

A machine learning-based phenotype for long COVID in children: an EHR-based study from the RECOVER program

Authors

Vitaly Lorman, PhD¹, Hanieh Razzaghi, MPH¹, Xing Song, PhD², Keith Morse, MD, MBA³, Levon Utidjian, MD¹, Andrea J. Allen, MS¹, Suchitra Rao, MBBS, MSCS⁴, Colin Rogerson, MD, MPH⁵, Tellen D. Bennett, MD, MS⁶, Hiroki Morizono, PhD⁷, Daniel Eckrich, MLIS⁸, Ravi Jhaveri, MD⁹, Yungui Huang, PhD, MBA¹⁰, Daksha Ranade, MPH, MBA¹¹, Nathan Pajor, MD, MS¹², Grace M. Lee, MD, MPH¹³, Christopher B. Forrest, MD, PhD¹, L. Charles Bailey, MD, PhD¹

Affiliations:

¹Applied Clinical Research Center, Children's Hospital of Philadelphia, Philadelphia, PA, United States

²Department of Health Management and Informatics, University of Missouri School of Medicine, Columbia, MO, United States

³ Division of Pediatric Hospital Medicine, Department of Pediatrics, Stanford University School of Medicine, Stanford, California, United States

⁴ Department of Pediatrics, University of Colorado School of Medicine and Children's Hospital of Colorado, Aurora, CO, United States

⁵ Division of Critical Care, Department of Pediatrics, Indiana University School of Medicine, Indianapolis, IN, United States

⁶ Departments of Biomedical Informatics and Pediatrics, University of Colorado School of Medicine and Children's Hospital Colorado, Aurora, CO, United States

⁷ Center for Genetic Medicine Research, Children's National Hospital, Washington DC, United States

⁸ Biomedical Research Informatics Center, Nemours Children's Health, Wilmington, DE, United States

⁹Division of Infectious Diseases, Ann & Robert H. Lurie Children's Hospital of Chicago, Chicago, IL, United States

¹⁰IT Research and Innovation, The Research Institute at Nationwide Children's Hospital, Columbus, OH, United States

¹¹ Research Informatics Department, Seattle Children's Hospital, Seattle, WA, United States

¹² Division of Pulmonary Medicine, Cincinnati Children's Hospital Medical Center and University of Cincinnati College of Medicine, Cincinnati, OH, United States

¹³ Division of Infectious Diseases, Department of Pediatrics, Stanford University School of Medicine, Stanford, CA, United States

Abstract

Background

As clinical understanding of pediatric Post-Acute Sequelae of SARS CoV-2 (PASC) develops, and hence the clinical definition evolves, it is desirable to have a method to reliably identify patients who are likely to have post-acute sequelae of SARS CoV-2 (PASC) in health systems data.

Methods and Findings

In this study, we developed and validated a machine learning algorithm to classify which patients have PASC (distinguishing between Multisystem Inflammatory Syndrome in Children (MIS-C) and non-MIS-C variants) from a cohort of patients with positive SARS- CoV-2 test results in pediatric health systems within the PEDSnet EHR network. Patient features included in the model were selected from conditions, procedures, performance of diagnostic testing, and medications using a tree-based scan statistic approach. We used an XGboost model, with hyperparameters selected through cross-validated grid search, and model performance was assessed using 5-fold cross-validation. Model predictions and feature importance were evaluated using Shapley Additive exPlanation (SHAP) values.

Conclusions

The model provides a tool for identifying patients with PASC and an approach to characterizing PASC using diagnosis, medication, laboratory, and procedure features in health systems data. Using appropriate threshold settings, the model can be used to identify PASC patients in health systems data at higher precision for inclusion in studies or at higher recall in screening for clinical trials, especially in settings where PASC diagnosis codes are used less frequently or less reliably. Analysis of how specific features contribute to the classification process may assist in gaining a better understanding of features that are associated with PASC diagnoses.

Funding Source: This research was funded by the National Institutes of Health (NIH) Agreement OT2HL161847-01 as part of the Researching COVID to Enhance Recovery (RECOVER) program of research.

Disclaimer: The content is solely the responsibility of the authors and does not necessarily represent the official views of the RECOVER Program, the NIH or other funders.

Introduction

While long-term consequences of SARS-CoV2 infection have been studied from the perspectives of both clinical manifestations and underlying mechanisms^{1,2}, formal definitions for Post-Acute Sequelae of SARS CoV-2 (PASC, or long COVID) are currently broad and necessarily nonspecific. Although advances have been made in describing features that characterize PASC, studies to date have suggested a heterogeneous presentation, particularly in children^{3,4}. This heterogeneity has made it difficult to form a definition for predicting or classifying children with PASC in the absence of patient-specific expert review, resulting in challenges in obtaining “gold standard” labels for cases and controls, which consequently poses a challenge to conducting large-scale research towards improving patient outcomes.

To address this gap, researchers have developed machine learning algorithms intended to classify adult patients with PASC in large clinical databases⁵. In contrast with explicit rule-based definitions, machine learning algorithms have the advantage of being able to detect complex patterns involving thousands of covariates⁶.

Diagnosing PASC in the pediatric population is particularly challenging due to large differences in clinical manifestations across the age spectrum, compounded by a paucity of pediatric research to date. Multisystem Inflammatory Syndrome in Children (MIS-C), a clinically severe illness which follows SARS-CoV-2 infection and satisfies the current time-based definition for PASC is considered a distinct entity in practice and does have a case definition which involves laboratory evidence of inflammation and involvement of at least two organ systems among SARS CoV-2-positive patients⁷. However, a fuller understanding of the subphenotypes of MIS-C, including the mechanisms by which they develop and their longer-term trajectories, is still emerging^{8,9}. In the case of non-MIS-C PASC, the variety of presentations, variability of methods for diagnosis, and treatment modalities presents an even greater challenge in identifying such patients from EHR data^{1-3,3,4,10,11}.

The goal of this study is to implement and validate a machine learning model to classify patients with a PASC diagnosis (including both MIS-C and non-MIS-C variants). Our model was trained to classify patients with a PASC diagnosis code in a cohort of SARS-CoV-2 positive patients, with the model’s utility found in its ability to detect patients who are likely to have had PASC based on a large collection of clinical features in settings where the diagnosis code was not reliably present. To define patient features for use in the model from large hierarchical vocabularies of clinical codes, we employed the tree-based scan statistic¹², a data mining tool previously used for pharmacovigilance, vaccine safety surveillance, and occupational disease surveillance. In our context, the tree-based scan statistic was used to detect clusters of diagnosis, medication, procedure, and diagnostic test codes which occur disproportionately often among PASC-diagnosed patients; it detects the level of threshold hierarchical granularity at which to cluster the codes, allowing for feature selection without expert clinical input regarding which features to use and how to cluster them. To our knowledge, this is a novel application of the tree-based scan statistic to selecting features from a hierarchical structure for use in machine learning models and may be of independent methodological interest.

Methods

Study population

This retrospective cohort study is part of the NIH Researching COVID to Enhance Recovery (RECOVER) Initiative, which seeks to understand, treat, and prevent PASC. For more information on RECOVER, visit

<https://recovercovid.org/>. The study population includes EHR data from the PEDSnet network and the following institutions: Children’s Hospital of Philadelphia, Cincinnati Children’s Hospital Medical Center, Children’s Hospital of Colorado, Ann & Robert H. Lurie Children’s Hospital of Chicago, Nationwide Children’s Hospital, Nemours Children’s Health System (in Delaware and Florida), Seattle Children’s Hospital, and Stanford Children’s Health. The Children’s Hospital of Philadelphia’s institutional review board designated this study as not human subjects research and the need for consent was waived. The PEDSnet RECOVER Database Version 202207 was used.

Identifying PASC diagnoses

We identified patients with MIS-C by the presence of ICD-10-CM code M35.81, ICD 10 codes U10 or U10.9, OHDSI extension code OMOP5042964, or an (EHR) interface term containing the string ‘MIS-C’ or both the strings ‘multisystem’ and ‘inflam’. We identified patients with non-MIS-C subphenotypes of PASC using the ICD-10-CM diagnosis code U09.9 (available October 1, 2021¹³) or by an interface term in the EHR containing the string ‘long covid’ or ‘post-acute’ and ‘covid’ or ‘sars’. Based on evidence¹⁴ and early CDC guidance¹⁵ that the non-COVID-specific ICD-10-CM code B94.8 (“Sequelae of other specified infectious and parasitic diseases”) was used as a placeholder for long COVID prior to the introduction of U09.9, we elected to include among our cases patients who had an occurrence of this code or similar nonspecific post-viral infection codes in the SNOMED vocabulary that were not attributable to other conditions, when such codes occurred following a positive SARS CoV-2 test result.

Composite PASC diagnosis was defined as the presence of any of the above codes, including those for MIS-C. Non-MIS-C PASC diagnosed patients were defined as the subset of patients with a PASC diagnosis who did not have a MIS-C diagnosis.

Cohort definition

Our cohort comprised patients less than 21 years old who had a positive SARS CoV-2 test (antigen, serology, or Reverse Transcriptase-Polymerase Chain Reaction (RT-PCR)), or PASC or MIS-C diagnosis at any point after January 1, 2021⁷. We used this cutoff date because it is when the MIS-C diagnosis code M35.81 first came into use. Because home viral testing became increasingly prevalent in 2022, and not all PASC-diagnosed patients in our cohort had a SARS CoV-2 test result, we did not require one. Further, for many PASC (including MIS-C)-diagnosed patients who did have a positive SARS CoV-2 test result, the test occurred on or near their earliest PASC diagnosis. Due to the resulting difficulties in capturing the true date of onset of SARS-CoV-2 infection in the EHR, we elected to impute the index date for all PASC-diagnosed patients by selecting a random date in the 28 to 90 days prior to their earliest PASC diagnosis. For non-PASC SARS-CoV-2 infected patients, we defined their index date as the date of first positive test result. We further required the presence of at least two visits following the index date for all patients in our cohort.

Feature selection

Our feature set consisted of person-level demographics, conditions, performance of diagnostic testing, procedures, and medications. Demographic variables included site, index date month and year, age at index date, sex, and race/ethnicity. Diagnosis codes used to define the presence of PASC (U09.9, M35.81, B94.8) were excluded from the feature set.

The remaining variables were constructed from clusters selected from clinical vocabulary hierarchies using the tree-based scan statistic computed by the TreeScan software. In this study, we employed the unconditional analysis based on a Bernoulli probability model at each node of the tree. This approach identifies branches of the tree at which outcomes belonged most disproportionately to patients with PASC as compared to COVID-positive patients without PASC. We refer the reader to the TreeScan literature for more details^{12,16}.

We used as input into the TreeScan software the SNOMED hierarchy for conditions, the RxNorm hierarchy for medications, the LOINC hierarchy for labs, and a union of the ICD-10-PCS, HCPCS, and CPT4 hierarchies for procedures. To ensure uniform follow-up time, the cohort we used for TreeScan feature selection was a subcohort of the full cohort of our study consisting of patients who had at least 3 months of follow-up. The code occurrences used as input to TreeScan were counted during the 28-180 days following the index date.

From the TreeScan output, we then selected the branches which were 1) significant at the $p < 0.05$ threshold and 2) had at least 50 occurrences for conditions, diagnostic tests, and procedures. For medications, we required 1000 occurrences given the per-dose structure of inpatient medication records. We also omitted from consideration cuts above the 3rd level of the respective vocabularies for conditions and above the 2nd level for medications and diagnostic tests to avoid selecting clusters that were too general.

To create the feature space for our model, six binary variables were constructed for each branch selected by TreeScan to indicate whether the feature was present for each of the following time periods relative to index date: -1 to 0 months, 0 to 1 months, 1 to 2 months, 2-3 months, and 3-6 months.

Model selection and evaluation

We used an XGBoost model with hyperparameters selected using cross-validated grid search. To distinguish between MIS-C and non-MIS-C PASC, the model was trained to classify patients as belonging to one of three classes (MIS-C, non-MIS-PASC, or no PASC) and model output for each patient consisted of one probability for each class such that the three probabilities sum to 1.

We used fivefold cross-validation to evaluate our model. The feature selection process was performed during each training step, and hence the dimension of the feature space varies slightly between the folds. Model performance was evaluated using a variety of metrics, including recall and precision at various probability thresholds, F1 scores, area under the Receiver Operating Characteristic (ROC) curve (AUROC), and area under the Precision-Recall (PR) curve (AUPR). Because both accuracy and AUROC can be misleading measures of model performance for imbalanced classification problems (in our case, cases form about 3% of our cohort), we favored AUPR as the primary metric to evaluate our models as it provides a concise summary of the recall-precision tradeoff across different levels of probability threshold¹⁷.

Feature importance

To evaluate our model's predictions and feature importance, we calculated SHapley Additive exPlanation (SHAP) values.^{18,19} SHAP values, a game-theoretic concept repurposed for machine learning, allow us to see how various features from our high dimensional feature space contribute to determining

the model's output for each patient. Feature importance ranking is based on the mean absolute value of SHAP values for each feature.

Results

Cohort

Our cohort comprised 87,398 children meeting eligibility criteria for SARS-CoV-2 infection and follow-up (Table 1). Of these, 1,049 had a diagnosis of MIS-C, and 1,215 were diagnosed with non-MIS-C PASC. Most patients entered the cohort between September 2021 and April 2022. Contributions from individual health systems varied from 1.7% to 26.8% of the cohort. Serology testing in the absence of viral test results was used to identify 2.3% of the controls. Consistent with code availability, a higher proportion of patients with MIS-C diagnoses entered the cohort in early 2021 than those with non-MIS-C PASC (Figure S1).

TreeScan-selected features

TreeScan selection resulted in 645 condition features, 1,102 diagnostic test features, 197 procedure features, and 112 medication features. The terms from the SNOMED-CT ontology for which TreeScan found the greatest enrichment in patients diagnosed with PASC are shown in Table S1a. Some terms reflect specific pathologic findings (e.g. myocarditis), especially cardiovascular processes, and particularly those consistent with MIS-C. However, many of the terms with highest risk ratios are internal nodes in the SNOMED-CT hierarchy, and reflect the sum of contributions from multiple more specific terms. This pattern reflects the heterogeneity of presentation, particularly in patients without MIS-C, such that no single specific diagnosis best distinguishes between cases and non-cases. Results from medication data (Table S1c), diagnostic testing (Table S1d), and procedures (Table S1b) again reflect cardiovascular therapies, as well as immunomodulation, and supportive care for acute illness.

Model performance

The final algorithm was an XGBoost²⁰ model with a 10,339-dimensional feature space consisting of features described above evaluated at 5 different time windows around time of SARS-CoV-2 infection (-1 to 0 months, 0 to 1 months, 1 to 2 months, 2-3 months, and 3-6 months).

Using five-fold cross-validation we found that our model classified whether patients had PASC of any kind with 91.7% AUPR, 99.0% AUROC (Figure 1), with 79.9% recall and 94.5% precision at the $p=0.5$ threshold. In classifying whether patients had MIS-C, the model achieved 93.9% AUPR and 99.8% AUROC, with 82.0% recall and 95.4% precision at the $p=0.5$ threshold. In classifying non-MISC PASC patients, the model performed with 81.0% AUPR and 98.1% AUROC, with 66.9% recall and 86.4% precision at the $p=0.5$ threshold.

Feature importance

While SHAP summaries^{18,19} do not provide precise data about feature importance, they do provide approximations of features' impact on model results. Results for the 3-outcome model are presented in Figure 2, aggregated by clinical domain. Because the model output is multinomial, the x axis is expressed in terms of log odds, rather than probabilities. Among demographic features, time of cohort entry was most impactful. Unsurprisingly, the clinical feature with the greatest positive result was the presence of an inflammatory diagnosis. Diagnosis features appeared more often as positive predictors, while

laboratory testing appeared more often as negative predictors. Additional SHAP summaries for specific data domains are presented in Figure S2.

Discussion

We have employed machine learning to identify potential patients with PASC based on EHR data collected during clinical care. As controls substantially outnumber cases in our cohort, accuracy and AUROC tend to be inflated, so we assessed our model's performance primarily in terms of the Precision-Recall curve. At the $p=0.5$ threshold, our model was a high-precision classifier for PASC in general as well as for MIS-C and non-MIS-C PASC separately. Indeed, the model was able to classify patients with up to 70% recall with near perfect precision, with a drop-off in precision below 80% beginning only around 85% recall. At this threshold, our model's high precision could be a valuable tool for identifying patients with PASC for study cohorts, particularly patients who did not receive a diagnosis due to unavailability of specific diagnosis codes, evolving clinical understanding of PASC, or less severe presentation. At lower probability thresholds, the model is potentially useful for identifying patients who may have features of PASC for further screening, such as might be done for clinical trial recruitment.

Review of feature contribution using SHAP value visualization reveals complex patterns, especially for non-MIS-C variants of PASC. In the more straightforward MIS-C case, the single diagnosis with consistent positive impact is ECG abnormalities; procedure codes show a similar pattern. Continuing this pattern, laboratory testing for inflammation and cardiac injury has a positive impact, while additional infection-related testing is more characteristic of unaffected patients. Aspirin, used only for a narrow set of indications in pediatrics, has the strongest positive influence among medications, with other immunomodulators associated with MIS-C (corticosteroid and immunoglobulin) appearing as well. SHAP data for the non-MIS-C group of patients reflects their greater heterogeneity. The skew toward later cohort entry despite masking of the PASC diagnosis code itself may reflect differences in PASC natural history, or in practice patterns for evaluation and treatment. Like children with MIS-C, lab utilization at the time of SARS-CoV-2 infection is not associated with caseness, but post-acute utilization is. However, no specific laboratory testing has a strong effect. Interestingly, several types of diagnostic imaging, but especially CT imaging during the acute phase of COVID19 may have positive predictive value. Among diagnoses, respiratory and neurologic codes in the early postacute period were observed to exert the greatest positive impact, consistent with respiratory medications, while aspirin and acetaminophen have negative impact, likely due to their role in MIS-C treatment. Coronavirus infection codes during the 1 to 2 month window suggests such codes are often still used for PASC-diagnosed patients during the post-acute period.

The model's difficulties with the remaining 15% of PASC patients were driven by patients with non-MIS-C PASC. The lower recall for these patients could be a result of several factors. First, because the U09.9 diagnosis code was implemented in October 2021, there is a shorter overall duration for appropriately labelled cases to accrue in the training data. Though we attempted to mitigate this by including as cases patients who received more general post-viral disorder codes following a positive COVID test, it is possible that these more general codes were not used consistently. Additionally, non-MIS-C PASC patients had, on average, less follow-up time than patients with MIS-C. A second factor is that MIS-C is a more sharply defined condition with similarities to Kawasaki disease, an existing clinical phenotype, and thus there are features in the data related to diagnostic test utilization that align with consensus definitions of MIS-C (*cf.* Figure S2e). By contrast, we see more reliance on co-occurring symptoms and

conditions to define non-MISC-PASC. These are more heterogeneous given the range of possible symptoms and presentations and likely include more characteristics that are poorly coded in discrete EHR data, such as fatigue or school difficulties.

We expect that our observed ceiling in recall at 85% may also reflect several limitations in the available training set. The primary challenge of using PASC diagnosis information in EHR data as a training standard, particularly for non-MIS-C PASC, is that the lack of a specific definition means the presence or absence of PASC-specific diagnosis codes does not constitute a gold standard for defining PASC. Moreover, patients who were seen with PASC and MIS-C may not have received the relevant diagnosis codes either due to the recent release of the codes or the early limited understanding of PASC, especially in the case of less severe sub-phenotypes. Thus, while diagnosis codes can serve as a starting point for assessing complications due to COVID-19, there is likely under-ascertainment of PASC which may result in false negative labels in our control population. The lack of a consensus definition may drive differences in clinical assignment of the code, and many patients may either be undiagnosed or never visit a clinician if the symptoms are not severe. On the other hand, false positives are also possible: many PASC features may reflect ongoing symptoms or conditions related to the acute COVID-19 infection, and therefore the lack of clear boundary between acute and post-acute symptoms and conditions may have led to patients in the acute phase of illness receiving a PASC diagnosis. Additionally, unlike adult care, there are relatively few institutional markers, such as dedicated long COVID-19 clinics, for patients with PASC, who receive their care in broader specialty settings such as cardiology or neurology. Lack of follow up for non-MIS-C PASC patients may also limit the model's ability to identify distinguishing characteristics. The SHAP value plots indicate that many of the predictive features of PASC occur within the first three months from diagnosis, which may reflect this limitation.

When compared with adults, PASC in children may present with a milder clinical course²¹, leading to underdiagnosis in the health care setting. Some effects of PASC may have been seen in school performance or extracurricular participation, neither of which may have been captured in the HER. Pediatric PASC may less frequently involve focal organ dysfunction that prompts specific utilization, such as laboratory testing or diagnostic imaging which is well-captured in EHR data. Additionally, the majority of PASC manifestations in children possibly occur shortly following acute infection and resolve more rapidly than in adults¹⁰. These factors both require a distinct approach to classifying children with PASC and warrant further investigation as clinical guidelines and definitions are developed.

To our knowledge, this is the first study that has applied machine learning to an exclusively pediatric population to identify patients with PASC. A recent study⁵ developed a machine learning phenotype in the adult population trained on classifying whether patients were seen at a long COVID clinic. The model achieved 0.92 AUROC with 0.85 precision and 0.86 recall, the latter being slightly greater than observed here. Differences between our model and theirs include our pediatric population, and the use of PASC diagnoses as an outcome rather than specialty clinic attendance, due to limited use of dedicated PASC clinics at pediatric institutions.

Of note, we believe this study is also the first to use TreeScan for clinical feature selection in a machine learning model. The sets of possible features in each domain number in the hundreds of thousands, making reduction of dimensionality a necessary part of model development. Further, it is difficult to identify the correct level of granularity at which concepts should be clustered to define model features. This can be addressed through clinician expertise but is a time-consuming process which may

incorporate bias stemming from individual clinicians' practices or the coding practices at their institutions. However, vocabularies such as SNOMED-CT or RxNorm have rich hierarchies based on physiology or chemical structure, respectively. TreeScan therefore presents an appealing alternative: it takes the hierarchical structure of clinical vocabularies into account and does not require pre-specifying the level of granularity at which features should be selected.

Our study does have additional limitations, many in common with other studies which use EHR data to investigate PASC, a new condition with a broad definition. First, our model was trained to classify whether patients had a PASC diagnosis, which does not include all patients having PASC. Patients with milder manifestations may have been less likely to receive a diagnosis and thus would be less likely to be detected by our model. Second, our model was trained on data from tertiary and quaternary pediatric health systems, and so it may reflect biases in access to care as well as clinicians' practices. Though a matched cohort design could be used to address some of these biases, it would run counter to our purpose of identifying patients likely to have PASC from routine EHR data. Third, while using data from a variety of sites may avoid overfitting the model to site-specific coding practices, there could be bias arising from coding practices across PEDSnet, as a network of pediatric health systems. A fourth potential limitation stems from our approach to imputing index dates for patients with PASC. Due to our difficulty in accurately capturing the initial SARS-CoV-2 infection date for such patients, we used a random date between 28 and 90 days prior to the first PASC diagnosis as the index date. However, this may not accurately reflect the date of the patient's SARS-CoV-2 infection and as a result, relevant features may appear in a different time window relative to infection than when they actually occurred. Given the short follow-up for many patients, this could lead to a model that heavily weighs data close in time to infection. Finally, as noted above, the limited availability of PASC and MIS-C diagnosis codes may have not only reduced follow-up but led to models that focus on manifestation associated with more recent viral variants.

There is substantial potential for future work as understanding of PASC develops and more data become available. Training on data with more cases could help address the model's potential limitation due to heterogeneity of PASC (particularly non-MIS-C PASC). Clinician review of false positives from the model could be used to determine the extent to which PASC-likely patients as determined by the model may have been undiagnosed, and review of false negatives could indicate which features the model is either misinterpreting or failing to detect. Further, our planned validation of the model on sites outside the PEDSnet network will help to evaluate the model's generalizability. Supplementing the model with other sources of clinical data such as unstructured chart notes could further improve model performance. Different features may be associated with PASC over different time periods relative to COVID infection. The tree temporal scan statistic has the potential to address this by picking out both features and time windows over which patients with PASC disproportionately have those features occur, and thereby construct a more refined feature space than the fixed windows we used.

Conclusion

We have applied machine learning methods to available EHR data for children with SARS-CoV-2 infection and subsequent diagnosis of PASC. Using appropriate threshold settings, the model can be used to identify PASC patients in health systems data at higher precision for inclusion in studies or at higher recall in screening for clinical trials, especially in settings where PASC diagnosis codes are used less frequently or less reliably. Analysis of how specific features contribute to the classification process may

assist in gaining a better understanding of features that are associated with PASC diagnoses. While additional work is required to improve identification of children with uncommon manifestations of PASC, the current classifier provides a valuable research tool, especially in cases where the scale or provenance of data make it infeasible to determine each patient's PASC status by direct report. It may also find use in health care delivery, to identify undiagnosed patients who may benefit from further screening to identify and ameliorate effects of PASC.

References

1. Fainardi V, Meoli A, Chiopris G, et al. Long COVID in Children and Adolescents. *Life Basel Switz*. 2022;12(2):285. doi:10.3390/life12020285
2. Thallapureddy K, Thallapureddy K, Zerda E, et al. Long-Term Complications of COVID-19 Infection in Adolescents and Children. *Curr Pediatr Rep*. 2022;10(1):11-17. doi:10.1007/s40124-021-00260-x
3. Rao S, Lee GM, Razzaghi H, et al. Clinical features and burden of post-acute sequelae of SARS-CoV-2 infection in children and adolescents: an exploratory EHR-based cohort study from the RECOVER program. *MedRxiv Prepr Serv Health Sci*. Published online May 25, 2022:2022.05.24.22275544. doi:10.1101/2022.05.24.22275544
4. Reese J, Blau H, Bergquist T, et al. Generalizable Long COVID Subtypes: Findings from the NIH N3C and RECOVER Programs. *MedRxiv Prepr Serv Health Sci*. Published online May 25, 2022:2022.05.24.22275398. doi:10.1101/2022.05.24.22275398
5. Pfaff ER, Girvin AT, Bennett TD, et al. Identifying who has long COVID in the USA: a machine learning approach using N3C data. *Lancet Digit Health*. 2022;4(7):e532-e541. doi:10.1016/S2589-7500(22)00048-6
6. Yang S, Varghese P, Stephenson E, Tu K, Gronsbell J. Machine learning approaches for electronic health records phenotyping: A methodical review. Published online April 27, 2022:2022.04.23.22274218. doi:10.1101/2022.04.23.22274218
7. HAN Archive - 00432 | Health Alert Network (HAN). Published September 21, 2021. Accessed August 18, 2022. <https://emergency.cdc.gov/han/2020/han00432.asp>
8. Algarni AS, Alamri NM, Khayat NZ, Alabdali RA, Alsubhi RS, Alghamdi SH. Clinical practice guidelines in multisystem inflammatory syndrome (MIS-C) related to COVID-19: a critical review and recommendations. *World J Pediatr*. 2022;18(2):83-90. doi:10.1007/s12519-021-00499-w
9. Mahmoud S, El-Kalliny M, Kotby A, El-Ganzoury M, Fouda E, Ibrahim H. Treatment of MIS-C in Children and Adolescents. *Curr Pediatr Rep*. 2022;10(1):1-10. doi:10.1007/s40124-021-00259-4
10. Borch L, Holm M, Knudsen M, Ellermann-Eriksen S, Hagstroem S. Long COVID symptoms and duration in SARS-CoV-2 positive children - a nationwide cohort study. *Eur J Pediatr*. 2022;181(4):1597-1607. doi:10.1007/s00431-021-04345-z
11. Ramakrishnan RK, Kashour T, Hamid Q, Halwani R, Tleyjeh IM. Unraveling the Mystery Surrounding Post-Acute Sequelae of COVID-19. *Front Immunol*. 2021;12:686029. doi:10.3389/fimmu.2021.686029
12. Kulldorff M, Fang Z, Walsh SJ. A tree-based scan statistic for database disease surveillance. *Biometrics*. 2003;59(2):323-331. doi:10.1111/1541-0420.00039
13. CDC. Healthcare Workers. Centers for Disease Control and Prevention. Published February 11, 2020. Accessed August 18, 2022. <https://www.cdc.gov/coronavirus/2019-ncov/hcp/clinical-care/post-covid-public-health-recs.html>

14. Coding Long COVID: Characterizing a new disease through an ICD-10 lens | medRxiv. Accessed August 18, 2022. <https://www.medrxiv.org/content/10.1101/2022.04.18.22273968v1>
15. CDC Announces Approval of ICD-10 Code for Post-Acute Sequelae of COVID-19. [aapmr.org](https://www.aapmr.org/members-publications/member-news/member-news-details/2021/07/20/cdc-announces-approval-of-icd-10-code-for-post-acute-sequelae-of-covid-19). Accessed August 22, 2022. <https://www.aapmr.org/members-publications/member-news/member-news-details/2021/07/20/cdc-announces-approval-of-icd-10-code-for-post-acute-sequelae-of-covid-19>
16. Wang SV, Maro JC, Baro E, et al. Data Mining for Adverse Drug Events With a Propensity Score-matched Tree-based Scan Statistic. *Epidemiol Camb Mass*. 2018;29(6):895-903. doi:10.1097/EDE.0000000000000907
17. Saito T, Rehmsmeier M. The Precision-Recall Plot Is More Informative than the ROC Plot When Evaluating Binary Classifiers on Imbalanced Datasets. *PLoS ONE*. 2015;10(3):e0118432. doi:10.1371/journal.pone.0118432
18. Lundberg SM, Lee SI. A Unified Approach to Interpreting Model Predictions. In: *Advances in Neural Information Processing Systems*. Vol 30. Curran Associates, Inc.; 2017. Accessed August 18, 2022. <https://proceedings.neurips.cc/paper/2017/hash/8a20a8621978632d76c43dfd28b67767-Abstract.html>
19. GPUtreeShap: massively parallel exact calculation of SHAP scores for tree ensembles [PeerJ]. Accessed August 18, 2022. <https://peerj.com/articles/cs-880/>
20. XGBoost | Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. Accessed August 18, 2022. <https://dl.acm.org/doi/10.1145/2939672.2939785>
21. Pellegrino R, Chiappini E, Licari A, Galli L, Marseglia GL. Prevalence and clinical presentation of long COVID in children: a systematic review. *Eur J Pediatr*. 2022;181(12):3995-4009. doi:10.1007/s00431-022-04600-x

Tables and Figures

Table 1: Cohort demographic and clinical characteristics

	Level	Overall	Non-PASC SARS-CoV-2 positive	MIS-C	Non MIS-C PASC
Number of patients		87,398	85,134	1,049	1,215
Age group (years) (%)	<1	12460 (14.3)	12397 (14.6)	28 (2.7)	35 (2.9)
	1-4	23449 (26.8)	23047 (27.1)	293 (27.9)	109 (9.0)
	12-15	14520 (16.6)	13927 (16.4)	180 (17.2)	413 (34.0)
	16-20	10289 (11.8)	9826 (11.5)	98 (9.3)	365 (30.0)
	5-11	26680 (30.5)	25937 (30.5)	450 (42.9)	293 (24.1)
Sex (%)	Female	41988 (48.0)	40901 (48.0)	421 (40.1)	666 (54.8)
	Male/other/unknown	45410 (52.0)	44233 (52.0)	628 (59.9)	549 (45.2)
Race/ethnicity (%)	Asian/PI	3702 (4.2)	3622 (4.3)	41 (3.9)	39 (3.2)
	Black/AA	20316 (23.2)	19950 (23.4)	232 (22.1)	134 (11.0)
	Hispanic	15215 (17.4)	14828 (17.4)	209 (19.9)	178 (14.7)
	Multiple	3832 (4.4)	3744 (4.4)	41 (3.9)	47 (3.9)
	Other/Unkn own	6678 (7.6)	6523 (7.7)	85 (8.1)	70 (5.8)
	White	37655 (43.1)	36467 (42.8)	441 (42.0)	747 (61.5)
Cohort entry period (%)	Jan-Apr 2021	15778 (18.1)	15438 (18.1)	249 (23.7)	91 (7.5)
	May-Aug 2021	9600 (11.0)	9181 (10.8)	226 (21.5)	193 (15.9)
	Sep-Dec 2021	24593 (28.1)	23724 (27.9)	400 (38.1)	469 (38.6)
	Jan-Apr 2022	28586 (32.7)	28096 (33.0)	128 (12.2)	362 (29.8)
	May-Jun 2022	8841 (10.1)	8695 (10.2)	46 (4.4)	100 (8.2)
Institution (%)	A	6330 (7.2)	6120 (7.2)	102 (9.7)	108 (8.9)
	B	8753 (10.0)	8478 (10.0)	152 (14.5)	123 (10.1)
	C	23388 (26.8)	22914 (26.9)	185 (17.6)	289 (23.8)
	D	4321 (4.9)	4189 (4.9)	92 (8.8)	40 (3.3)
	E	6509 (7.4)	6281 (7.4)	117 (11.2)	111 (9.1)
	F	4727 (5.4)	4612 (5.4)	45 (4.3)	70 (5.8)
	G	1503 (1.7)	1383 (1.6)	66 (6.3)	54 (4.4)
	H	16663 (19.1)	16380 (19.2)	101 (9.6)	182 (15.0)
	I	15204 (17.4)	14777 (17.4)	189 (18.0)	238 (19.6)

It is made available under a [CC-BY-NC-ND 4.0 International license](#) .

Index event type (%)	Antigen test positive	6030 (6.9)	6030 (7.1)	0 (0.0)	0 (0.0)
	MIS-C diagnosis	1049 (1.2)	0 (0.0)	1049 (100.0)	0 (0.0)
	Non-MIS-C PASC diagnosis	1215 (1.4)	0 (0.0)	0 (0.0)	1215 (100.0)
	PCR test positive	77164 (88.3)	77164 (90.6)	0 (0.0)	0 (0.0)
	Serology test positive	1940 (2.2)	1940 (2.3)	0 (0.0)	0 (0.0)

Table 2: Machine learning model performance in identifying patients with PASC (MIS-C or non-MIS-C variants)

Caption: The table below displays several performance metrics for each of the three outcomes (PASC (any), non MIS-C PASC, and MIS-C). In addition to composite performance statistics, we report the area under the precision-recall curve (AUPR) and area under the receiver operating characteristic curve (AUROC). The curves themselves are shown and described in Figure 1.

	PASC (any)	MIS-C	Non MIS-C PASC
Accuracy (at p=0.5 threshold)	0.994	0.997	0.995
AUPR	0.917	0.939	0.810
AUROC	0.990	0.998	0.981
F1 score (at p=0.5 threshold)	0.866	0.882	0.754
Precision (at p=0.5 threshold)	0.945	0.954	0.864
Recall (at p=0.5 threshold)	0.799	0.820	0.669

Figure 1: Receiver Operating Characteristic (ROC) and Precision-Recall (PR) curves describing model performance in identifying PASC with (MIS-C or non-MIS-C variants)

Caption: For each of the three outcomes (PASC (any), non MIS-C PASC, and MIS-C) the Receiver Operating Characteristic (ROC) curves and Precision-Recall (PR) curves are estimated and plotted 5 times, once for each cross-validation fold.

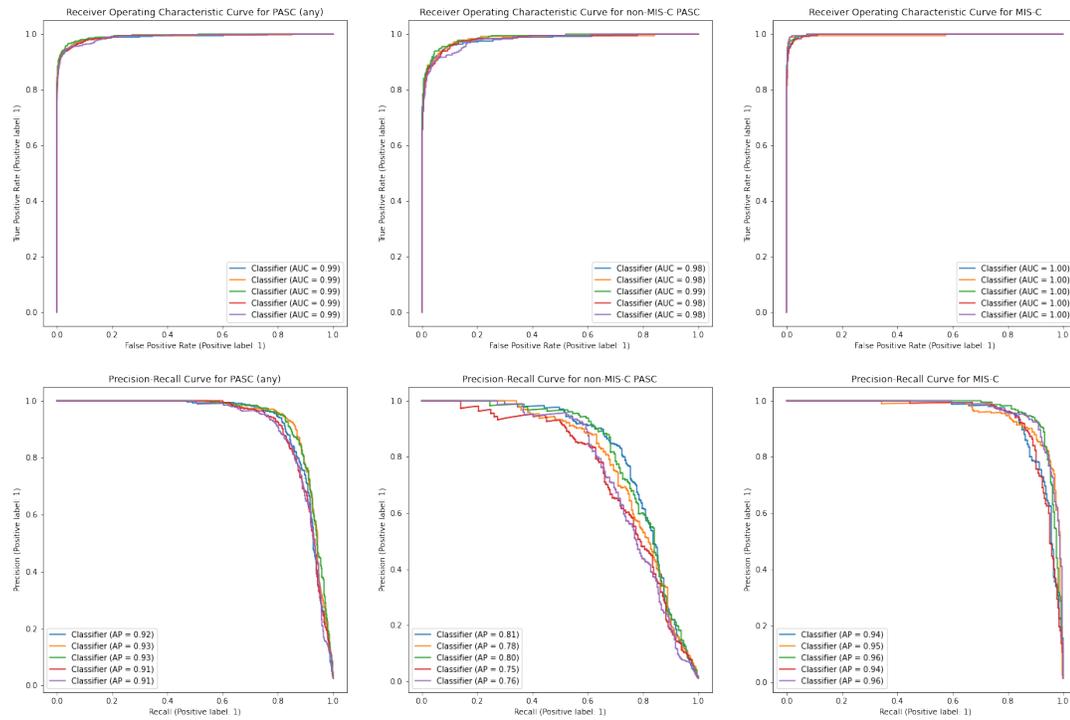


Figure 2 SHapley Additive exPlanation (SHAP) values for top administrative and clinical model features by class in predicting non-MIS-C PASC

Caption: The plots show the most significant features as determined by the sum of SHAP value magnitudes over all samples. For each feature, SHAP values for each patient are plotted, with color representing the feature value (e.g. red if feature was present and blue if absent in case of a binary variable). For the SHAP values pictured, the x axis is interpreted as change in log odds (in particular, SHAP values are not confined to be between -1 and 1).

