



BMJ Open Protocol for developing a personalised prediction model for viral suppression among under-represented populations in the context of the COVID-19 pandemic

Jiajia Zhang,^{1,2} Xueying Yang ,³ Sharon Weissman,⁴ Xiaoming Li,^{2,3} Bankole Olatosi ^{2,5}

To cite: Zhang J, Yang X, Weissman S, *et al.* Protocol for developing a personalised prediction model for viral suppression among under-represented populations in the context of the COVID-19 pandemic. *BMJ Open* 2023;**13**:e070869. doi:10.1136/bmjopen-2022-070869

► Prepublication history for this paper is available online. To view these files, please visit the journal online (<http://dx.doi.org/10.1136/bmjopen-2022-070869>).

Received 07 December 2022
Accepted 19 April 2023



© Author(s) (or their employer(s)) 2023. Re-use permitted under CC BY-NC. No commercial re-use. See rights and permissions. Published by BMJ.

For numbered affiliations see end of article.

Correspondence to

Dr Bankole Olatosi;
olatosi@mailbox.sc.edu

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’ federal campaign launched in 2019. Under-represented 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 under-represented people living with HIV (PLWH) due to interruptions in healthcare access and other worsened socioeconomic and environmental conditions. However, biomedical research rarely includes under-represented populations, resulting in biased algorithms. This proposal targets a broadly defined under-represented HIV population. It aims to develop a personalised viral suppression prediction model using machine learning (ML) techniques by incorporating multilevel factors using All of Us (AoU) data.

Methods and analysis This cohort study will use data from the AoU research programme, which aims to recruit a broad, diverse group of US populations historically under-represented in biomedical research. The programme harmonises data from multiple sources on an ongoing basis. It has recruited ~4800 PLWH with a series of self-reported survey data (eg, Lifestyle, Healthcare Access, COVID-19 Participant Experience) and relevant longitudinal electronic health records data. We will examine the change in viral suppression and develop personalised viral suppression prediction due to the impact of the COVID-19 pandemic using ML techniques, such as tree-based classifiers (classification and regression trees, random forest, decision tree and eXtreme Gradient Boosting), support vector machine, naïve Bayes and long short-term memory.

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.

INTRODUCTION

Sustained viral suppression, an indicator of long-term treatment success and mortality reduction,¹ is one of four strategic areas of the

STRENGTHS AND LIMITATIONS OF THIS STUDY

- ⇒ The diverse group of populations recruited in the All of Us research programme enables us to have a large representative sample of under-represented populations in biomedical research and reduce algorithmic bias.
- ⇒ The data integration from multiple data sources in the All of Us research programme for cohort analyses allows us to robustly evaluate the viral suppression prediction for the under-represented population with a long follow-up.
- ⇒ The machine learning-based approach to developing personalised prediction for viral suppression has the benefit of accurately modelling 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.

‘Ending the HIV Epidemic (EtHE): A Plan for America’² federal campaign launched in 2019. According to the Centers for Disease Control and Prevention (CDC) national surveillance data, approximately 66% of all people living with HIV (PLWH) were virally suppressed in the USA.³ The COVID-19 pandemic uniquely affects PLWH and has a mixed impact on viral suppression across different countries or settings. In the USA, 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

under-represented in biomedical research (hereafter referred to as ‘under-represented population’).⁹ HIV and COVID-19 both have a disproportionate impact on under-represented 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 Nations¹¹ report has indicated that increases in food costs and market stockpiling during the COVID-19 pandemic have had the most harmful impact on under-represented populations. Those with stigmatised or marginalised 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 under-represented 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 are not designed for a specific research purpose. Thus, it has a limited capacity to recruit an adequate sample of under-represented populations due to their historically limited access to specialty care and academic medical centres that serve as the primary sources for EHR data. Consequently, it poses more challenges in understanding the viral suppression among under-represented populations, particularly those facing the COVID-19 pandemic.

The All of Us (AoU) research programme is an ongoing national, historic effort supported by the NIH. The cohort in AoU includes a broadly diverse group of the US population, with more than 50% of the participants from racial and ethnic minority groups and more than 80% from populations historically under-represented in biomedical research (eg, sex orientation, socioeconomic status, geographical location, physical disability). Therefore, this protocol aims to target under-represented populations using AoU data, which includes ~4800 PLWH with a series of self-reported survey data (eg, Lifestyle, Physical Measurement, Healthcare Access, COVID-19 Participant Experience (COPE)) 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 under-represented HIV population by harnessing the AoU big data resources.

Aim 2: Develop personalised viral suppression prediction models using machine learning (ML) techniques by incorporating COVID-19 interruption, antiretroviral

therapy history, pre-existing conditions (comorbidities), psychological well-being (eg, depression, resilience), healthcare utilisation and social, environmental factors in AoU.

A deeper understanding of the impact of the pandemic on viral suppression among under-represented PLWH populations is essential to promote health equity and better direct clinical management and guideline development. The proposed personalised viral suppression prediction can provide data-driven evidence on tailored HIV treatment strategies for different under-represented populations, particularly during unexpected interruptions like the COVID-19 pandemic. Thus, the results could facilitate the clinical identification of PLWH among under-represented populations with poor viral control, provide them with tailored HIV care management and eventually serve towards the goal of EtHE in the USA. 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 (eg, age, race and gender) and overall health, will be extracted from ‘The Basics’ survey. The COVID-19-related experiences (levels 2 and 5) refer to the impact of the pandemic on their health and psychosocial well-being, such as social support, depression, anxiety, drug and alcohol abuse, and resilience, will be extracted from ‘COPE’ and ‘Lifestyle’ surveys. The neighbourhood-level factors (level 3), including the neighbourhood economic environment (eg, poverty, education, health insurance coverage) and healthcare access (type of healthcare facility, structural barriers to healthcare access), will be defined from the ‘healthcare access and utilisation’ survey. With the appropriate data management/preprocessing, we will examine the change in viral suppression and develop the personalised viral suppression prediction due to the impact of the COVID-19 pandemic using ML techniques, which will have translational potential to inform future HIV care among under-represented populations.

Data sources

Overview of the AoU programme

The AoU research programme seeks to recruit persons in demographic categories that have been and continue to be under-represented in biomedical research; such persons typically have relatively poor access to good healthcare.¹³ AoU opened for enrolment in May 2018, and the inclusion criteria

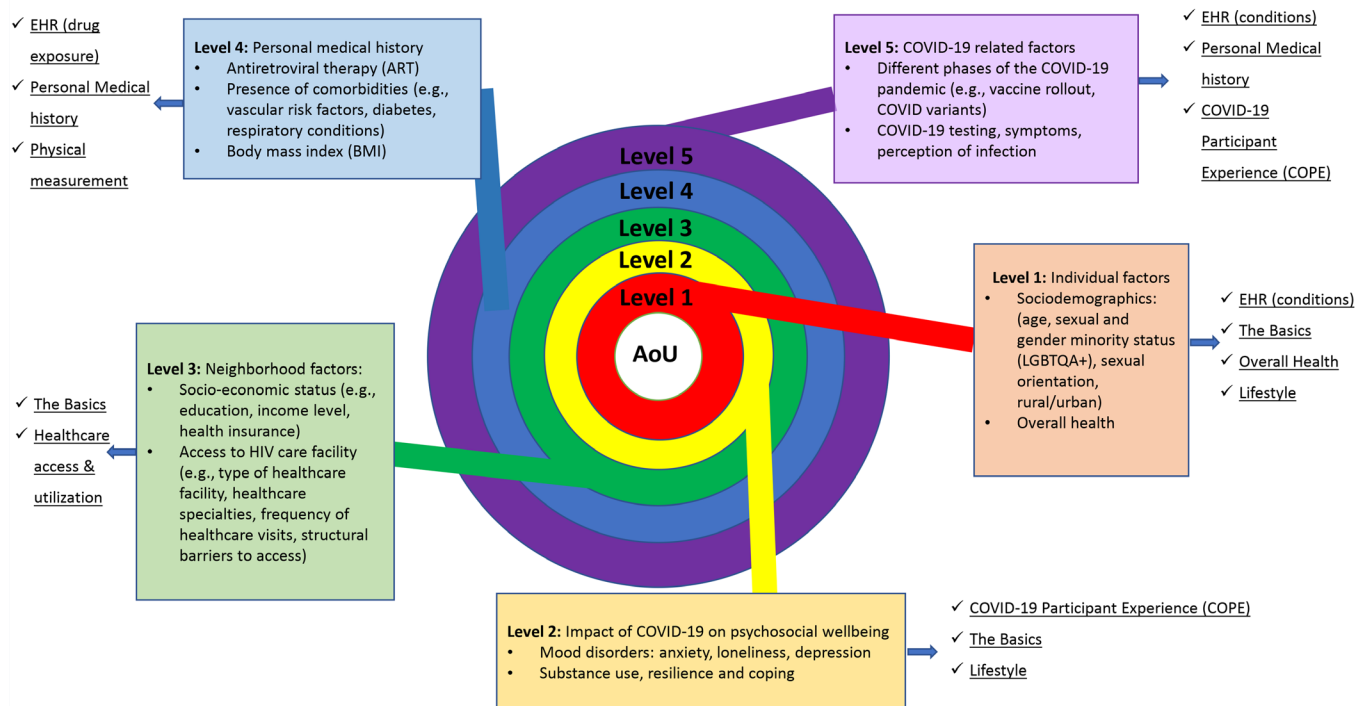


Figure 1 Multilevel factors from multidomains in All of Us programme.

are age ≥ 18 years with the capacity to provide consent. The recruitment methods and scientific rationale for AoU have been described previously.¹³ Through 19 November 2021, AoU has harmonised 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 harmonising data on an ongoing basis. Each participant completed informed consent for sharing their EHR data with the Data and Research Centre 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 harmonisation. 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 analyse a centralised version of the AoU data.

'HIV and COVID-19' project in the AoU researcher workbench platform

AoU research programme data in its final format, after harmonisation 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 V.3.0 (Python Software Foundation) programming languages within Jupyter Notebooks. The purpose of our workspace is twofold: (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 (eg, a disease). Computable phenotypes are essential for analysing 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 (eg, diabetes) is the base instance, combined with all possible feature representations in data (eg, International Classification of Diseases codes (ICD) codes for diabetes+insulin; or ICD codes for diabetes+haemoglobin 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

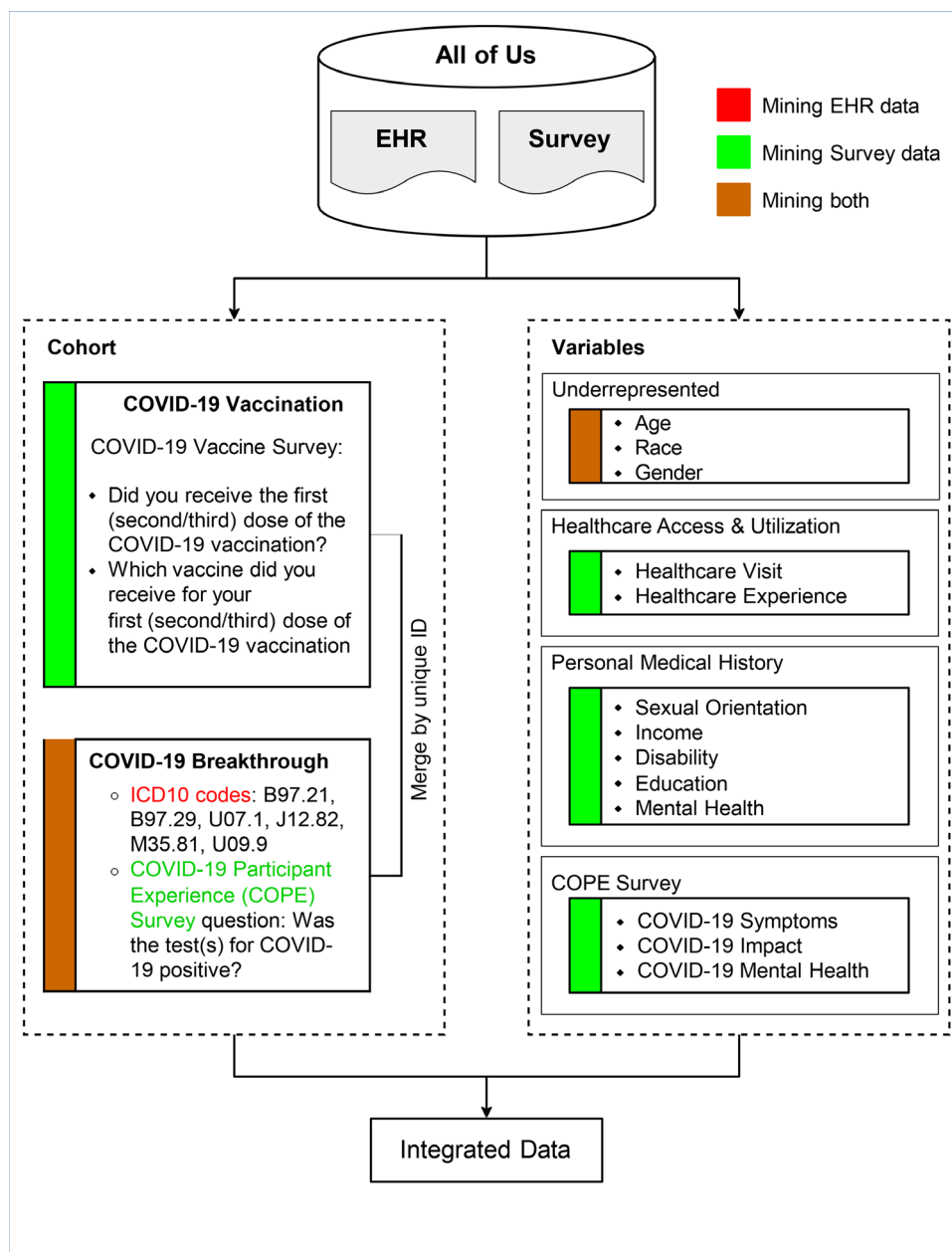


Figure 2 Flow chart for data extraction and integration.

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 (eg, HIV population) and relevant variables (eg, 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 Systemised Nomenclature of Medicine-Clinical Terms diagnosis codes (the standardised vocabulary in AoU sourced from corresponding 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 V.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 organisations—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 characterisation 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 (eg, condition domain, drug domain) to best identify individuals with a particular clinical entity (‘best of breed’) (eg, HIV

Table 1 Characteristics of under-represented population of people living with HIV in All of US programme data

Characteristics	N (%)	Characteristics	N (%)
Data from EHR and survey		Data from survey only	
Total HIV population	4794 (100%)	Sex/gender (n=1080)	
Age		LGBTQIA+, no	291 (26.94)
<75 years	4619 (96.35)	LGBTQIA+, yes	789 (73.06)
≥75 years	175 (3.65)	Education (n=1067)	
Race		High school degree or more	977 (90.46)
White	1232 (25.70)	Less than a high school degree	90 (8.33)
Black or African American	2448 (51.06)	Household Income (n=980)	
Asian	29 (0.60)	>US\$35 000	621 (63.37)
Other/unknown	1085 (22.63)	<US\$35 000	359 (36.63)
COVID-19 infection		Physical disability (n=1069)	
Yes	402 (8.39)	No	958 (89.62)
No	4392 (91.61)	Yes	111 (10.38)

EHR, electronic health record; LGBTQIA+, Gay, Lesbian, Bisexual, Transgender, Queer, Intersex, and Asexual people collectively.

infection). In EHR data, we will define HIV by documentation of any of the following: (1) HIV condition (ICD, 9th/10th Revision (ICD-9/10) diagnostic codes, ICD-9/10 procedure codes) in the ‘condition’ domain; (2) HIV-related laboratory results (eg, HIV antibody) in the ‘Labs &Measurements’ domain or (3) HIV-related medications (eg, 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 healthcare 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 sets¹⁹ based on all the above information will be built for the HIV cohort (see [table 1](#) for summary characteristics).

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 January 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 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

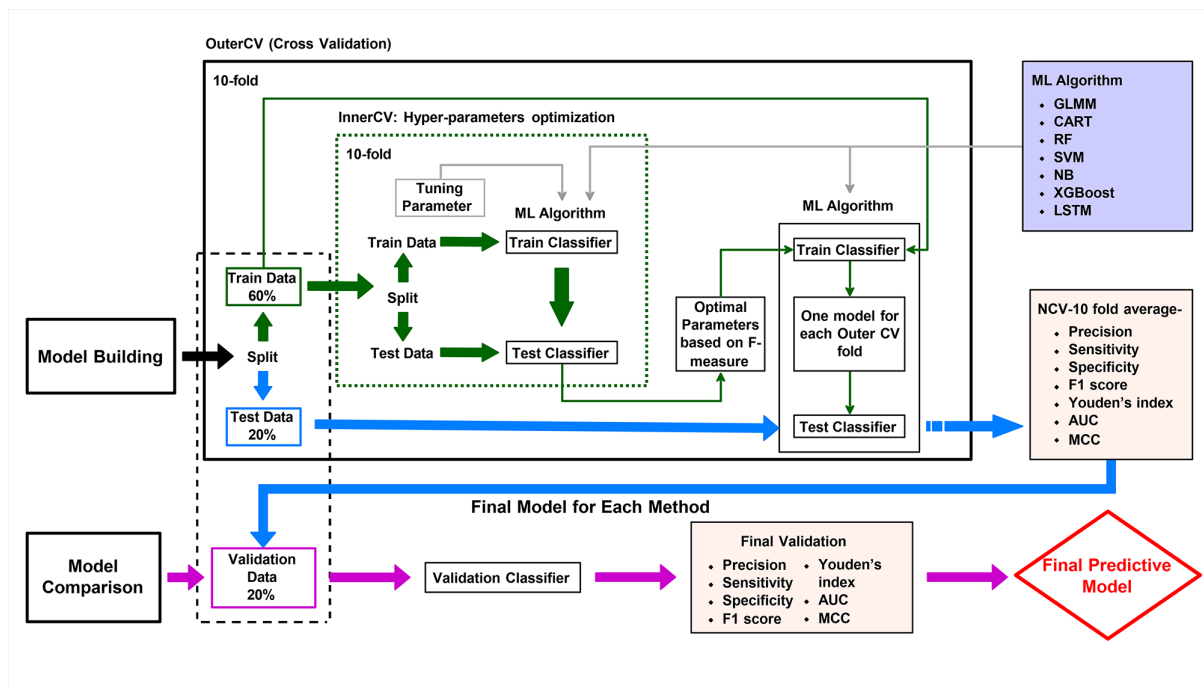


Figure 3 Machine learning pipeline and relative data flow.

by the COVID-19 status/time periods. The absolute CD4 cell count will be treated as a continuous variable and a categorical variable (categorised into <200, 200–500, >500 cells/mm³). 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 (eg, NRTI-based, NNRTI-based, PI-based or multiclass regimen with three or more classes of ART) or (3) specific drug regimens (eg, 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 healthcare. 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 1 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 organised into 12 disease categories.²³ We will use a combination of self-reported responses to the 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), hyperlipidaemia (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 healthcare 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 utilisation

The healthcare utilisation information is extracted from the 'healthcare access and utilisation' survey data. It includes health insurance, type of healthcare facility visits (eg, urgent care, emergency room), healthcare specialties (eg, nurse practitioner, physician assistant, mental health professional), frequency of healthcare visits, patient-provider communication, structural barriers of healthcare access (eg, lack of transportation, long distance to

a healthcare provider, the affordability of medical cost), compromised adherence due to unaffordability, and stigmatised 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 under-represented population will be summarised (mean, SD, counts) and compared using the t-test, analysis of variance test or χ^2 test as appropriate. If test assumptions are not satisfied, non-parametric 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 generalised linear mixed regression with different prespecified correlation matrix as appropriate such as autoregression covariance matrices and choose the best model based on Quasi information criterion (QIC) to evaluate the differences in the probability of viral suppression between prepandemic and peripandemic periods (using March 2020 as a time cut-off, when the first COVID-19 case was reported in the USA) adjusting for key demographic characteristics (eg, under-represented population) and other variables. The model will be built sequentially by (1) including the characteristics of under-represented individuals only for the crude model, (2) adding the COVID-19 indicators, (3) the interaction between the under-represented population and COVID-19 status and (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 Akaike Information Criterion (AIC) or Bayesian Information Criterion (BIC) criteria. Depending on the sample size of subset of interest in the integrated data, we could (1) conduct a stratified analysis for each under-represented population using similar generalised linear mixed regression models and (2) add the interaction term between under-represented population and COVID-19 pandemic indicator. We will use forest plots will be used to display the regression results.

Personalised 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, healthcare utilisation, HIV markers, COVID-19 infection/interruption and other socialenvironmental 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 under-represented 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 (generalised linear mixed model), tree-based classifiers (classification and regression trees, random forest,²⁴ decision tree, and eXtreme Gradient Boosting²⁵), support vector machine,²⁶ naïve Bayes, and long short-term memory. The input feature includes all information extracted from the integrated dataset. To account for time-dependent variables (ie, 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-validation²⁷ 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 hyperparameters 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 10-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, Area Under the Curve (AUC) and Matthew's 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 state-wide population database of all PLWH 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 26 October 2022. A deeper understanding of the impact of the pandemic on viral suppression among under-represented PLWH populations is essential to promote health equity and better direct clinical management and guideline development. The proposed personalised viral suppression prediction can provide data-driven evidence on tailored HIV treatment strategies for different under-represented populations, particularly during unexpected interruptions like the COVID-19 pandemic. Thus, the results could facilitate the clinical identification of PLWH among under-represented populations with poor viral control, provide them with tailored HIV care management, and eventually serve towards EtHE in the USA.

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 capitalise 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 (eg, 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 policy-makers in the USA at local (SC Department of Health and Environmental Control (DHEC), Prisma Health) and national levels (CDC) through various policy forums, policy papers and special presentations.

Author affiliations

¹Department of Epidemiology and Biostatistics, University of South Carolina Arnold School of Public Health, Columbia, South Carolina, USA

²South Carolina SmartState Center for Healthcare Quality, University of South Carolina Arnold School of Public Health, Columbia, South Carolina, USA

³Health Promotion Education and Behavior, University of South Carolina Arnold School of Public Health, Columbia, South Carolina, USA

⁴Department of Internal Medicine, School of Medicine, University of South Carolina, Columbia, South Carolina, USA

⁵Health Services Policy and Management, University of South Carolina Arnold School of Public Health, Columbia, South Carolina, USA

Contributors 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 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.

Competing interests None declared.

Patient and public involvement Patients and/or the public were not involved in the design, or conduct, or reporting, or dissemination plans of this research.

Patient consent for publication Not applicable.

Provenance and peer review Not commissioned; externally peer reviewed.

Open access This is an open access article distributed in accordance with the Creative Commons Attribution Non Commercial (CC BY-NC 4.0) license, which permits others to distribute, remix, adapt, build upon this work non-commercially, and license their derivative works on different terms, provided the original work is properly cited, appropriate credit is given, any changes made indicated, and the use is non-commercial. See: <http://creativecommons.org/licenses/by-nc/4.0/>.

ORCID iDs

Xueying Yang <http://orcid.org/0000-0001-6788-0688>

Bankole Olatosi <http://orcid.org/0000-0002-8295-8735>

REFERENCES

- Lee JS, Cole SR, Richardson DB, *et al*. Incomplete viral suppression and mortality in HIV patients after antiretroviral therapy initiation. *AIDS* 2017;31:1989–97.
- Services USDoHH. Ending the HIV epidemic: a plan for America. 2019. Available: <https://www.hhs.gov/blog/2019/02/05/ending-the-hiv-epidemic-a-plan-for-america.html>
- CDC U. Monitoring selected national HIV prevention and care objectives by using HIV surveillance data United States and 6 dependent areas, 2019: tables. n.d. Available: <https://www.cdc.gov/hiv/library/reports/hiv-surveillance/vol-26-no-2/content/tables.html>
- 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;34:2328–31.
- 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;26:2825–9.
- 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;35:517–9.
- 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;18:31.
- 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* 2022;26:1095–109.
- Mapes BM, Foster CS, Kusnoor SV, *et al*. Diversity and inclusion for the all of US research program: a scoping review. *PLoS One* 2020;15:e0234962.
- CDC U. Estimated HIV incidence and prevalence in the United States, 2015–2019, and US census bureau, quick facts—United states. n.d. Available: <https://www.cdc.gov/hiv/pdf/library/reports/surveillance/cdc-hiv-surveillance-supplemental-report-vol-26-1.pdf>
- United Nations. Shared responsibility, global solidarity: responding to the socioeconomic impacts of COVID-19. n.d. Available: https://www.un.org/sites/un2.un.org/files/sg_report_socio-economic_impact_of_covid19.pdf
- Chenneville T, Gabbidon K, Hanson P, *et al*. The impact of COVID-19 on HIV treatment and research: a call to action. *Int J Environ Res Public Health* 2020;17:12.
- Denny JC, Rutter JL, Goldstein DB, *et al*. The "All of US" research program. *N Engl J Med* 2019;381:668–76.
- 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.
- All of US public data browser. View survey questions and answers. n.d. Available: <https://databrowser.researchallofus.org/survey/family-health-history>
- PCORnet. PCORnet® COVID-19 common data model launched, enabling rapid capture of insights on patients infected with the novel coronavirus. n.d. Available: <https://pcor.net.org/news/pcor-net-covid-19-common-data-model-launched-enabling-rapid-capture-of-insights/>

- 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: <https://loinc.org/sars-cov-2-and-covid-19/> [Accessed 21 Jun 2022].
- 19 ATLAS. Atlas OHDSI concept sets. Available: <http://atlas-covid19.ohdsi.org/#/home> [Accessed 27 May 2022].
- 20 Goldenholz DM, Sun H, Ganglberger W, *et al.* Sample size analysis for machine learning clinical validation studies. *Biomedicines* 2023;11:685.
- 21 Figueroa RL, Zeng-Treitler Q, Kandula S, *et al.* Predicting sample size required for classification performance. *BMC Med Inform Decis Mak* 2012;12:8.
- 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. *Clin Infect Dis* 2022;74:584–90.
- 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. *J Am Med Inform Assoc* 2022;29:1131–41.
- 24 Breiman L. Random forests. *Mach Learn* 2001;45:5–32.
- 25 Friedman JH. Greedy function approximation: a gradient boosting machine. *Ann Statist* 2001;29:1189–232.
- 26 Cortes C, Vapnik V. Support-vector networks. *Mach Learn* 1995;20:273–97.
- 27 Yurduseven K, Babal YK, Celik E, *et al.* Multiple sclerosis biomarker candidates revealed by cell-type-specific interactome analysis. *OMICS* 2022;26:305–17.