prolonged symptoms following their initial exposure to SARS-CoV-2. Post-acute sequelae of SARS-CoV-2/COVID-19 (PASC), also referred to as post-COVID conditions (PCCs) and long COVID,[1,2] have emerged as a complex issue, subject to ongoing scientific and political debate globally. A growing body of evidence suggests that PASC infection affects multi-organ systems.[3-9]

Despite extensive efforts to characterize and evaluate risks for PASC, a scalable, universally accepted algorithm for identifying patients who may be suffering from PASC is lacking. In the US, the current diagnostic codes (e.g., ICD-10 code U09.9: PCC) lack sensitivity and specificity for accurately identifying afflicted patients.[10-13] Zhang et al.[14] demonstrated that U09.9 (ICD-10 code) could be an untrustworthy proxy for gauging long COVID, where the positive predictive value (PPV) ranged from 40%–65%, based on the PASC reference definition. In addition, Pfaff et al. found a notable tilt in the demographic makeup of patients diagnosed with U09.9 toward women, White, non-Hispanic individuals, along with those residing in regions characterized by low poverty rates, high educational attainment, and ample access to medical services.[11] The lack of a robust phenotyping algorithm for identifying patients with PASC has hindered effective enrollment for large-scale clinical studies of potential therapies.

Further, standard approaches for measuring relative effects or associations aimed at discerning conditions exhibiting elevated relative risks among an exposed group (in this case, patients with COVID-19) are insufficient to identify individuals afflicted with PASC. For instance, shortness of breath has been extensively documented as a PASC.[15] Nevertheless, not all instances of shortness of breath observed in individuals with a history of COVID-19 denote a PASC, as such symptoms may be attributable to pre-existing conditions such as heart failure or asthma. These challenges demand particular scrutiny to enable the development of a robust algorithm for the characterization of PASC with real-world data.[14]

The World Health Organization (WHO) characterizes PASC as a diagnosis of exclusion, which offers a practical basis for identifying those suffering from PASC (henceforth referred to as long-haulers). The WHO defines PASC as the continuation or development of new symptoms 3 months after the initial infection, lasting at least 2 months with no other explanation.[16,17] Passively collected longitudinal data from electronic health records (EHRs) provide a cost-effective option for enriching cohort definitions and identifying at-risk individuals.

In this study, we introduce a reproducible precision phenotyping algorithm for the PASC based on the WHO's definition of PASC as a diagnosis of exclusion, with clinical data from EHRs. In this case-control study, our algorithm first identifies conditions associated with SARS-CoV-2 (similar to prior studies[18-21]; Figure 1). Second, using a specialized temporal pattern mining algorithm (tSPM),[22] our algorithm takes an extra step to distinguish those sequelae that cannot be explained by the patient's past medical history and infection-associated chronic condition (IACC). This algorithm adds a novel personalized exclusion by temporal association step to the standard risk studies, resulting in the largest validated computationally curated cohort of long-haulers (Figure 2). We provide a fully executable environment that can be directly applied and/or tuned to any healthcare organization with ICD-10 diagnosis and procedure codes.

Our method for identifying PASC boasts superior precision (vs. U09.9), accurately gauging the prevalence of this condition without downplaying its significance. Further, compared to the conventional U09.9 diagnosis code, our approach exhibits less bias in identifying patients with PASC across demographic groups, offering a more nuanced understanding of patients with long COVID. Further, our approach enables addressing temporal questions on the recurrence and sequence of PASC following various episodes of COVID-19 infection.

In addition to introducing the precision phenotyping approach, we provide an in-depth analysis outlining the clinical attributes, encompassing identified lingering effects by organ, comorbidity profiles, and temporal differences in the risk of PASC. The comprehensive PASC cohort resulting from our precision phenotyping algorithm will enable deep dives into the multifaceted expressions of long COVID through genetic, metabolomic, and clinical inquiries. This surpasses the constraints of earlier cohort studies, which were hampered by limited size and outcome data.

RESULTS

There were 85,364 COVID-19 cases, 170,497 (matched 1:2) post-pandemic controls, and 39,817 participants in the pre-pandemic control group who met our EHR longitudinal continuity threshold. The COVID-19 group had a mean age of 53.6 years (SD: 17.2), and 62.6% of participants were female. The post- and pre-pandemic control groups had mean ages of 53.7 (SD: 17.2) and 54.7 years (SD: 17.8), and 62.5% and 63.2% of participants were female, respectively (Table 1). All patients with an infection were followed up for 12 months after their last infection episode.

We evaluated 434 Clinical Classifications Software Refined (CCSR) categories for PASC, 66 of which remained in the final output after applying the diagnosis of exclusion (Table S2). For benchmarking, we reviewed clinical charts from 862 randomly selected patients with a U09.9 ICD-10 diagnosis code. 17 charts were reviewed by two separate reviewers to evaluate inter-rater reliability. There was 88% concordance and a Cohen's kappa of 0.76 between reviewers. In 671 (77.8%), the PASC diagnosis code was true. Figure S1 illustrates the bootstrap validation estimates of the PPVs across the cumulative correlations. After minimizing the type I error, the PPV/precision of our PASC phenotyping algorithm was 79.9%– i.e., a 2.7% improvement in precision over U09.9. Figure S2 illustrates an entry from the long-haulers list output, representing a hypothetical patient with PASC.

In the study period, 6,340 patients had a record of U09.9 in the Mass General Brigham (MGB) clinical data repository, though only 3,970 (62.6%) had a positive COVID-19 record. 80.1% of patients with a U09.9 diagnosis record were White, 4.83% were Black or African American, 2.7% were Hispanic, and 67.5% were female. Over 62% (3,970) of the patients with a U09.9 record had a record indicating COVID-19 infection (diagnosis code or positive PCR test).

The precision PASC phenotyping algorithm identified 24,360 patients (28.5% of the cases) as long-haulers having at least one PASC. According to our algorithm, over 13% of COVID-19 patients had more than one PASC, with 4% afflicted by more than three. 71.4% of long-haulers were White, 10.4% Black, 6.6% Hispanic, and 64.5% female (Table S5). Of the patients with a U09.9 diagnosis code, 2,104 had both medical record(s) indicating COVID-19 infection(s) and met our data continuity criteria. Our algorithm picked up 73.8% of these patients.

In Figures 3 and 4, we plot the distribution of PASC by organ (see Table S3 for more details). Of the 85,364 patients with COVID-19 (cases), 10,044 (11.8% of the cases and 41.2% of the long-haulers) experienced systemic post-COVID sequelae, including edema, generalized pain, sleep disorders, change in weight or nutrition, and malaise and fatigue.

More rare systemic complications included dizziness, dysphonia, hyperhidrosis, and sexual dysfunction in the form of low libido. About 7,000 patients had cardiovascular PASC, including palpitations, changes in heart rate (tachy/bradycardia), chest pain, dysrhythmia, changes in blood pressure (hyper/hypotension), and, more rarely, coronary atherosclerosis, heart failure, and myocarditis.

Over 4,000 patients (4.7% cases and 16.4% of long-haulers) experienced long-term gastrointestinal (GI) problems, including prolonged abdominal pain, nausea, and vomiting, changes in bowel habits (constipation/diarrhea), esophageal disorders and dysphagia, and biliary and liver-associated symptoms. Less than 1 in 25 patients with a past SARS-CoV-2 infection, comprising 13.6% of long-haulers, experienced respiratory sequelae, including upper/lower respiratory disease, shortness of breath, cough, pneumonia, respiratory arrest, pulmonary embolism, chronic obstructive pulmonary disease (COPD) exacerbations, pleural effusion, or atelectasis. About 3,000 patients experienced neurological PASC, including pain syndromes and polyneuropathies, nervous system disorders, abnormal cognition, gait, smell, and taste, and headache.

Skin PASC was observed in 1,972 patients (8.1% of longhaulers), including skin disorders, prolonged rashes, paresthesia, and swelling. Genitourinary PASC affected 1,827 patients, including urinary tract infections, hematuria, renal disorders and sometimes failure, and other genitourinary signs and symptoms. Similarly, endocrine sequelae affected 1,826 patients, including problems with lipid, glucose (i.e., diabetes), and thyroid hormone (i.e., hypothyroidism) regulations. Mental health sequelae, including anxiety, fear-related disorders, and depression, affected 1,653 patients. Musculoskeletal PASC, including spondylopathy or arthropathy, was observed in 1,639 patients. More rarely but still notably, around 1% of long-haulers experienced visual or hearing loss and gynecological or pelvic PASC.

After correcting demographics and comorbidities (Figure 5), we found that women have significantly higher odds of developing systemic, GI, neurological, skin, mental-health-related, musculoskeletal, and gynecological PASC than men. Odds of respiratory, renal/genitourinary (GU), endocrine, and eyes and ears PASC were not significantly different at p value <0.05 (odds ratios [ORs] in Table S4). We also found statistically significant reduced odds among Hispanic patients in 50% of the organ categories. Asian patients were more likely to develop endocrine PASC and less likely to have cardiovascular and systemic PASC than White patients. The increased odds of PASC among Black patients were statistically significant in skin, renal/GU, neurologic, cardiovascular, and systemic organs. Gynecological, skin, and psychiatric PASC had considerably higher odds among patients under 45 of age, whereas PASC problems related to eyes and ears, endocrine musculoskeletal, and renal/GU systems were more likely among older patients with COVID-19.

Most sequelae emerge within the first 3 months following COVID-19 infection and persist for 2 months or longer (Figure 6). This temporal pattern is more pronounced in respiratory, cardiovascular, neurology, general, and GI problems. Specific sequelae, such as respiratory or renal failure, are more likely to occur right after the acute phase of the infection. In contrast, weight gain, diabetic nephropathy, or hypertensive heart disease tend to occur later on. Some PASC, such as autoimmune disorders, seem to be bimodal, increasing in frequency both immediately after and 4 months after COVID infection. Other sequelae seem to have a steadily uniform frequency in occurrence over the first 6 months and then decrease afterward, such as pain syndromes or mental health signs and symptoms.

DISCUSSION

We presented a precision phenotyping algorithm for identifying patients with PASC. Despite significant efforts to characterize and assess the risks associated with PASC, the development of a scalable and universally accepted algorithm for identifying affected patients remains a challenge.

The conventional approach in long COVID studies, which depends on the new recording of a diagnosis code following COVID-19 infection, falls short in reliably identifying patients with PASCs. Addressing these challenges is crucial for the development of a robust algorithm capable of accurately characterizing PASC with real-world data. Standard risk studies identifying health outcomes with higher relative risks among patients with COVID-19 are insufficient for identifying long-haulers. PASC can be explained by a prior comorbidity and/or procedure in some patients. Relying on the WHO’s definition, our approach takes the prevalent PASC risk studies to the next level by adding an extra step: diagnosis of exclusion. From all patients infected with SARS-CoV-2, our algorithm distinguishes those with PASC that were not attributable to another documented pre-existing condition. For instance, recorded cases of shortness of breath in individuals with a COVID-19 history may not necessarily signify PASC, as these symptoms could be attributed to pre-existing conditions like heart failure or asthma.

Another method used for identifying patients with PASC is through using diagnostic codes, such as ICD-10 code U09.9 (PCC) in the US, but these lack the requisite sensitivity and specificity, leading to potential biases and inaccuracies. Compared to relying solely on the U09.9 diagnosis code to identify patients with long COVID, our method boasts superior precision, accurately gauges the prevalence of PASC without underestimating long COVID, and exhibits less bias across demographic groups.

The ICD-10 code currently in use for patients with PASC, U09.9, is not precise enough in identifying affected patients. Zhang et al. showed that relying on U09.9 as a stand-in for PASC can be unreliable, with its PPV/precision fluctuating between 40% and 65%, depending on the PASC reference definition.[14] Our chart reviews of nearly 900 patients indicated a PPV of 77.8% for the U09.9 diagnosis code. As such, our precision PASC phenotyping algorithm provided a 2.7% improvement in precision (79.9%). This demonstrates that the accuracy of our algorithm is at least equal to the only available diagnosis code that is supposed to be used in clinical care to identify patients with PASCs.

Our algorithm also outperformed the U09.9 in providing a more accurate estimate of long COVID prevalence, suggesting that the latter may be falling short in capturing the true scope of the condition. Based on estimates from the National Center for Health Statistics[23] derived from data spanning June 1,2022, to October 2,2023, the prevalence of PASC in Massachusetts is 24.0%. Our algorithm indicated a raw estimate of 28.5%. With a PPV/precision of 79.9%, our adjusted prevalence estimate is 22.8% (28.5% * 79.9%). Meanwhile, in the study period, 6,340 patients had a record of U09.9 in the MGB clinical data repository, which, considering the 77.8% precision, would lead to an adjusted estimated number of under 5,000 patients with PASC.

In general, bias is a significant issue in EHR diagnosis codes. It has been reported that the demographic composition of patients diagnosed with U09.9 leans toward females and White, non-Hispanic patients.[11] We found similar biases: 80.1% of patients with a U09.9 diagnosis code were White, 4.83% were Black or African American, 2.7% were Hispanic, and 67.5% were female. According to the US census, 69.6% of the Massachusetts population is White, 51% is female, 9.5% is Black or African American, and 13.1% is Hispanic.[24] Our algorithm also provides a more unbiased distribution of patients with PASC across race, gender, and ethnicity compared to the U09.9 diagnosis code. 71.4% of long-haulers identified by our precision phenotyping algorithm were White, 10.4% were Black or African American, 6.6% were Hispanic, and 64.5% were female.

We also provided an in-depth analysis outlining the clinical attributes, encompassing identified lingering effects by organ, comorbidity profiles, and temporal differences in the risk of PASC. Of the 24,360 long-haulers in our cases, less than half had multiple PASC, which could be in the same infection episode or subsequent infections. The approach to identifying cohorts with PASC offers the highest precision to date. For example, we identified PASC linked with different episodes of COVID-19 infections. We found that having a prior PASC increases the chances of having more PASC in subsequent infections. This could be due to long-haulers’ inability to mount an appropriate, timely immune response to clear the COVID-19 infection each time, leading to greater susceptibility to developing PASC in subsequent infections.

Most of the PASC we identified in this study (systemic symptoms including malaise and fatigue, sleep problems) have also been reported by prior research. However, our algorithm also allowed us to go a step further by identifying rarer PASC such as vision or hearing loss, loss of libido resulting in sexual dysfunction, gynecological complications, and diabetic complications in various organs, such as diabetic nephropathy, neuropathy, or vasculopathy. Further, our estimates are more realistic, as we exclude sequelae that can be explained at the patient level. For example, we find that only 1% of our COVID-19 cases suffered from long-term malaise and fatigue that can be attributed to a SARS-CoV-2 infection episode compared to much higher rates reported in other studies.[25,26]

Our precision phenotyping enabled us to discover statistically significant differences in the odds of developing PASC among racial and ethnic groups and across organ systems, which are not well studied. For example, gynecological, skin, and psychiatric PASC had considerably higher odds among patients under 45, whereas PASC problems related to eyes and ears, endocrine musculoskeletal, and renal/GU systems were more likely among older patients with COVID-19. Women had higher odds of PASC in 8 organs than men. Asian (vs. White) and Hispanic (vs. non-Hispanic) patients had lower odds of developing PASC (mainly in the cardiovascular system), and Black (vs. White) patients had greater odds of PASC regardless of comorbidity and age. Black patients' odds were statistically higher primarily in the skin, renal/GU, neurologic, cardiovascular, and systemic organs.

With the programs and data we offer alongside this publication, this cohort can now be curated in any healthcare system with reliable longitudinal diagnosis and procedure data on their patients. Access to large cohorts of long-haulers enriched with the precision offered by this algorithm will offer unprecedented opportunities to stratify patients who are at risk for post-COVID sequelae, identify genetic risk factors for PASC, study the possible impacts of therapeutics and immunizations, and diversify recruitment for clinical studies on PASC.

Compared to the conventional U09.9 diagnosis code, our method for identifying PASC boasts superior precision and exhibits less bias, accurately gauging the prevalence of this condition without downplaying its significance, offering a more nuanced understanding of patients with long COVID. The comprehensive PASC cohort resulting from our precision phenotyping algorithm will enable deep dives into the multifaceted expressions of long COVID through genetic, metabolomic, and clinical inquiries bolstered by robust statistical prowess, which surpasses the constraints of earlier PASC cohort studies due to limited size and outcome data.

The National Academies of Sciences, Engineering, and Medicine (NASEM) has recently proposed a broad definition that aims to be inclusive by characterizing long COVID as an IACC with a continuous, relapsing and remitting, or progressive disease state that affects one or more organ systems.[27] This definition is fully compatible with the attention mechanism implemented in this study, as temporal associations are used to both include and exclude conditions on the individual and population levels.

Limitations of the study

Using structured clinical data from real-world settings for studying PASC signs and symptoms may be limiting, as this information is often better documented in clinical notes. As we demonstrated in this study, structured diagnosis codes capture an array of signs and symptoms. We picked the CCSR categories to work with a manageable and clinically meaningful grouping of conditions, signs, and symptoms. This allowed us to reduce the computational costs of running our algorithms and facilitate implementation in diverse settings. We traced the CCSR categories to the data entries for further clinical interpretations and analyses.

Our reliance on structured data may have resulted in an underestimation of PASC. However, it has been shown that the transitive sequential patterns of the events stored in clinical data can elevate signal detection from structured data, compensating for the possible loss of information.[28] Future studies can incorporate timestamped signs and symptoms from clinical notes.

Another limitation of this study is that we did not capture the possible worsening of a prior condition, which could be characterized as a PASC. For example, COVID-19 could lead to prolonged COPD exacerbation; however, if the patient had prior episodes of COPD exacerbations before the COVID infection, then this likely has been removed per our diagnosis of exclusion. Identification of such possible PASC will be complex and require the inclusion of information on the severity of records over time. As a potential solution to this limitation, future research could explore the application of our algorithm to various sub-cohorts of patients with specific chronic conditions (e.g., type II diabetes, COPD). This approach could help determine the consistency of PASC manifestations and the specific types of PASC across these different patient populations. Finally, we separated COVID-19 variants using the date of infection rather than genetic data; thus, interpretations of the variant analysis should be cautiously approached.

Implementing the algorithm to curate similar cohorts is contingent on the availability of a true or estimated date of infection, presenting a limitation due to the declining frequency of COVID-19 testing. To address this limitation, future research could expand on this work by developing precision definitions for PASC phenotypes. These definitions could then be applied retrospectively to create "postdiction"[21] algorithms for identifying individuals who may have had SARS-CoV-2 infections in the past. Additionally, future studies should focus on identifying common temporal patterns of utilization and illness in patients with PASC. These patterns can be tested in a larger cohort of patients without a known COVID-19 infection, followed by chart reviews to determine if these patients also exhibit PASC.

The attention mechanism we developed in this study relied on the bootstrap tuning approach for identifying thresholds for exclusion by temporal association. Our tuning approach only relied on optimizing PPVs, as we could only validate cases with PASC. Future work can further expand this concept by evaluating the generalizability of exclusion thresholds for different concepts and temporal windows, increasing the validation samples, and incorporating additional validation criteria, such as negative predictive values.

STAR★METHODS

EXPERIMENTAL MODEL AND STUDY PARTICIPANT DETAILS

We conducted a retrospective case-control study that included (1) patients with a clinical record indicating COVID-19 infection between 03/2020 and 06/2023 (cases) and (2) a non-COVID control group from the pandemic era, and (3) a viral infection control group from the pre-pandemic era (Figure 1). Cases were patients with at least one confirmed SARS-CoV-2 infection, captured by a positive Polymerase Chain Reaction (PCR) test or a recorded diagnosis. For controls, we identified two distinct groups. The first, termed post-pandemic (post-2020) group, comprised patients with neither a record of SARS-CoV-2 infection nor any indication of probable infection. Concurrently, pre-pandemic controls included patients with a viral infection in 2018 (see Table S1 for the inclusion/exclusion criteria). We mandated at least a year of follow-up data from the last infection record for the cases and pre-pandemic controls. For cases and both controls, we extracted historical data from 3 years prior to the index date (i.e., 2017 for the pandemic cohorts and 2015 for the pre-pandemic controls).

Data—Use of patient data in this study was approved by the Mass General Brigham Institutional Review Board (protocol 2020P001063). We utilized electronic health record

(EHR) data from 14 hospitals and 20 community health centers within the Mass General Brigham (MGB) integrated healthcare system in Massachusetts.

Participants information on sex, age, race and ethnicity was self-reported. Information on gender and socioeconomic status was not collected. We selected post-pandemic controls by propensity score matching them (1:2) with cases on sex, race, age, and Charlson comorbidity index scores. We did not match the pre-pandemic viral controls because the number of patients in that group was smaller than the cases.

We extracted clinical data from the MGB Research Patient Data Repository.[29] To define our cohorts (i.e., cases and controls), we used the cohort discovery platform from the Evolve to Next-Gen ACT (ENACT) Network,[30-33] which provides a harmonized ontology within the i2b2 clinical research platform.[34,35] We mapped the ENACT ontology to the Clinical Classifications Software Refined (CCSR),[36] developed as part of the Healthcare Cost and Utilization Project (HCUP), aggregating International Classification of Diseases, 10th Revision, Clinical Modification/Procedure Coding System (ICD-10-CM/PCS) codes into clinically meaningful categories. To ensure utility for PASC research, we made minor adaptations to the CCSR categories based on clinical expertise. A list of the CCSR categories and corresponding ICD-10-CM/PCS codes is publicly available.[37]

We utilized the loyalty score metric from the ENACT network to enrich each cohort for continuity. The loyalty score[38,39] uses proxies of routine care and healthcare utilization metrics to compute a per-patient score for data completeness. By selecting patients with minimal missing data, we enriched our research cohorts, ensuring data completeness necessary for accurate modeling.[39] As a result, only adult patients (>18 years of age) were included in the study.

We identified SARS-CoV-2 infections from medical records indicating either a positive Polymerase Chain Reaction (PCR) test or an ICD-10 diagnosis code. A single patient may have multiple records for one infection episode. Following the methods of Azhir et al. (2023) and Strasser et al. (2022)[40,41] we organized infection into temporal clusters, where records within 90 days from the first infection were grouped as one event to distinguish episodes of SARS-CoV-2 infections for each patient.

METHOD DETAILS

PASC definition—We developed a computational implementation of the WHO definition of PASC[16,17] using a sequential pattern mining approach our team had previously developed for mining temporal representations from clinical data. The transitive Sequential Pattern Mining algorithm, tSPM, mines temporal sequences of time-stamped events in the clinical data with a transitivity property.[42]

We implemented the WHO definition using a high-performance program for transitive temporal representation mining, tSPM+,[22] illustrated in Figure 2 (in paper). tSPM+ mines transitive sequences of medical records that begin or end with a pre-specified set of clinical concepts (i.e., CCSR categories) and are not sparse (with a given sparsity parameter). For each transitive sequence $a \rightarrow b$, tSPM+ also computes the temporal duration, d_{ab}, between the two elements of the sequence $\{a,b\}$, measured in months.

I - Identifying candidate PASC, *J*—We began by identifying a subset of the CCSR categories that meet WHO’s first two criteria – i.e., continuation or development of new symptoms after initial recovery from an acute SARS-CoV-2 episode, with these symptoms lasting for at least 2 months. Drawing exclusively from the case data, we mined the tSPM+ sequences that began with a COVID-19 infection, $C \in \{c_1,\ldots,c_5\}$, where c_i is the i th COVID-19 infection episode. We considered five infections as the maximum number of infections possible for a patient within our cases (the selection of this parameter was informed by our data). A CCSR concept $J \in \{j_1,\ldots,j_n\}$ would qualify as a candidate for PASC if it were 1) not sparse, 2) a diagnosis, sign, or symptom (clinically determined), 3) featured in more than one transitive sequence of $C \rightarrow J$ (i.e., J appeared in the clinical records more than once ensuing an episode of COVID-19 infection), 4) continuous for more than or equal to 2 months (i.e., the delta duration between the first $C \rightarrow J$ and last $C \rightarrow J$, during the 12-month observation window after the given infection C), and 5) observed in more than 0.1 percent of cases. Upon satisfying these criteria, J was selected for the next consideration step as a PASC.

Figure 2 illustrates the implementation of WHO’s PASC definition as a diagnosis of exclusion.

II- attention mechanism—To prepare for computational implementation of the diagnosis of exclusion, we computed temporal associations for the previously identified J using the data from the entire study population29,30 (encompassing cases and the two control groups), using the Spearman rank-order correlations, rho (ρ), within a temporal bucketing scheme. tSPM+ mined all transitive sequences that ended with the vector of candidate J s. We computed $\rho_{XJ} \mid d_{XJ}$ across four temporal buckets of 1) 0–14 days (to capture the immediate acute phase), 2) 15–30 days, 3) 30–90 days, and 4) >90 days. Spearman’s rank-order correlation is a nonparametric measure of the strength and direction of the monotonic relationship between two variables measured at ordinal, interval, or ratio levels.[43,44] In addition to being nonparametric, the Spearman correlations are suitable for this work as they can measure monotonic relationships where the value of variables can move in the same relative direction, but not necessarily at a constant rate, which can capture both linear and more complex associations.[45,46] In addition to the correlation coefficients (rho), we computed p-values for statistical significance, applying the Holm–Bonferroni[47] adjustment for multiple comparisons.

Given that the attention mechanism was applied to both cases and controls, this phase mirrors the endpoint of other PASC studies: isolating clusters of signs, symptoms, and problems statistically associated with a SARS-CoV-2 infection.

III - Diagnosis of exclusion—We developed a computational model to operationalize the diagnosis of the exclusion concept using the tSPM+ program in conjunction with the temporal attention vector. Initially, we used the temporal correlations to identify the J s associated with COVID-19 infection. We used the adjusted correlation p-value <0.05 for this step. We then applied tSPM+ to the clinical data from cases to mine transitive sequences ending with the subset of Js or a COVID-19 infection episode. Using the mined non-sparse sequences, we excluded a given J if it was explainable by another concept X (i.e., a CCSR category) recorded prior (i.e., $x \rightarrow j$), based on the temporal association stored in the attention mechanism, which considered the duration between x and j. For example, we excluded shortness of breath (SOB) after a given episode of COVID-19 infection if it had a statistically significant correlation with a correlation coefficient greater than a specific threshold (to be defined in step V) with another record (e.g., Asthma) from the patient's records prior to the given record of SOB. It is important to note that the attention mechanism was obtained from the entire dataset, including cases and both control groups.

To implement this exclusion by sequential association, we started from the first episode of COVID-19 infection for each patient and removed sequences of $C \rightarrow J$, when J could be explained by a prior x. However, identifying a cut-off threshold for correlation coefficient could be challenging as they can be condition-specific (dependent on both x, which could be a prior chronic/acute condition, a procedure, or an abnormal lab result, and J that can be an acute or chronic problem). To address this complexity, we implemented the diagnosis of exclusion in a set of experiments at different cut-off thresholds for correlation coefficients ranging from 0.3 to 0.9, applied in addition to statistical significance at adjusted $p < 0.05$.

The output of the diagnosis of exclusion included, for each patient that had at least one PASC, the COVID-19 infection episode, infection date, the PASC, J, the number of months after the infection when J was recorded, and the correlation coefficient for that condition to not be excluded. We then mapped the selected Js (i.e., CCSR categories) to the associated clinical concept recorded for the patient and grouped them into clinically meaningful categories and by organ for reporting.

Chart reviews—We performed an extensive chart review to explore the unstructured data in the clinical notes and to label patients who truly have PASC. The inclusion criteria for chart reviews included patients with at least an ICD-10 code U09.9 withing the study period. Eight clinical faculty developed chart review guidelines, and two clinical nurses trained in this specific chart review process reviewed the charts between April and November 2023 (see Document S2 Method S1 for more details).

Validation and tuning—To validate the results and estimate the cut-off threshold for the exclusion by temporal association, we leveraged the 309 chart-reviewed patients with confirmed PASC. We computed positive predictive values (PPV) via bootstrapping.

Maximizing PPV: In a 100 random sampling with replacement, we: (1) split the chart reviews into tuning and validation sets at a 50-50 ratio, (2) for each j, estimate the optimized correlation coefficient's cut-off threshold that maximized PPV on the tuning dataset, (3) estimate the overall PPV in the output file with the optimized threshold against the held-out validation set. This step resulted in removing records J with multiple correlation coefficients from the long-haulers list.

Minimizing type I error: Once we curated the updated list of long-haulers, to minimize false positive detection rates (type I error), we utilized clinical expertise to identify a subset of the J from the long-haulers list that may not contribute to the PPV. This may be possible in patients with multiple PASCs. We removed each of the suspected conditions from the long-haulers list. We recalculated the positive predictive value in the overall chart review set to eliminate suspected Js that can be removed without reducing the PPV in the overall long-haulers list.

QUANTIFICATION AND STATISTICAL ANALYSIS

We computed the Spearman rank-order correlations[48] and p-values with the Holm–Bonferroni[47] adjustment for multiple comparisons to measure temporal associations rho (ρ) between clinical concepts in the four temporal buckets. We fitted the logistic regression models with a binomial logit link function to separately evaluate the development of organ-specific PASC, using age, sex, race, ethnicity, and Charlson's comorbidity score as predictors. A significance level of 0.05 was used to evaluate statistical significance.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

ACKNOWLEDGMENTS

This study has been supported by grants from the National Institutes of Health, National Institute of Allergy and Infectious Diseases (NIAID) R01AI165535, National Heart, Lung, and Blood Institute (NHLBI) OT2HL161847, and National Center for Advancing Translational Sciences (NCATS) UL1 TR003167, UL1 TR001881, and U24TR004111. J.H.'s work was partially funded by a fellowship within the IFI programme of the German Academic Exchange Service (DAAD) and by the Federal Ministry of Education and Research (BMBF), as well by the German Research Foundation (426671079).

REFERENCES

1. HHS (2022). Long COVID terms and definitions development explained (COVID.gov). https://www.covid.gov/longcovid/definitions.
2. HHS (2022). National Research Action Plan on Long COVID (Department of Health and Human Services, Office of the Assistant Secretary for Health).
3. Raveendran AV, Jayadevan R, and Sashidharan S (2021). Long COVID: An overview. Diabetes Metabol. Syndr 15, 869–875.
4. Crook H, Raza S, Nowell J, Young M, and Edison P (2021). Long covid—mechanisms, risk factors, and management. BMJ 374, n1648. 10.1136/bmj.n1648. [PubMed: 34312178]
5. Smallwood M. (2023). The Future of Long COVID: A Threatcasting Approach (Springer Nature).
6. Medinger G, and Altmann D (2022). The Long Covid Handbook (Random House).
7. Dagliati A, Strasser ZH, Hossein Abad ZS, Klann JG, Wagholikar KB, Mesa R, Visweswaran S, Morris M, Luo Y, Henderson DW, et al. (2023). Characterization of long COVID temporal sub-phenotypes by distributed representation learning from electronic health record data: a cohort study. eClinicalMedicine 64, 102210. 10.1016/j.eclinm.2023.102210. [PubMed: 37745021]
8. Subramanian A, Nirantharakumar K, Hughes S, Myles P, Williams T, Gokhale KM, Taverner T, Chandan JS, Brown K, Simms-Williams N, et al. (2022). Symptoms and risk factors for long COVID in non-hospitalized adults. Nat. Med 28, 1706–1714. [PubMed: 35879616]
9. Davis HE, McCorkell L, Vogel JM, and Topol EJ (2023). Long COVID: major findings, mechanisms and recommendations. Nat. Rev. Microbiol 21, 133–146. [PubMed: 36639608]

10. O'Hare AM, Vig EK, Iwashyna TJ, Fox A,Taylor JS, Viglianti EM, Butler CR, Vranas KC, Helfand M, Tuepker A, et al. (2022). Complexity and Challenges of the Clinical Diagnosis and Management of Long COVID. JAMA Netw. Open 5, e2240332. [PubMed: 36326761]
11. Pfaff ER, Madlock-Brown C, Baratta JM, Bhatia A, Davis H, Girvin A, Hill E, Kelly E, Kostka K, Loomba J, et al. (2023). Coding long COVID: characterizing a new disease through an ICD-10 lens. BMC Med. 21, 58. [PubMed: 36793086]
12. Duerlund LS, Shakar S, Nielsen H, and Bodilsen J (2022). Positive Predictive Value of the ICD-10 Diagnosis Codefor Long-COVID. Clin. Epidemiol 14, 141–148. [PubMed: 35177935]
13. Ioannou GN, Baraff A, Fox A, Shahoumian T, Hickok A, O'Hare AM, Bohnert ASB, Boyko EJ, Maciejewski ML, Bowling CB, et al. (2022). Rates and Factors Associated With Documentation of Diagnostic Codes for Long COVID in the National Veterans Affairs Health Care System. JAMA Netw. Open 5, e2224359. [PubMed: 35904783]
14. Zhang HG, Honerlaw JP, Maripuri M, Samayamuthu MJ, Beaulieu-Jones BR, Baig HS, L'Yi S, Ho Y-L, Morris M, Panickan VA, et al. (2023). Potential pitfalls in the use of real-world data for studying long COVID. Nat. Med 29, 1040–1043. [PubMed: 37055567]
15. Wirth KJ, and Scheibenbogen C (2022). Dyspnea in Post-COVID Syndrome following Mild Acute COVID-19 Infections: Potential Causes and Consequences for a Therapeutic Approach. Medicina 58, 419. 10.3390/medicina58030419. [PubMed: 35334595]
16. WHO (2023). Coronavirus disease (COVID-19): Post COVID-19 condition. https://www.who.int/news-room/questions-and-answers/item/coronavirus-disease-(covid-19)-post-covid-19-condition.
17. Soriano JB, Murthy S, Marshall JC, Relan P, and Diaz JV; WHO Clinical Case Definition Working Group on Post-COVID-19 Condition (2022). A clinical case definition of post-COVID-19 condition by a Delphi consensus. Lancet Infect. Dis 22, e102–e107. [PubMed: 34951953]
18. Bowe B, Xie Y, and Al-Aly Z (2023). Postacute sequelae of COVID-19 at 2 years. Nat. Med 29, 2347–2357. [PubMed: 37605079]
19. Xie Y, Xu E, Bowe B, and Al-Aly Z (2022). Long-term cardiovascular outcomes of COVID-19. Nat. Med 28, 583–590. [PubMed: 35132265]
20. Al-Aly Z, Xie Y, and Bowe B (2021). High-dimensional characterization of post-acute sequelae of COVID-19. Nature 594, 259–264. [PubMed: 33887749]
21. Estiri H, Strasser ZH, Brat GA, Semenov YR, Consortium for Characterization of COVID-19 by EHR 4CE; Patel CJ, and Murphy SN (2021). Evolving phenotypes of non-hospitalized patients that indicate long COVID. BMC Med. 19, 249. [PubMed: 34565368]
22. Hügel J, Sax U, Murphy SN, and Estiri H (2023). tSPM+; a high-performance algorithm for mining transitive sequential patterns from clinical data. Preprint at arXiv [cs.LG] 888, 888. 10.48550/arXiv.2309.05671.
23. National Center for Health Statistics. U.S. Census Bureau (2023). Long COVID. Household Pulse Survey. https://www.cdc.gov/nchs/covid19/pulse/long-covid.htm.
24. United States Census Bureau > Communications Directorate - Center for New Media(2023). QuickFacts: Massachusetts. https://www.census.gov/quickfacts/fact/table/MA.
25. Ceban F, Ling S, Lui LMW, Lee Y, Gill H, Teopiz KM, Rodrigues NB, Subramaniapillai M, Di Vincenzo JD, Cao B, et al. (2022). Fatigue and cognitive impairment in Post-COVID-19 Syndrome: A systematic review and meta-analysis. Brain Behav. Immun 101, 93–135. [PubMed: 34973396]
26. Jason LA, and Dorri JA (2022). ME/CFS and Post-Exertional Malaise among Patients with Long COVID. Neurol. Int 15, 1–11. [PubMed: 36648965]
27. National Academies of Sciences, Engineering, and Medicine (2024). In A Long COVID Definition: A Chronic, Systemic Disease State with Profound Consequences Fineberg H,V, Brown L, Worku T, and Goldowitz I, eds. (The National Academies Press).
28. Estiri H, Strasser ZH, and Murphy SN (2021). High-throughput phenotyping with temporal sequences. J. Am. Med. Inf. Assoc 28, 772–781.
29. Nalichowski R, Keogh D, Chueh HC, and Murphy SN (2006). Calculating the Benefits of a Research Patient Data Repository. AMIA Annu. Symp. Proc 2006, 1044. [PubMed: 17238663]
30. Visweswaran S, Becich MJ, D'Itri VS, Sendro ER, MacFadden D, Anderson NR, Allen KA, Ranganathan D, Murphy SN, Morrato EH, et al. (2018). Accrual to Clinical Trials (ACT):

A Clinical and Translational Science Award Consortium Network. JAMIA Open 1, 147–152. [PubMed: 30474072]

31. Visweswaran S, Samayamuthu MJ, Morris M, Weber GM, MacFadden D, Trevvett P, Klann JG, Gainer VS, Benoit B, Murphy SN, et al. (2021). Development of a Coronavirus Disease 2019 (COVID-19) Application Ontology for the Accrual to Clinical Trials (ACT) network. JAMIA Open 4, ooab036. [PubMed: 34113801]
32. Lenert LA, Zhu V, Jennings L, McCauley JL, Obeid JS, Ward R, Hassanpour S, Marsch LA, Hogarth M, Shipman P, et al. (2022). Enhancing research data infrastructure to address the opioid epidemic: the Opioid Overdose Network (O2-Net). JAMIA Open 5, ooac055. [PubMed: 35783072]
33. Morrato EH, Lennox LA, Sendro ER, Schuster AL, Pincus HA, Humensky J, Firestein GS, Nadler LM, Toto R, and Reis SE (2020). Scale-up of the Accrual to Clinical Trials (ACT) network across the Clinical and Translational Science Award Consortium: a mixed-methods evaluation of the first 18 months. J. Clin. Transl. Sci 4, 515–528. [PubMed: 33948228]
34. Murphy SN, Weber G, Mendis M, Gainer V, Chueh HC, Churchill S, and Kohane I (2010). Serving the enterprise and beyond with informatics for integrating biology and the bedside (i2b2). J. Am. Med. Inf. Assoc 17, 124–130.
35. Murphy SN, Mendis ME, Berkowitz DA, Kohane I, and Chueh HC (2006). Integration of clinical and genetic data in the i2b2 architecture. AMIA Annu. Symp, 1040 (Proc.).
36. Healthcare Cost and Utilization Project (HCUP). (2024). Clinical Classifications Software Refined (CCSR). www.hcup-us.ahrq.gov/toolssoftware/ccsr/ccs_refined.jsp.
37. CLAI (2023). CCSR categories and corresponding ICD-10-CM/PCS codes. (Github).
38. Lin KJ, Singer DE, Glynn RJ, Murphy SN, Lii J, and Schneeweiss S (2018). Identifying Patients With High Data Completeness to Improve Validity of Comparative Effectiveness Research in Electronic Health Records Data. Clin. Pharmacol. Ther 103, 899–905. [PubMed: 28865143]
39. Klann JG, Henderson DW, Morris M, Estiri H, Weber GM, Visweswaran S, and Murphy SN (2023). A broadly applicable approach to enrich electronic-health-record cohorts by identifying patients with complete data: a multisite evaluation. J. Am. Med. Inf. Assoc 30, 1985–1994. 10.1093/jamia/ocad166.
40. Azhir A, Strasser ZH, Murphy SN, and Estiri H (2023). Severity of COVID-19-Related Illness in Massachusetts, July 2021 to December 2022. JAMA Netw. Open 6, e238203. [PubMed: 37052921]
41. Strasser ZH, Greifer N, Hadavand A, Murphy SN, and Estiri H (2022). Estimates of SARS-CoV-2 Omicron BA.2 Subvariant Severity in New England. JAMA Netw. Open 5, e2238354. [PubMed: 36282501]
42. Estiri H, Vasey S, and Murphy SN (2020).Transitive Sequential Pattern Mining for Discrete Clinical Data. In Artificial Intelligence in Medicine (Springer International Publishing)), pp. 414–424.
43. Astivia OLO, and Zumbo BD (2017). Population models and simulation methods: The case of the Spearman rank correlation. Br. J. Math. Stat. Psychol 70, 347–367. [PubMed: 28140458]
44. Corder GW, and Foreman DI (2014). Nonparametric Statistics: A Step-by-step Approach (John Wiley & Sons).
45. Ramsey PH (1989). Critical values for Spearman's rank order correlation. J. Educ. Stat 14, 245–253.
46. Schober P, Boer C, and Schwarte LA (2018). Correlation Coefficients: Appropriate Use and Interpretation. Anesth. Analg 126, 1763–1768. [PubMed: 29481436]
47. Holm S. (1979). A Simple Sequentially Rejective Multiple Test Procedure. Scand. Stat. Theory Appl 6, 65–70.
48. Spearman C. (1904). The Proof and Measurement of Association between Two Things. Am. J. Psychol 15, 72–101.

Highlights

- Precision PASC phenotyping algorithm identifies long COVID with attention mechanism
- Incorporates both infection-association chronic condition and diagnosis of exclusion
- Outperforms U09.9 in precision and reduces bias in long COVID identification
- Captures rare long COVID symptoms, including vision loss and diabetic complications

CONTEXT AND SIGNIFICANCE

Identifying cohorts of patients with post-acute sequelae of COVID-19 (PASC), or long COVID, using real-world data is complex. The absence of precise definitions for PASC poses significant challenges in clinical research and patient care. Utilizing electronic health records from a large integrated healthcare system, Azhir et al. developed a precision phenotyping algorithm incorporating a novel attention mechanism that accounts for both infection-related chronic conditions and differential diagnoses. This approach demonstrated superior accuracy in identifying PASC cases compared to the existing ICD-10-CM code U09.9 while also mitigating demographic biases in diagnosis. The implications are profound, offering a refined tool for constructing research cohorts to explore the genetics and metabolomics of long COVID, thereby enhancing the health systems’ capacity to manage it.

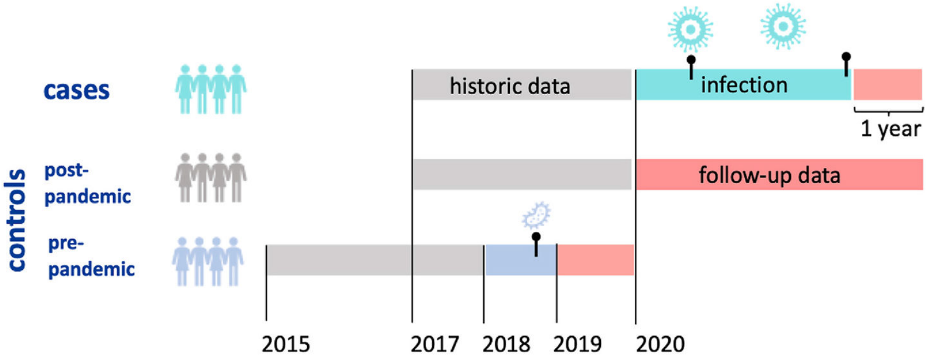


Figure 1. Criteria for selection of cases and controls

Cases are infected at least once by SARS-CoV-2, whereas the pre-pandemic controls had a viral infection in 2018.

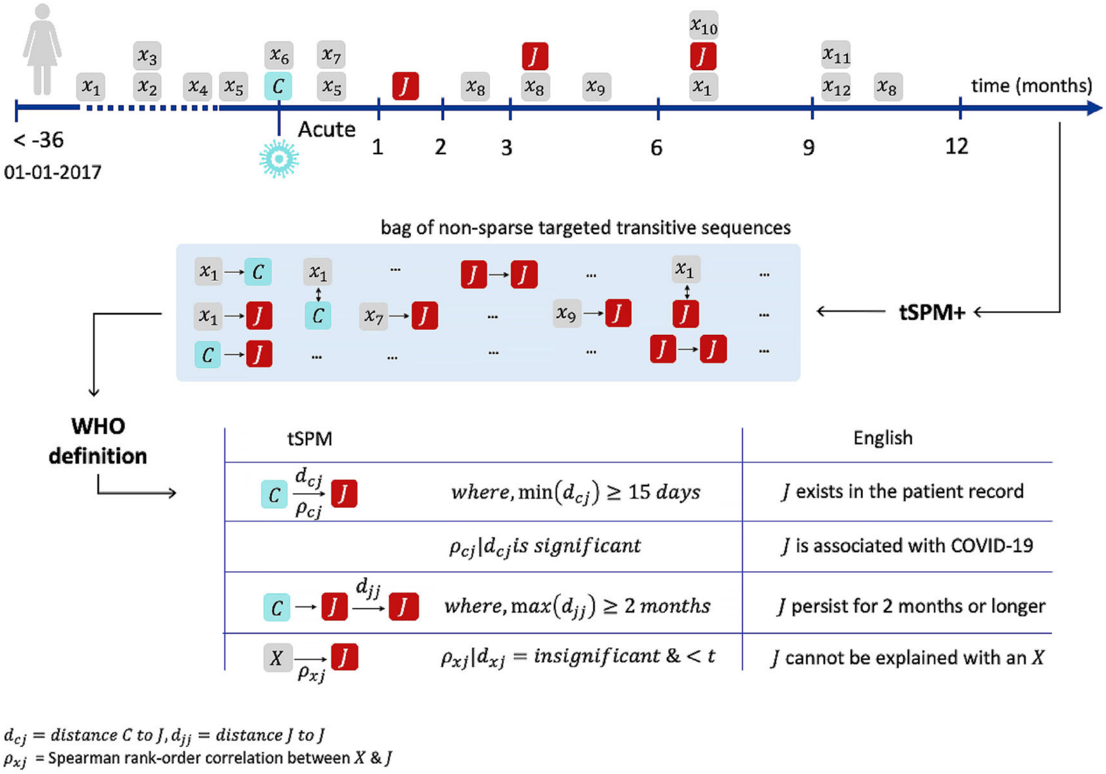

x_{10}
x_3 x_6 x_7 J J x_{11}
x_1 x_2 x_4 x_5 C x_5 J x_8 x_8 x_9 x_1 x_{12} x_8
time (months)
< -36
01-01-2017
Acute
1 2 3 6 9 12
bag of non-sparse targeted transitive sequences
$x_1 \rightarrow C$
$x_1 \rightarrow J$
$C \rightarrow J$
$x_1 \leftrightarrow C$
$x_7 \rightarrow J$
$J \rightarrow J$
$x_9 \rightarrow J$
$x_1 \leftrightarrow J$
$J \rightarrow J$
tSPM+
WHO definition
tSPM
English
$C \xrightarrow[\rho_{cj}]{d_{cj}} J$ where, $\min(d_{cj}) \geq 15$ days
J exists in the patient record
$\rho_{cj}|d_{cj}$ is significant
J is associated with COVID-19
$C \rightarrow J \xrightarrow{d_{jj}} J$ where, $\max(d_{jj}) \geq 2$ months
J persist for 2 months or longer
$X \xrightarrow[\rho_{xj}]{} J$ $\rho_{xj}|d_{xj}$ = insignificant & < t
J cannot be explained with an X
d_{cj} = distance C to J, d_{jj} = distance J to J
ρ_{xj} = Spearman rank-order correlation between X & J

Figure 2. Identifying PASC with the tSPM+ algorithm and WHO definition

After temporarily ordering clinical records, we mined transitive sequences. We then identified a subset of the CCSR categories meeting WHO's criteria for PASC: new or continued symptoms post-acute SARS-CoV-2 episode lasting at least 2 months. Using case data, we extracted tSPM+ sequences beginning with a COVID-19 infection ($C \in \{c_1, \ldots, c_5\}$), with a maximum of five infections per patient. A CCSR concept $J \in \{j_1, \ldots, j_n\}$ qualified as a PASC candidate if it was not sparse, clinically significant, appeared in more than one $C \rightarrow J$ sequence, persisted for ≥2 months, and was observed in >0.1% of cases. Satisfying these criteria, J was considered for further PASC evaluation to operationalize the diagnosis of exclusion. Initially, temporal correlations (adjusted p value < 0.05) identified Js associated with COVID-19 infection. We then applied tSPM+ to mine transitive sequences ending with Js or a COVID-19 episode. Non-sparse sequences were excluded if a given J was explainable by another concept X (e.g., a CCSR category) recorded prior ($x \rightarrow j$), based on the temporal association and duration between x and j. For example, shortness of breath (SOB) post-COVID-19 was excluded if it significantly correlated with a prior record (e.g., asthma). The attention mechanism used data from all cases and control groups.

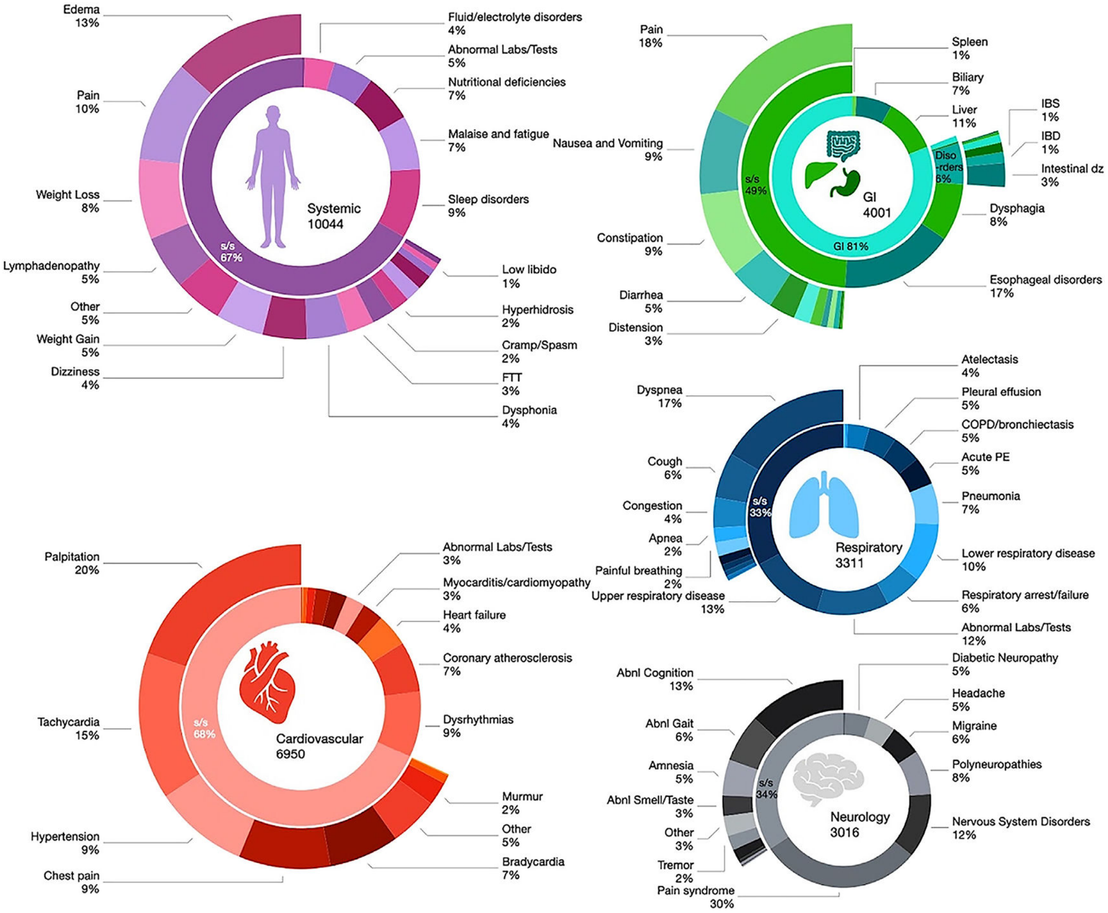


Figure 3. Distribution of PASC by organ

This figure focuses on the systemic, cardiovascular, neurology, respiratory, and gastrointestinal (GI) PASC. Underlying issues are presented as the percentage of patients afflicted in each category. s/s, signs and symptoms; Abnl, abnormal; FTT, failure to thrive; IBS, inflammatory bowel syndrome; IBD, inflammatory bowel disease; dz, disease.

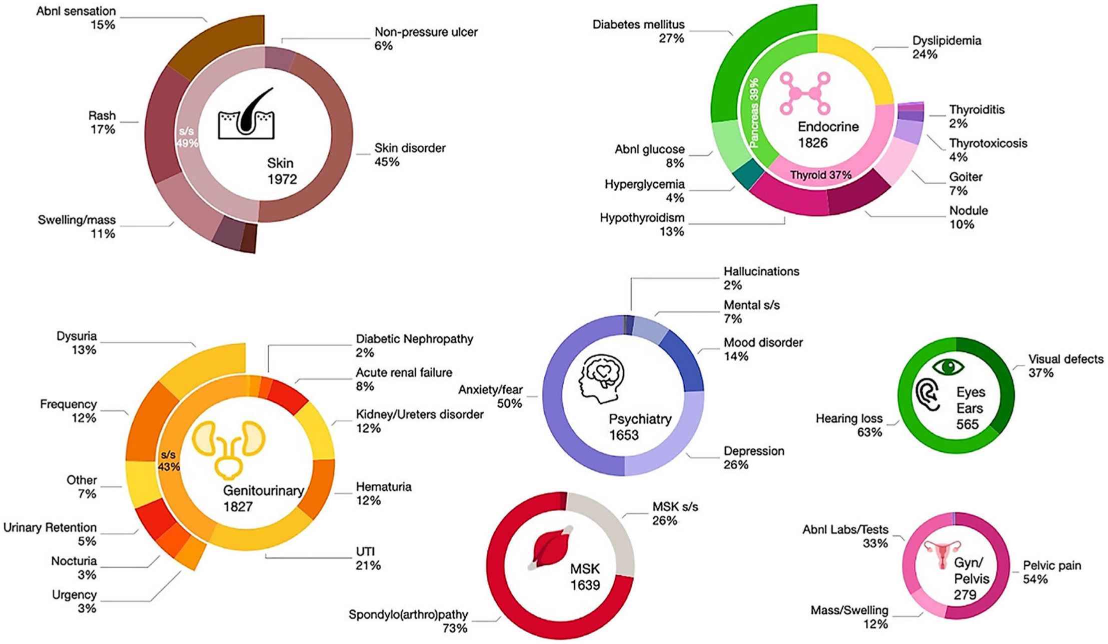


Figure 4. Distribution of PASC by organ

This figure focuses on the genitourinary, skin, endocrine, psychiatry, eyes and ears, musculoskeletal (MSK), and gynecology (GYN)/pelvis PASC. Underlying issues are presented as the percentage of patients afflicted in each category. s/s, signs and symptoms; Abnl, abnormal.

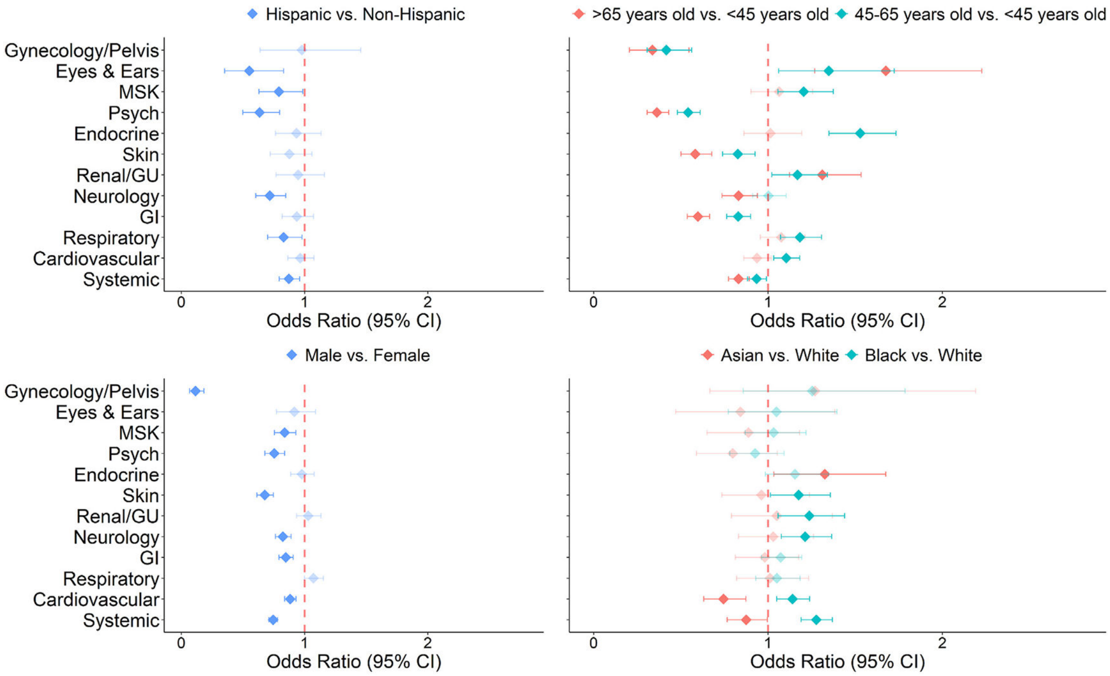


Figure 5. Odds of developing PASC by organ and demographics

95% confidence intervals of each odds ratio is demonstrated with bars around each estimate.

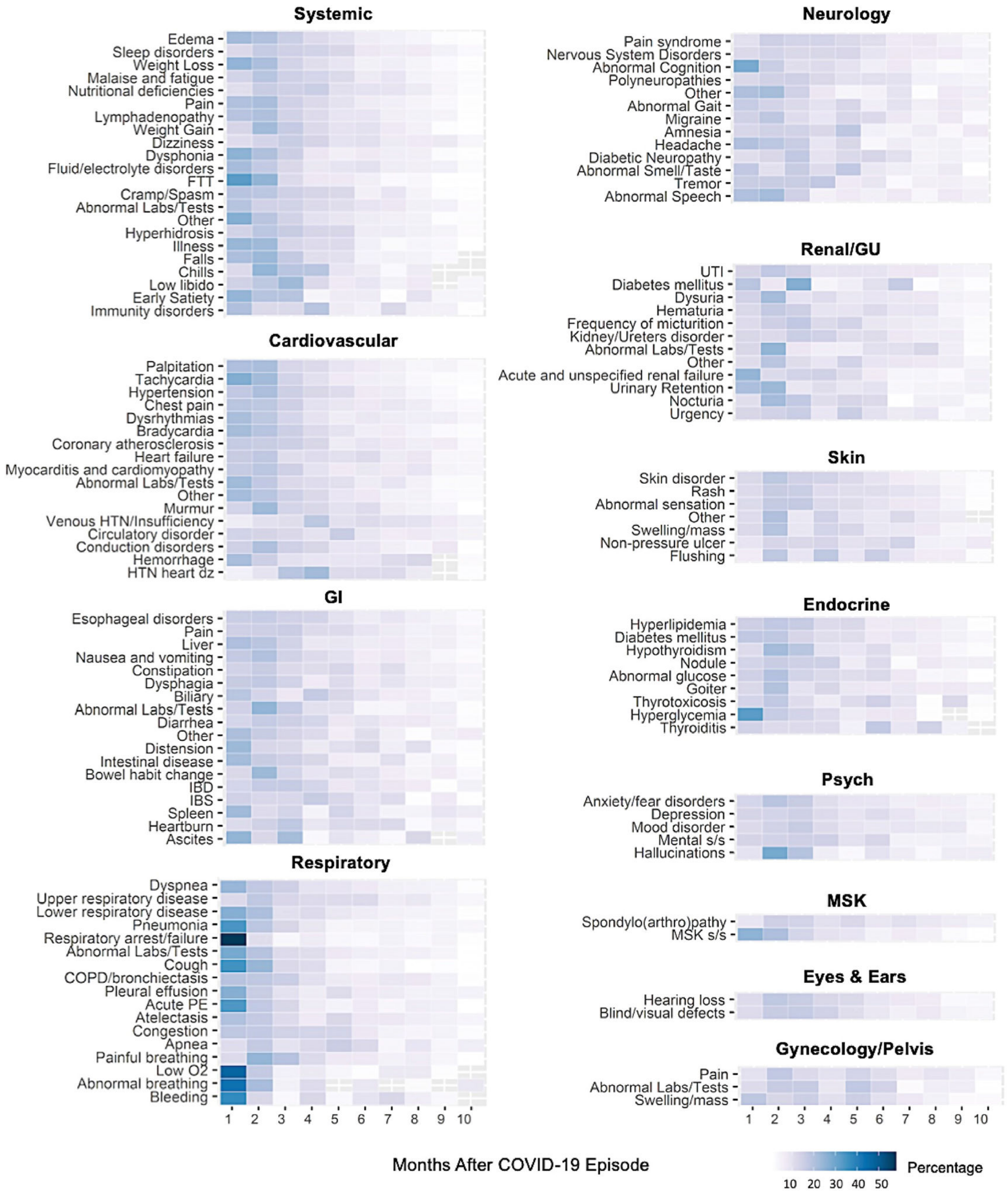


Figure 6. Temporal presentation of the PASC over time

Table 1.

Summary statistics of the study population

	COVID-19 cases	Post-pandemic controls	Pre-pandemic controls
Number of patients	85,364	170,497	39,817
Age, mean	53.6	53.7	54.7
Female, %	62.6	62.5	63.2
Charlson, mean	2.2	2.2	2.0
Date, range	January 1, 2017–June 8, 2023	January 1, 2017–June 8, 2023	January 1, 2015–January 1, 2020
Race(%)			
White	60,935 (71.4)	122,646 (71.9)	29,687 (74.2)
Other	10,243 (12.0)	19,799 (11.6)	3,415 (8.5)
Black	8,409 (9.9)	16,688 (9.8)	3,273 (8.2)
Unknown	2,845 (3.3)	5,757 (3.4)	2,185 (5.5)
Asian	2,933 (3.4)	5,607 (3.3)	1,257 (3.1)
Hispanic ethnicity	6,110 (7.2)	8,840 (5.2)	1,950 (4.9)

Key Resource Table

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Software and algorithms		
long_covid_ai_scripts	This paper	https://doi.org/10.5281/zenodo.13835061
long_covid_ai_implementation_guide	This paper	https://doi.org/10.5281/zenodo.13835071