De-black-boxing health AI: demonstrating reproducible machine learning computable phenotypes using the N3C-RECOVER Long COVID model in the *All of Us* data repository

Emily R. Pfaff [1], Andrew T. Girvin[2], Miles Crosskey[3], Srushti Gangireddy[4], Hiral Master [5], Wei-Qi Wei[4], V. Eric Kerchberger [6], Mark Weiner [7], Paul A. Harris[4], Melissa Basford[5], Chris Lunt[8], Christopher G. Chute [9], Richard A. Moffitt[10], and Melissa Haendel [11]; on behalf of the N3C and RECOVER Consortia

[1]Department of Medicine, University of North Carolina at Chapel Hill School of Medicine, Chapel Hill, North Carolina, USA, [2]Palantir Technologies, Denver, Colorado, USA, [3]CoVar Applied Technologies, Durham, North Carolina, USA, [4]Department of Biomedical Informatics, Vanderbilt University Medical Center, Nashville, Tennessee, USA, [5]Vanderbilt Institute for Clinical and Translational Research, Vanderbilt University Medical Center, Nashville, Tennessee, USA, [6]Department of Medicine, Division of Allergy, Pulmonary & Critical Care Medicine, Vanderbilt University Medical Center, Nashville, Tennessee, USA, [7]Department of Medicine, Weill Cornell Medicine, New York, USA, [8]National Institutes of Health, Bethesda, Maryland, USA, [9]Johns Hopkins Schools of Medicine, Public Health, and Nursing. Baltimore, Maryland, USA, [10]Departments of Hematology and Medical Oncology and Biomedical Informatics, Emory University, Atlanta, Georgia, USA and [11]Department of Biomedical Informatics, University of Colorado Anschutz Medical Campus, Denver, Colorado, USA

Corresponding Author: Emily R. Pfaff, PhD, MS, Department of Medicine, University of North Carolina at Chapel Hill School of Medicine, 160 N Medical Drive, Chapel Hill, NC 27599, USA; epfaff@email.unc.edu

Received 18 January 2023; Revised 28 March 2023; Editorial Decision 22 April 2023; Accepted 24 April 2023

ABSTRACT

Machine learning (ML)-driven computable phenotypes are among the most challenging to share and reproduce. Despite this difficulty, the urgent public health considerations around Long COVID make it especially important to ensure the rigor and reproducibility of Long COVID phenotyping algorithms such that they can be made available to a broad audience of researchers. As part of the NIH Researching COVID to Enhance Recovery (RECOVER) Initiative, researchers with the National COVID Cohort Collaborative (N3C) devised and trained an ML-based phenotype to identify patients highly probable to have Long COVID. Supported by RECOVER, N3C and NIH's *All of Us* study partnered to reproduce the output of N3C's trained model in the *All of Us* data enclave, demonstrating model extensibility in multiple environments. This case study in ML-based phenotype reuse illustrates how open-source software best practices and cross-site collaboration can de-black-box phenotyping algorithms, prevent unnecessary rework, and promote open science in informatics.

Key words: electronic health records, machine learning, phenotype, SARS-CoV-2

INTRODUCTION

Post-acute sequelae of SARS-CoV-2 infection (PASC) and Long COVID (hereafter referred to collectively as Long COVID) have been recognized as potentially debilitating conditions associated with COVID-19 infection since the Spring of 2020, and have attracted significant research attention and funding in that time. However, a firm clinical definition of Long COVID continues to be elusive. The World Health Organization (WHO) published a consensus definition in 2021,[1] but it has not been universally accepted; its breadth, non-specificity, and overlap with other conditions makes it difficult to apply in clinical practice or research.[2] This definitional uncertainty impacts clinical care, but also affects the accuracy with which we can use data to ascertain cases for retrospective or prospective research, surveillance, or clinical decision support.

In 2021, we used the National COVID Cohort Collaborative (N3C) electronic health record (EHR) data repository of over 16M patients across >230 clinical sites to develop a machine learning (ML) model to identify potential Long COVID patients.[3] We trained the model to recognize Long COVID using the records of patients who had sought or been referred to care at a Long COVID specialty clinic. We have since updated the model to train on data from patients with a "U09.9" International Classification of Diseases-10-Clinical Modification (ICD-10-CM) diagnosis code ("Post COVID-19 condition, unspecified"), which was released for use in the United States on October 1, 2021. This model has been used in studies[4–6] as 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/.

Ideal computable phenotypes are standardized, shareable, and machine-readable artifacts that allow reproducible patient cohort identification,[7] all characteristics that promote rigor, reproducibility, and transparency to enable translational science, improved clinical research outcomes, and clinical decision support. However, ML-driven phenotypes are among the most challenging to share and reproduce due to their complexity, often-extensive feature engineering pipelines, and underlying assumptions about both the data and the data modeling that make translation to another environment or site challenging. Despite this difficulty, the urgent public health considerations around Long COVID make it especially important to ensure the rigor and reproducibility of Long COVID phenotyping algorithms such that they can be made available to a broad base of researchers, institutions, and clinical settings. Without a generalizable computable phenotype for Long COVID, we cannot electronically identify cohorts of Long COVID patients in clinical data repositories. Without this ability, otherwise "unlabeled" patients may be left out of opportunities to join clinical trials, and retrospective researchers will lack the ability to investigate Long COVID, its risk factors, and its outcomes at a population level. Through RECOVER, N3C and NIH's *All of Us* study[8] partnered to reproduce the output of N3C's trained model in the *All of Us* data enclave, demonstrating extensibility in multiple environments.

Here, we describe our efforts to translate the N3C model in a second, massive multi-institutional research data environment, despite the challenges presented by a newly defined disease. We believe these principles and lessons learned can be applied more broadly to promote reproducibility and transparency of ML-based computable phenotypes, thus realizing the potential of this increasingly prominent method in clinical informatics.

MATERIALS AND METHODS

Study design and base population

To model Long COVID, we used EHR data integrated and harmonized inside the secure N3C Data Enclave to identify healthcare utilization patterns and clinical features among patients with COVID-19. The methods for patient identification, data acquisition, ingestion, and harmonization into the N3C Enclave have been described previously.[9,10] Our ML-based Long COVID phenotype has also been previously described,[3] with some details repeated below to promote understanding of the current work. Detailed information on updates made to the original model since initial publication is available in Supplemental Methods.

We define our base population ($n = 2\,465\,242$, as of N3C data release v87) as any adult patient (age ≥18 years) with either a COVID-19 diagnosis code (U07.1) or a positive SARS-CoV-2 PCR or antigen test, for whom at least 145 days have passed since COVID-19 index date, and who have had at least one healthcare encounter between 45 and 300 days from their COVID-19 index date. See Supplemental Figure S1 for a visualization of these criteria. "COVID-19 index date" is defined as the earliest date of a positive indicator for a patient. For patients with multiple positive tests or diagnosis codes, we select the date of the first positive test as the index.

The original model was trained on patients who were seen at a Long COVID specialty clinic, as there was no official ICD-10-CM code for PASC or Long COVID until October 1, 2021. At present, however, the ICD-10-CM code U09.9 ("Post COVID-19 condition, unspecified") has been available for use for just over a year. Due to its greater specificity and larger sample size, the current model uses N3C patients that qualify for all of the above inclusion criteria *and* have a U09.9 diagnosis code as training and test data ($n = 7221$ for training and $n = 1653$ for test).

Feature engineering for the updated model proceeded much the same as the original model, accounting for demographics, healthcare visit details, medical conditions, and new prescriptions in each patient's analysis window. We used the Python package XGBoost to construct the model, using 200 features in total. This is a smaller number of features than the original model's 924; in testing we found that performance is not impacted by limiting to the top 200 features, making this change desirable from the standpoint of computational efficiency, shareability, and model explainability. Categorical features were one-hot encoded. Age and healthcare visit rates were treated as continuous variables, and diagnoses and prescription drugs were modeled as binary features. Model hyperparameters were tuned using GridSearchCV (scikit-learn), with 10-fold cross-validation, set to optimize the area under the receiver operating characteristic curve (AUROC). We trained each model using 10-fold cross-validation, repeated 5 times.[3]

Translating the model to the *All of Us* EHR data repository

Because N3C is built using the Observational Medical Outcomes Partnership (OMOP) common data model, the N3C model can be run against any other OMOP database, aiding with transparency, reproducibility, and external validation efforts. Like N3C, the NIH *All of Us* (AoU) study collects EHR data from over 50 healthcare provider organizations in the OMOP format, and is an effective test bed for our model. Participants over 18 years of age are enrolled in

AoU after an informed consent process from a direct volunteer platform or healthcare provider organizations, which compose the AoU Research Program network. A detailed description of AoU has been published elsewhere.[8] For this study, we used the AoU Controlled Tier Dataset version 6 (C2022Q2R2 Curated Data Repository) available to registered users on AoU's Researcher Workbench, a secure cloud-based platform. This dataset includes longitudinal EHR data from participants who were enrolled from May 30, 2018 to January 1, 2022.

The AoU team used N3C's open-source model code[11] to replicate N3C's model in their environment, using AoU data. The N3C and AoU data teams met on a weekly basis for 12 weeks in order to plan, share knowledge, and troubleshoot during implementation. Required efforts included programming language translation (from N3C's PySpark and Spark SQL to AoU's Python [pandas] and Google BigQuery), comparing base population characteristics, and aligning assumptions about the underlying data and their meaning. Links to the AoU-translated version of N3C's code are available in Supplemental Methods.

RESULTS

When the model is run on the N3C population qualifying for the base inclusion criteria ($n = 2\,465\,242$; see Materials and Methods), each patient is assigned a predicted probability of Long COVID. We replicated this process on the AoU population qualifying for the base inclusion criteria ($n = 8998$, out of approximately 258 000 AoU participants with available EHR data). The version of the AoU data (C2022Q2R2) used for this work contains 40 patients with a U09.9 code, 30 of whom pass the initial inclusion criteria. The *All of Us* team ran N3C's pre-trained model in the AoU data without retraining, thus enabling us to assess the performance of the model developed using the N3C data. The distribution of predicted probabilities of Long COVID across the populations of both repositories is shown in Figure 1.

Interestingly, the AoU data have a greater proportion of patients with the highest scores, and a lower proportion of patients with the lowest scores. Differences in the underlying data, in both size and context, likely contribute to these differences (see Discussion).

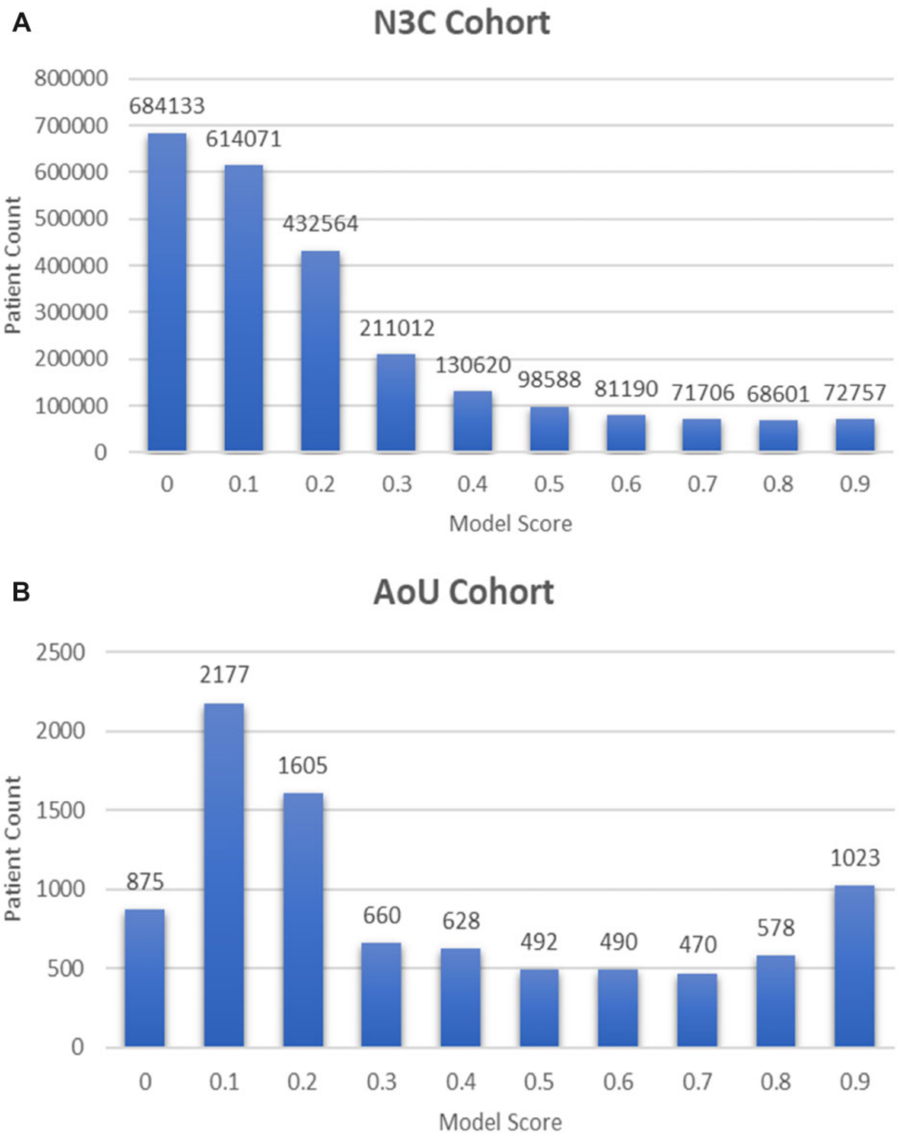

Figure 1. Distribution of predicted probability of Long COVID across the (A) N3C and (B) AoU base populations. About 16.0% of qualifying SARS-CoV-2 positive N3C patients and 34.0% of qualifying SARS-CoV-2 positive AoU patients have predicted probabilities ≥0.5.

Table 1. Demographic breakdown of the N3C population scored by the model, stratified by model score

	Model score <0.50 *n* = 2 071 869	Model score between 0.50 and 0.75 *n* = 216 949	Model score >0.75 *n* = 176 424
Sex (%)			
Female	1 217 052 (58.7)	146 706 (67.6)	120 840 (68.5)
Male	853 811 (41.2)	70 214 (32.4)	55 563 (31.5)
Other/unknown	1006 (0.0)	29 (0.0)	21 (0.0)
Race (%)			
Asian	41 915 (2.0)	4537 (2.1)	3394 (1.9)
Black	277 395 (13.4)	34 604 (16.0)	29 388 (16.7)
Hawaiian/Pac Isldr.	3974 (0.2)	487 (0.2)	371 (0.2)
White	1 494 592 (72.2)	153 956 (71.0)	124 452 (70.6)
Other	64 561 (3.1)	3789 (1.7)	2915 (1.7)
Unknown	188 114 (9.1)	19 405 (9.0)	15 724 (8.9)
Ethnicity (%)			
Hispanic/Latino	232 263 (11.2)	24 139 (11.1)	18 645 (10.6)
Not Hispanic/Latino	1 557 107 (75.2)	162 284 (74.8)	130 641 (74.0)
Unknown	282 499 (13.6)	30 526 (14.1)	27 138 (15.4)
Age group (%)			
18–45	1 031 578 (49.8)	67 355 (31.0)	48 245 (27.3)
46–65	666 967 (32.2)	89 430 (41.2)	76 105 (43.1)
66+	373 324 (18.0)	60 164 (27.7)	52 074 (29.5)

Table 2. Demographic breakdown of the AoU population scored by the model, stratified by model score

	Model score <0.50 *n* = 5937	Model score between 0.50 and 0.75 *n* = 1218	Model score >0.75 *n* = 1843
Sex (%)			
Female	4188 (70.5)	747 (61.3)	1093 (59.3)
Male	1595 (26.9)	443 (36.4)	702 (38.1)
Other/unknown	154 (2.6)	28 (2.3)	48 (2.6)
Race (%)			
Asian	100–130	<20	20–40
Black	1249 (21.0)	280 (23.0)	465 (25.2)
Hawaiian/Pac Isldr.	<20	<20	<20
White	2879 (48.5)	538 (44.2)	776 (42.1)
Other	125 (2.1)	29 (2.4)	32 (1.7)
Unknown	1553 (26.2)	353 (29.0)	544 (29.5)
Ethnicity (%)			
Hispanic/Latino	1421 (23.9)	322 (26.4)	510 (27.7)
Not Hispanic/Latino	4270 (71.9)	850 (69.8)	1249 (67.8)
Unknown	246 (4.1)	46 (3.8)	84 (4.6)
Age group (%)			
18–45	1929 (32.5)	311 (25.5)	415 (22.5)
46–65	2515 (42.4)	538 (44.2)	848 (46.0)
66+	1493 (25.1)	369 (30.3)	580 (31.5)

Note: Counts displayed as a range are required to comply with AoU's policy preventing recalculation of small cell sizes.

Tables 1 and 2 show a demographic breakdown of the model-eligible base population of both repositories, stratified by binned predicted probabilities.

Note the distinct demographic differences between N3C and AoU patients. N3C sites submit data for *all* SARS-CoV-2-positive patients in their EHR warehouses, along with matched controls, whereas AoU is a consented patient cohort whose enrollment aims specifically emphasize diversity.[12]

Figure 2 shows the Shapley values for feature importance during the training of the updated N3C ML model. Important features include patient age, dyspnea, fatigue, and other diagnosis and medication information available within the EHR. Table 3 compares the results of running the trained model in the N3C and AoU data, respectively.

DISCUSSION

This work resulted in successful translation of the N3C Long COVID ML-based phenotype to the AoU environment. Through our teams' collaborative work, we also gained an understanding of the complexities of sharing machine learning-based phenotypes for

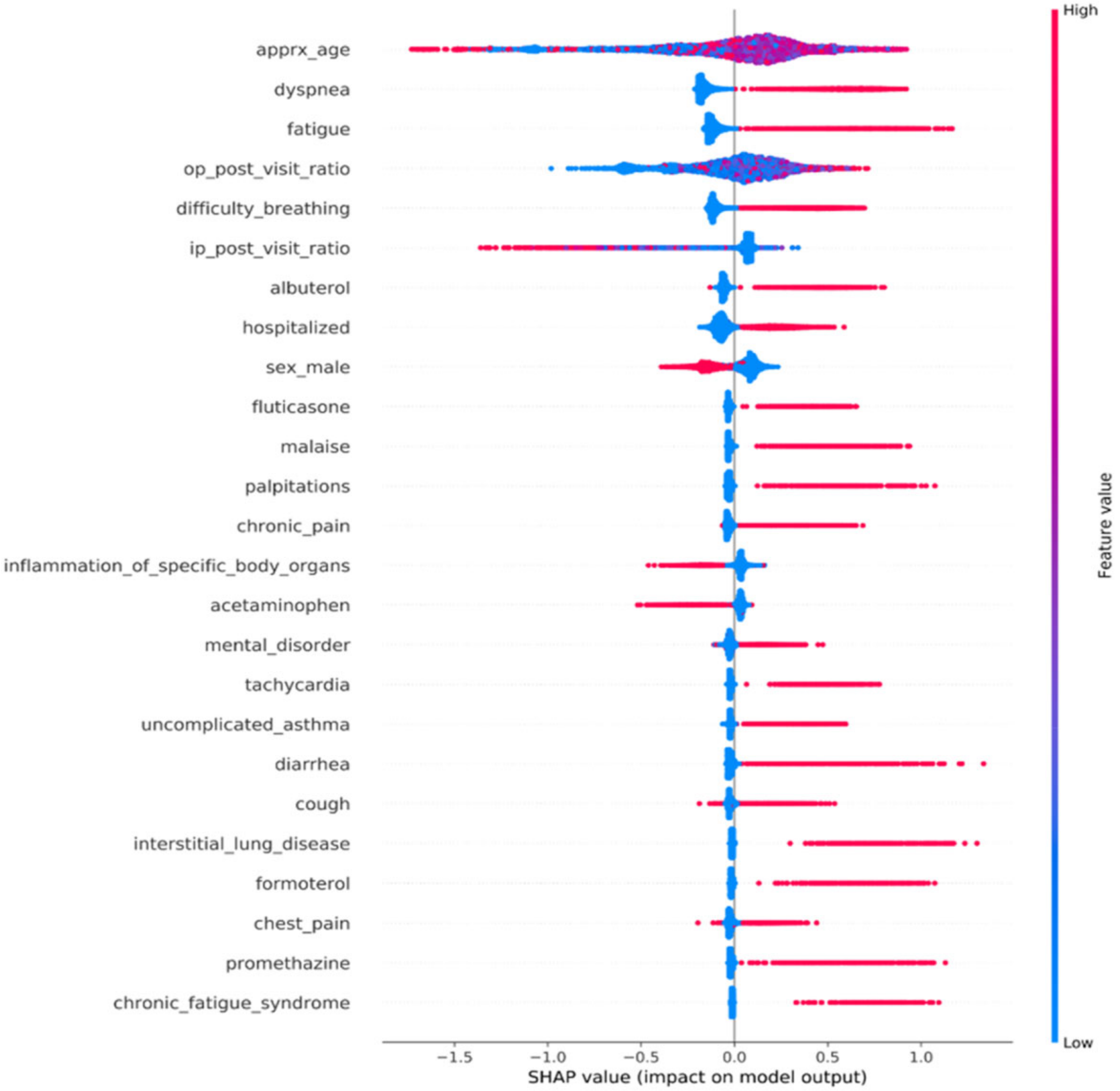

Figure 2. Shapley plot of the top 25 features of the updated N3C model. The color of each point represents the importance of each feature in determining the predicted probability for a given patient. Features with points to the left of the center line are more likely to contribute to a lower predicted probability of Long COVID; features with points to the right are more likely to contribute to a higher predicted probability.

reuse, and developed methods for overcoming many of those challenges.

Challenges and opportunities in sharing ML models

Open-source software is a key component of open science, but its existence alone does not equate to reusability. Numerous examples—and calls to action—exist regarding the need to de-black-box artificial intelligence and ML algorithms.[13,14] For computable phenotypes, even seemingly simple rule-based phenotypes are complex compendiums of codes from different sources such as ICD-10 and CPT, inclusion/exclusion criteria, and scripts that make it challenging to reliably execute across systems. A recent manuscript has documented the significant lack of interoperability across studies.[15]

In order to truly promote reuse, code must be clearly documented and thoroughly commented, ideally with sharing in mind from the start. For this reason, the GitHub repository for the N3C Long COVID phenotype[11] includes features such as README files in each subfolder, code and folders organized in numbered steps, and heavy commenting. However, running a shared phenotype from start to finish with no errors still does not guarantee faithful translation. Rather, the context and meaning of the underlying data must be known and described along with the code. In the context of an ML-based phenotype, the challenges are even greater than in a rule-based context. Not only do local code and common data model mappings vary, but there are a much greater number of computational resources utilized. Further, it is almost certain that populations from different institutions will be dissimilar, requiring understanding of inherent differences and selection biases in order to properly interpret and contextualize results. Our respective teams' weekly meetings were used to convey this additional context, and proved to be an ideal venue to convey complexities that would be difficult to anticipate when writing documentation.

Our teams also overcame multiple challenges in the technical aspects of model translation, including converting N3C's code from PySpark and Spark SQL to Python (pandas) and Google BigQuery syntax. Skilled programmers from both the N3C and AoU teams were required to execute this process accurately. Once AoU

Table 3. Comparison of results across N3C and AoU data

Metric	N3C	All of Us
n qualifying for model inclusion criteria	2 077 866	8998
n qualifying for model inclusion criteria with U09.9 label	13 990	30
AUROC	0.83	0.72
Number of features	200	161

developers translated N3C's code, the translated versions were uploaded to N3C's GitHub via pull requests, enabling others with the same translation needs to leverage AoU's work.

Comparing model results in AoU versus N3C

Though both N3C and AoU use the OMOP data model, our underlying populations differ significantly (see Tables 1 and 2). Moreover, the number of patients with recorded SARS-CoV-2 infections in the AoU database is much smaller than N3C (8998 versus 2 077 866), and the population with a U09.9 diagnosis code is smaller still (30 versus 13 990). Because N3C's model was pre-trained on N3C's larger population, the fact that the AoU has a much smaller eligible cohort did not impact AoU's ability to run the model—however, the differences in cohort size may explain some of the differences we see in our results. A relative lack of patients with low scores in the AoU cohort could be a reflection of the overall increased health system engagement of the typical AoU affirmative enrollee, compared to the all-comer waiver of consent cohort reflected in N3C. This may also explain the higher numbers of AoU patients with high scores, as outpatient utilization rates are an important model feature and may have outsize influence among the care-engaged AoU enrollees. This could be validly interpreted as a mark against the generalizability of the N3C ML model to a *consented* cohort, revealed via this translation exercise.

Another major difference between N3C and AoU's data was the presence or absence of the model's top features. Because the model was trained on N3C data, the importance of the top 200 features (as measured by Shapley value) was determined from the variables present in the N3C data. Even with a shared data model, there is no guarantee that all features from the training data will be present in another data repository. In our case, AoU's cohort of 8998 eligible patients lacked coverage for 39 features of N3C's top 200. This may be the result of (1) coding idiosyncrasies among contributing sites, which differ between N3C and AoU; (2) absence of low-prevalence concepts in the smaller AoU cohort; or (3) the shorter pandemic-era time window available in AoU. Two missing features, "tachycardia" and "diarrhea," are among the top 25 most important features from the N3C model (see Figure 2)—these and other missing features may have also been a contributor to result differences.

A limitation of this work is its restriction to adult patients. We fully recognize the burden of Long COVID on the pediatric population—however, Long COVID appears to present differently in children, and these distinctions likely necessitate one or more separate models.[16] We should note that this exercise in ML-based phenotype reuse is not, and was not intended to be a validation of the accuracy of the phenotype. ML-based computable phenotypes present challenges with performance assessment,[17] and the challenge is even greater in the case of a new disease like Long COVID, where few concrete diagnostic guidelines or gold standards exist. For this effort, performance assessment was particularly challenging due to the small number of U09.9 patients in the AoU dataset. Long COVID in particular introduces additional complexities, as the list of possible Long COVID symptoms is lengthy, heterogeneous, and has significant overlap with many other conditions.[1] Validation of this phenotype will be the subject of future work, requiring chart review and alignment with emerging biomarkers.

CONCLUSION

Through this effort, we have demonstrated the transfer of an ML-based phenotype from one multi-institutional data repository to another. This work generated a set of principles applicable to other ML translation efforts including: (1) Leverage open-source code and a common data model that is shared among all participants; (2) Convene small teams integrating methods and programming experts from each participating group, and encourage those teams to have regular working sessions during the translation and testing process; and (3) Document code far above the bare minimum, including plenty of detail about assumptions, data cleaning steps, and derived variables and well-written, stepwise instructions. This workflow should be translatable to other phenotyping use cases, and will hopefully encourage more research teams to decrease rework and promote open science by sharing phenotyping and other data manipulation code in this manner.

FUNDING

This research was funded by the National Institutes of Health (NIH) Agreement OTA OT2HL161847 as part of the Researching COVID to Enhance Recovery (RECOVER) research program, as well as CD2H—The National COVID Cohort Collaborative (N3C) U24TR002306.

The *All of Us* Research Program is supported by the National Institutes of Health, Office of the Director: Regional Medical Centers (1 OT2 OD026549; 1 OT2 OD026554; 1 OT2 OD026557; 1 OT2 OD026556; 1 OT2 OD026550; 1 OT2 OD 026552; 1 OT2 OD026553; 1 OT2 OD026548; 1 OT2 OD026551; 1 OT2 OD026555; IAA: AOD21037, AOD22003, AOD16037, AOD21041), Federally Qualified Health Centers (HHSN 263201600085U), Data and Research Center (5 U2C OD023196), Biobank (1 U24 OD023121), The Participant Center (U24 OD023176), Participant Technology Systems Center (1 U24 OD023163), Communications and Engagement (3 OT2 OD023205; 3 OT2 OD023206), and Community Partners (1 OT2 OD025277; 3 OT2 OD025315; 1 OT2 OD025337; 1 OT2 OD025276).

AUTHOR CONTRIBUTIONS

Manuscript drafting: ERP, ATG, MC, HM, and MH. Data analysis: ERP, ATG, MC, SG, HM, and WQW. Program leadership: ERP, RM, CGC, MH, PAH, MB, and CL. Final manuscript approval: ERP, ATG, MC, SG, HM, WQW, EK, PAH, MB, CL, CGC, RM, and MH.

SUPPLEMENTARY MATERIAL

Supplementary material is available at *Journal of the American Medical Informatics Association* online.

ACKNOWLEDGMENTS

This study is part of the NIH Researching COVID to Enhance Recovery (RECOVER) Initiative, which seeks to understand, treat, and prevent the post-acute sequelae of SARS-CoV-2 infection (PASC). For more information on RECOVER, visit https://recovercovid.org/. We would like to thank the National Community Engagement Group (NCEG), all patient, caregiver, and community representatives, and all the participants enrolled in the RECOVER Initiative.

The analyses described in this publication were conducted with data or tools accessed through the NCATS N3C Data Enclave (covid.cd2h.org/enclave). This research was possible because of the patients whose information is included within the data from participating organizations (covid.cd2h.org/dtas) and the organizations and scientists (covid.cd2h.org/duas) who have contributed to the on-going development of this community resource. The N3C data transfer to NCATS is performed under Johns Hopkins University Reliance Protocol # IRB00249128 or individual site agreements with NIH. The N3C Data Enclave is managed under the authority of the NIH; information can be found at https://ncats.nih.gov/n3c/resources. The work was performed under DUR RP-5677B5.

Authorship was determined using ICMJE recommendations. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health, N3C, or RECOVER.

The *All of Us* Research Program would not be possible without the partnership of its participants.

We also acknowledge the following institutions whose data are released or pending in N3C:

Advocate Health Care Network—UL1TR002389: The Institute for Translational Medicine (ITM) • Boston University Medical Campus—UL1TR001430: Boston University Clinical and Translational Science Institute • Brown University—U54GM115677: Advance Clinical Translational Research (Advance-CTR) • Carilion Clinic—UL1TR003015: iTHRIV Integrated Translational health Research Institute of Virginia • Charleston Area Medical Center—U54GM104942: West Virginia Clinical and Translational Science Institute (WVCTSI) • Children's Hospital Colorado—UL1TR002535: Colorado Clinical and Translational Sciences Institute • Columbia University Irving Medical Center—UL1TR001873: Irving Institute for Clinical and Translational Research • Duke University—UL1TR002553: Duke Clinical and Translational Science Institute • George Washington Children's Research Institute—UL1TR001876: Clinical and Translational Science Institute at Children's National (CTSA-CN) • George Washington University—UL1TR001876: Clinical and Translational Science Institute at Children's National (CTSA-CN) • Indiana University School of Medicine—UL1TR002529: Indiana Clinical and Translational Science Institute • Johns Hopkins University—UL1TR003098: Johns Hopkins Institute for Clinical and Translational Research • Loyola Medicine—Loyola University Medical Center • Loyola University Medical Center—UL1TR002389: The Institute for Translational Medicine (ITM) • Maine Medical Center—U54GM115516: Northern New England Clinical & Translational Research (NNE-CTR) Network • Massachusetts General Brigham—UL1TR002541: Harvard Catalyst • Mayo Clinic Rochester—UL1TR002377: Mayo Clinic Center for Clinical and Translational Science (CCaTS) • Medical University of South Carolina—UL1TR001450: South Carolina Clinical & Translational Research Institute (SCTR) • Montefiore Medical Center—UL1TR002556: Institute for Clinical and Translational Research at Einstein and Montefiore • Nemours—U54GM104941: Delaware CTR ACCEL Program • NorthShore University HealthSystem—UL1TR002389: The Institute for Translational Medicine (ITM) • Northwestern University at Chicago—UL1TR001422: Northwestern University Clinical and Translational Science Institute (NUCATS) • OCHIN—INV-018455: Bill and Melinda Gates Foundation grant to Sage Bionetworks • Oregon Health & Science University—UL1TR002369: Oregon Clinical and Translational Research Institute • Penn State Health Milton S. Hershey Medical Center—UL1TR002014: Penn State Clinical and Translational Science Institute • Rush University Medical Center—UL1TR002389: The Institute for Translational Medicine (ITM) • Rutgers, The State University of New Jersey—UL1TR003017: New Jersey Alliance for Clinical and Translational Science • Stony Brook University—U24TR002306 • The Ohio State University—UL1TR002733: Center for Clinical and Translational Science • The State University of New York at Buffalo—UL1TR001412: Clinical and Translational Science Institute • The University of Chicago—UL1TR002389: The Institute for Translational Medicine (ITM) • The University of Iowa—UL1TR002537: Institute for Clinical and Translational Science • The University of Miami Leonard M. Miller School of Medicine—UL1TR002736: University of Miami Clinical and Translational Science Institute • The University of Michigan at Ann Arbor—UL1TR002240: Michigan Institute for Clinical and Health Research • The University of Texas Health Science Center at Houston—UL1TR003167: Center for Clinical and Translational Sciences (CCTS) • The University of Texas Medical Branch at Galveston—UL1TR001439: The Institute for Translational Sciences • The University of Utah—UL1TR002538: Uhealth Center for Clinical and Translational Science • Tufts Medical Center—UL1TR002544: Tufts Clinical and Translational Science Institute • Tulane University—UL1TR003096: Center for Clinical and Translational Science • University Medical Center New Orleans—U54GM104940: Louisiana Clinical and Translational Science (LA CaTS) Center • University of Alabama at Birmingham—UL1TR003096: Center for Clinical and Translational Science • University of Arkansas for Medical Sciences—UL1TR003107: UAMS Translational Research Institute • University of Cincinnati—UL1TR001425: Center for Clinical and Translational Science and Training • University of Colorado Denver, Anschutz Medical Campus—UL1TR002535: Colorado Clinical and Translational Sciences Institute • University of Illinois at Chicago—UL1TR002003: UIC Center for Clinical and Translational Science • University of Kansas Medical Center—UL1TR002366: Frontiers: University of Kansas Clinical and Translational Science Institute • University of Kentucky—UL1TR001998: UK Center for Clinical and Translational Science • University of Massachusetts Medical School Worcester—UL1TR001453: The UMass Center for Clinical and Translational Science (UMCCTS) • University of Minnesota—UL1TR002494: Clinical and Translational Science Institute • University of Mississippi Medical Center—U54GM115428: Mississippi Center for Clinical and Translational Research (CCTR) • University of Nebraska Medical Center—U54GM115458: Great Plains IDeA-Clinical & Translational Research • University of North Carolina at Chapel Hill—UL1TR002489: North Carolina Translational and Clinical Science Institute • University of Oklahoma Health Sciences Center—U54GM104938: Oklahoma Clinical and Translational Science Institute (OCTSI) • University of Rochester—UL1TR002001: UR Clinical & Translational Science Institute • University of Southern California—UL1TR001855: The Southern California Clinical and Translational Science Institute (SC CTSI) • University of Vermont—U54GM115516: Northern New England Clinical & Translational Research (NNE-CTR) Network • University of Virginia—UL1TR003015: iTHRIV Integrated Translational health Research Institute of Virginia • University of Washington—UL1TR002319: Institute of Translational Health Sciences • University of Wisconsin-Madison—UL1TR002373: UW Institute for Clinical and Translational Research • Vanderbilt University Medical Center—UL1TR002243: Vanderbilt Institute for Clinical and Translational Research • Virginia Commonwealth University—UL1TR002649: C. Kenneth and Dianne Wright Center for Clinical and Translational Research • Wake Forest University Health Sciences—UL1TR001420: Wake Forest Clinical and Translational Science Institute • Washington University in St. Louis—UL1TR002345:

Institute of Clinical and Translational Sciences • Weill Medical College of Cornell University—UL1TR002384: Weill Cornell Medicine Clinical and Translational Science Center • West Virginia University—U54GM104942: West Virginia Clinical and Translational Science Institute (WVCTSI) • Icahn School of Medicine at Mount Sinai—UL1TR001433: ConduITS Institute for Translational Sciences • University of California, Davis—UL1TR001860: UCDavis Health Clinical and Translational Science Center • University of California, Irvine—UL1TR001414: The UC Irvine Institute for Clinical and Translational Science (ICTS) • University of California, Los Angeles—UL1TR001881: UCLA Clinical Translational Science Institute • University of California, San Diego—UL1TR001442: Altman Clinical and Translational Research Institute • University of California, San Francisco—UL1TR001872: UCSF Clinical and Translational Science Institute • Arkansas Children's Hospital—UL1TR003107: UAMS Translational Research Institute • Baylor College of Medicine—None (Voluntary) • Cincinnati Children's Hospital Medical Center—UL1TR001425: Center for Clinical and Translational Science and Training • Loyola University Chicago—UL1TR002389: The Institute for Translational Medicine (ITM) • Medical College of Wisconsin—UL1TR001436: Clinical and Translational Science Institute of Southeast Wisconsin • MetroHealth—None (Voluntary) • NYU Langone Medical Center—UL1TR001445: Langone Health's Clinical and Translational Science Institute • Ochsner Medical Center—U54GM104940: Louisiana Clinical and Translational Science (LA CaTS) Center • Regenstrief Institute—UL1TR002529: Indiana Clinical and Translational Science Institute • University of Florida—UL1TR001427: UF Clinical and Translational Science Institute.

CONFLICT OF INTEREST STATEMENT

None declared.

DATA AVAILABILITY STATEMENT

N3C/RECOVER: The N3C Data Enclave is managed under the authority of the NIH; information can be found at ncats.nih.gov/n3c/resources. Enclave data are protected, and can be accessed for COVID-related research with an approved (1) IRB protocol and (2) Data Use Request (DUR). A detailed accounting of data protections and access tiers is found in.[1] Enclave and data access instructions can be found at https://covid.cd2h.org/for-researchers; all code used to produce the analyses in this manuscript is available within the N3C Enclave to users with valid login credentials to support reproducibility.

All of Us: To ensure privacy of participants, *All of Us* Research Program data used for this study are available to approved researchers following registration, completion of ethics training, and attestation of a data use agreement through the *All of Us* Research Workbench platform, which can be accessed via https://workbench.researchallofus.org/login.

REFERENCES

1. A Clinical Case Definition of Post COVID-19 Condition by a Delphi Consensus, 6 October 2021. 2021. https://www.who.int/publications/i/item/WHO-2019-nCoV-Post_COVID-19_condition-Clinical_case_definition-2021.1. Accessed July 18, 2022.
2. Ledford H. *How Common is Long COVID? Why Studies Give Different Answers*. London: Nature Publishing Group UK; 2022. doi:10.1038/d41586-022-01702-2.
3. 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: e532–41.
4. Hill E, Mehta H, Sharma S, *et al.* Risk factors associated with post-acute sequelae of SARS-CoV-2 in an EHR cohort: a national COVID cohort collaborative (N3C) analysis as part of the NIH RECOVER program. medRxiv. 2022;2022.08.15.22278603. doi:10.1101/2022.08.15.22278603.
5. Daniel Brannock M, Chew RF, Preiss AJ, *et al.* Long COVID risk and pre-COVID vaccination: an EHR-based cohort study from the recover program. medRxiv. 2022;2022.10.06.22280795. doi:10.1101/2022.10.06.22280795.
6. Sidky H, Sahner DK, Girvin AT, *et al.* Assessing the effect of selective serotonin reuptake inhibitors in the prevention of post-acute sequelae of COVID-19. medRxiv. 2022;2022.11.09.22282142. doi:10.1101/2022.11.09.22282142.
7. Mo H, Thompson WK, Rasmussen LV, *et al.* Desiderata for computable representations of electronic health records-driven phenotype algorithms. *J Am Med Inform Assoc* 2015; 22: 1220–30.
8. The All of Us Research Program Investigators. The 'all of us' research program. *N Engl J Med* 2019; 381: 668–76.
9. Pfaff ER, Girvin AT, Gabriel DL, *et al.* Synergies between centralized and federated approaches to data quality: a report from the national COVID cohort collaborative. *J Am Med Inform Assoc* 2021; 29: 609–18.
10. Haendel MA, Chute CG, Bennett TD, *et al.* The National COVID Cohort Collaborative (N3C): rationale, design, infrastructure, and deployment. *J Am Med Inform Assoc* 2021; 28: 427–43.
11. NCTraCSIDSci/n3c-longcovid. GitHub. https://github.com/NCTraCSIDSci/n3c-longcovid. Accessed July 26, 2022.
12. Mapes BM, Foster CS, Kusnoor SV, *et al.* Diversity and inclusion for the all of us research program: a scoping review. *PLoS One* 2020; 15: e0234962.
13. Savage N. *Breaking into the Black Box of Artificial Intelligence*. London: Nature Publishing Group UK; 2022. doi:10.1038/d41586-022-00858-1
14. [No title]. ACM Digital Library. https://doi.org/10.1145/3457607. Accessed December 13, 2022.
15. Brandt PS, Kho A, Luo Y, *et al.* Characterizing variability of electronic health record-driven phenotype definitions. *J Am Med Inform Assoc* 2022; 30: 427–37
16. Lorman V, Razzaghi H, Song X, *et al.* A machine learning-based phenotype for long COVID in children: an EHR-based study from the RECOVER program. medRxiv. 2022;2022.12.22.22283791. doi:10.1101/2022.12.22.22283791.
17. Bekker J, Davis J. Learning from positive and unlabeled data: a survey. *Mach Learn* 2020; 109: 719–60.