




# BMJ Open Curating a knowledge base for individuals with coinfection of HIV and SARS-CoV-2: a study protocol of EHR-based data mining and clinical implementation

Chen Liang <sup>1,2</sup>, Sharon Weissman,<sup>2,3</sup> Bankole Olatosi <sup>1,2</sup>, Eric G Poon,<sup>4</sup> Michael E Yarrington <sup>4</sup>, Xiaoming Li <sup>2,5</sup>

**To cite:** Liang C, Weissman S, Olatosi B, *et al.* Curating a knowledge base for individuals with coinfection of HIV and SARS-CoV-2: a study protocol of EHR-based data mining and clinical implementation. *BMJ Open* 2022;**12**:e067204. doi:10.1136/bmjopen-2022-067204

► Prepublication history for this paper is available online. To view these files, please visit the journal online (<http://dx.doi.org/10.1136/bmjopen-2022-067204>).

Received 05 August 2022  
Accepted 25 August 2022



© Author(s) (or their employer(s)) 2022. 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 Chen Liang;  
[cliang@mailbox.sc.edu](mailto:cliang@mailbox.sc.edu)

## ABSTRACT

**Introduction** Despite a higher risk of severe COVID-19 disease in individuals with HIV, the interactions between SARS-CoV-2 and HIV infections remain unclear. To delineate these interactions, multicentre Electronic Health Records (EHR) hold existing promise to provide full-spectrum and longitudinal clinical data, demographics and sociobehavioural data at individual level. Presently, a comprehensive EHR-based cohort for the HIV/SARS-CoV-2 coinfection has not been established; EHR integration and data mining methods tailored for studying the coinfection are urgently needed yet remain underdeveloped.

**Methods and analysis** The overarching goal of this exploratory/developmental study is to establish an EHR-based cohort for individuals with HIV/SARS-CoV-2 coinfection and perform large-scale EHR-based data mining to examine the interactions between HIV and SARS-CoV-2 infections and systematically identify and validate factors contributing to the severe clinical course of the coinfection. We will use a nationwide EHR database in the USA, namely, National COVID Cohort Collaborative (N3C). Ultimately, collected clinical evidence will be implemented and used to pilot test a clinical decision support prototype to assist providers in screening and referral of at-risk patients in real-world clinics.

**Ethics and dissemination** The study was approved by the institutional review boards at the University of South Carolina (Pro00121828) as non-human subject study. Study findings will be presented at academic conferences and published in peer-reviewed journals. This study will disseminate urgently needed clinical evidence for guiding clinical practice for individuals with the coinfection at Prisma Health, a healthcare system in collaboration.

## INTRODUCTION

The COVID-19 pandemic has cast a heavy burden on individuals with HIV infection. Based on data from 15 522 hospitalised patients with the coinfection of HIV and SARS-CoV-2 from 24 countries, a recent WHO report for the first time confirmed that HIV is a key risk factor for severe COVID-19.<sup>1</sup>

## STRENGTHS AND LIMITATIONS OF THIS STUDY

- ⇒ This study will be among the first that systematically integrates HIV viral suppression status, antiretroviral therapy (ART) adherence, vaccination, sociobehavioural and social determinants of health with full-spectrum clinical characteristics for individuals with HIV/SARS-CoV-2 coinfection.
- ⇒ Our methods can explain the role of temporal dependency among patients' underlying conditions, comorbidities, ART adherence, vaccine exposure and received therapeutics in individuals' heterogeneous responses to the coinfection.
- ⇒ Our methods support real-time prediction of coinfecting individuals' clinical outcomes, disease progression, prognosis, and risk factors of adverse events.
- ⇒ The proposed methods are highly innovative in that they are designed to extract temporal sequences and temporal properties of every clinical event from Electronic Health Records and are fully capable of embedding the temporal data into machine learning models.
- ⇒ The study only includes pilot validity and usability testing of the proposed clinical decision support prototype due to limited time for an exploratory/developmental study.

The severity of COVID-19 in individuals with HIV is correlated with certain comorbidities (eg, type 2 diabetes mellitus, cardiovascular diseases, obesity, chronic obstructive pulmonary diseases, chronic kidney diseases, and some cancers) in which some comorbidities are more prevalent in people living with HIV (PLWH). Individuals with low CD4<sup>+</sup> T-cell count (eg, <200 cells/ $\mu\text{L}$ <sup>2</sup> or <500<sup>3</sup> cells/ $\mu\text{L}$ ) and unsuppressed viral load, and prolonged antiretroviral therapy (ART) exposure are associated with severe clinical course. These clinical facts are further complicated by the disrupted HIV healthcare services (eg, access

to HIV testing, ART and distribution of pre-exposure prophylaxis and post-exposure prophylaxis).<sup>4</sup>

Despite a generally high risk of severe COVID-19 clinical course in PLWH, the interactions between SARS-CoV-2 and HIV infections remain unclear. First, several contradictory findings suggested the predominant role of comorbidities in severity of COVID-19 regardless of HIV infection.<sup>5–8</sup> Second, risk factors for the severe clinical course of the coinfection are undetermined because individuals with the same or similar severity level of COVID-19 show different clinical characteristics.<sup>4</sup> Third, the role of ART adherence and HIV viral suppression status in the context of COVID-19 exposure is undetermined. These unsolved problems are attributed by several data and methodological gaps. For example, most existing studies are based on small-sample and single-centre cohorts. Temporal sequences and patterns of clinical events (eg, underlying conditions, comorbidities, diagnoses, ART-related visits and treatments) are understudied, which diminish the opportunities for understanding the aetiology of multifaceted HIV-associated comorbidities, their natural history and their interactions with the current coinfection. Critical data components such as adherence to HIV treatment, viral suppression, social determinants of health (SDOH), COVID-19 vaccination and sociobehavioural patterns (eg, substance use/dependence) are closely related to disparities in HIV and SARS-CoV-2 infections but are understudied in part due to the challenges in Electronic Health Records (EHR) data integration and phenotyping. EHR hold existing promise to provide full-spectrum and longitudinal clinical data, demographics and sociobehavioural data at the individual level. However, currently we do not have a comprehensive EHR-based cohort for individuals with HIV/SARS-CoV-2 coinfection; EHR integration and data mining tailored for studying the coinfection are urgently needed but are not yet developed.

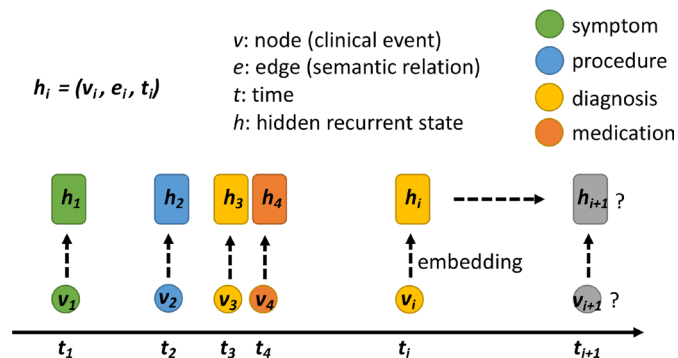
The overarching goal of this exploratory/developmental study is to establish an EHR-based cohort for individuals with HIV/SARS-CoV-2 coinfection and perform large-scale EHR-based data mining to examine the interactions between HIV and SARS-CoV-2 infections and systematically identify and validate factors contributing to the severe clinical course of the coinfection. Ultimately, collected clinical evidence will be implemented and used to pilot test a clinical decision support (CDS) prototype to assist providers in screening and referral of at-risk patients in real-world clinics. We will approach this goal by pursuing the following tasks. First, we will extract comprehensive phenotypic traits (ie, clinical characteristics, demographics, sociobehavioural patterns) and their temporal series and patterns from structured and unstructured EHR—National COVID Cohort Collaborative (N3C).<sup>9</sup> To extract and model temporal series and patterns of phenotypic traits, we will incorporate biomedical ontologies to develop a graphical model of EHR. Second, we will examine patterns and sequences of phenotypic traits for their predictive

ability in clinical outcomes and disease prognosis. Major phenotypic traits to be examined include demographics, underlying conditions, comorbidities, CD4<sup>+</sup> counts, viral suppression, ART procedures and medications, laboratory results for immune components and viral presence, treatments (eg, procedures and medications), SDOH, and sociobehavioural patterns. We will develop machine learning models to explore real-time predictive associations between these phenotypic traits and poor clinical outcomes and prognosis, including outcomes of the acute phase of COVID-19 and postacute sequelae of SARS-CoV-2 infection (PASC).<sup>10</sup> Third, we will develop and pilot test a CDS prototype that delivers collected clinical evidence to providers through the Epic EHR system at Prisma Health. Predictive associations generated from the second task will be presented for providers to assist in screening patients at high risk of severe COVID-19 course. Outcomes to be measured include (1) the rate of identification and referral of individuals at high risk of poor clinical outcomes, (2) the rate of successful referral and clinical actions and (3) system usability. The proposed study protocol will result in (1) a comprehensive knowledge base that details risk factors of severe clinical outcomes and disease prognosis in individuals with HIV/SARS-CoV-2 coinfection and (2) a prototype CDS that can identify patients at high risk and provide actionable clinical decisions. This work will provide time-sensitive public health implications: clinical evidence for interactions between HIV and SARS-CoV-2 infections is desperately needed. This proposed EHR-based data mining offers a rapid and empirically grounded approach to collecting such evidence and to informing the design of prospective clinical trials that can focus on inflammatory pathways, biophysiological evidence of the coinfection and sociobehavioural determinants.

## METHODS AND ANALYSIS

### Data description

We will use EHR from N3C. As of 8/2022, N3C has aggregated 15.2 million patients (5.8 million COVID-19 patients) from 50 states.<sup>11</sup> The EHR are normalised by the Observational Health Data Sciences and Informatics (OHDSI)'s Observational Medical Outcomes Partnership (OMOP) Common Data Model (CDM).<sup>12</sup> EHR variables (individual level) span demographics, encounters, medical/social history, diagnoses, procedures, medication prescriptions, medication exposure (eg, vaccines), laboratory tests/results, etc. Clinical notes have already been annotated in the CDM.<sup>13</sup> N3C has epidemiological and community data (population level) including temporal COVID-19 burden, vaccination, viral variance, health systems data and geospatial data. N3C has pre-pandemic data since 2018 and peripandemic data to the present. All data were deidentified and updated every 2 weeks. We have the highest level of data access, which allows using patients' residential ZIP codes and dates of clinical events. As of August 2021, N3C has at least 13 000



**Figure 1** Electronic Health Records model design.

adults have been diagnosed and/or have had laboratory-confirmed HIV. A pilot study shows that patients with coinfection have a higher risk of hospitalisation and mortality as compared with COVID-19 patients without HIV infection.<sup>14</sup>

### EHR data modeling

We will remodel the EHR data extracted from N3C. Because individuals' longitudinal health records are stored at distributed locations in EHR. Many clinical events do not have explicit and/or complete temporal information. Raw EHR data as such are of little value for understanding questions such as why individuals with certain coinfection present different clinical outcomes and disease prognosis.

We will locate relevant phenotypic traits from N3C by curating OMOP CDM concept sets, a procedure called electronic phenotyping. Using these concept sets, we will then retrieve and integrate individuals' phenotypic traits from N3C. At last, we will retrieve temporal information for clinical events and establish a graphical model<sup>15</sup> to represent clinical events and their temporal information.

### EHR phenotyping

Phenotyping is the process of identifying cohorts and variables from raw EHR.<sup>16</sup> Because we use OMOP CDM-normalised EHR, phenotyping is the process of finding 'OMOP concepts' that correspond to specific cohorts (eg, all patients with ART-related visits) or variables (eg, myocardial infarction).<sup>17</sup> An OMOP concept is a unique identifier that is mapped to diverse medical codes that have the same semantic meaning but may be from different medical nomenclatures and EHR systems.

We will use standard phenotyping procedures. The logical procedures are as follows: (1) To identify existing OMOP concept sets (available in OHDSI's Atlas system and N3C) that can be used with minor revision. (2) If no appropriate concept sets available, we will follow the generic phenotyping procedures<sup>16 18</sup> and curate OMOP CDM concepts from the Athena vocabulary repository, which allows retrieval of OMOP CDM concepts.<sup>19</sup> (3) To validate the revised or newly developed concept sets by

using EHR chart review that is performed independently by two domain experts.<sup>16</sup>

### Data retrieval and integration

To link external epidemiological and community-level data with individual-level EHR data, we will use individual patients' residential ZIP/county as the reference. We will use data imputation algorithms to impute missing values because population-towards-individual data integration automatically creates missing values. For existing missing values in EHR, we will infer missing values based on semantic relationships of OMOP CDM concepts. For geographical locations, we will use cities, counties and states to infer locations at appropriate levels. For the rest of the missing values, we will use multiple imputation methods. Specifically, we will use selection models or pattern-mixture models for systematic missingness. We will also selectively use mean/median imputation, principal component analysis, singular value decomposition, k-nearest neighbour, least squares, expectation maximisation and random forest.<sup>20 21</sup> Data retrieval and integration will be implemented using SQL, R, Python and PySpark, whichever appropriate.

### Graphical EHR model

We will develop a customised graphical model to represent the temporal relations among patients' demographic, clinical and sociobehavioural data. A Graphical model encodes clinical events as nodes and their semantic relations (including temporal relations) as edges. **Figure 1** shows the proposed design of the model. General modelling procedures include: first, we will retrieve time stamps of clinical events. Some clinical events have explicit time stamps as captured by structured EHR. Many others do not (eg, patient-reported symptoms as documented in clinical notes, trimester and/or gestational age). For those that do not have explicit time stamps, we will infer the time of the event based on neighbouring EHR data. For example, symptom onsets could be found in clinical notes (eg, admission, discharge); trimester and gestational age could be estimated by gestation-related diagnoses (eg, Z3A: weeks of gestation), procedures (eg, ultrasound procedures) and clinical notes (eg, last menstrual period).<sup>22 23</sup> Second, we will represent temporal information of clinical events by modelling the occurrences of an event and the instantaneous impact of the event. The occurrences of an event are resulted from the first step. The instantaneous impact of an event is formulated using exponential kernel functions and association rules based on clinical observation. Intuitively, two clinical events with a long interval in between would have less impact on one another, but this can be overwritten by events that hold special clinical meanings. Third, we will create recurrent states for every clinical event by embedding the event, its semantic relations (ie, edge in a Graph) including the instantaneous impact of an event, and time. These recurrent states will form a recurrent layer to be

**Table 1** Patient state according to WHO clinical progression scale

| Patient state                  | Clinical characteristics   | Severity score |
|--------------------------------|--|----------------|
| Uninfected                     | Uninfected, no viral RNA detected  | 0              |
| Ambulatory mild disease        | Asymptomatic, viral RNA detected   | 1              |
|                                | Symptomatic, independent   | 2              |
|                                | Symptomatic, assistance needed   | 3              |
| Hospitalised, moderate disease | Hospitalised, no oxygen therapy  | 4              |
|                                | Hospitalised, oxygen by mask or nasal prongs   | 5              |
| Hospitalised, severe disease   | Hospitalised, oxygen by non-invasive ventilation (NIV) or high flow  | 6              |
|                                | Intubation and mechanical ventilation, $pO_2/FiO_2 \geq 150$ or $SpO_2/FiO_2 \geq 200$                               | 7              |
|                                | Mechanical ventilation, $pO_2/FiO_2 < 150$ ( $SpO_2/FiO_2 < 200$ ) or vasopressors                                   | 8              |
|                                | Mechanical ventilation, $pO_2/FiO_2 < 150$ and vasopressors, dialysis, or extracorporeal membrane oxygenation (ECMO) | 9              |
| Dead                           | Dead   | 10             |

used for training machine learning models, which will be discussed later.

### Machine learning modeling

We will use supervised machine learning to examine patterns and sequences of phenotypic traits for their predictive ability in clinical outcomes and disease prognosis. Existing studies conclude differently on clinical outcomes among individuals with coinfection as well as the factors correlated with these clinical outcomes. We provide two hypotheses. Hypothesis 1: individual patients respond differently to the coinfection. Hypothesis 2: patients' clinical outcomes and disease prognosis are attributed by the temporal dynamics of clinical events. If these hypotheses are successfully tested, we will be able to delineate the impact of coinfection on individuals' clinical outcomes. Therefore, we will customise recurrent neural network (RNN) models to be used for predicting clinical outcomes and disease prognosis in real time by learning about patients' retrospective EHR at the individual level (personalised) as time progresses. We adopt RNN for its unique advantage in capturing temporal dependencies of data.<sup>24</sup> Trained RNN models will be tested for their performance where the best-performed model will be identified for identifying patterns/sequences of phenotypic traits predictive of clinical outcomes and prognosis.

### Cohort

Based on the estimated >13 000 patients with the coinfection in our dataset, we will blend in controls (COVID-19 patients without HIV) for each output variable using a match ratio of 1:2, stratified by sex, race/ethnicity and age. For a possible occasion of small sample, for example, death cases ( $n < 1000$  with coinfection), the alternative strategy is to create synthetic cases to impute and balance the sample.

### Machine learning input

A complete and longitudinal health history together with linked external epidemiological and community-level

data will be included as the input of machine learning models by which the models can learn from the input to predict individuals' in-time clinical outcomes and disease prognosis. We will include but are not limited to the following phenotypic traits: demographics, SDOH, diagnoses, underlying conditions, vitals, laboratory tests, procedures, medication prescriptions/dispensing, medication exposure (eg, vaccine) and annotated clinical notes. Because there is no gold standard measure for ART adherence, we will use 'multiple measures' to estimate levels of ART adherence.<sup>25</sup> Multiple measures include medication events (inpatient dispensing), HIV-1 RNA copies (laboratory results) and medication adherence data from clinical notes. For those with complete medication adherence data, we use the proportion of days covered to categorise ART adherence levels (eg, <50%, 50%–80%, 80%–85%, 85%–90%,  $\geq 90\%$ ).<sup>26</sup> We will categorise antiviral medications into integrase inhibitor-based, non-nucleoside reverse transcriptase inhibitor-based, protease inhibitor-based and other regimens. The approach to measuring ART adherence using EHR has limitations, but the limitations can be mitigated by the well-presented and large-scale national data. We will collect both  $CD4^+$  counts as an indicator of existing damage and plasma HIV-1 RNA copies as an indicator of projected disease progression.

### Machine learning output

Clinical outcome measures as machine learning output include inpatient admissions, length of stay (LOS), ICU admission, ICU LOS, comorbidities and primary discharge diagnosis.<sup>27</sup> In addition to the measures within the acute phase of COVID-19, we will also use symptoms, diagnoses, comorbidities and PASC-associated readmissions as outcome and prognosis measures for individuals in the postacute phase.



## Box 1 Key clinical outcome measures

Organ dysfunction  
 ⇒ Murray score  
 ⇒ Sequential organ failure assessment score, multiple organ dysfunction score  
 ⇒ Acute coronary syndrome; arrhythmias  
 ⇒ Delirium  
 Comorbidities  
 ⇒ Pulmonary, cardiovascular, renal, neurological, etc  
 Secondary infection  
 ⇒ Bacterial, viral  
 Biochemical parameters  
 ⇒ C reactive protein, D-dimers, IL-6, and ferritin serum concentrations, and leucocyte counts  
 Radiological findings  
 ⇒ CT scan of the chest, X-ray of the chest  
 Duration of intervention  
 ⇒ Inpatient admission, length of stay (LOS)  
 ⇒ ICU admission, ICU LOS  
 ⇒ Ventilation  
 ⇒ Organ support or hospital-free days  
 Pregnancy outcomes  
 ⇒ Preterm delivery, miscarriage  
 ⇒ Fetal status  
 ⇒ Severe maternal morbidity measures  
 Mortality  
 ⇒ All-cause mortality at hospital discharge  
 Quality of life  
 ⇒ Longer-term survival and primary diagnoses for readmission (postacute phase)

### Model design

We will use RNN as the machine learning architecture to learn from patients' longitudinal EHR and make the prediction of current and future clinical outcomes and disease prognosis. We adopt RNN models because this neural network architecture is specialised for capturing temporal dependency among event sequences. The RNN models will be trained to learn from model input and to make the prediction of model output. With respect to the embedding, we will include standard long short-term memory (LSTM) as well as phased LSTM and other variants wherever appropriate.<sup>24</sup> We will use the bag-of-pattern matrix as the baseline embedding method to be compared against LSTM, in which this baseline method does not fully consider temporal dependency. To test against RNN, we will use Support Vector Machine (SVM) as the benchmark algorithm, in which SVM is a well-performed kernel-based algorithm<sup>28</sup> but does not take full advantage of temporal dependency (hypothesis 2) and personalised health records (hypothesis 1). Because the nature of our machine learning output is binary variables, the proposed machine learning tasks are essentially binary classification tasks. We will use Python for machine learning modelling.

### Model evaluation (internal validity)

To test the effectiveness of the prediction model, we will use 10-fold cross validation. With respect to evaluation

metrics, we will use the F score, precision, recall, and the area under the receiver operating characteristic (AUC) to assess the models' predictive performance. We expect the F score, assuming balanced data, to reach a minimum of 0.8. If trained models fail to meet this expectation, alternative strategies include manually adding features hand-picked by researchers after error analysis of models.

### CDS system

Based on the automatic clinical outcomes and disease prognosis prediction model, we will design and implement a CDS prototype in collaboration with Prisma Health clinics. The proposed CDS prototype is anticipated to assist providers in screening and identifying patients who are at high risk of worse COVID-19 clinical outcomes (see [table 1](#)), and worse disease prognosis, including individuals with PASC. Specifically, the CDS will identify individuals with a high risk of disease progression from their current clinical state (eg, not hospitalised, hospitalised, postacute phase) by learning from the trained machine learning models. The effectiveness of the CDS demonstrates the external validity of the internally validated predictive model and will be assessed by (1) appropriate identification for at-risk individuals, (2) appropriate clinical actions and (3) CDS system usability.

### CDS workflow

The proposed CDS is a hybrid of knowledge-based and non-knowledge-based system.<sup>29</sup> It has (1) a machine learning-based prediction module (non-knowledge based) for identifying high-risk patients and (2) a provider-curated medical logic module (knowledge-based) for generating clinical actions for identified high-risk patients. The CDS testing takes place in a retrospective way (ie, using retrospective EHR).

### Cohort definition and data collection

Using retrospective EHR data (2-year baseline~2023) from Prisma Health's Epic system, we will first group the existing PLWH who have COVID-19 (sampling  $n > 500$ ) based on their state at the point of CDS screening. The patient states include (1) ambulatory patients with COVID-19, (2) hospitalised patients for COVID-19 with moderate disease, (3) hospitalised patients for COVID-19 with severe disease and (4) post-acute phase of COVID-19 (ie, from beyond 4 weeks after symptom onset).<sup>30</sup> See [table 1](#) for definitions of states 1–3 based on WHO's clinical progression scale for COVID-19.

### Prediction module

For patients in each state, we will use the trained machine learning model to learn from previous medical records and predict worsening clinical outcomes as time progresses (ie, acute, and postacute phases every 3 months). The prediction will include primary COVID-19 clinical outcomes ([Box 1](#)) developed by the WHO Working Group on the Clinical Characterisation and Management of COVID-19.<sup>31</sup>

### Medical logic module

Patients identified by the CDS to have an increased risk of worse clinical outcomes will be reviewed and discussed by two providers who are specialised in HIV and COVID-19. First, the providers will generate gold-standard judgement on whether a patient is correctly identified by the prediction module, which later will be used for assessing the effectiveness of CDS. Second, the providers will generate appropriate clinical actions on chart review. These clinical actions will be made up to date with the 'NIH Guidance for COVID-19 and People with HIV', including treatment options based on cohorts and risk factors, medication reconciliation considering ART regimens, consultation with specialists for multiorgan system complications and PASC, referrals and outreach.<sup>32</sup> Providers' decision-making processes will be programmed using Arden Syntax (V.3) or Clinical Quality Language<sup>33</sup> in the knowledge base, which is determined by specific EHR data model.

### Effectiveness of CDS (external validity)

There are two evaluation metrics: (1) appropriate identification for individuals at high risk for adverse clinical outcomes (Box 1) by comparing model-identified cases against the gold standard generated from chart review. We will use F measure (>0.8), AUC, precision and recall for assessment; (2) appropriate clinical actions using a quasiexperimental design. We will compare outcomes of patients who naturally used the medical logic module-suggested care against those who did not (n=100 each). The outcomes include but are not limited to readmissions (eg, same day, 7, 14, 30 days), healthcare utilisation (eg, LOS, emergency room (ER)/observation visits, ICU admission). We will use mixed-effect generalised regression models to estimate model effectiveness wherever appropriate.

### Usability testing

We will assess CDS usability by adopting the 'think aloud' protocol.<sup>34</sup> The two providers from Prisma Health will participate in the test. Each one will be presented with randomly selected EHR (n=5 at-risk cases+n=5 control cases) along with the CDS output. In each case, participants will be instructed to verbalise their reasoning procedures (eg, phenotypic traits from EHR that can be used in the reasoning, logic flow) towards identifying at-risk patients and corresponding clinical decisions. Sessions are audio recorded and will then be coded (eg, by content, understandability, navigation, workflow, visibility and usability) independently by two researchers for downstream analyses.

### Patient and public involvement

No patient involved.

### Ethics and dissemination

The study was approved by the institutional review boards at the University of South Carolina (Pro00121828) as non-human subject study.

This study will result in a comprehensive knowledge base that documents clinical outcomes and disease prognosis for individuals with the coinfection, their risk factors (eg, underlying conditions, ART adherence, comorbidities, sociobehavioural) and their responses to therapeutics. This study will also result in a prototype CDS that can identify patients at high risk of worsening clinical outcomes and prognosis in real time. These results are generalisable and will form a foundation for developing comprehensive real-world CDS systems for implementation in state-wide and national HIV and COVID-19 clinics.

Study findings will be presented at academic conferences and published in peer-reviewed journals. This study will disseminate urgently needed clinical evidence for guiding clinical practice for individuals with the coinfection at Prisma Health.

### Author affiliations

<sup>1</sup>Department of Health Services Policy and Management, University of South Carolina, Columbia, South Carolina, USA

<sup>2</sup>Big Data Health Science Center, University of South Carolina, Columbia, South Carolina, USA

<sup>3</sup>Department of Internal Medicine, University of South Carolina, Columbia, South Carolina, USA

<sup>4</sup>Department of Medicine, Duke University, Durham, North Carolina, USA

<sup>5</sup>Department of Health Promotion Education and Behavior, University of South Carolina, Columbia, South Carolina, USA

**Contributors** CL conceived the study design and drafted the manuscript. CL completed preliminary data collection. SW, BO, EG-CP, MY and XL contributed critical edits to the manuscript. All authors reviewed and approved the manuscript.

**Funding** Research reported in this publication was supported by the National Institute of Allergy and Infectious Diseases of the National Institutes of Health under Award Number R21AI170171. 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; peer reviewed for ethical and funding approval prior to submission.

**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

Chen Liang <http://orcid.org/0000-0002-9803-9880>

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

Michael E Yarrington <http://orcid.org/0000-0003-3186-1519>

Xiaoming Li <http://orcid.org/0000-0002-5555-9034>

### REFERENCES

- 1 World Health Organization. *Clinical Features and Prognostic Factors of COVID-19 in People Living with HIV Hospitalized with Suspected or Confirmed SARS-CoV-2 Infection*, 2021.
- 2 Dandachi D, Geiger G, Montgomery MW. Characteristics, comorbidities, and outcomes in a multicenter registry of patients with human immunodeficiency virus and coronavirus disease 2019. *Clin Infect Dis* 2020.

- 3 Braunstein SL, Lazar R, Wahnich A, *et al.* COVID-19 infection among people with HIV in New York City: a population-level analysis of matched surveillance data. *Clin Infect Dis* 2020.
- 4 Eisinger RW, Lerner AM, Fauci AS. Human Immunodeficiency Virus/AIDS in the Era of Coronavirus Disease 2019: A Juxtaposition of 2 Pandemics. *The Journal of Infectious Diseases*. *Published online* 2021.
- 5 Cooper TJ, Woodward BL, Alom S. COVID-19) outcomes in HIV/AIDS patients: a systematic review. *HIV Med* 2019;20(2):567–77.
- 6 Calza L, Bon I, Tadolini M, *et al.* COVID-19 in patients with HIV-1 infection: a single-centre experience in northern Italy. *Infection* 2021;49:333–7.
- 7 Costenaro P, Minotti C, Barbieri E, *et al.* SARS-CoV-2 infection in people living with HIV: a systematic review. *Rev Med Virol* 2021;31:1–12.
- 8 Park LS, Rentsch CT, Sigel K. COVID-19 in the largest us HIV cohort AIDS, 2020: 23rd.
- 9 Haendel MA, Chute CG, Gersing K. The National COVID cohort collaborative (N3C): rationale, design, infrastructure, and deployment. *J Am Med Inform Assoc* . 2020.
- 10 Deer RR, Rock MA, Vasilevsky N, *et al.* Characterizing long COVID: deep phenotype of a complex condition. *EBioMedicine* 2021;74:103722.
- 11 Datavent. COVID-19 research database. Available: <https://covid19researchdatabase.org/> [Accessed 20 Feb 2021].
- 12 OHDSI community. Observational health data sciences and informatics common data model
- 13 N3C. COVID-19 clinical data Warehouse data dictionary
- 14 Yang X, Zhang J, Guo S. The role of HIV infection in the clinical spectrum of COVID-19: a population-based cohort analysis based on us national COVID cohort collaborative (N3C) Enclave data. Available at SSRN:3860395.
- 15 Liu C, Wang F, Hu J. *Temporal phenotyping from longitudinal electronic health records: a graph based framework proceedings of the 21th ACM SIGKDD International Conference on knowledge discovery and data mining*, 2015: 705–14.
- 16 Banda JM, Seneviratne M, Hernandez-Boussard T, *et al.* Advances in electronic phenotyping: from rule-based definitions to machine learning models. *Annu Rev Biomed Data Sci* 2018;1:53–68.
- 17 Richesson RL, Hammond WE, Nahm M, *et al.* Electronic health records based phenotyping in next-generation clinical trials: a perspective from the NIH health care systems Collaboratory. *J Am Med Inform Assoc* 2013;20:e226–31.
- 18 Weng C, Shah NH, Hripcsak G. Deep phenotyping: embracing complexity and temporality-Towards scalability, portability, and interoperability. *J Biomed Inform* 2020;105:103433.
- 19 OHDSI Athena standard vocabularies. Available: <https://www.ohdsi.org/analytic-tools/athena-standardized-vocabularies/> [Accessed 01 Sep 2021].
- 20 Luo Y. Evaluating the state of the art in missing data imputation for clinical data. *Brief Bioinform* 2022;23:bbab489.
- 21 Li J, Yan XS, Chaudhary D, *et al.* Imputation of missing values for electronic health record laboratory data. *NPJ Digit Med* 2021;4:1–14.
- 22 Lyu T, Liang C, Liu J, *et al.* Temporal events detector for pregnancy care (TED-PC): a rule-based algorithm to infer gestational age and delivery date from electronic health records of pregnant women with and without COVID-19. *SSRN Journal* 2022:220502933.
- 23 Liu J, Hung P, Liang C, *et al.* Multilevel determinants of racial/ethnic disparities in severe maternal morbidity and mortality in the context of the COVID-19 pandemic in the USA: protocol for a concurrent triangulation, mixed-methods study. *BMJ Open* 2022;12:e062294.
- 24 Goodfellow I, Bengio Y, Courville A. *Deep learning*. Vol 1. MIT press Cambridge, 2016.
- 25 Castillo-Mancilla JR, Haberer JE. Adherence measurements in HIV: new advancements in pharmacologic methods and real-time monitoring. *Curr HIV/AIDS Rep* 2018;15:49–59.
- 26 Byrd KK, Hou JG, Hazen R, *et al.* Antiretroviral adherence level necessary for HIV viral suppression using real-world data. *J Acquir Immune Defic Syndr* 2019;82:245–51.
- 27 Lavery AM, Preston LE, Ko JY, *et al.* Characteristics of Hospitalized COVID-19 Patients Discharged and Experiencing Same-Hospital Readmission - United States, March-August 2020. *MMWR Morb Mortal Wkly Rep* 2020;69:1695–9.
- 28 Murphy KP. Machine learning: a probabilistic perspective. *MIT press* 2012.
- 29 Shiffman RN, Wright A. Evidence-Based clinical decision support. *Yearb Med Inform* 2013;22:120–7.
- 30 Nalbandian A, Sehgal K, Gupta A. Post-Acute COVID-19 syndrome. *Nat Med* 2021:1–15.
- 31 MarshallJC, MurthyS, DiazJ, *et al.* A minimal common outcome measure set for COVID-19 clinical research. *Lancet Infect Dis* 2020;20:e192–7.
- 32 Guidelines Working Groups of the NIH Office of AIDS Research Advisory Council. *Guidance for COVID-19 and people with HIV*, 2019.
- 33 Hripcsak G, Clayton P, Pryor T. *The Arden syntax for medical logic modules*. In: *Proceedings Symposium on Computer Applications in Medical Care*, 1990: 200–4.
- 34 Li AC, Kannry JL, Kushniruk A, *et al.* Integrating usability testing and think-aloud protocol analysis with "near-live" clinical simulations in evaluating clinical decision support. *Int J Med Inform* 2012;81:761–72.