BMJ Open

Personalized Prediction of Viral Suppression among Underrepresented Population: protocol for a longitudinal cohort study

Journal:	*BMJ Open*
Manuscript ID	bmjopen-2022-070869
Article Type:	Protocol
Date Submitted by the Author:	07-Dec-2022
Complete List of Authors:	Zhang, Jiajia; University of South Carolina Arnold School of Public Health, Department of Epidemiology and Biostatistics; University of South Carolina Arnold School of Public Health, South Carolina SmartState Center for Healthcare Quality Yang, Xueying; University of South Carolina Arnold School of Public Health, Health Promotion Education and Behavior Olatosi, Bankole; University of South Carolina Arnold School of Public Health, Health Services Policy and Management; University of South Carolina Arnold School of Public Health, South Carolina SmartState Center for Healthcare Quality Weissman, Sharon; University of South Carolina, Department of Internal Medicine, School of Medicine Li, Xiaoming; University of South Carolina Arnold School of Public Health; University of South Carolina Arnold School of Public Health, South Carolina SmartState Center for Healthcare Quality
Keywords:	COVID-19, HIV & AIDS < INFECTIOUS DISEASES, PUBLIC HEALTH, STATISTICS & RESEARCH METHODS

Personalized Prediction of Viral Suppression among Underrepresented Population: protocol for a longitudinal cohort study

Jiajia Zhang [1,2], Xueying Yang [1,3]*, Bankole Olatosi [1,4], Sharon Weissman [1,5], Xiaoming Li [1,3]

[1] South Carolina SmartState Center for Healthcare Quality, Arnold School of Public Health, University of South Carolina, Columbia, SC, USA, 29208

[2] Department of Epidemiology and Biostatistics, Arnold School of Public Health, University of South Carolina, Columbia, SC, USA, 29208

[3] Department of Health Promotion, Education and Behavior, Arnold School of Public Health, University of South Carolina, Columbia, SC, USA, 29208

[4] Department of Health Services Policy and Management, Arnold School of Public Health, University of South Carolina, Columbia, SC, USA, 29208

[5] Department of Internal Medicine, School of Medicine, University of South Carolina, Columbia, SC, USA, 29208

*Corresponding author:

Bankole Olatosi PhD,

915 Greene Street, Suite 303B, Columbia, SC 29016

Email: olatosi@mailbox.sc.edu Telephone:+18037779865

Abstract: 340

Word Count: 3518

Abstract

Introduction Sustained viral suppression, an indicator of long-term treatment success and mortality reduction, is a strategic area for the "*Ending the HIV Epidemic (EtHE)*: *A Plan for America*". Underrepresented populations, like racial/ethnic minority populations, sexual/gender minority groups, and disadvantaged populations, are disproportionately affected by HIV and subsequently experience virological failure. COVID-19 pandemic interruptions to healthcare might magnify the risk of incomplete viral suppression among underrepresented Persons living with HIV (PLWH). Unfortunately, underrepresented populations are rarely included in biomedical research resulting in biased findings. This proposal targets a broadly defined underrepresented HIV population. It aims to develop a personalized viral suppression prediction model using machine learning techniques by incorporating multilevel factors using All of Us (AoU) data.

Methods and analysis This cohort study will use data from AoU Research Program, which recruits a diverse group of populations historically underrepresented in biomedical research. The AoU Research Program harmonizes data from multiple sources regularly. It currently has ~ 4800 PLWH with a series of self-reported survey data (e.g., Lifestyle, Healthcare Access, COVID-19 Participant Experience) and relevant longitudinal EHR data (laboratory and medication). We will examine changes in viral suppression and develop personalized viral suppression predictions due to the impact of the COVID-19 pandemic using machine learning techniques (ML), like tree-based classifiers (Classification and Regression Trees (CART), Random Forest (RF), Decision Tree, and eXtreme Gradient Boosting [XGBoost]), Support Vector Machine (SVM), Naïve Bayes (NB), and Long short-term memory (LSTM). The prediction accuracy performance for each method will be evaluated using a confusion matrix.

Ethics and dissemination This study was approved by the Institutional Review Board at the University of South Carolina (Pro00124806) as a Non-Human Subject study. Study findings will be published in peer-reviewed journals and disseminated at national and international conferences and through social media.

Keywords: HIV/AIDS, Viral Suppression, COVID-19, Machine Learning

Article Summary

Strengths and limitations of this study

- The diverse group of populations recruited in the All of Us research program enables us access and examine a representative large sample of underrepresented populations in biomedical research to reduce algorithmic bias.
- The data integration from multiple data sources in All of Us research program for cohort analyses allows us to generate a robust evaluation of the viral suppression prediction for underrepresented population with a long follow up.
- The machine-learning based approach to developing personalized prediction for viral suppression has the benefit of accurately modeling a different data structure many risk factors.
- We expect missing data in both electronic health record data and survey results; therefore, caution may be needed when interpreting the risk prediction results.

Introduction

Sustained viral suppression, an indicator of long-term treatment success and mortality reduction,[1] is one of four strategic areas of the "*Ending the HIV Epidemic (EtHE)*: *A Plan for America*"[2] federal campaign launched in 2019. According to the Centers for Disease Control and Prevention national surveillance data, approximately 66% of all people living with HIV (PLWH) were virally suppressed in the United States (US).[3] The COVID-19 pandemic is affecting PLWH in unique ways and has a mixed impact on viral suppression across different countries or settings. In the US, a few studies revealed the decreased probability of viral suppression due to the negative impact of the pandemic,[4, 5] but one study in San Francisco did not report the same findings.[6] Similar inconsistent results were shown in European and Asian studies.[7, 8] The mixed results might be caused by small sample sizes, lack of sample diversity, and/or insufficient phenotypic data.

Individuals with inadequate access to medical care, low household incomes, low education attainment, and racial or sexual and gender minorities are often underrepresented in biomedical research (hereafter as 'underrepresented population').[9] HIV and COVID-19 both have disproportionate impact on underrepresented populations. For instance, 45% of new HIV infections were among gay and bisexual men under the age of 35 and 26% among Black gay and bisexual men.[10] Moreover, these vulnerable populations experience a more striking virological failure.[3] The United Nations' [11] report has indicated that increases in food costs and market stockpiling during the COVID-19 pandemic have had the most harmful impact on underrepresented populations. Among them, those with stigmatized or marginalized intersecting identities often experience the highest HIV burden, including men who have sex with men, transgender women, people who inject drugs, commercial sex workers, and youths, who account for a third of all new HIV infections.[12] Thus, the pandemic might magnify the risk of incomplete viral suppression among the underrepresented PLWH population due to interruptions in health care access and other worsened socio-economic and environmental conditions.

The increasing availability of electronic health records (EHRs) has presented the opportunity to discover new knowledge via extensive data linkage and integration. However, as a real-world clinical routine data, EHR data is not designed for a specific research purpose. Thus, it has limited capacity to recruit an adequate sample of underrepresented populations due to their historically limited access to specialty care and academic medical centers that serve as the primary sources for EHR data. Consequently, it poses more challenges in understanding the viral suppression among underrepresented populations particularly facing the COVID-19 pandemic.

The All of Us (AoU) Research Program is an ongoing national, historic effort supported by the NIH. The cohort in AoU includes a broad diverse group of the US population with more than 50% of the

participants from racial and ethnic minority groups and more than 80% are from populations historically underrepresented in biomedical research (e.g., sex orientation, socioeconomic status, geographical location, physical disability). The variables collected include longitudinal observations of clinical, environmental, lifestyle, and genetic data. Therefore, this protocol aims to target underrepresented populations using AoU data, which includes ~4800 PLWH with a series of self-reported survey data (e.g., Lifestyle, Physical Measurement, Healthcare Access, COVID-19 Participant Experience) and relevant longitudinal EHR data (laboratory and medication). With the data integration, the current exploratory study has the following specific aims:

Aim 1: Examine the impact of the COVID-19 pandemic on the viral suppression among a broadly defined underrepresented HIV population by harnessing the AoU big data resources.

Aim 2: Develop personalized viral suppression prediction models using machine learning techniques by incorporating COVID-19 interruption, antiretroviral therapy history, preexisting conditions (comorbidities), psychological wellbeing (e.g., depression, resilience), healthcare utilization, and social environmental factors in AoU.

The availability of comprehensive phenotypic data and Researcher Workbench in AoU platform fully assure the transparency and reproducibility of the proposed project. A deeper understanding of the impact of the pandemic on viral suppression among underrepresented PLWH populations is essential to promote health equity and better direct clinical management and guideline development. The proposed personalized viral suppression prediction can provide data driven evidence on tailored HIV treatment strategies to different underrepresented populations, particularly in the face of the unexpected interruptions like the COVID-19 pandemic. Thus, the results could facilitate the clinical identification of PLWH among underrepresented populations with poor viral control, provide them with tailored HIV care management, and eventually serve towards the goal of ending the HIV epidemic in the US.

Methods and Analysis

Overview of the Study Design

To guide our proposed research, we have developed a conceptual framework (**Figure 1**) that depicts how we harness the comprehensive phenotypic data from different domains of AoU Researcher Workbench to achieve the Specific Aims. The cohort building and outcomes will be defined from EHR data and survey data. For example, the intrapersonal factors (Level 1) including demographic characteristics (e.g., age, race, and gender) and overall health will be extracted from *"The Basics"* survey. The COVID-19 related experiences (Level 2 & 5), referred to the impact of the pandemic on their health and psychosocial wellbeing, such as social support, depression, anxiety, drug and alcohol abuse, and resilience, will be extracted from *"COVID-19 Participant Experience (COPE)"* and *"Lifestyle"* surveys. The neighborhood

level factors (Level 3) including the neighborhood economic environment (e.g., poverty, education, health insurance coverage) and health care access (type of healthcare facility, structural barriers to healthcare access) will be defined from *"Healthcare access & utilization"* survey. With the appropriate data management/preprocessing, we will examine the change in viral suppression and develop the personalized viral suppression prediction due to the impact of the COVID-19 pandemic using machine learning techniques (ML), which will have translational potential to inform future HIV care among underrepresented populations.

Data sources

Overview of the AoU program

The AoU Research Program seeks to recruit persons in demographic categories that have been and continue to be underrepresented in biomedical research; such persons typically have relatively poor access to good health care.[13] AoU opened for enrollment in May 2018 and the inclusion criteria are age ≥18 years and has capacity to provide consent. The recruitment methods and scientific rationale for AoU have been described previously.[13-14] Until November 19, 2021, AoU has harmonized from over 340 institutional sites contributing data for about 331,360 participants using the Observational Medical Outcomes Partnership (OMOP) Common Data Model. We anticipate an ample size to conduct the proposed analysis since AoU is harmonizing data on an ongoing basis. Each participant completed informed consent for sharing their EHR data with the Data and Research Center and provided survey responses across different domains. Each participating institutional site contributes demographics, medications, laboratory tests, diagnoses, and vital status to the central data repository for data harmonization. The AoU protocol and materials have been approved by a dedicated institutional review board, the AoU Institutional Review Board. Deidentified data were shared through the AoU Researcher Workbench (www.allofus.nih.gov) for analyses through institutional data use agreements. All analyses will be conducted within a secure informatic workspace provided by the National Institutes of Health that allows users to access and analyze a centralized version of the AoU data.

"HIV and COVID-19" Project in the AoU Researcher Workbench Platform

AoU Research Program data in its final format, after harmonization and refinement, are referred to as a curated dataset. Three different levels of information are available: Public tier, registered tier, and controlled tire. We have obtained access to data at the registered tier. Following the AoU instructions, we have created a project entitled as "HIV and COVID-19" in the AoU Researcher Workbench platform, a cloud-based platform that enables researchers to cluster participants into cohorts, select certain health information within each cohort, and perform direct analysis and query using R (R Foundation for Statistical Computing) and Python 3.0 (Python Software Foundation) programming languages within

Jupyter Notebooks. The purpose of our Workspace is 2-fold: 1) Cohort building: to determine the data inclusion and exclusion criteria for HIV cohort building (computable phenotype) and create and maintain a set of scripts to execute the computable phenotype and extract relevant data for this cohort; and 2) Model building: to examine the impact of COVID-19 on HIV and its potential predictors and build the prediction model for viral suppression.

Cohort Building and Data Extraction in AoU. Given our understanding of disease signs and symptoms, we will define computable phenotypes that can accurately identify both study cohort (e.g., HIV population) and relevant variables (e.g., COVID-19 infection) from their EHR data and survey data (**Figure 2**). The EHR data derived from captured data including billing codes and encounter records will be used to cluster participants into disease cohorts based on Systemized Nomenclature of Medicine—Clinical Terms diagnosis codes (the standardized vocabulary in AoU sourced from corresponding International Classification of Diseases codes [ICD]), whereas some other data will be derived from survey responses. Examples of the surveys can be found through the publicly available Data Browser.[15] Both survey data and EHR data are mapped to the OMOP common data model version 5.2. The data extraction will be performed on EHR domains and survey results that are available via the AoU Researcher Workbench.

HIV Cohort. To build the HIV cohort, we will adopt the existing inclusion criteria and code sets from a number of organizations-for example, PCORnet,[16] OHDSI,[17] LOINC[18] etc.--- into a "best-of-breed" phenotype and extract data from both EHR and survey questionnaires. In EHR data, HIV will be defined by documentation of any of the following: (1) HIV condition (*ICD, Ninth/Tenth Revision (ICD-9/10)* diagnostic codes, ICD-9/10 procedure codes) in the '*Condition*' domain; (2) HIV-related laboratory results (e.g., HIV antibody) in the '*Labs &Measurements'* domain; or (3) HIV-related medications (e.g., Tenofovir disoproxil) excluding pre-exposure prophylaxis in the '*Drug Exposures*' domain. In the survey data, HIV will be defined by answering affirmatively to the following questions: "*Has a doctor or health care provider ever told you that you have or had any of the following infectious disease?*" or "*Are you currently prescribed medications and/or receiving treatment for HIV/AIDS?" i*n the "*Personal Medical History*" survey. Individuals who answered Yes to "*Infectious Disease Condition: HIV/AIDS*" or "*HIV/AIDS Currently*" will be counted as HIV population. Patients who meet at least one of these inclusion criteria in either EHR data or survey data and patients who meet all of these inclusion criteria will be calculated and compared with other nationwide initiatives to develop precision rule-based algorithms for use in data analysis. A template of concept sets [19] based on all the above information will be built for HIV cohort.

COVID-19 Cohort. AoU study participants in all 50 US states provided blood specimens since January

2020 for COVID-19 testing. Similar to defining HIV population, COVID-19 patients will be identified by both EHR data and survey data. *In the EHR data*, the COVID-19 positive cases will be defined as patients with any encounter on or after 1/1/2020 with either: 1) a positive result for one of a set of a priori-defined SARS-CoV-2 laboratory tests (SARS-CoV-2 immunoglobulin G (IgG) antibodies with the Abbott Architect SARS-CoV-2 IgG enzyme-linked immunosorbent assay (ELISA) and the EUROIMMUN SARS-CoV-2 ELISA in a sequential testing algorithm). Until March 2020, over 24,000 samples tested for COVID-19 antibodies and showed the high sensitivities and specificities (~99%-100%)[20]; or 2) one or more diagnosis codes from the ICD-10 or SNOMED tables, or 3) one or more diagnosis codes from ICD-10 procedure codes. *In the survey data*, COVID-19 infection will be defined by answering affirmatively to the following questions: "*Do you think you have had COVID-19?*" in the *"COVID-19 Participant Experience (COPE)"* survey. Individuals who answered Yes to this question will be considered have potential COVID-19 infection. Similar precision rule-based algorithms described in HIV cohort building will be developed to ensure the accuracy of the cohort definition.

Variable Definitions.

AoU uses several means to collect longitudinal health data, including continuous abstraction of EHR data in the form of billing codes, laboratory and medication data, radiology reports, and narrative content and linkage with other data sources.

Viral Suppression and other HIV Related Factors. The historical VL measure will be extracted from the '*Labs &Measurements'* domain. HIV VL will be classified into: <200 copies/ml (virally suppressed) and ≥200 copies/ml (incomplete viral suppression) and stratified by the COVID-19 status/time periods. The absolute CD4 cell count will be treated as continuous variable as well as categorical variables (categorized into <200, 200-500, >500 cells/mm^3). The patients' antiretroviral therapy record will be extracted from drug exposure domain in EHR data as well as the responses from personal medical history survey data. The available ART medications will be examined as: 1) any drug use; 2) drug classes (e.g., NRTI-based, NNRTI-based, PI-based, or multi-class regimen with 3 or more classes of ART); or 3) specific drug regimens (e.g., Tenofovir disoproxil) as appropriate depending on data availability.

Baseline Health Surveys**.** Initial surveys include information on sociodemographic characteristics, overall health, lifestyle, and substance use (smoking and drinking problem), with subsequent modules covering personal and family medical history and access to health care. Per-protocol measurements include blood pressure, heart rate, weight, height, body-mass index, and hip and waist circumferences.

COPE Survey for COVID-19**.** The COPE survey asked questions about the impact of COVID-19 on participants' mental health, well-being, and everyday life. The survey was deployed six times between May 2020 and February 2021 to help researchers understand how COVID-19 impacted participants over time. The COPE survey includes information of COVID-19 related symptoms, self-reported perception of

COVID-19 infection, COVID-19 testing, COVID-19 related impact, such as anxiety and mood disorders, general well-being, social support status, stress, physical activity, loneliness, substance use, resilience, and discrimination. In addition, it also collects the health basics include pregnancy status, health insurance coverage, and marital status. Until June 2022, over 99,000 participants completed the COPE survey one or more times.

Medical History**.** The AoU medical history survey includes self-report questionnaire that asks about diagnoses to over 150 medical conditions organized into 12 disease categories.[21] We will use a combining self-reported responses to the past medical history survey and data from diagnosis codes in the EHR data to ascertain the presence of all comorbidities, such as vascular risk factors, including hypertension (OMOP code 316866), hyperlipidemia (OMOP code 432867), and type 2 diabetes mellitus (OMOP code 201826), and used self-reported data from the lifestyle survey to ascertain smoking status. Individuals with comorbidities will also be defined by answering affirmatively to either of the following questions: "*Has a doctor or health care provider ever told you that you have or had any of the following circulatory conditions/respiratory conditions/ cancers/digestive conditions/kidney conditions?"* In addition, we will use data from '*physical measurements'* to calculate the body mass index.

Healthcare Utilization**.** The healthcare utilization information is extracted from the "*Healthcare access & Utilization*" survey data. It includes health insurance, type of health care facility visit (e.g., urgent care, emergency room), healthcare specialties (e.g., nurse practitioner, physician assistant, mental health professional), frequency of healthcare visits, patient-provider communication, structural barriers of healthcare access (e.g., lack of transportation, long distance to healthcare provider, affordance of medical cost), compromised adherence due to unaffordability, and stigmatized environment.

Statistical analysis

Association Analysis. We will conduct the data cleaning and management for the integrated analysis and then conduct the correlation analysis. The distributions of demographic variables for the HIV cohort with respect to the underrepresented population will be summarized (mean, standard deviation, counts), and compared using the t-test, ANOVA test, or chi-square test as appropriate. If test assumptions are not satisfied, nonparametric tests (Wilcoxon rank test and Kruskal-Wallis Test) will be applied. The box plot and heat map will be used to depict the difference of continuous measures over time and a bar graph will be applied to the categorical measures. We will employ generalized linear mixed regression with different pre-specified correlation matrix as appropriate such as autoregression covariance matrix and choose the best model based on QIC to evaluate the differences of the probability of viral suppression between pre- and peri-pandemic periods (using March 2020 as a time cutoff, when the first COVID-19 case was reported in the US) adjusting key demographic characteristics (e.g., underrepresented population) and

other variables. The model will be built sequentially by 1) including the characteristics of underrepresented individuals only for the crude model; 2) add the COVID-19 indicators; 3) the interaction between the underrepresented population and COVID-19 status, 4) stepwise selection all variables. The lasso regression will be used if the standard stepwise selection could not work due to the high dimension of risk factors. The best model will be selected based on AIC or BIC criteria. Depending on the sample size of subset of interest in the integrated data, we could 1) conduct the stratify analysis for each underrepresented population using similar generalized linear mixed regression models, and 2) add the interaction term between underrepresented population and COVID-19 pandemic indicator. The forest plots will be used to demonstrate the regression results.

*Personalized Prediction Model***.** ML techniques are predominantly target prediction performance of single-subject outcomes. Given the multiple input features, such as sex orientation, antiretroviral therapy, comorbidities, health care utilization, HIV markers, COVID-19 infection/interruption, and other social-environmental factors, several most common and popular supervised ML algorithms will be trained to predict viral suppression in the context of the COVID-19 pandemic, to get the highest achievable prediction performance for underrepresented populations. We will investigate and evaluate the performance of several well-known ML algorithms to classify the individuals at higher risk of virological failure.

ML Algorithms. We will split the unique patient IDs into training IDs (60%), testing IDs (20%) and validation IDs (20%). The training set, testing set, and validation set will be entries with corresponding training IDs, testing IDs and validation IDs. Training set and testing set will be used to train predictive models and metrics of predictive performance will be calculated based on validation set. More specifically, we will consider the traditional Logistic Regression technique (generalized linear mixed model [GLMM]), tree-based classifiers (Classification and Regression Trees (CART), Random Forest (RF),[22] Decision Tree, and eXtreme Gradient Boosting [XGBoost][23]), Support Vector Machine (SVM),[24] Naïve Bayes (NB), and Long short-term memory (LSTM). The input feature includes all information extracted from the integrated dataset. To account for time-dependent variables (i.e., VL indicators, comorbidities, and substance use), we will consider the time lag for prediction purpose such as 1-, 3-, and 5-month as appropriate. We will apply these seven common ML approaches for different time windows accordingly.

For the potential unbiased comparison of each distinct learning algorithm, we will use a nested cross-validation (NCV)[25] workflow followed by final validation on the validation data set and then compare seven methods based on their predictive accuracy (**Figure 3**). Fine-tuning of the specific hyper-parameters of each algorithm will be performed automatically in an inner cross-validation loop (innerCV) nested inside an outer cross-validation loop (outerCV), which will be used for the proper estimation of

each predictive model. The best hyper parameters are determined based on F measure. To preserve class ratio in each split of the training data, a ten-fold stratified CV will be applied to both inner and outer loops. The validation data will be used to assess each method based on multiple measures using a confusion matrix.

Accuracy Evaluation. All the ML algorithms will be compared for prediction accuracy based on the validation data set. We will examine performance and prediction accuracy using the mean precision (positive predictive value), sensitivity (recall, true positive rate), specificity (true negative rate), F1 score, Youden's index, AUC and Matthews Correlation Coefficient (MCC). The optimal threshold of Youden's index or AUC can be determined through sensitivity, specificity, and MCC. Data with high Youden's index or AUC values near 1 indicate its high chance of correct classification, whereas low Youden's index and AUC values of models near 0 indicate a higher probability of making incorrect classifications.

Patient and public involvement

None.

Ethics and dissemination

The study was approved by the Institutional Review Boards at the University of South Carolina (Pro00124806) as a Non-Human Subject study on 10/26/2022. A deeper understanding of the impact of the pandemic on viral suppression among underrepresented PLWH populations is essential to promote health equity and better direct clinical management and guideline development. The proposed personalized viral suppression prediction can provide data driven evidence on tailored HIV treatment strategies to different underrepresented populations, particularly in the face of the unexpected interruptions like the COVID-19 pandemic. Thus, the results could facilitate the clinical identification of PLWH among underrepresented populations with poor viral control, provide them tailored HIV care management, and eventually serve towards the goal of ending the HIV epidemic in the US.

We will publish the findings in peer-reviewed scientific journals and present the study findings at national and international professional conferences and through appropriate social media outlets. We will capitalize on social media and professional networks that can increase the reach and accessibility of findings, such as open access publication, webinars, files and videos available on websites and publicly available channels (e.g., YouTube), to increase visibility and impact of the scientific publications and presentations. The dissemination efforts of this project will extend beyond the scientific arena and also target our stakeholders in healthcare system and policy makers in the US at local (SC DHEC, Prisma

Health) and national levels (CDC) through various policy forums, policy papers, and special presentations.

Data Statement

The data for All of Us can be accessed through the All of Us Research Program at the National Institutes of Health (NIH). The Data and Research Center houses the database of information provided by *All of Us* participants. With information from one million participants, the center focuses on ensuring the data is organized and secure. The center also manages researchers' access to and use of that data, helping to build a strong community of citizen scientists and researchers from leading health care research institutions, industries, and community colleges. See https://allofus.nih.gov/funding-and-program-partners/data-and-research-center for more details.

Article Summary

Strengths and limitations of this study

- The diverse group of populations recruited in the All of Us research program enables us access and examine a representative large sample of underrepresented populations in biomedical research to reduce algorithmic bias.
- The data integration from multiple data sources in All of Us research program for cohort analyses allows us to generate a robust evaluation of the viral suppression prediction for underrepresented population with a long follow up.
- The machine-learning based approach to developing personalized prediction for viral suppression has the benefit of accurately modeling a different data structure many risk factors.
- We expect missing data in both electronic health record data and survey results; therefore, caution may be needed when interpreting the risk prediction results.

Funding statement This work was supported by the U.S. Department of Health and Human Services, National Institutes of Health, National Institute of Allergy And Infectious Diseases [grant number R01AI164947-02S1]. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

Conflict of interest The authors declare that there is no conflict of interest.

Figure legend/caption

Figure 1 Multi-level factors from multi-domains in All of Us program

Figure 2 Flowchart for data extraction and integration

Figure 3 Machine learning pipeline and relative data flow

Author Contribution BO and JZ is the principal investigator of this project and led the study design. XY contributed to the conception and design of the study. XY led the writing of this protocol manuscript. SW and XL contributed significantly to the editing of this manuscript. All authors reviewed and provided comments to improve the manuscript. All authors contributed to the editing and final approval of the protocol.

References

1. Lee JS, Cole SR, Richardson DB, et al. Incomplete viral suppression and mortality in HIV patients after antiretroviral therapy initiation. AIDS **2017**; 31(14): 1989-97.
2. Services USDoHH. Ending the HIV Epidemic: A plan for America. available at: https://wwwhhsgov/blog/2019/02/05/ending-the-hiv-epidemic-a-plan-for-americahtml **2019**.
3. CDC U. Monitoring Selected National HIV Prevention and Care Objectives By Using HIV Surveillance Data United States and 6 Dependent Areas, 2019: Tables. Available at: https://www.cdc.gov/hiv/library/reports/hiv-surveillance/vol-26-no-2/content/tables.html. Accessed June 10.
4. Spinelli MA, Hickey MD, Glidden DV, et al. Viral suppression rates in a safety-net HIV clinic in San Francisco destabilized during COVID-19. AIDS **2020**; 34(15): 2328-31.
5. Norwood J, Kheshti A, Shepherd BE, et al. The Impact of COVID-19 on the HIV Care Continuum in a Large Urban Southern Clinic. AIDS Behav **2022**.
6. Hickey MD, Imbert E, Glidden DV, et al. Viral suppression during COVID-19 among people with HIV experiencing homelessness in a low-barrier clinic-based program. AIDS **2021**; 35(3): 517-9.
7. Izzo I, Carriero C, Gardini G, et al. Impact of COVID-19 pandemic on HIV viremia: a single-center cohort study in northern Italy. AIDS Res Ther **2021**; 18(1): 31.
8. Matsumoto S, Nagai M, Luong DAD, et al. Evaluation of SARS-CoV-2 Antibodies and the Impact of COVID-19 on the HIV Care Continuum, Economic Security, Risky Health Behaviors, and Mental Health Among HIV-Infected Individuals in Vietnam. AIDS Behav **2021**.
9. 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(7): e0234962.
10. CDC U. Estimated HIV incidence and prevalence in the United States, 2015–2019, and US Census Bureau, Quick Facts—United States. Available at: https://www.cdc.gov/hiv/pdf/library/reports/surveillance/cdc-hiv-surveillance-supplemental-report-vol-26-1.pdf.
11. United Nations. Shared Responsibility, Global Solidarity: Responding to the Socioeconomic Impacts of COVID-19. . Available at: https://www.un.org/sites/un2.un.org/files/sg_report_socio-economic_impact_of_covid19.pdf Accessed June 27.
12. Chenneville T, Gabbidon K, Hanson P, Holyfield C. The Impact of COVID-19 on HIV Treatment and Research: A Call to Action. Int J Environ Res Public Health **2020**; 17(12).
13. Investigators AoURP. The "All of Us" research program. New England Journal of Medicine **2019**; 381(7): 668-76.
14. Acosta JN, Leasure AC, Both CP, et al. Cardiovascular Health Disparities in Racial and Other Underrepresented Groups: Initial Results From the All of Us Research Program. J Am Heart Assoc **2021**; 10(17).
15. All of Us Public Data Browser. View survey questions and answers. Available at: https://databrowser.researchallofus.org/survey/family-health-history.

16. PCORnet. PCORnet® COVID-19 Common Data Model Launched, Enabling Rapid Capture Of Insights On Patients Infected With The Novel Coronavirus. Available at: https://pcornet.org/news/pcornet-covid-19-common-data-model-launched-enabling-rapid-capture-of-insights/. Accessed June 21.
17. Burn E, You SC, Sena AG, et al. An international characterisation of patients hospitalised with COVID-19 and a comparison with those previously hospitalised with influenza. medRxiv **2020**.
18. LOINC. SARS-CoV-2 and COVID-19 related LOINC terms. Available at: https://loinc.org/sars-cov-2-and-covid-19/. Accessed June 21.
19. ATLAS. Atlas OHDSI concept sets. . Available at: http://atlas-covid19.ohdsi.org/#/home Accessed May 27.
20. Althoff KN, Schlueter DJ, Anton-Culver H, et al. Antibodies to severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) in All of Us Research Program participants, 2 January to 18 March 2020. Clinical Infectious Diseases **2022**; 74(4): 584-90.
21. Sulieman L, Cronin RM, Carroll RJ, et al. Comparing medical history data derived from electronic health records and survey answers in the All of Us Research Program. Journal of the American Medical Informatics Association **2022**.
22. Breiman L. Random forests. Machine learning **2001**; 45(1): 5-32.
23. Friedman JH. Greedy function approximation: a gradient boosting machine. Annals of statistics **2001**: 1189-232.
24. Cortes C, Vapnik V. Support-vector networks. Machine Learning **1995**; 20(3): 273-97.
25. Yurduseven K, Babal YK, Celik E, Kerman BE, Kurnaz IA. Multiple Sclerosis Biomarker Candidates Revealed by Cell-Type-Specific Interactome Analysis. OMICS **2022**; 26(5): 305-17.

Figure 1: Multi-level Factors from Multi-Domains in All of Us Program

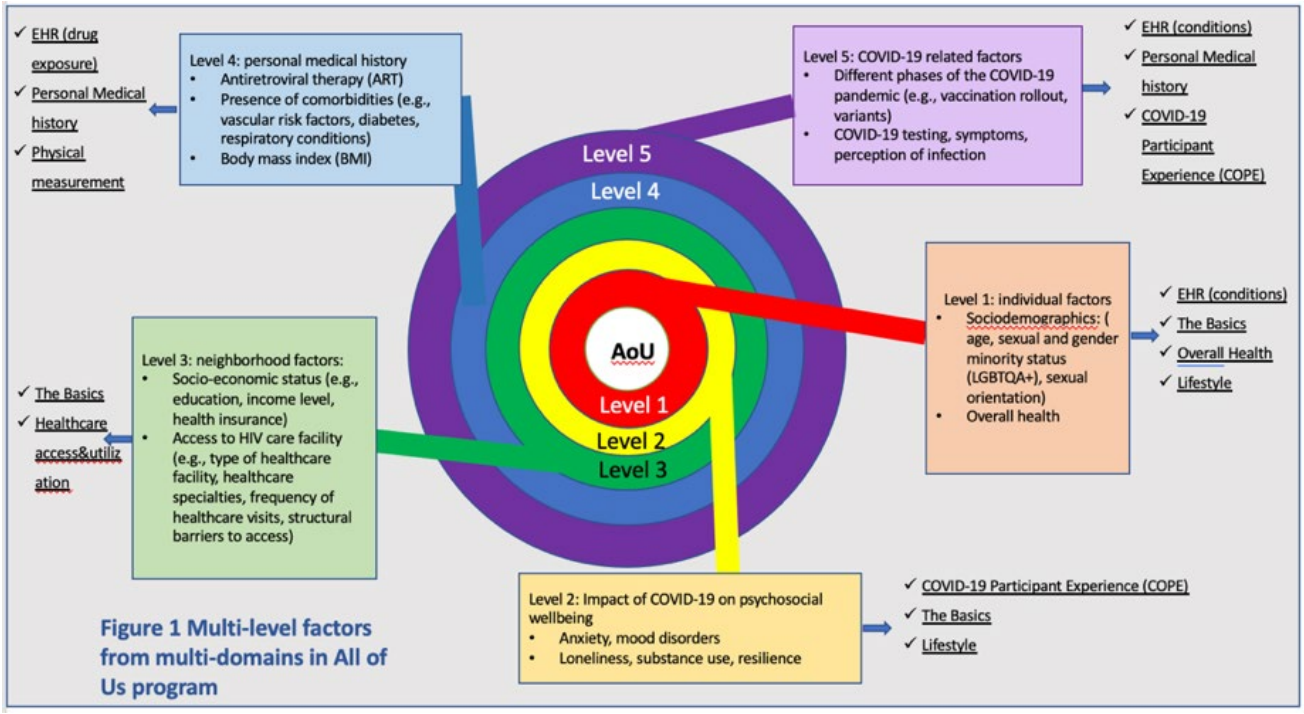

Figure 2: Flowchart for All of Us Data Extraction and Integration

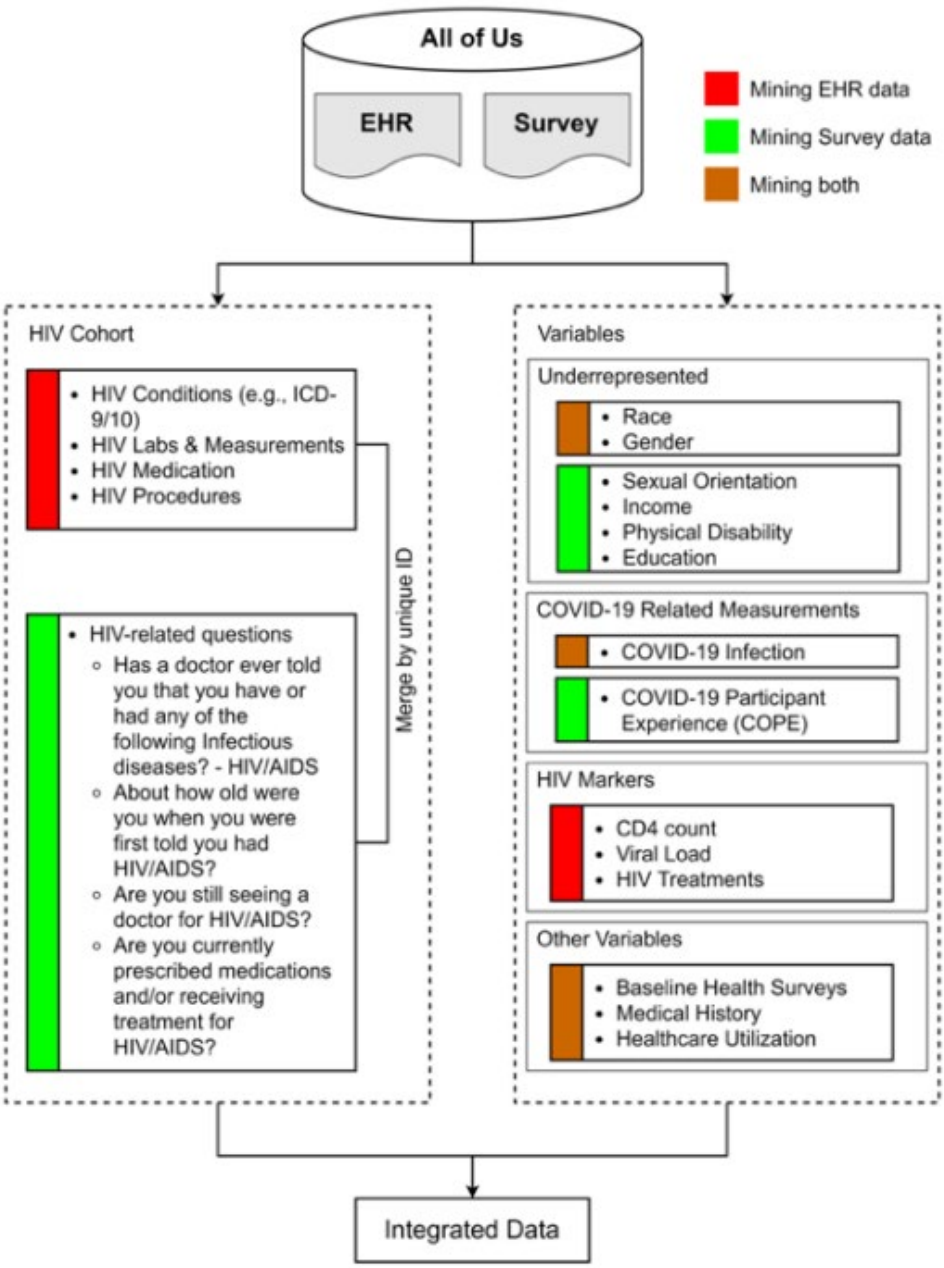

Figure 3: Flowchart for All of Us Data Extraction and Integration

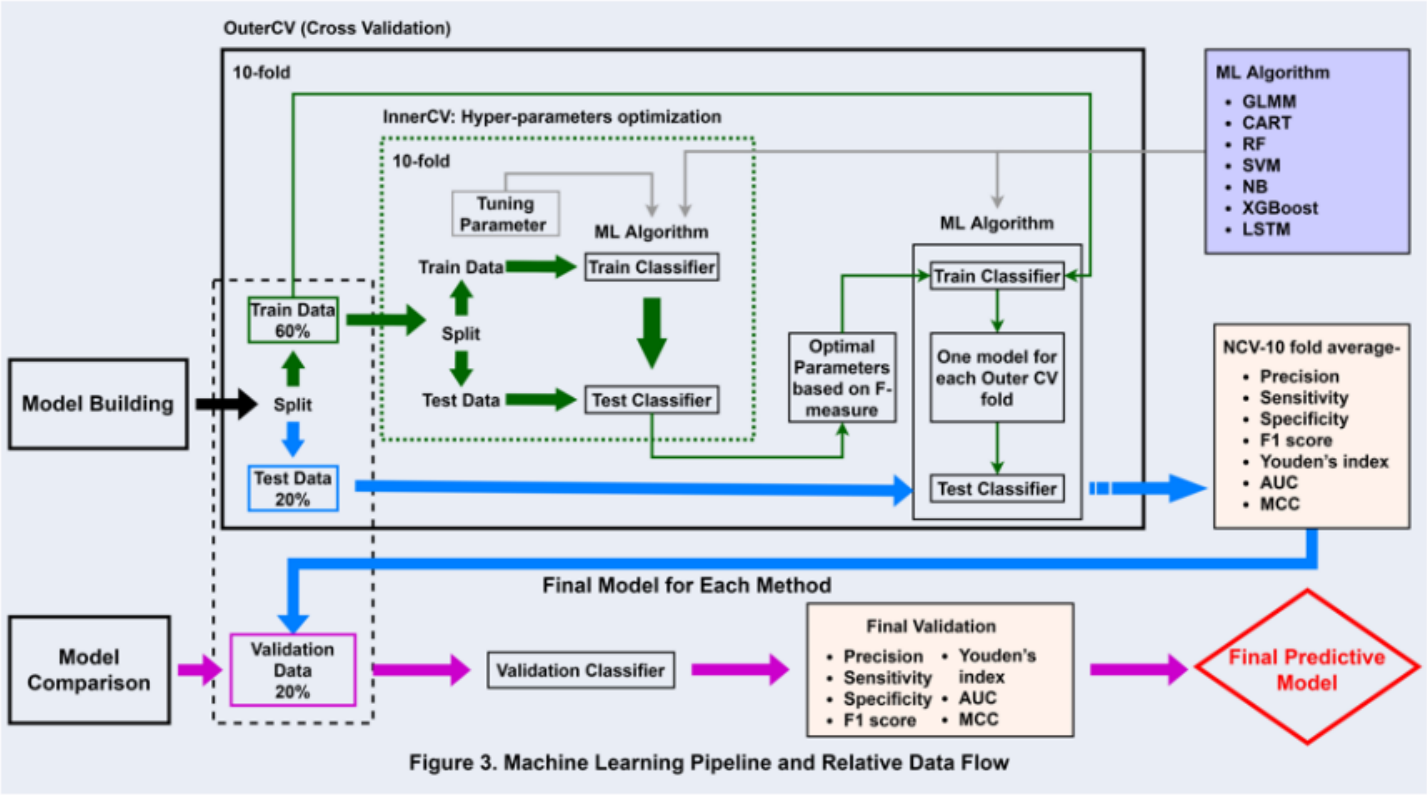

Figure 3. Machine Learning Pipeline and Relative Data Flow

STROBE Statement—Checklist of items that should be included in reports of ***cohort studies***

	Item No	Recommendation	Page No
Title and abstract	1	(*a*) Indicate the study's design with a commonly used term in the title or the abstract	1
		(*b*) Provide in the abstract an informative and balanced summary of what was done and what was found	2
Introduction			
Background/rationale	2	Explain the scientific background and rationale for the investigation being reported	4
Objectives	3	State specific objectives, including any prespecified hypotheses	5
Methods			
Study design	4	Present key elements of study design early in the paper	5
Setting	5	Describe the setting, locations, and relevant dates, including periods of recruitment, exposure, follow-up, and data collection	6
Participants	6	(*a*) Give the eligibility criteria, and the sources and methods of selection of participants. Describe methods of follow-up	6-7
		(*b*) For matched studies, give matching criteria and number of exposed and unexposed	
Variables	7	Clearly define all outcomes, exposures, predictors, potential confounders, and effect modifiers. Give diagnostic criteria, if applicable	8
Data sources/ measurement	8*	For each variable of interest, give sources of data and details of methods of assessment (measurement). Describe comparability of assessment methods if there is more than one group	6
Bias	9	Describe any efforts to address potential sources of bias	9-11
Study size	10	Explain how the study size was arrived at	5
Quantitative variables	11	Explain how quantitative variables were handled in the analyses. If applicable, describe which groupings were chosen and why	9-11
Statistical methods	12	(*a*) Describe all statistical methods, including those used to control for confounding	9-11
		(*b*) Describe any methods used to examine subgroups and interactions	
		(*c*) Explain how missing data were addressed	
		(*d*) If applicable, explain how loss to follow-up was addressed	
		(*e*) Describe any sensitivity analyses	
Results			
Participants	13*	(a) Report numbers of individuals at each stage of study—eg numbers potentially eligible, examined for eligibility, confirmed eligible, included in the study, completing follow-up, and analysed	N/A
		(b) Give reasons for non-participation at each stage	
		(c) Consider use of a flow diagram	
Descriptive data	14*	(a) Give characteristics of study participants (eg demographic, clinical, social) and information on exposures and potential confounders	N/A
		(b) Indicate number of participants with missing data for each variable of interest	
		(c) Summarise follow-up time (eg, average and total amount)	
Outcome data	15*	Report numbers of outcome events or summary measures over time	N/A

Main results	16	(*a*) Give unadjusted estimates and, if applicable, confounder-adjusted estimates and their precision (eg, 95% confidence interval). Make clear which confounders were adjusted for and why they were included (*b*) Report category boundaries when continuous variables were categorized (*c*) If relevant, consider translating estimates of relative risk into absolute risk for a meaningful time period	N/A
Other analyses	17	Report other analyses done—eg analyses of subgroups and interactions, and sensitivity analyses	N/A
Discussion			**N/A**
Key results	18	Summarise key results with reference to study objectives	
Limitations	19	Discuss limitations of the study, taking into account sources of potential bias or imprecision. Discuss both direction and magnitude of any potential bias	
Interpretation	20	Give a cautious overall interpretation of results considering objectives, limitations, multiplicity of analyses, results from similar studies, and other relevant evidence	
Generalisability	21	Discuss the generalisability (external validity) of the study results	
Other information			
Funding	22	Give the source of funding and the role of the funders for the present study and, if applicable, for the original study on which the present article is based	13

*Give information separately for exposed and unexposed groups.

Note: An Explanation and Elaboration article discusses each checklist item and gives methodological background and published examples of transparent reporting. The STROBE checklist is best used in conjunction with this article (freely available on the Web sites of PLoS Medicine at http://www.plosmedicine.org/, Annals of Internal Medicine at http://www.annals.org/, and Epidemiology at http://www.epidem.com/). Information on the STROBE Initiative is available at http://www.strobe-statement.org.

BMJ Open

Protocol for Developing A Personalized Prediction Model for Viral Suppression among Underrepresented Populations in the context of the COVID-19 Pandemic.

Journal:	*BMJ Open*
Manuscript ID	bmjopen-2022-070869.R1
Article Type:	Protocol
Date Submitted by the Author:	25-Mar-2023
Complete List of Authors:	Zhang, Jiajia; University of South Carolina Arnold School of Public Health, Department of Epidemiology and Biostatistics; University of South Carolina Arnold School of Public Health, South Carolina SmartState Center for Healthcare Quality Yang, Xueying; University of South Carolina Arnold School of Public Health, Health Promotion Education and Behavior Weissman, Sharon; University of South Carolina, Department of Internal Medicine, School of Medicine Li, Xiaoming; University of South Carolina Arnold School of Public Health; University of South Carolina Arnold School of Public Health, South Carolina SmartState Center for Healthcare Quality Olatosi, Bankole; University of South Carolina Arnold School of Public Health, Health Services Policy and Management; University of South Carolina Arnold School of Public Health, South Carolina SmartState Center for Healthcare Quality
Primary Subject Heading:	HIV/AIDS
Secondary Subject Heading:	Public health
Keywords:	COVID-19, HIV & AIDS < INFECTIOUS DISEASES, PUBLIC HEALTH, STATISTICS & RESEARCH METHODS

Protocol for Developing A Personalized Prediction Model for Viral Suppression among Underrepresented Populations in the context of the COVID-19 Pandemic

Jiajia Zhang [1,2], Xueying Yang [1,3], Sharon Weissman [1,5], Xiaoming Li [1,3] Bankole Olatosi [1,4*]

[1] South Carolina SmartState Center for Healthcare Quality, Arnold School of Public Health, University of South Carolina, Columbia, SC, USA, 29208

[2] Department of Epidemiology and Biostatistics, Arnold School of Public Health, University of South Carolina, Columbia, SC, USA, 29208

[3] Department of Health Promotion, Education and Behavior, Arnold School of Public Health, University of South Carolina, Columbia, SC, USA, 29208

[4] Department of Health Services Policy and Management, Arnold School of Public Health, University of South Carolina, Columbia, SC, USA, 29208

[5] Department of Internal Medicine, School of Medicine, University of South Carolina, Columbia, SC, USA, 29208

*Corresponding author:

Bankole Olatosi, Email: olatosi@mailbox.sc.edu

Abstract: 299

Word limit: 3561 (limit 4000)

Abstract

Introduction Sustained viral suppression, an indicator of long-term treatment success and mortality reduction, is one of four strategic areas of the "*Ending the HIV Epidemic (EtHE)*" federal campaign launched in 2019. Underrepresented populations, like racial or ethnic minority populations, sexual and gender minority groups, and socioeconomically disadvantaged populations, are disproportionately affected by HIV and experience a more striking virological failure. The COVID-19 pandemic might magnify the risk of incomplete viral suppression among underrepresented Persons living with HIV (PLWH) due to interruptions in healthcare access and other worsened socioeconomic and environmental conditions. However, biomedical research rarely includes underrepresented populations, resulting in biased algorithms. This proposal targets a broadly defined underrepresented HIV population. It aims to develop a personalized viral suppression prediction model using machine learning techniques by incorporating multilevel factors using All of Us (AoU) data.

Methods and analysis This cohort study will use data from the AoU Research Program, which aims to recruit a broad, diverse group of U.S. populations historically underrepresented in biomedical research. The program harmonizes data from multiple sources on an ongoing basis. It has recruited ~ 4800 PLWH with a series of self-reported survey data (e.g., Lifestyle, Healthcare Access, COVID-19 Participant Experience) and relevant longitudinal EHR data. We will examine the change in viral suppression and develop personalized viral suppression prediction due to the impact of the COVID-19 pandemic using machine learning techniques (ML), such as tree-based classifiers (Classification and Regression Trees (CART), Random Forest (RF), Decision Tree, and eXtreme Gradient Boosting [XGBoost]), Support Vector Machine (SVM), Naïve Bayes (NB), and Long short-term memory (LSTM).

Ethics and dissemination The Institutional Review Board approved the study at the University of South Carolina (Pro00124806) as a Non-Human Subject study. Findings will be published in peer-reviewed journals and disseminated at national and international conferences and through social media.

Keywords: HIV/AIDS, Viral Suppression, COVID-19, Machine Learning

Strengths and limitations of this study

- The diverse group of populations recruited in the All of Us research program enables us to have a large representative sample of underrepresented populations in biomedical research and reduce algorithmic bias.
- The data integration from multiple data sources in the All of Us research program for cohort analyses allows us to robustly evaluate the viral suppression prediction for the underrepresented population with a long follow-up.
- The machine-learning based approach to developing personalized prediction for viral suppression has the benefit of accurately modeling a different data structure of many risk factors.
- We expect missing data in both electronic health record data and survey results; thus, caution may be needed when interpreting the risk prediction results.

Introduction

Sustained viral suppression, an indicator of long-term treatment success and mortality reduction,[1] is one of four strategic areas of the "*Ending the HIV Epidemic (EtHE)*: *A Plan for America*"[2] federal campaign launched in 2019. According to the Centers for Disease Control and Prevention national surveillance data, approximately 66% of all people living with HIV (PLWH) were virally suppressed in the United States (US).[3] The COVID-19 pandemic uniquely affects PLWH and has a mixed impact on viral suppression across different countries or settings. In the US, a few studies revealed the decreased probability of viral suppression due to the negative impact of the pandemic[4, 5], but one study in San Francisco did not report the same findings.[6] Similar inconsistent results were reported in European and Asian studies.[7, 8] The mixed results might be caused by small sample sizes, lack of sample diversity, and/or insufficient phenotypic data.

Individuals with inadequate access to medical care, low household incomes, low education attainment, and racial or sexual and gender minorities are often underrepresented in biomedical research (hereafter referred to as 'underrepresented population').[9] HIV and COVID-19 both have a disproportionate impact on underrepresented populations. For instance, 45% of new HIV infections were among gay and bisexual men under 35 years and 26% among Black gay and bisexual men.[10] Moreover, these vulnerable populations experience a more striking virological failure.[3] The United Nations [11] report has indicated that increases in food costs and market stockpiling during the COVID-19 pandemic have had the most harmful impact on underrepresented populations. Those with stigmatized or marginalized intersecting identities often experience the highest HIV burden, including men who have sex with men, transgender women, people who inject drugs, commercial sex workers, and youths, who account for a third of all new HIV infections.[12] Thus, the pandemic might magnify the risk of incomplete viral suppression among the underrepresented PLWH population due to interruptions in healthcare access and other worsened socioeconomic and environmental conditions.

The increasing availability of electronic health records (EHRs) has presented the opportunity to discover new knowledge via extensive data linkage and integration. However, as a real-world clinical routine data source, EHR data is not designed for a specific research purpose. Thus, it has a limited capacity to recruit an adequate sample of underrepresented populations due to their historically limited access to specialty care and academic medical centers that serve as the primary sources for EHR data. Consequently, it poses more challenges in understanding the viral suppression among underrepresented populations, particularly those facing the COVID-19 pandemic.

The All of Us (AoU) Research Program is an ongoing national, historic effort supported by the NIH. The cohort in AoU includes a broadly diverse group of the U.S. population, with more than 50% of the

participants from racial and ethnic minority groups and more than 80% from populations historically underrepresented in biomedical research (e.g., sex orientation, socioeconomic status, geographical location, physical disability). Therefore, this protocol aims to target underrepresented populations using AoU data, which includes ~4800 PLWH with a series of self-reported survey data (e.g., Lifestyle, Physical Measurement, Healthcare Access, COVID-19 Participant Experience) and relevant longitudinal EHR data (laboratory and medication). The variables collected include longitudinal observations of clinical, environmental, lifestyle, and genetic data. With the data integration, the current exploratory study has the following specific aims:

Aim 1: Examine the impact of the COVID-19 pandemic on viral suppression among a broadly defined underrepresented HIV population by harnessing the AoU big data resources.

Aim 2: Develop personalized viral suppression prediction models using machine learning techniques by incorporating COVID-19 interruption, antiretroviral therapy history, preexisting conditions (comorbidities), psychological wellbeing (e.g., depression, resilience), healthcare utilization, and social, environmental factors in AoU.

A deeper understanding of the impact of the pandemic on viral suppression among underrepresented PLWH populations is essential to promote health equity and better direct clinical management and guideline development. The proposed personalized viral suppression prediction can provide data-driven evidence on tailored HIV treatment strategies for different underrepresented populations, particularly during unexpected interruptions like the COVID-19 pandemic. Thus, the results could facilitate the clinical identification of PLWH among underrepresented populations with poor viral control, provide them with tailored HIV care management, and eventually serve towards the goal of ending the HIV epidemic in the U.S. The availability of comprehensive phenotypic data and Researcher Workbench in AoU platform fully ensures the transparency and reproducibility of the proposed project.

Methods and Analysis

Overview of the Study Design

To guide our proposed research, we have developed a conceptual framework (**Figure 1**) that depicts how we harness the comprehensive phenotypic data from different domains of AoU Researcher Workbench to achieve the Specific Aims. The cohort building and outcomes will be defined from EHR data and survey data. For example, the intrapersonal factors (Level 1), including demographic characteristics (e.g., age, race, and gender) and overall health, will be extracted from *"The Basics"* survey. The COVID-19 related experiences (Levels 2 & 5) refer to the impact of the pandemic on their health and psychosocial wellbeing, such as social support, depression, anxiety, drug and alcohol abuse, and resilience, will be extracted from *"COVID-19 Participant Experience (COPE)"* and *"Lifestyle"* surveys. The neighborhood-level factors (Level 3), including the neighborhood economic environment (e.g., poverty,

education, health insurance coverage) and healthcare access (type of healthcare facility, structural barriers to healthcare access), will be defined from the *"Healthcare access & utilization"* survey. With the appropriate data management/preprocessing, we will examine the change in viral suppression and develop the personalized viral suppression prediction due to the impact of the COVID-19 pandemic using machine learning techniques (ML), which will have translational potential to inform future HIV care among underrepresented populations.

Data sources

Overview of the AoU program

The AoU Research Program seeks to recruit persons in demographic categories that have been and continue to be underrepresented in biomedical research; such persons typically have relatively poor access to good health care.[13] AoU opened for enrollment in May 2018, and the inclusion criteria are age ≥18 years with the capacity to provide consent. The recruitment methods and scientific rationale for AoU have been described previously.[13] Through November 19, 2021, AoU has harmonized data from over 340 institutional sites contributing data for about 331,360 participants using the Observational Medical Outcomes Partnership (OMOP) Common Data Model. We anticipate an ample size to conduct the proposed analysis since AoU is harmonizing data on an ongoing basis. Each participant completed informed consent for sharing their EHR data with the Data and Research Center and provided survey responses across different domains. Each participating institutional site contributes demographics, medications, laboratory tests, diagnoses, and vital status to the central data repository for data harmonization. A dedicated institutional review board, the AoU Institutional Review Board, has approved the AoU protocol and materials. Deidentified data were shared through the AoU Researcher Workbench (www.allofus.nih.gov) for analyses through institutional data use agreements. All analyses will be conducted within a secure informatic workspace provided by the National Institutes of Health that allows users to access and analyze a centralized version of the AoU data.

"HIV and COVID-19" Project in the AoU Researcher Workbench Platform

AoU Research Program data in its final format, after harmonization and refinement, are referred to as a curated dataset. Three different levels of information are available: Public tier, registered tier, and controlled tier. We have obtained access to data at the registered tier. Following the AoU instructions, we have created a project entitled "HIV and COVID-19" in the AoU Researcher Workbench platform. This is a cloud-based platform that enables researchers to cluster participants into cohorts, select certain health information within each cohort, and perform direct analysis and query using R (R Foundation for Statistical Computing) and Python 3.0 (Python Software Foundation) programming languages within Jupyter Notebooks. The purpose of our Workspace is 2-fold: 1) Cohort building: to determine the data inclusion and exclusion criteria for HIV cohort building (computable phenotype) and create and maintain

a set of scripts to execute the computable phenotype and extract relevant data for this cohort; and 2) Model building: to examine the impact of COVID-19 on HIV and its potential predictors and build the prediction model for viral suppression.

Cohort Building and Data Extraction in AoU.

In biomedical research, a phenotype is an observable manifestation of a clinical entity (e.g., a disease). Computable phenotypes are essential for analyzing large clinical observational data. The development process for computable phenotypes occurs iteratively by identifying and refining concepts from controlled healthcare terminologies (also known as "concept sets"). The concept/disease condition (e.g., diabetes) is the base instance, combined with all possible feature representations in data (e.g., ICD codes for diabetes + insulin; or ICD codes for diabetes + Hemoglobin A1c). The combination of all possible "concept/s" and feature/logic representations of the concept (AND, OR, NOT) allows the computer to interpret or determine the right computable phenotype automatically for further analyses.[14] Given our understanding of disease signs and symptoms, we will define computable phenotypes that can accurately identify both the study cohort (e.g., HIV population) and relevant variables (e.g., COVID-19 infection) from EHR data and survey data (**Figure 2**). The EHR data derived from captured data including billing codes and encounter records will be used to cluster participants into disease cohorts based on Systemized Nomenclature of Medicine - Clinical Terms diagnosis codes (the standardized vocabulary in AoU sourced from corresponding International Classification of Diseases codes [ICD]). In contrast, other data will be extracted from survey responses. Examples of the surveys can be found through the publicly available Data Browser.[15] We will map survey and EHR data to the OMOP common data model version 5.2. We will extract data from the EHR domains and available survey results via the AoU Researcher Workbench.

HIV Cohort. To build the HIV cohort, we will adopt the existing inclusion criteria and code sets from several organizations - for example, PCORnet,[16] OHDSI,[17] LOINC[18], etc. into a “best-of-breed” phenotype and extract data from both EHR and survey questionnaires. The best-of-breed phenotypic characterization approach helps identify and document diversity within and between distinct traits of subjects (known as “breeds”).[14] In practice, we apply phenotyping algorithms, which map to various domains (e.g., condition domain, drug domain) to best identify individuals with a particular clinical entity (“best of breed”) (e.g., HIV infection). In EHR data, we will define HIV by documentation of any of the following: (1) HIV condition (*ICD, Ninth/Tenth Revision (ICD-9/10)* diagnostic codes, ICD-9/10 procedure codes) in the ‘*Condition*’ domain; (2) HIV-related laboratory results (e.g., HIV antibody) in the ‘*Labs &Measurements’* domain; or (3) HIV-related medications (e.g., Tenofovir disoproxil) excluding pre-exposure prophylaxis in the ‘*Drug Exposures*’ domain. In the survey data, we will define HIV based on affirmative answers to the following questions: “*Has a doctor or health care provider ever told you that you have or had any of the following infectious diseases?*” or “*Are you currently prescribed*

medications and/or receiving treatment for HIV/AIDS?" in the "*Personal Medical History*" survey. Individuals who answered Yes to "*Infectious Disease Condition: HIV/AIDS*" or "*HIV/AIDS Currently*" will be counted as the HIV population. Patients who meet at least one of these inclusion criteria in either EHR data or survey data and those who meet all of these inclusion criteria will be calculated and compared with other national initiatives to develop precision rule-based algorithms for data analysis. A template of concept sets [19] based on all the above information will be built for the HIV cohort. (See Table 1 for summary characteristics).

Table 1: Characteristics of Underrepresented Population of Persons Living With HIV in All of Us Program Data

Characteristics	N (%)		Characteristics	N (%)
Data from EHR and survey			**Data from survey only**	
Total HIV	4794		Sex/gender (n=1080)	
Age			LGBTQIA+, no	291
<75 y	4619		LGBTQIA+, yes	789
≥75 y	175		Education (n=1067)	
Race			High school degree	977
White	1232		Less than a high	90 (8.33)
Black or African	2448		Household Income	
Asian	29		>$35,000 US dollars	621
Other/unknown	1085 (22.63)		<$35,000 US dollars	359 (36.63)
COVID-19 infection			Physical Disability	
Yes	402		No	958
No	4392		Yes	111

Note we estimate the final sample size will be greater than 1000 even after we exclude the missing and other unknown information.[20] As mentioned in Figueroa et al., 560 trained samples are adequate to achieve a mean average and root mean squared error below 0.01 based on supervised learning.[21] That means using 60% of data for training the model should be adequate for supervised learning. We will use the remaining data for testing and validation (See Figure 3).

COVID-19 Cohort. AoU study participants in all 50 US states have provided blood specimens since January 2020 for COVID-19 testing. Similar to defining the HIV population, COVID-19 patients will be identified using EHR data and survey data. *In the EHR data*, the COVID-19 positive cases will be defined as patients with any encounter on or after 1/1/2020 with either: 1) a positive result for one of a set of a priori defined SARS-CoV-2 laboratory tests (SARS-CoV-2 immunoglobulin G (IgG) antibodies with the Abbott Architect SARS-CoV-2 IgG enzyme-linked immunosorbent assay (ELISA) and the

EUROIMMUN SARS-CoV-2 ELISA in a sequential testing algorithm). Through March 2020, over 24,000 samples tested for COVID-19 antibodies and showed high sensitivities and specificities (~99%-100%)[22]; or 2) one or more diagnosis codes from the ICD-10 or SNOMED tables, or 3) one or more diagnosis codes from ICD-10 procedure codes. *In the survey data*, COVID-19 infection will be defined by answering affirmatively to the following questions: "*Were you tested for COVID-19?*" and "*Was the test(s) for COVID-19 positive?*" in the *"COVID-19 Participant Experience (COPE)"* survey. Individuals who answered Yes to this question will be considered to have potential COVID-19 infection. We will apply similar precision rule-based algorithms described in HIV cohort building will be developed to ensure the accuracy of the cohort definition.

Variable Definitions.

AoU uses several means to collect longitudinal health data, including continuous abstraction of EHR data in the form of billing codes, laboratory and medication data, radiology reports, and narrative content and linkage with other data sources.

Viral Suppression and other HIV Related Factors. The historical VL measure will be extracted from the '*Labs &Measurements'* domain. HIV VL will be classified into: <200 copies/ml (virally suppressed) and ≥200 copies/ml (incomplete viral suppression) and stratified by the COVID-19 status/time periods. The absolute CD4 cell count will be treated as a continuous variable and a categorical variable (categorized into <200, 200-500, >500 cells/mm^3). The patients' antiretroviral therapy records will be extracted from drug exposure domain in EHR data and the responses from personal medical history survey data. The available ART medications will be examined as 1) any drug use; 2) drug classes (e.g., NRTI-based, NNRTI-based, PI-based, or multi-class regimen with 3 or more classes of ART); or 3) specific drug regimens (e.g., Tenofovir disoproxil) as appropriate depending on data availability.

Baseline Health Surveys**.** Initial surveys include information on sociodemographic characteristics, overall health, lifestyle, and substance use (smoking and alcohol use), with subsequent modules covering personal and family medical history and access to health care. Per-protocol measurements include blood pressure, heart rate, weight, height, body mass index, and hip and waist circumferences.

COPE Survey for COVID-19**.** The COPE survey asked questions about the impact of COVID-19 on participants' mental health, well-being, and everyday life. The survey was deployed six times between May 2020 and February 2021 to help researchers understand how COVID-19 impacted participants over time. The COPE survey includes information on COVID-19 related symptoms, self-reported perception of COVID-19 infection, COVID-19 testing, COVID-19 related impact, such as anxiety and mood disorders, general well-being, social support status, stress, physical activity, loneliness, substance use, resilience, and discrimination. In addition, it also collects the health basics include pregnancy status, health insurance coverage, and marital status. Through June 2022, over 99,000 participants completed the

COPE survey one or more times, with over 1000 PLWH represented.

Medical History**.** The AoU medical history survey includes a self-report questionnaire about diagnoses of over 150 medical conditions organized into 12 disease categories.[23] We will use a combination of self-reported responses to the past medical history survey and data from diagnosis codes in the EHR data to ascertain the presence of all comorbidities, such as cardiovascular risk factors, including hypertension (OMOP code 316866), hyperlipidemia (OMOP code 432867), and type 2 diabetes mellitus (OMOP code 201826), and use self-reported data from the lifestyle survey to ascertain smoking status. Individuals with comorbidities will also be defined by answering affirmatively to either of the following questions: "*Has a doctor or health care provider ever told you that you have or had any of the following circulatory conditions/respiratory conditions/ cancers/digestive conditions/kidney conditions?"* In addition, we will use data from '*physical measurements'* to calculate the body mass index.

Healthcare Utilization**.** The healthcare utilization information is extracted from the "*Healthcare access & Utilization*" survey data. It includes health insurance, type of healthcare facility visits (e.g., urgent care, emergency room), healthcare specialties (e.g., nurse practitioner, physician assistant, mental health professional), frequency of healthcare visits, patient-provider communication, structural barriers of healthcare access (e.g., lack of transportation, long distance to a healthcare provider, the affordance of medical cost), compromised adherence due to unaffordability, and stigmatized environment.

Statistical analysis

Association Analysis. We will conduct the data cleaning and management for the integrated analysis and then conduct the correlation analysis. The distributions of demographic variables for the HIV cohort with respect to the underrepresented population will be summarized (mean, standard deviation, counts), and compared using the t-test, ANOVA test, or chi-square test as appropriate. If test assumptions are not satisfied, nonparametric tests (Wilcoxon rank test and Kruskal-Wallis Test) will be applied. The box plot and heat map will depict the difference between continuous measures over time, and a bar graph will be applied to the categorical measures. We will employ generalized linear mixed regression with different pre-specified correlation matrix as appropriate such as autoregression covariance matrices and choose the best model based on QIC to evaluate the differences in the probability of viral suppression between pre- and peri-pandemic periods (using March 2020 as a time cutoff, when the first COVID-19 case was reported in the US) adjusting for key demographic characteristics (e.g., underrepresented population) and other variables. The model will be built sequentially by 1) including the characteristics of underrepresented individuals only for the crude model; 2) adding the COVID-19 indicators; 3) the interaction between the underrepresented population and COVID-19 status, 4) stepwise selection of all variables. The lasso regression will be used if the standard stepwise selection cannot work due to the high

dimension of risk factors. The best model will be selected based on AIC or BIC criteria. Depending on the sample size of subset of interest in the integrated data, we could 1) conduct a stratified analysis for each underrepresented population using similar generalized linear mixed regression models, and 2) add the interaction term between underrepresented population and COVID-19 pandemic indicator. We will use forest plots will be used to display the regression results.

*Personalized Prediction Model***.** ML techniques predominantly target the prediction performance of single-subject outcomes. Given the multiple input features, such as sex orientation, antiretroviral therapy, comorbidities, health care utilization, HIV markers, COVID-19 infection/interruption, and other social-environmental factors, several most common and popular supervised ML algorithms will be trained to predict viral suppression in the context of the COVID-19 pandemic, to get the highest achievable prediction performance for underrepresented populations. We will investigate and evaluate the performance of several well-known ML algorithms to classify individuals at higher risk of virological failure.

ML Algorithms. We will split the unique patient IDs into training IDs (60%), testing IDs (20%), and validation IDs (20%). The training, testing, and validation sets will be entries with corresponding training IDs, testing IDs and validation IDs. The training and testing sets will be used to train predictive models, and predictive performance metrics will be calculated based on the validation set. More specifically, we will consider the traditional Logistic Regression technique (generalized linear mixed model [GLMM]), tree-based classifiers (Classification and Regression Trees (CART), Random Forest (RF),[24] Decision Tree, and eXtreme Gradient Boosting [XGBoost][25]), Support Vector Machine (SVM),[26] Naïve Bayes (NB), and Long short-term memory (LSTM). The input feature includes all information extracted from the integrated dataset. To account for time-dependent variables (i.e., VL indicators, comorbidities, and substance use), we will consider the time lag for a prediction purpose such as 1-, 3-, and 5 months as appropriate. We will apply these seven common ML approaches for different time windows accordingly.

For the potential unbiased comparison of each distinct learning algorithm, we will use a nested cross-validation (NCV)[27] workflow followed by final validation on the validation data set and then compare seven methods based on their predictive accuracy (**Figure 3**). The validation data will be used to assess each method based on multiple measures using a confusion matrix. Fine-tuning of the specific hyper-parameters of each algorithm will be performed automatically in an inner cross-validation loop (innerCV) nested inside an outer cross-validation loop (outerCV), which will be used for the proper estimation of each predictive model. The best hyperparameters are determined based on the F measure. To preserve the class ratio in each split of the training data, a ten-fold stratified CV will be applied to inner and outer loops.

Accuracy Evaluation. All the ML algorithms will be compared for prediction accuracy based on the validation data set. We will examine performance and prediction accuracy using the mean precision (positive predictive value), sensitivity (recall, true positive rate), specificity (true negative rate), F1 score, Youden's index, AUC and Matthews Correlation Coefficient (MCC). The optimal threshold of Youden's index or AUC can be determined through sensitivity, specificity, and MCC. Data with high Youden's index or AUC values near 1 indicate a high chance of correct classification, whereas low Youden's index and AUC values of models near 0 indicate a higher probability of making incorrect classifications. Plans for external validation include using a comprehensive statewide population database of all Persons Living with HIV in South Carolina. Second, we will also leverage a patient engagement studio specific for HIV to validate findings with PLWH and HIV care providers.

Patient and public involvement

None.

Ethics and dissemination

The Institutional Review Boards approved the study at the University of South Carolina (Pro00124806) as a Non-Human Subject study on 10/26/2022. A deeper understanding of the impact of the pandemic on viral suppression among underrepresented PLWH populations is essential to promote health equity and better direct clinical management and guideline development. The proposed personalized viral suppression prediction can provide data-driven evidence on tailored HIV treatment strategies for different underrepresented populations, particularly during unexpected interruptions like the COVID-19 pandemic. Thus, the results could facilitate the clinical identification of PLWH among underrepresented populations with poor viral control, provide them with tailored HIV care management, and eventually serve towards ending the HIV epidemic in the U.S.

We will publish the findings in peer-reviewed scientific journals and present the study findings at national and international professional conferences and through appropriate social media outlets. We will capitalize on social media and professional networks that can increase the reach and accessibility of findings, such as open-access publications, webinars, files, and videos available on websites and publicly available channels (e.g., YouTube), to increase the visibility and impact of the scientific publications and presentations. The dissemination efforts of this project will extend beyond the scientific arena and also target our stakeholders in the healthcare system and policymakers in the U.S. at local (SC DHEC, Prisma Health) and national levels (CDC) through various policy forums, policy papers, and special presentations.

Author Contribution (change later) BO and JZ is the principal investigator of this project and led the study design. XY contributed to the conception and design of the study. XY led the writing of this protocol manuscript. SW and XL contributed significantly to the editing of this manuscript. All authors reviewed and provided comments to improve the manuscript. All authors contributed to the editing and final approval of the protocol.

Funding statement This work was supported by the U.S. Department of Health and Human Services, National Institutes of Health, National Institute of Allergy And Infectious Diseases [grant number R01AI164947-02S1]. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

Conflict of interest The authors declare that there is no conflict of interest.

Figure legend/caption

Figure 1 Multilevel factors from multi-domains in All of Us program

Figure 2 Flowchart for data extraction and integration

Figure 3 Machine learning pipeline and relative data flow

References

1. Lee JS, Cole SR, Richardson DB, et al. Incomplete viral suppression and mortality in HIV patients after antiretroviral therapy initiation. AIDS **2017**; 31(14): 1989-97.
2. Services USDoHH. Ending the HIV Epidemic: A plan for America. available at: https://wwwhhsgov/blog/2019/02/05/ending-the-hiv-epidemic-a-plan-for-americahtml **2019**.
3. CDC U. Monitoring Selected National HIV Prevention and Care Objectives By Using HIV Surveillance Data United States and 6 Dependent Areas, 2019: Tables. Available at: https://www.cdc.gov/hiv/library/reports/hiv-surveillance/vol-26-no-2/content/tables.html. Accessed June 10.
4. Spinelli MA, Hickey MD, Glidden DV, et al. Viral suppression rates in a safety-net HIV clinic in San Francisco destabilized during COVID-19. AIDS **2020**; 34(15): 2328-31.
5. Norwood J, Kheshti A, Shepherd BE, et al. The Impact of COVID-19 on the HIV Care Continuum in a Large Urban Southern Clinic. AIDS Behav **2022**.
6. Hickey MD, Imbert E, Glidden DV, et al. Viral suppression during COVID-19 among people with HIV experiencing homelessness in a low-barrier clinic-based program. AIDS **2021**; 35(3): 517-9.
7. Izzo I, Carriero C, Gardini G, et al. Impact of COVID-19 pandemic on HIV viremia: a single-center cohort study in northern Italy. AIDS Res Ther **2021**; 18(1): 31.
8. Matsumoto S, Nagai M, Luong DAD, et al. Evaluation of SARS-CoV-2 Antibodies and the Impact of COVID-19 on the HIV Care Continuum, Economic Security, Risky Health Behaviors, and Mental Health Among HIV-Infected Individuals in Vietnam. AIDS Behav **2021**.
9. 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(7): e0234962.
10. CDC U. Estimated HIV incidence and prevalence in the United States, 2015–2019, and US Census Bureau, Quick Facts—United States. Available at: https://www.cdc.gov/hiv/pdf/library/reports/surveillance/cdc-hiv-surveillance-supplemental-report-vol-26-1.pdf.
11. United Nations. Shared Responsibility, Global Solidarity: Responding to the Socioeconomic Impacts of COVID-19. . Available at: https://www.un.org/sites/un2.un.org/files/sg_report_socio-economic_impact_of_covid19.pdf Accessed June 27.
12. Chenneville T, Gabbidon K, Hanson P, Holyfield C. The Impact of COVID-19 on HIV Treatment and Research: A Call to Action. Int J Environ Res Public Health **2020**; 17(12).
13. Investigators AoURP. The "All of Us" research program. New England Journal of Medicine **2019**; 381(7): 668-76.
14. Rodriguez VA, Tony S, Thangaraj P, et al. Phenotype Concept Set Construction from Concept Pair Likelihoods. In: AMIA Annual Symposium Proceedings: American Medical Informatics Association, 2020:1080.
15. All of Us Public Data Browser. View survey questions and answers. Available at: https://databrowser.researchallofus.org/survey/family-health-history.

16. PCORnet. PCORnet® COVID-19 Common Data Model Launched, Enabling Rapid Capture Of Insights On Patients Infected With The Novel Coronavirus. Available at: https://pcornet.org/news/pcornet-covid-19-common-data-model-launched-enabling-rapid-capture-of-insights/. Accessed June 21.
17. Burn E, You SC, Sena AG, et al. An international characterisation of patients hospitalised with COVID-19 and a comparison with those previously hospitalised with influenza. medRxiv **2020**.
18. LOINC. SARS-CoV-2 and COVID-19 related LOINC terms. Available at: https://loinc.org/sars-cov-2-and-covid-19/. Accessed June 21, 2022.
19. ATLAS. Atlas OHDSI concept sets. . Available at: http://atlas-covid19.ohdsi.org/#/home Accessed May 27, 2022.
20. Goldenholz DM, Sun H, Ganglberger W, Westover MB. Sample size analysis for machine learning clinical validation studies. Biomedicines **2023**; 11(3): 685.
21. Figueroa RL, Zeng-Treitler Q, Kandula S, Ngo LH. Predicting sample size required for classification performance. BMC medical informatics and decision making **2012**; 12: 1-10.
22. Althoff KN, Schlueter DJ, Anton-Culver H, et al. Antibodies to severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) in All of Us Research Program participants, 2 January to 18 March 2020. Clinical Infectious Diseases **2022**; 74(4): 584-90.
23. Sulieman L, Cronin RM, Carroll RJ, et al. Comparing medical history data derived from electronic health records and survey answers in the All of Us Research Program. Journal of the American Medical Informatics Association **2022**.
24. Breiman L. Random forests. Machine learning **2001**; 45(1): 5-32.
25. Friedman JH. Greedy function approximation: a gradient boosting machine. Annals of statistics **2001**: 1189-232.
26. Cortes C, Vapnik V. Support-vector networks. Machine Learning **1995**; 20(3): 273-97.
27. Yurduseven K, Babal YK, Celik E, Kerman BE, Kurnaz IA. Multiple Sclerosis Biomarker Candidates Revealed by Cell-Type-Specific Interactome Analysis. OMICS **2022**; 26(5): 305-17.

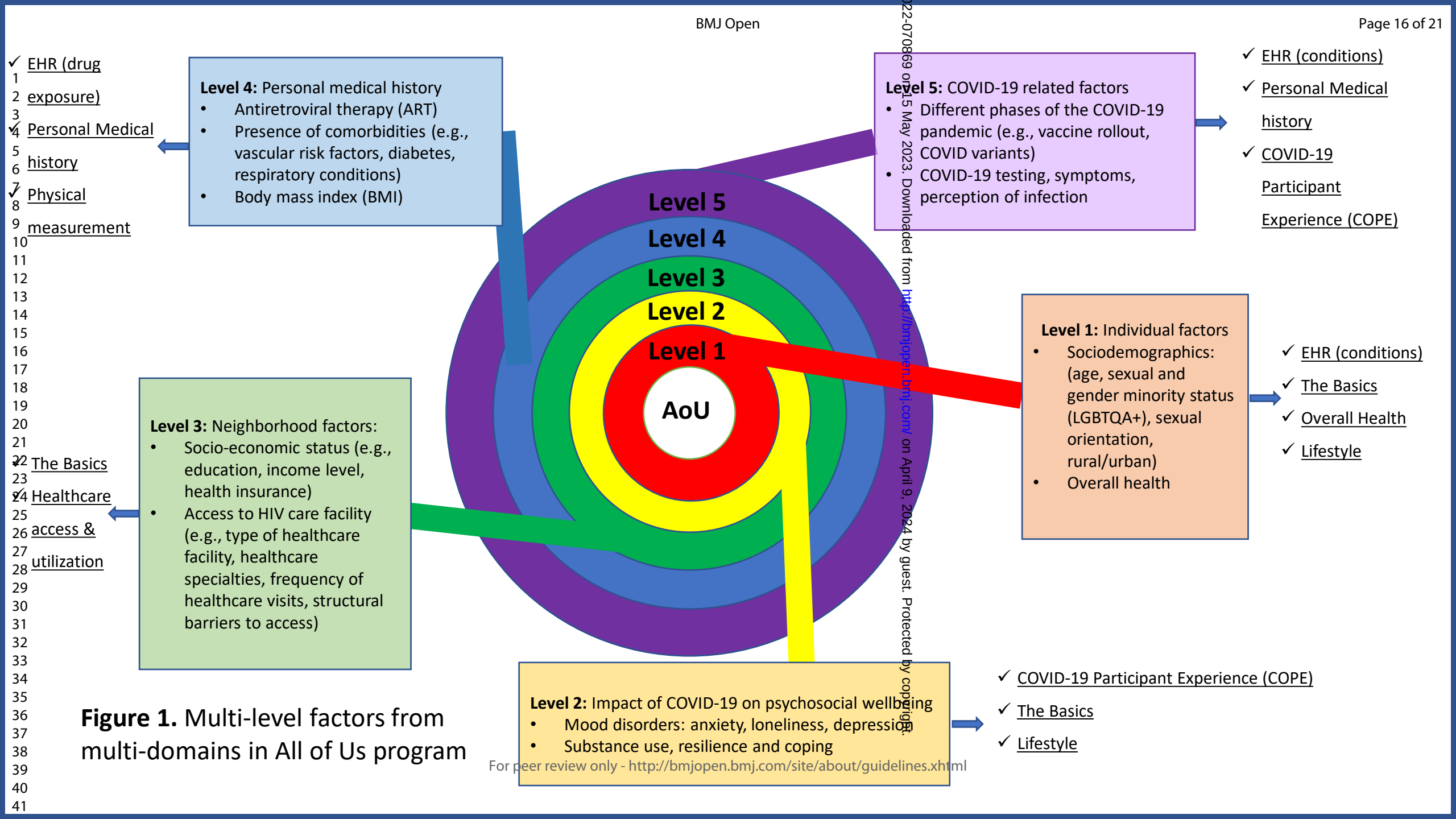

Figure 1. Multi-level factors from multi-domains in All of Us program

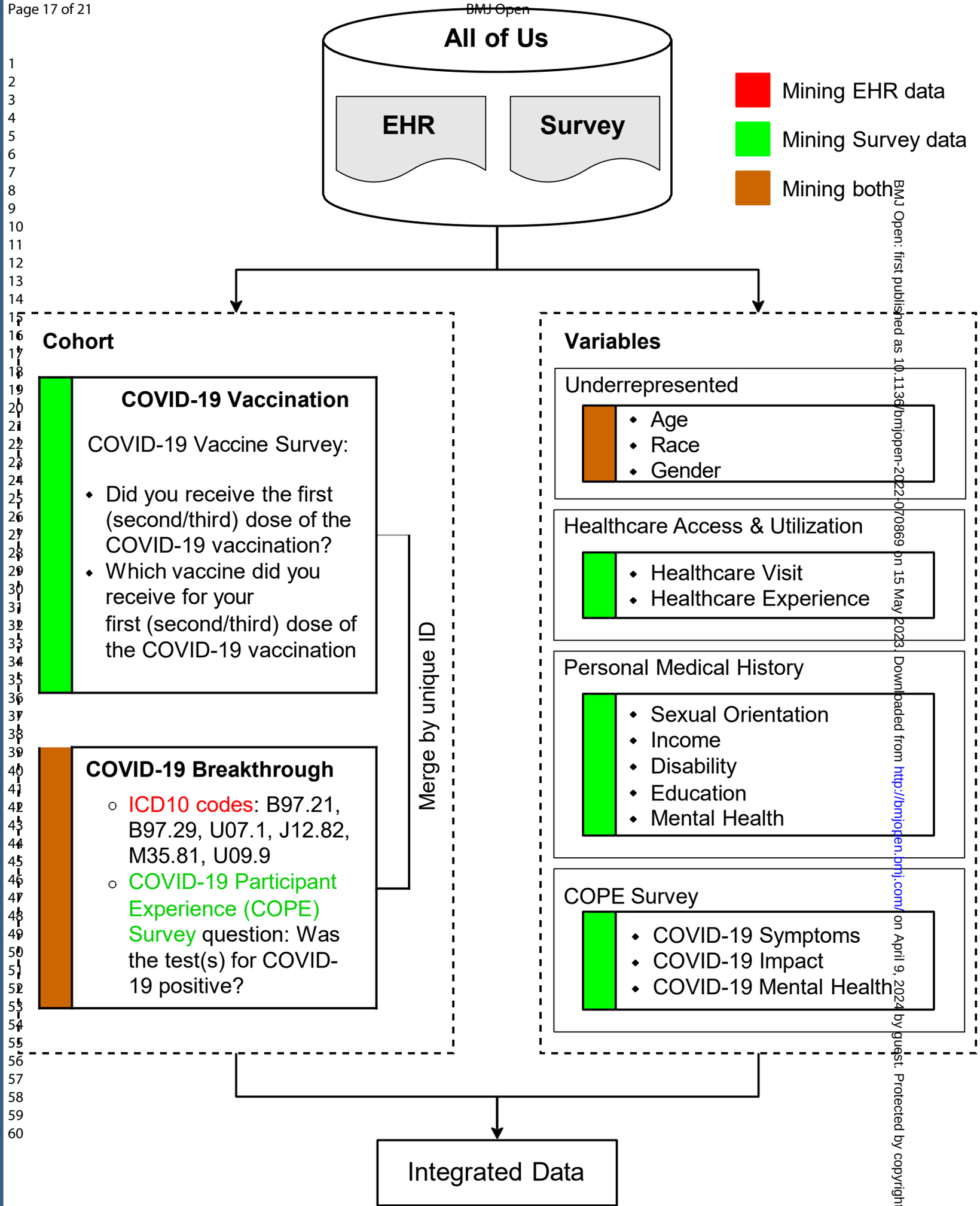

Figure 2. Flowchart for Data Extraction and Integration

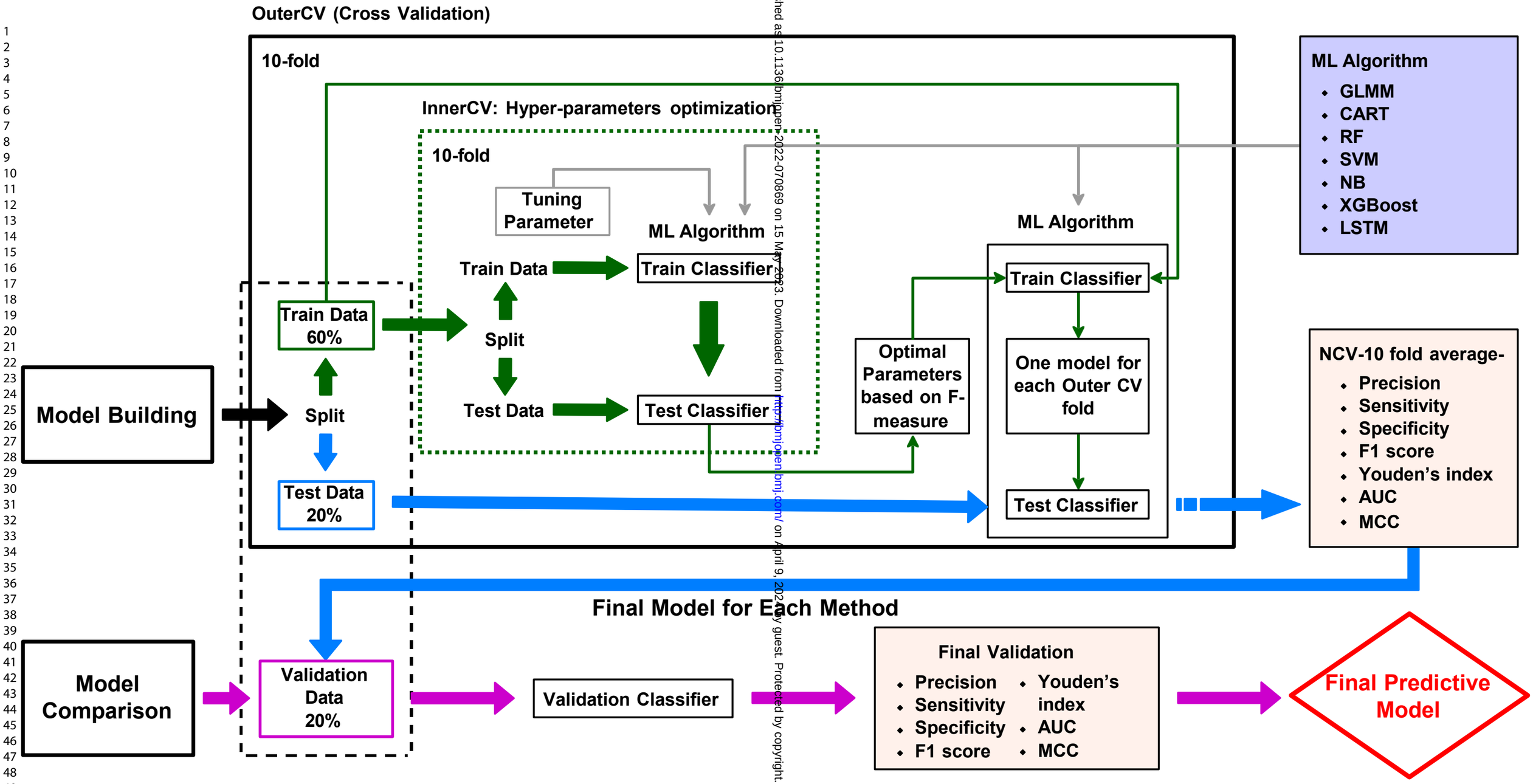

Figure 3. Machine Learning Pipeline and Relative Data Flow

STROBE Statement—Checklist of items that should be included in reports of ***cohort studies***

	Item No	Recommendation	Page No
Title and abstract	1	(*a*) Indicate the study's design with a commonly used term in the title or the abstract	1
		(*b*) Provide in the abstract an informative and balanced summary of what was done and what was found	2
Introduction			
Background/rationale	2	Explain the scientific background and rationale for the investigation being reported	4
Objectives	3	State specific objectives, including any prespecified hypotheses	5
Methods			
Study design	4	Present key elements of study design early in the paper	5
Setting	5	Describe the setting, locations, and relevant dates, including periods of recruitment, exposure, follow-up, and data collection	6
Participants	6	(*a*) Give the eligibility criteria, and the sources and methods of selection of participants. Describe methods of follow-up	6-7
		(*b*) For matched studies, give matching criteria and number of exposed and unexposed	
Variables	7	Clearly define all outcomes, exposures, predictors, potential confounders, and effect modifiers. Give diagnostic criteria, if applicable	8
Data sources/ measurement	8*	For each variable of interest, give sources of data and details of methods of assessment (measurement). Describe comparability of assessment methods if there is more than one group	6
Bias	9	Describe any efforts to address potential sources of bias	9-11
Study size	10	Explain how the study size was arrived at	5
Quantitative variables	11	Explain how quantitative variables were handled in the analyses. If applicable, describe which groupings were chosen and why	9-11
Statistical methods	12	(*a*) Describe all statistical methods, including those used to control for confounding	9-11
		(*b*) Describe any methods used to examine subgroups and interactions	
		(*c*) Explain how missing data were addressed	
		(*d*) If applicable, explain how loss to follow-up was addressed	
		(*e*) Describe any sensitivity analyses	
Results			
Participants	13*	(a) Report numbers of individuals at each stage of study—eg numbers potentially eligible, examined for eligibility, confirmed eligible, included in the study, completing follow-up, and analysed	N/A
		(b) Give reasons for non-participation at each stage	
		(c) Consider use of a flow diagram	
Descriptive data	14*	(a) Give characteristics of study participants (eg demographic, clinical, social) and information on exposures and potential confounders	N/A
		(b) Indicate number of participants with missing data for each variable of interest	
		(c) Summarise follow-up time (eg, average and total amount)	
Outcome data	15*	Report numbers of outcome events or summary measures over time	N/A

Main results	16	(*a*) Give unadjusted estimates and, if applicable, confounder-adjusted estimates and their precision (eg, 95% confidence interval). Make clear which confounders were adjusted for and why they were included (*b*) Report category boundaries when continuous variables were categorized (*c*) If relevant, consider translating estimates of relative risk into absolute risk for a meaningful time period	N/A
Other analyses	17	Report other analyses done—eg analyses of subgroups and interactions, and sensitivity analyses	N/A
Discussion			**N/A**
Key results	18	Summarise key results with reference to study objectives	
Limitations	19	Discuss limitations of the study, taking into account sources of potential bias or imprecision. Discuss both direction and magnitude of any potential bias	
Interpretation	20	Give a cautious overall interpretation of results considering objectives, limitations, multiplicity of analyses, results from similar studies, and other relevant evidence	
Generalisability	21	Discuss the generalisability (external validity) of the study results	
Other information			
Funding	22	Give the source of funding and the role of the funders for the present study and, if applicable, for the original study on which the present article is based	13

*Give information separately for exposed and unexposed groups.

Note: An Explanation and Elaboration article discusses each checklist item and gives methodological background and published examples of transparent reporting. The STROBE checklist is best used in conjunction with this article (freely available on the Web sites of PLoS Medicine at http://www.plosmedicine.org/, Annals of Internal Medicine at http://www.annals.org/, and Epidemiology at http://www.epidem.com/). Information on the STROBE Initiative is available at http://www.strobe-statement.org.