Investigating the Impact of Temporal Labeling of Emergency Department Visits for COVID-19: Comparing Healthcare Disparities Analyses Using Comprehensive, Single-Site Data with National COVID Cohort Collaborative (N3C) Data

Madeleine Jones
School of Data Science
University of Virginia

Aubrey Winger
School of Data Science
University of Virginia

Christian Wernz
Data Science
UVA Health

Jonathan Michel
Data Science
UVA Health

Sihang Jiang
Dept of Sys & Inf Eng
University of Virginia

Andrea Zhou
iTHRIV
University of Virginia

Ebony Jade Hilton, MD
Anesthesiology & CC Medicine
UVA Health

Mikhail Zemmel
Research & Clinical Trial Analytics
UVA Health

Saurav Sengupta
School of Data Science
University of Virginia

Kierah Barnes, DNP
Department of Surgery
UVA Health

Johanna Loomba
iTHRIV
University of Virginia

Donald E. Brown
School of Data Science
University of Virginia

N3C Consortium
Consortial contributors
are in the process of being documented

***Abstract*—National COVID Cohort Collaborative (N3C) enclave provides health researchers with a rich dataset from 76 contributing clinical sites. However, the harmonized data lacks certain details available in sites' local electronic health records (EHRs), such as the principal diagnosis code for reported emergency department (ED) and inpatient (IP) visits. This means a principal diagnosis of COVID-19 can only be inferred by applying a time relationship between the visit dates and the record of infection and diagnosis. The purpose of this study is to perform a single-site sensitivity analysis modeled after an N3C study examining potential race-ethnicity based bias in hospitalization decisions during COVID-19 related ED visits. The analytic pipeline was first run in N3C, then reproduced locally with N3C data fields from a single-site, and finally run a third time using the additional principal diagnosis data. We find the effects of patient comorbidities and race-ethnicity groups on direct IP admittance to be consistent among the three cohorts with varying levels of statistical significance due to different sample sizes.**

I. Introduction

The coronavirus disease of 2019 (COVID-19) is an illness caused by the highly transmissible betacoronavirus severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2). Between February 2020 and September 2021, the CDC estimates that 146.6 million individuals were infected by COVID-19 in the United States, resulting in approximately 921,000 deaths [1] . Several initiatives have sought to utilize the variety of data collected from the pandemic in order to further assess COVID-19, including improving our knowledge on potential risk factors, preventative measures, and long-term health consequences [2]. The National COVID Cohort Collaborative (N3C) enclave is a secure analytic environment that contains coded data from health records for over 18 million patients from 76 contributing clinical sites. Of these patients, over 7 million are confirmed to be COVID positive [3]. Within the N3C enclave, a data pipeline has been constructed that applies generalized estimating equation (GEE) modeling to assess the hospitalization outcomes of COVID-positive patients in the emergency department. This modeling was implemented to investigate inequalities in COVID-19 patient admissions from the ED and IP based on demographic features [4] [5]. The goal of this paper is to determine if the N3C enclave data processing pipeline can be applied to patient data collected from a single clinical site, and how the results of single-site GEE modeling compare to the results seen in the enclave. The enclave's primary limitation is that the data does not include patient principal diagnosis, and thus a principal diagnosis of COVID-19 can only be approximated by temporal means. Single-site data, however, includes principal diagnosis information, as well as information that indicates whether a patient was directly admitted from the ED. Therefore, this work aims to evaluate the pipeline's performance on single-site data by incorporating the principal diagnosis and reason for ED visit/hospitalization information.

II. Related Work

Many statistical and machine learning models have been created to investigate the disparities in treatment among COVID-19 patients. Biological vulnerabilities such as gender,

TABLE I
BREAKDOWN OF DEMOGRAPHICS, COMBORBIDITIES, AND HOSPITALIZATION STATISTICS BY COHORT

	Number of Patients		
	UVA Data (Temporal Diagnosis) 9841 Patients	**UVA Data (Principal Diagnosis) 759 Patients**	**N3C Enclave Data 681889 Patients**
Demographics			
Age (65 or Older)	2606 (26.5%)	220 (29.0%)	147239 (21.6%)
Gender			
Male	4666 (47.4%)	390 (51.4%)	308646 (45.3%)
Female	5175 (52.6%)	369 (48.6)	373243 (54.7%)
Race/Ethnicity			
Asian NH	604 (6.2%)	23 (3.0%)	19645 (2.9%)
Black/African American NH	2689 (27.3%)	209 (27.5%)	186571 (27.3%)
White NH	6548 (66.5%)	527 (69.5%)	475673 (69.8%)
Comorbidities Prior or During COVID Index Date			
Complicated Diabetes	956 (9.7%)	144 (19.0%)	63132 (9.3%)
HIV Positive	42 (0.4%)	4 (0.5%)	3267 (0.5%)
Congestive Heart Failure	373 (3.8%)	28 (3.7%)	31048 (4.6%)
Malignant Cancer	761 (7.7%)	93 (12.2%)	39928 (5.9%)
Obesity	3061 (31.1%)	383 (50%)	224002 (33.9%)
Chronic Lung Disease	1406 (14%)	158 (20.8%)	104717 (15.4%)
Hospitalization after ED			
Overall	2175 (22%)	342 (45%)	104627 (15.3%)
By Race/Ethnicity			
Asian NH	60 (9.9%)	6 (26.1%)	3935 (20.0%)
Black/African American NH	533 (19.8%)	97 (46.4%)	20754 (11.1%)
White NH	1582 (24.2%)	239 (45.3%)	79938 (16.8%)

age, obesity, and certain medical comorbidities have all been identified as contributing to severe disease complications and death [2]. Additionally, research has found that minority individuals, especially those of African American, American Indian or Alaska Native, and Hispanic or Latino descent, were of the highest risk to contract SARS-CoV-2, and to develop COVID-19 illness that results in a fatality [6]. Another study conducted in 2022 with 43,222 adult veterans using a multivariable random effects logistic regression model found that African American patients had lower odds of receiving COVID-19 specific treatments [7]. A similar logistic regression study on 505,992 patients concluded that African Americans have a higher mortality rate with COVID-19 that is not explainable by differences in age, comorbidities, or social determinant of health variables [4]. The work described in this paper is a direct extension of previous modeling in the N3C enclave conducted to analyze disparities between race-ethnicity groups in patient hospitalization after a COVID-associated ED visit [5]. Generalized estimating equations general linear modeling (GEE GLM) were used to account for unknown correlations between health care sites and predictor variables. Analysis of the cohort revealed that Asian had a higher probability of being hospitalized after an ED visit and African American patients had a lower probability of being hospitalized after an ED visit as compared to White patients. The work presented in this paper builds upon the previous modeling by applying similar GLM models to single-site data.

III. DATA

The data for this study comes from two sources. The first is the National COVID Cohort Collaborative (N3C), a secure enclave that currently provides controlled access to harmonized clinical data from 76 sites in order to advance COVID-19 research and effectively prepare for subsequent global health emergencies [3]. N3C hosts a high performing cloud computing platform, referred to as the N3C enclave, which allows researchers to conduct analyses on patient data in a secure environment. This study, which utilizes patient data from January 2018 to December 2022, was cleaned and prepared through the N3C Logic Liaison's COVID-19 Diagnosed or Lab Confirmed Patients template, a PySpark pipeline that is available in the N3C Knowledge Store and also publicly in GitHub [8]. Per N3C community practices for identifying the first COVID-19 associated ED or patient admission was tagged as COVID-related if ED or hospital admission occurred in the 16 days after or 1 day prior to a positive PCR or AG test. Additionally, ED to IP visits flags are generated using both temporal admission data and the macro-visit patient table, since the actual encounter type is not included in N3C patient data. The data from the N3C enclave formed the first cohort used in this analysis with a total of 681,889 patients (Table I). The second source of data was a single clinical site which is a data contributing member to N3C, the University of Virginia (UVA) Health System. This dataset, consisting of the same Observational Medical Outcomes Partnership (OMOP) data tables for UVA

COVID-positive patients that are submitted to N3C, was de-identified by a local data steward and uploaded to Rivanna, UVA's secure computing platform. In order to transform the UVA sourced data into cohort tables matching the enclave format, the original N3C Logic Liaison PySpark pipeline was converted to Python to be used outside the enclave. The Python pipeline (described in Section IV) can be used to analyze any standardized clinical OMOP data. The UVA clinical data was successfully augmented through the Python pipeline and retained the N3C-defined temporal tagging of COVID-related ED and hospital visits. This temporally tagged UVA Health data formed the second cohort used in this analysis with 9,841 total patients. Finally, because the UVA patient data included records of COVID-19 principal diagnosis and ED to IP admission, this additional data was used to generate the COVID-related ED and IP visit tags without relying on temporal inference. This data augmentation reveals how true principal diagnosis information narrows the pool of COVID-related ED to IP visits. The patient visits tagged were those with both COVID-19 as a principal diagnosis and dated within the CDC defined timeframe of the COVID-19 positivity index date. This ensures that the analyses remain analogous in that tagged visits occur at the time of the initial COVID infection, and are not related to a re-infection. The UVA data tagged using principal diagnosis information formed the third cohort used in this analysis with a total of 759 patients (Table I). For both of the cohorts using UVA patient data, ED to immediate IP admission is determined using the recorded encounter type. Since the patient records for UVA provide an actual encounter type it does not have to be approximated, unlike in the N3C enclave. These two cohorts also fall under UVA Health institutional policy, so analyzing groups with less than 20 patients is permissible. All N3C data used has at least 20 patients in each group, per N3C policy.

IV. Methods

The N3C Logic Liaison pipeline is a series of functions that are used to create a patient-level summary table of COVID-positive patients. The pipeline generates default derived variables for 29 pre-COVID, during-COVID, and post-COVID comorbidities and critical events. Figure 1 shows the architecture of the N3C Logic Liaison pipeline. Figure 2 shows the implementation of the pipeline to obtain cohorts from the data. Several study-specific variables were created to indicate various ED outcomes of interest. All COVID-19 associated ED visits (based on the analysis specific definition) and ED visits that resulted in immediate hospitalization were tagged for use in modeling patient outcomes. Variables were also created for each comorbidity in the cohorts, with a 1 indicating the presence of the comorbidity in the patient and a 0 indicating the absence. The presence of comorbidities and the proportion of racial and ethnic groups within each of the three patient cohorts were compared. Comorbidities of interest were determined based on those used in previous modeling work [5]. Across the three cohorts, the Black/African

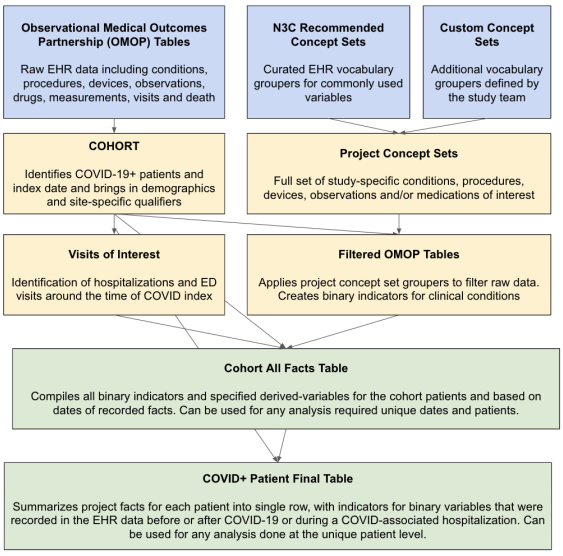

Fig. 1. The Logic Liaison Pyspark pipeline and Python pipeline are identical. The Python pipeline was adapted only for Cohort 3 at the Visits of Interest node to create COVID-associated ED and IP visits based on principal diagnosis rather than temporal approximation.

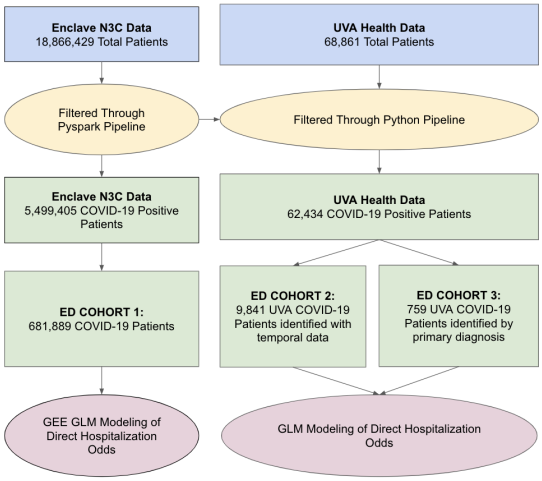

Fig. 2. The pipelines were applied to their respective data sources to create the ED Cohorts to be used for modeling.

American non-Hispanic population has the highest average number of comorbidities per patient (Figure 3). Further, the UVA cohort tagged with principal diagnosis has the highest average comorbidity number per patient (Figure 3). Finally, the proportions of racial-ethnic groups and the distributions of age, gender, and comorbidities across the racial-ethnic groups are alike for all three cohorts (Table I) (Figure 4).

V. Modeling

GLM modeling is used to predict if patients are admitted directly to the hospital from the ED on the three COVID-positive cohorts: N3C enclave patients, UVA patients tagged temporally, and UVA patients tagged with principal diag-

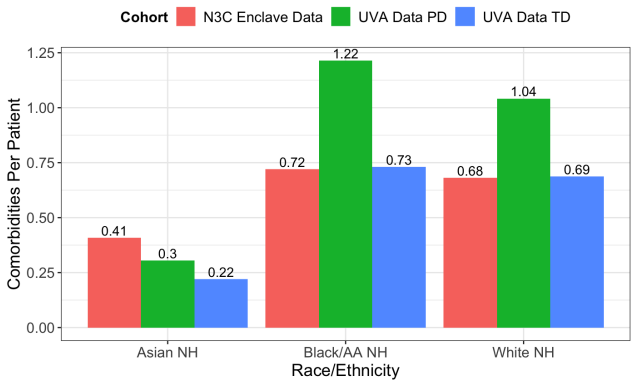

Fig. 3. Average Comorbidities per Patient across Cohorts and Racial/Ethnic Groups. PD indicates principal diagnosis tagging. TD indicates temporal inference tagging.

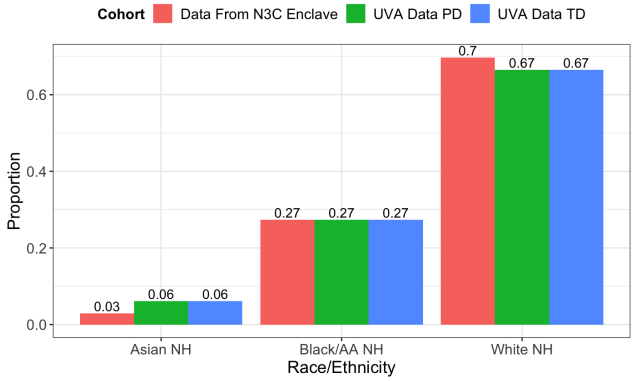

Fig. 4. Proportion of Race/Ethnicity Groups per Cohort. PD indicates principal diagnosis tagging. TD indicates temporal inference tagging.

nosis. The N3C enclave cohort uniquely implements GEE GLM to account for site clustering in the data. GLM is utilized because it allows the response variable to have an arbitrary distribution, rather than imposing a normality assumption. Since the ED to hospitalization outcome is binary, a Bernouilli distribution is best suited for this analysis. Additionally, GLM allows an arbitrary function of the response variable to vary linearly with the predictors. In this case, a logit link function is used [9]. The GLM parameters include patient comorbidities, age, gender, and racial-ethnic group. The comorbidities remained consistent with those identified in the previous modeling work, and included diabetes, HIV infection, congestive heart failure, malignant cancer, obesity, and chronic lung disease [5]. Racial-ethnic groups are used as model parameters to determine whether they play a significant role in patient IP admission; however, the groups 'American Indian or Alaskan Native' and 'Native Hawiian or Other Pacific Islander' were both removed because the few patient records are not representative of the larger ethnic group. Only patients with records of age, gender, and ethnic/racial group are included in the model. Forest plots of the parameters' odds ratios were created to investigate significance. The odds ratios statistically quantify the association between direct admittance and each variable (Figure 5 6 7). An odds ratio of 1 indicates that the likelihood of direct IP admittance is not changed by a variable's presence or absence in a patient. The odds ratios are assessed using a significance level of 0.05, or equivalently, 95 percent confidence in the variables' estimated odds intervals. Additionally, a Pearson's Chi-squared test is used to assess whether the proportions of direct IP admittance across racial-ethnic groups are significantly different for the two UVA cohorts. The ethnic-racial groups analyzed are the same as those in the GEE modeling: Black/African American Non-Hispanic, Asian Non-Hispanic, and White Non-Hispanic.

VI. Results

The odds ratios for the three GEE models' variables are assessed to determine whether age, gender, racial-ethnic group, or comorbidities are significant in a patient's direct admittance from the ED to the hospital. As expected, the confidence intervals and significance levels for each of the models' variables are indicative of the respective cohort sizes. The model for N3C enclave data has the greatest variable significance due to its large cohort size (Figure 7). The model for UVA data tagged temporally has moderate variable significance due to its intermediate cohort size (Figure 5). The model for UVA data tagged with principal diagnosis has limited variable significance due to its small cohort size (Figure 5).

Across the three cohorts, the point-estimates for each variable are consistent. The variables which increase the likelihood of direct IP admission are being male or older in age, or having any of the comorbidities (malignant cancer, HIV infection, diabetes, congestive heart failure, and chronic lung disease). The Black/ African American non-Hispanic racial-ethnic group has decreased likelihood of direct IP admission. The only difference between models is in the Asian non-Hispanic ethnic-racial group which has increased odds of direct IP admittance in the N3C enclave cohort and decreased odds in the UVA cohorts. The Pearson's Chi-squared test further investigates the proportion of direct IP admittance among racial and ethnic groups. Similar to the GEE model results, direct IP admittance varies significantly among the three racial-ethnic groups for the UVA cohort tagged temporally, but not for the UVA cohort tagged with principal diagnosis. For the UVA cohort tagged temporally, there is less than a 0.1 percent chance of obtaining such varying proportions among racial-ethnic group direct IP admittance if they are truly all equivalent.

VII. Conclusions and Future Work

The work described in this paper demonstrates that the large-scale pipeline in N3C built to process millions of COVID-19 patient records across multiple sites can successfully be converted from PySpark to Python and applied to data from a single clinical site. When the pipeline is run on single-site data from the University of Virginia Health system

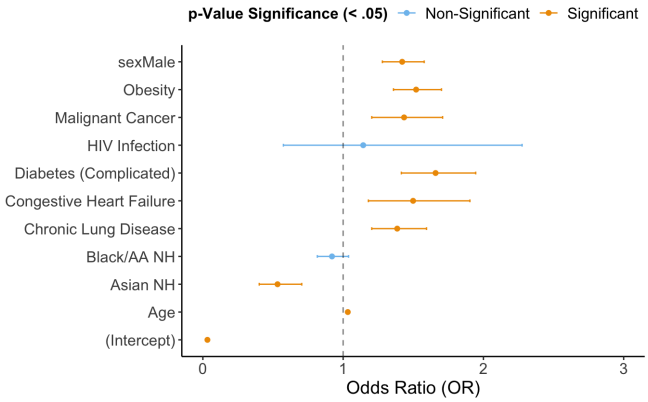

Fig. 5. Hospitalization after ED: Predictor Odds Ratios for UVA ED Cohort (Temporal Diagnosis)

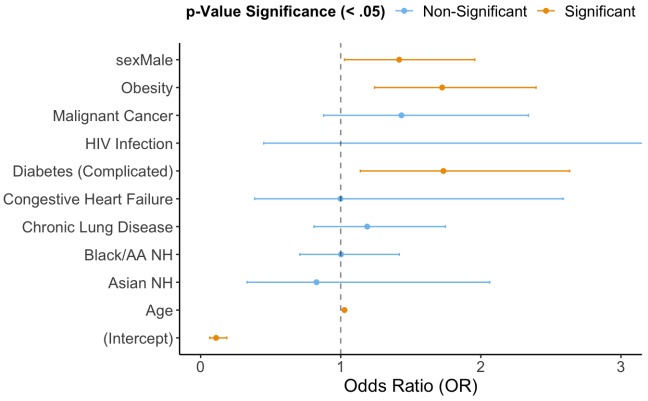

Fig. 6. Hospitalization after ED: Predictor Odds Ratios for UVA ED Cohort (Principal Diagnosis)

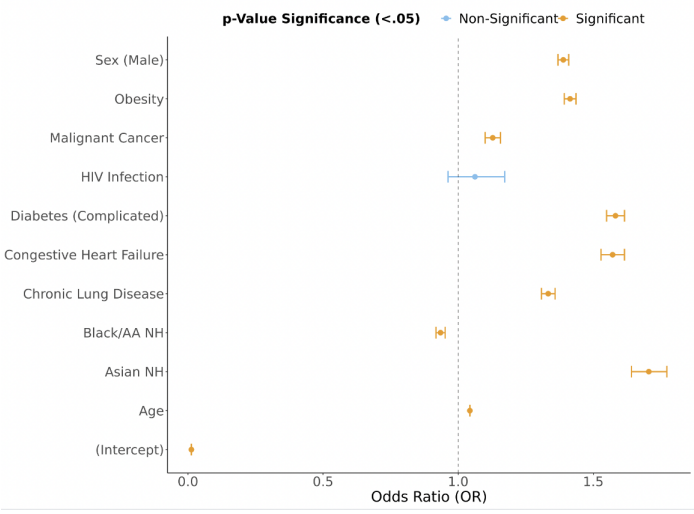

Fig. 7. Hospitalization after ED: Predictor Odds Ratios for N3C ED Cohort

and GEE GLM modeling is implemented, the impact of patient characteristics and comorbidities on direct IP admission is consistent with those in the N3C enclave. The pipeline has further been adapted to utilize information, such as principal diagnosis and ED to IP admittance, which is available in local electronic health records. This data augmentation has refined the pipeline by removing temporal approximations of COVID-19 diagnosis, subsequently decreasing the number of COVID-positive patient visits. The impact of patient characteristics and comorbidities on direct IP admittance are consistent between the single-site UVA COVID-19 ED cohorts tagged temporally and tagged using principal diagnosis. For all three cohorts, patients who are male, older, obese, or have malignant cancer, HIV infection, diabetes, congestive heart failure, or chronic lung disease are more likely to be directly admitted. Patients who are Black/African American non-Hispanic are less likely to be admitted. This is in spite of Black/African American non-Hispanic patients having the greatest number of comorbidities on average for each of the cohorts. Differences in statistical significance among the variables are due to the smaller sample size of the single-site data. This is evidenced by all of the parameters aside from HIV positivity having significance in the N3C enclave analysis. Overall, while the effects of patient characteristics and comorbidities can be speculated using single-site data, the vast amount of information in the N3C enclave is necessary to determine whether the effects are significant. Therefore, while single-site data provide important data enrichments that the enclave data currently lacks, the enclave data provide the necessary data volume to derive statistically significant findings. If contributing sites could provide primary diagnosis data to N3C, analyses investigating COVID-related ED and hospital visits could be refined to leverage this data in the larger multi-center dataset. Future work can evaluate groups of interest including patients with social determinants of health variables, as well as patients with dementia, pregnant patients, and patients that are experiencing housing insecurity. Further, aside from the current analysis of direct IP admittance, the pipeline can be utilized to analyze disparities in treatment types and subsequent patient outcomes for hospitalized patients. Overall, this pipeline is a tool that can be utilized to investigate factors contributing to COVID-19 patient outcomes for OMOP data sources outside of the N3C enclave.

Acknowledgments

N3C Attribution

The analyses described in this report were conducted with data or tools accessed through the NCATS N3C Data Enclave https://covid.cd2h.org and N3C Attribution & Publication Policy v 1.2-2020-08-25b supported by NCATS U24 TR002306, Axle Informatics Subcontract: NCATS-P00438-B, and iTHRIV Integrated Translational health Research Institute of Virginia UL1TR003015. This research was possible because of the patients whose

information is included within the data and the organizations (https://ncats.nih.gov/n3c/resources/data-contribution/data-transfer-agreement-signatories) and scientists who have contributed to the on-going development of this community resource. [3]

Disclaimer

The N3C Publication committee confirmed that this manuscript MSID:1023.31 is in accordance with N3C data use and attribution policies; however, this content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health or the N3C program.

IRB

The N3C data transfer to NCATS is performed under a 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.

Individual Acknowledgements For Core Contributors

We gratefully acknowledge the following core contributors to N3C: Adam B. Wilcox, Adam M. Lee, Alexis Graves, Alfred (Jerrod) Anzalone, Amin Manna, Amit Saha, Amy Olex, Andrea Zhou, Andrew E. Williams, Andrew Southerland, Andrew T. Girvin, Anita Walden, Anjali A. Sharathkumar, Benjamin Amor, Benjamin Bates, Brian Hendricks, Brijesh Patel, Caleb Alexander, Carolyn Bramante, Cavin Ward-Caviness, Charisse Madlock-Brown, Christine Suver, Christopher Chute, Christopher Dillon, Chunlei Wu, Clare Schmitt, Cliff Takemoto, Dan Housman, Davera Gabriel, David A. Eichmann, Diego Mazzotti, Don Brown, Eilis Boudreau, Elaine Hill, Elizabeth Zampino, Emily Carlson Marti, Emily R. Pfaff, Evan French, Farrukh M Koraishy, Federico Mariona, Fred Prior, George Sokos, Greg Martin, Harold Lehmann, Heidi Spratt, Hemalkumar Mehta, Hongfang Liu, Hythem Sidky, J.W. Awori Hayanga, Jami Pincavitch, Jaylyn Clark, Jeremy Richard Harper, Jessica Islam, Jin Ge, Joel Gagnier, Joel H. Saltz, Joel Saltz, Johanna Loomba, John Buse, Jomol Mathew, Joni L. Rutter, Julie A. McMurry, Justin Guinney, Justin Starren, Karen Crowley, Katie Rebecca Bradwell, Kellie M. Walters, Ken Wilkins, Kenneth R. Gersing, Kenrick Dwain Cato, Kimberly Murray, Kristin Kostka, Lavance Northington, Lee Allan Pyles, Leonie Misquitta, Lesley Cottrell, Lili Portilla, Mariam Deacy, Mark M. Bissell, Marshall Clark, Mary Emmett, Mary Morrison Saltz, Matvey B. Palchuk, Melissa A. Haendel, Meredith Adams, Meredith Temple-O'Connor, Michael G. Kurilla, Michele Morris, Nabeel Qureshi, Nasia Safdar, Nicole Garbarini, Noha Sharafeldin, Ofer Sadan, Patricia A. Francis, Penny Wung Burgoon, Peter Robinson, Philip R.O. Payne, Rafael Fuentes, Randeep Jawa, Rebecca Erwin-Cohen, Rena Patel, Richard A. Moffitt, Richard L. Zhu, Rishi Kamaleswaran, Robert Hurley, Robert T. Miller, Saiju Pyarajan, Sam G. Michael, Samuel Bozzette, Sandeep Mallipattu, Satyanarayana Vedula, Scott Chapman, Shawn T. O'Neil, Soko Setoguchi, Stephanie S. Hong, Steve Johnson, Tellen D. Bennett, Tiffany Callahan, Umit Topaloglu, Usman Sheikh, Valery Gordon, Vignesh Subbian, Warren A. Kibbe, Wenndy Hernandez, Will Beasley, Will Cooper, William Hillegass, Xiaohan Tanner Zhang. Details of contributions available at covid.cd2h.org/core-contributors.

REFERENCES

[1] M. K. Steele, A. Couture, C. Reed, D. Iuliano, M. Whitaker, H. Fast, A. J. Hall, A. MacNeil, B. Cadwell, K. J. Marks *et al.*, "Estimated number of covid-19 infections, hospitalizations, and deaths prevented among vaccinated persons in the us, december 2020 to september 2021," *JAMA Network Open*, vol. 5, no. 7, pp. e2 220 385–e2 220 385, 2022.

[2] S. Garg, L. Kim, M. Whitaker, A. O'Halloran, C. Cummings, R. Holstein, M. Prill, S. J. Chai, P. D. Kirley, N. B. Alden *et al.*, "Hospitalization rates and characteristics of patients hospitalized with laboratory-confirmed coronavirus disease 2019—covid-net, 14 states, march 1–30, 2020," *Morbidity and mortality weekly report*, vol. 69, no. 15, p. 458, 2020.

[3] M. A. Haendel, C. G. Chute, T. D. Bennett, D. A. Eichmann, J. Guinney, W. A. Kibbe, P. R. Payne, E. R. Pfaff, P. N. Robinson, J. H. Saltz *et al.*, "The national covid cohort collaborative (n3c): rationale, design, infrastructure, and deployment," *Journal of the American Medical Informatics Association*, vol. 28, no. 3, pp. 427–443, 2021.

[4] L. Golestaneh, J. Neugarten, M. Fisher, H. H. Billett, M. R. Gil, T. Johns, M. Yunes, M. H. Mokrzycki, M. Coco, K. C. Norris *et al.*, "The association of race and covid-19 mortality," *EClinicalMedicine*, vol. 25, p. 100455, 2020.

[5] E. Assefa, S. Scott-Dixon, D. Hanson, T. Musa, S. Sharma, C. Madlock-Brown, J. Phuong, J. Loomba, and D. Brown, "Racial/ethnic disparities in covid treatment/hospitalization: Social determinants & special populations," Dec 2022. [Online]. Available: https://doi.org/10.18130/mv6g-pg57

[6] K. Mackey, C. K. Ayers, K. K. Kondo, S. Saha, S. M. Advani, S. Young, H. Spencer, M. Rusek, J. Anderson, S. Veazie *et al.*, "Racial and ethnic disparities in covid-19–related infections, hospitalizations, and deaths: a systematic review," *Annals of internal medicine*, vol. 174, no. 3, pp. 362–373, 2021.

[7] A. D. Castro, F. B. Mayr, V. B. Talisa, O. S. Shaikh, S. B. Omer, S. Yende, and A. A. Butt, "Variation in clinical treatment and outcomes by race among us veterans hospitalized with covid-19," *JAMA Network Open*, vol. 5, no. 10, pp. e2 238 507–e2 238 507, 2022.

[8] J. Loomba, S. Ljazouli, A. Zhou, S. Sharma, E. Asseffa, and E. French, "Logic liaison confirmed covid positive template," 2023. [Online]. Available: https://github.com/National-COVID-Cohort-Collaborative/Logic-Liaison-Confirmed-COVID-Positive-Template

[9] J. W. Hardin and J. M. Hilbe, *Generalized estimating equations*. CRC press, 2012.