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

Journal:	*BMJ Open*
Manuscript ID	bmjopen-2022-067204
Article Type:	Protocol
Date Submitted by the Author:	05-Aug-2022
Complete List of Authors:	Liang, Chen; University of South Carolina, Department of Health Services Policy and Management; University of South Carolina Weissman, Sharon; University of South Carolina, Olatosi, Bankole; University of South Carolina Arnold School of Public Health, Health Services, Policy and Management Poon, Eric ; Duke University Yarrington, Michael; Duke University, Li, Xiaoming; University of South Carolina Arnold School of Public Health
Keywords:	COVID-19, HIV & AIDS < INFECTIOUS DISEASES, Health informatics < BIOTECHNOLOGY & BIOINFORMATICS

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[1,2], Sharon Weissman[2,3], Bankole Olatosi[1,2], Eric Poon[4], Michael Yarrington[4], Xiaoming Li[2,5]

[1] Department of Health Services Policy and Management, Arnold School of Public Health, University of South Carolina
[2] Big Data Health Science Center, University of South Carolina
[3] Department of Internal Medicine, School of Medicine, University of South Carolina
[4] Department of Medicine, School of Medicine, Duke University
[5] Department of Health Promotion Education and Behaviors, Arnold School of Public Health, University of South Carolina

Correspondence:
Chen Liang, PhD
915 Greene St, Suite 347
Columbia, SC
cliang@mailbox.sc.edu
803 777 1836

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, multi-center Electronic Health Records (EHR) hold existing promise to provide full-spectrum and longitudinal clinical data, demographics, and socio-behavioral data at individual-level. Presently, an EHR-based cohort for the HIV/SARS-SoV-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 United States, namely, National COVID Cohort Collaborative (N3C). 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.

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.

Keywords

Acquired Immunodeficiency Syndrome, COVID-19, Electronic Health Records, Data Mining, Decision Support Systems

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, socio-behavioral, and social determinants of health (SDOH) with full-spectrum clinical characteristics for individuals with the 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.
- Out methods support in-time prediction of coinfected individuals' clinical outcomes, disease progression, and prognosis, and their risk factors.
- The proposed method is highly innovative in that it is designed to extract temporal sequences and temporal properties of every clinical event from EHR and is fully capable of embedding the temporal data into machine learning models.
- The study does not include comprehensive validity and usability testing of the proposed CDS prototype due to limited time for an exploratory/developmental study.

Introduction

The COVID-19 pandemic has cast a heavy burden on individuals with HIV infection. Based on data from 15,522 hospitalized patients with the coinfection of HIV and SARS-CoV-2 from 24 countries, a recent World Health Organization (WHO) report for the first time confirmed that HIV is a key risk factor for severe COVID-19 [1]. The severity of COVID-19 in individuals with HIV is correlated with certain comorbidities (e.g., 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 $CD4^+$ T-cell count (e.g., <200 cells/μL [2] or <500 [3] cells/μL) 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 (e.g., access to HIV testing, ART, and distribution of PrEP and PEP). [4]

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 [5–8]. 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. [4] 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-center cohorts. Temporal sequences and patterns of clinical events (e.g., underlying conditions, comorbidities, diagnoses, ART-related visits and treatments) are understudied, which diminish the opportunities for understanding the etiology of multi-faceted 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 socio-behavioral patterns (e.g., substance use/dependence) are closely related to disparities in HIV and SARS-CoV-2 infections but are understudied in part due to challenges in Electronic Health Records (EHR) data integration and phenotyping. The EHR hold existing promise to provide full-spectrum and longitudinal clinical data, demographics, and socio-behavioral data at the individual-level. However, we currently do not have an 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 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 (i.e., clinical characteristics, demographics, socio-behavioral patterns) and their temporal series and patterns from structured and unstructured EHR – National COVID Cohort Collaborative (N3C) [9]. 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, $CD4^+$ counts, viral suppression, ART procedures and medications, laboratory results for immune components and viral presence, treatments (e.g., procedures and medications), SDOH, and socio-behavioral patterns. We will develop machine-learning models to explore real-time predictive associations between these phenotypic trains and poor clinical outcomes and prognosis, including outcomes of the acute phase and post-acute sequelae of SARS-CoV-2 infection (PASC) [10]. 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, 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, 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 socio-behavioral determinants.

Methods and Analysis

Data Description

We will use EHR from National COVID Cohort Collaborative (N3C). As of 3/2022, N3C has aggregated 12.4 million patients (4.5 million COVID-19 patients) from 50 states [11]. The EHR are normalized by the Observational Health Data Sciences and Informatics (OHDSI)'s Observational Medical Outcomes Partnership (OMOP) Common Data Model (CDM) [12]. EHR variables (individual level) span demographics, encounters, medical/social history, diagnoses, procedures, medication prescriptions, medication exposure (e.g., vaccines), laboratory tests/results, etc. Clinical notes have already been annotated in the CDM [13]. 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 peri-pandemic data to the present. All data were de-identified 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 of 2021, N3C has at least 13 thousand adults have been diagnosed and/or have had laboratory-confirmed HIV. A pilot study shows that patients with coinfection have a high risk of hospitalization and mortality as compared to COVID-19 patients without HIV infection [14].

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 phenotypic traits from N3C. At last, we will retrieve temporal information for clinical events and establish a graphical model [15] to represent clinical events and their temporal information.

EHR phenotyping. Phenotyping is the process of identifying cohorts and variables from raw EHR [16]. Because we use OMOP CDM-based EHR, phenotyping is the process of finding "OMOP concepts" that correspond to specific cohorts (e.g., all patients with ART-related visits) or variables (e.g., myocardial infarction) [17]. An OMOP concept is a unique identifier that is mapped to diverse medical codes that have the same semantic meaning but maybe from different coding systems 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 standardized phenotyping procedures [16,18] and curate OMOP concepts from the Athena vocabulary repository, which allows retrieval of OMOP concepts [19]. 3) To validate the revised or newly developed concept sets by using EHR chart review that is performed independently by two clinical experts [16].

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-to-individual-level data integration and missing values widely exist in EHR. For missing values in EHR, we will infer missing values based on semantic relationships of OMOP 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 (PCA), singular value decomposition (SVD), k-nearest neighbor (kNN), least squares, expectation-maximization, and random forest. [20,21] Data retrieval and integration will be implemented using SQL, R, and PySpark.

Graphical EHR model. We will develop a customized Graphical model to represent the temporal relations among patients' demographic, clinical, and socio-behavioral 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 modeling 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 (e.g., 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 neighboring EHR data. For example, symptom onsets could be found in clinical notes (e.g., admission, discharge); trimester

and gestational age could be estimated by gestation-related diagnoses (e.g., Z3A: weeks of gestation), procedures (e.g., ultrasound procedures), and clinical notes (e.g., last menstrual period). Second, we will represent temporal information of clinical events by modeling 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 unique clinical meanings. Third, we will create recurrent states for every clinical event by embedding the event, its semantic relations (i.e., edge in a Graph) including the instantaneous impact of an event, and time. These recurrent states will form a recurrent layer to be 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 customize 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 (personalized) as time progresses. We adopt RNN for its unique advantage in capturing temporal dependencies of data [22]. 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 thousand 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 occasions 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 from N3C: demographics, SDOH, diagnoses, underlying conditions, vitals, laboratory tests, procedures, medication prescriptions/dispensing, medication exposure (e.g., 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 [23]. 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 (PDC) to categorize ART

adherence levels (e.g., < 50, 50-80, 80-85, 85-90, ≥90%) [24]. We will categorize antiviral medications (ARV) into Integrase inhibitor (INSTI)-based, non-nucleoside reverse transcriptase inhibitor (NNRTI)-based, protease inhibitor (PI)-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 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 [25]. 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 post-acute 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 specialized 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 [22]. 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 [26] but does not take full advantage of temporal dependency (hypothesis 2) and personalized 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 modeling.

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 handpicked by researchers after error analysis of models.

Clinical Decision Support System

Based on the automatic clinical outcomes and disease prognosis prediction model, we will design and implement a clinical decision support (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 post-acute sequelae of SARS-CoV-2 infection (PASC)]. Specifically, the CDS will identify individuals with a high risk of disease progression from their current clinical state (e.g., not hospitalized, hospitalized, post-acute 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.

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
Hospitalized, moderate disease	Hospitalized, no oxygen therapy	4
	Hospitalized, oxygen by mask or nasal prongs	5
Hospitalized, severe disease	Hospitalized, oxygen by NIV or high flow	6
	Intubation and mechanical ventilation, pO_2/FiO_2 ≥150 or SpO_2/FiO_2 ≥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 ECMO	9
Dead	Dead	10

CDS workflow. The proposed CDS is a hybrid of knowledge-based and non-knowledge-based system [27]. 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 (i.e., 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) hospitalized patients for COVID-19 with moderate disease, 3) hospitalized patients for COVID-19 with severe disease, and 4) post-acute phase of COVID-19 (i.e., from beyond 4 weeks after symptom onset) [28]. 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 (i.e., acute, and post-acute phases every 3 months. The prediction will include primary COVID-19 clinical outcomes (***Table 2***) developed by the WHO working group on the Clinical Characterisation and Management of COVID-19 [29].

Table 2. 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 • Chest CT scan, chest x-ray
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 (SMM) measures
Mortality • All-cause mortality at hospital discharge
Quality of life • Longer term survival and primary diagnoses for readmission (post-acute phase)

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 specialized in HIV and COVID-19. First, the providers will generate gold-standard judgment on whether a patient is correctly identified by the prediction module, which will later be used for assessing the effectiveness of CDS. Second, the providers will generate appropriate clinical actions upon 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 multi-organ system complications and PASC, referrals, and outreach [30]. Providers' decision-making processes will be programmed using Arden Syntax (v3) or Clinical Quality Language [31] in the knowledge base, which is determined by specific Epic 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 (***Table 2***) 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 quasi-experimental 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 (e.g., same day, 7-, 14-, 30- days), healthcare utilization (e.g., LOS, ER/observation visits, ICU admission). We will use mixed-effect generalized regression models to estimate model effectiveness wherever appropriate.

Usability testing. We will assess CDS usability by adopting the "think aloud" protocol [32]. 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 verbalize their reasoning procedures (e.g., phenotypic traits from EHR that can be used in the reasoning, logic flow) towards identifying the at-risk patients and corresponding clinical decisions. Sessions are audio-recorded and will then be coded (e.g., 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 (e.g., underlying conditions, ART adherence, comorbidities, socio-behavioral), 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 generalizable 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.

Authors' Contributions

CL conceived the study design and drafted the manuscript. CL completed the preliminary data analysis. SW, BO, EP, MY, and XL contributed critical edits to the manuscript. All authors reviewed and approved the manuscript.

Funding Statement

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 Interest Statement

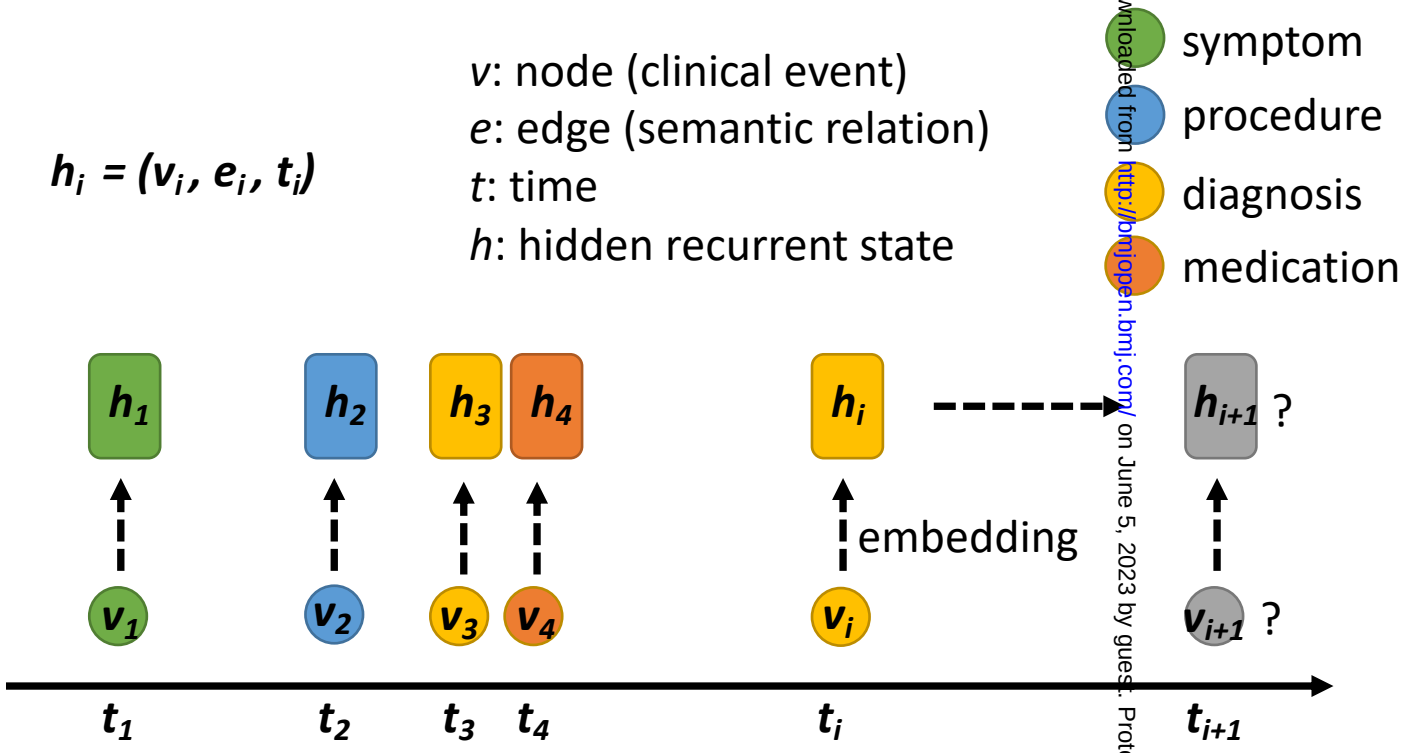
None declared.

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, et al. Characteristics, Comorbidities, and outcomes in a multicenter registry of patients with human immunodeficiency virus and coronavirus disease 2019. *Clin Infect Dis*. Published online 2020.
3. Braunstein SL, Lazar R, Wahnich A, Daskalakis DC, Blackstock OJ. COVID-19 infection among people with HIV in New York City: A population-level analysis of linked surveillance data. *Clinical Infectious Diseases: An Official Publication of the Infectious Diseases Society of America*. Published online 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, Harky A. Coronavirus disease 2019 (COVID-19) outcomes in HIV/AIDS patients: a systematic review. *HIV Med*. 2020;21(9):567-577.
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(2):333-337.
7. Costenaro P, Minotti C, Barbieri E, Giaquinto C, Donà D. SARS-CoV-2 infection in people living with HIV: a systematic review. *Reviews in Medical Virology*. 2021;31(1):1-12.
8. Park LS, Rentsch CT, Sigel K, et al. COVID-19 in the largest US HIV cohort. *AIDS*. 2020;2020:23rd.
9. Haendel MA, Chute CG, Gersing K. The National COVID Cohort Collaborative (N3C): rationale, design, infrastructure, and deployment. *Journal of the American Medical Informatics Association*. Published online 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. Accessed February 20, 2021. https://covid19researchdatabase.org/
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, Olatosi B, Weissman S, Li X. 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*. Published online 2021.
15. Liu C, Wang F, Hu J, Xiong H. Temporal phenotyping from longitudinal electronic health records: A graph based framework. In: *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ; 2015:705-714.
16. Banda JM, Seneviratne M, Hernandez-Boussard T, Shah NH. 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. *Journal of the American Medical Informatics Association*. 2013;20(e2):e226-e231.
18. Weng C, Shah NH, Hripcsak G. Deep phenotyping: Embracing complexity and temporality—Towards scalability, portability, and interoperability. *Journal of Biomedical Informatics*. 2020;105:103433.
19. OHDSI Athena standard vocabularies. Accessed September 1, 2021. https://www.ohdsi.org/analytic-tools/athena-standardized-vocabularies/
20. Luo Y. Evaluating the state of the art in missing data imputation for clinical data. *Briefings in Bioinformatics*. 2022;23(1):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):1-14.
22. Goodfellow I, Bengio Y, Courville A, Bengio Y. *Deep Learning*. Vol 1. MIT press Cambridge; 2016.
23. Castillo-Mancilla JR, Haberer JE. Adherence measurements in HIV: new advancements in pharmacologic methods and real-time monitoring. *Current Hiv/aids Reports*. 2018;15(1):49-59.
24. 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(3):245.
25. 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. *Morbidity and Mortality Weekly Report*. 2020;69(45):1695.
26. Murphy KP. *Machine Learning: A Probabilistic Perspective*. MIT press; 2012.
27. Shiffman RN, Wright A. Evidence-based clinical decision support. *Yearb Med Inform*. 2013;22(01):120-127.
28. Nalbandian A, Sehgal K, Gupta A, et al. Post-acute COVID-19 syndrome. *Nat Med*. Published online 2021:1-15.
29. Marshall JC, Murthy S, Diaz J, et al. A minimal common outcome measure set for COVID-19 clinical research. *The Lancet Infectious Diseases*. 2020;20(8):e192–e197.
30. Guidelines Working Groups of the NIH Office of AIDS Research Advisory Council. Guidance for COVID-19 and People with HIV.
31. Hripcsak G, Clayton P, Pryor T, Haug P, Wigertz O, der Lei J. The Arden syntax for medical logic modules. In: *Proceedings. Symposium on Computer Applications in Medical Care*. ; 1990:200-204.
32. 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(11):761-772.

Figure 1. EHR model design.



CONSORT 2010 checklist of information to include when reporting a randomised trial*

Section/Topic	Item No	Checklist item	Reported on page No
Title and abstract			
	1a	Identification as a randomised trial in the title	NA
	1b	Structured summary of trial design, methods, results, and conclusions (for specific guidance see CONSORT for abstracts)	2
Introduction			
Background and objectives	2a	Scientific background and explanation of rationale	4-5
	2b	Specific objectives or hypotheses	4-5
Methods			
Trial design	3a	Description of trial design (such as parallel, factorial) including allocation ratio	NA
	3b	Important changes to methods after trial commencement (such as eligibility criteria), with reasons	NA
Participants	4a	Eligibility criteria for participants	5, 8
	4b	Settings and locations where the data were collected	5
Interventions	5	The interventions for each group with sufficient details to allow replication, including how and when they were actually administered	NA
Outcomes	6a	Completely defined pre-specified primary and secondary outcome measures, including how and when they were assessed	11
	6b	Any changes to trial outcomes after the trial commenced, with reasons	NA
Sample size	7a	How sample size was determined	8
	7b	When applicable, explanation of any interim analyses and stopping guidelines	NA
Randomisation:			
Sequence generation	8a	Method used to generate the random allocation sequence	NA
	8b	Type of randomisation; details of any restriction (such as blocking and block size)	NA
Allocation concealment mechanism	9	Mechanism used to implement the random allocation sequence (such as sequentially numbered containers), describing any steps taken to conceal the sequence until interventions were assigned	NA
Implementation	10	Who generated the random allocation sequence, who enrolled participants, and who assigned participants to interventions	NA
Blinding	11a	If done, who was blinded after assignment to interventions (for example, participants, care providers, those	NA

		assessing outcomes) and how	
	11b	If relevant, description of the similarity of interventions	NA
Statistical methods	12a	Statistical methods used to compare groups for primary and secondary outcomes	NA
	12b	Methods for additional analyses, such as subgroup analyses and adjusted analyses	NA
Results			
Participant flow (a diagram is strongly recommended)	13a	For each group, the numbers of participants who were randomly assigned, received intended treatment, and were analysed for the primary outcome	NA
	13b	For each group, losses and exclusions after randomisation, together with reasons	NA
Recruitment	14a	Dates defining the periods of recruitment and follow-up	NA
	14b	Why the trial ended or was stopped	NA
Baseline data	15	A table showing baseline demographic and clinical characteristics for each group	NA
Numbers analysed	16	For each group, number of participants (denominator) included in each analysis and whether the analysis was by original assigned groups	8
Outcomes and estimation	17a	For each primary and secondary outcome, results for each group, and the estimated effect size and its precision (such as 95% confidence interval)	NA
	17b	For binary outcomes, presentation of both absolute and relative effect sizes is recommended	NA
Ancillary analyses	18	Results of any other analyses performed, including subgroup analyses and adjusted analyses, distinguishing pre-specified from exploratory	11
Harms	19	All important harms or unintended effects in each group (for specific guidance see CONSORT for harms)	NA
Discussion			
Limitations	20	Trial limitations, addressing sources of potential bias, imprecision, and, if relevant, multiplicity of analyses	NA
Generalisability	21	Generalisability (external validity, applicability) of the trial findings	9-11
Interpretation	22	Interpretation consistent with results, balancing benefits and harms, and considering other relevant evidence	9-11
Other information			
Registration	23	Registration number and name of trial registry	NA
Protocol	24	Where the full trial protocol can be accessed, if available	NA
Funding	25	Sources of funding and other support (such as supply of drugs), role of funders	12

*We strongly recommend reading this statement in conjunction with the CONSORT 2010 Explanation and Elaboration for important clarifications on all the items. If relevant, we also recommend reading CONSORT extensions for cluster randomised trials, non-inferiority and equivalence trials, non-pharmacological treatments, herbal interventions, and pragmatic trials. Additional extensions are forthcoming: for those and for up to date references relevant to this checklist, see www.consort-statement.org.

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

Journal:	*BMJ Open*
Manuscript ID	bmjopen-2022-067204.R1
Article Type:	Protocol
Date Submitted by the Author:	24-Aug-2022
Complete List of Authors:	Liang, Chen; University of South Carolina, Department of Health Services Policy and Management; University of South Carolina Weissman, Sharon; University of South Carolina, Olatosi, Bankole; University of South Carolina Arnold School of Public Health, Health Services, Policy and Management Poon, Eric ; Duke University Yarrington, Michael; Duke University, Li, Xiaoming; University of South Carolina Arnold School of Public Health
Primary Subject Heading:	Infectious diseases
Secondary Subject Heading:	Health informatics, HIV/AIDS, Respiratory medicine
Keywords:	COVID-19, HIV & AIDS < INFECTIOUS DISEASES, Health informatics < BIOTECHNOLOGY & BIOINFORMATICS

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[1,2,3], Sharon Weissman[2,4], Bankole Olatosi[1,2,3], Eric Poon[5], Michael Yarrington[5], Xiaoming Li[2,3,6]

[1] Department of Health Services Policy and Management, Arnold School of Public Health, University of South Carolina
[2] Big Data Health Science Center, University of South Carolina
[3] South Carolina SmartState Center for Healthcare Quality, University of South Carolina
[4] Department of Internal Medicine, School of Medicine, University of South Carolina
[5] Department of Medicine, School of Medicine, Duke University
[6] Department of Health Promotion Education and Behaviors, Arnold School of Public Health, University of South Carolina

Correspondence:
Chen Liang, PhD
915 Greene St, Suite 347
Columbia, SC
cliang@mailbox.sc.edu
803 777 1836

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, multi-center Electronic Health Records (EHR) hold existing promise to provide full-spectrum and longitudinal clinical data, demographics, and socio-behavioral data at individual-level. Presently, a comprehensive EHR-based cohort for the HIV/SARS-SoV-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 United States, namely, National COVID Cohort Collaborative (N3C). 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.

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.

Keywords

Acquired Immunodeficiency Syndrome, COVID-19, Electronic Health Records, Data Mining, Decision Support Systems

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, socio-behavioral, and social determinants of health (SDOH) 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.
- Out methods support real-time prediction of coinfected individuals' clinical outcomes, disease progression, and 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 EHR and is fully capable of embedding the temporal data into machine learning models.
- The study does not include comprehensive validity and usability testing of the proposed Clinical Decision Support (CDS) prototype due to limited time for an exploratory/developmental study.

Introduction

The COVID-19 pandemic has cast a heavy burden on individuals with HIV infection. Based on data from 15,522 hospitalized patients with the coinfection of HIV and SARS-CoV-2 from 24 countries, a recent World Health Organization (WHO) report for the first time confirmed that HIV is a key risk factor for severe COVID-19 [1]. The severity of COVID-19 in individuals with HIV is correlated with certain comorbidities (e.g., 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 $CD4^+$ T-cell count (e.g., <200 cells/μL [2] or <500 [3] cells/μL) 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 [e.g., access to HIV testing, ART, and distribution of pre-exposure prophylaxis (PrEP) and post-exposure prophylaxis (PEP)]. [4]

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 [5–8]. 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. [4] 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-center cohorts. Temporal sequences and patterns of clinical events (e.g., underlying conditions, comorbidities, diagnoses, ART-related visits and treatments) are understudied, which diminish the opportunities for understanding the etiology of multi-faceted 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 socio-behavioral patterns (e.g., 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 socio-behavioral 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 (i.e., clinical characteristics, demographics, socio-behavioral patterns) and their temporal series and patterns from structured and unstructured

EHR – National COVID Cohort Collaborative (N3C) [9]. 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, $CD4^+$ counts, viral suppression, ART procedures and medications, laboratory results for immune components and viral presence, treatments (e.g., procedures and medications), SDOH, and socio-behavioral 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 post-acute sequelae of SARS-CoV-2 infection (PASC) [10]. 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, 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 socio-behavioral determinants.

Methods and Analysis

Data Description

We will use EHR from National COVID Cohort Collaborative (N3C). As of 8/2022, N3C has aggregated 15.2 million patients (5.8 million COVID-19 patients) from 50 states [11]. The EHR are normalized by the Observational Health Data Sciences and Informatics (OHDSI)'s Observational Medical Outcomes Partnership (OMOP) Common Data Model (CDM) [12]. EHR variables (individual level) span demographics, encounters, medical/social history, diagnoses, procedures, medication prescriptions, medication exposure (e.g., vaccines), laboratory tests/results, etc. Clinical notes have already been annotated in the CDM [13]. 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 peri-pandemic data to the present. All data were de-identified 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 of 2021, N3C has at least 13 thousand adults have been diagnosed and/or have had laboratory-confirmed HIV. A pilot study shows that patients with coinfection have a high risk of hospitalization and mortality as compared to COVID-19 patients without HIV infection [14].

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 [15] to represent clinical events and their temporal information.

EHR phenotyping. Phenotyping is the process of identifying cohorts and variables from raw EHR [16]. Because we use OMOP CDM-normalized EHR, phenotyping is the process of finding "OMOP concepts" that correspond to specific cohorts (e.g., all patients with ART-related visits) or variables (e.g., myocardial infarction) [17]. 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 [16,18] and curate OMOP CDM concepts from the Athena vocabulary repository, which allows retrieval of OMOP CDM concepts [19]. 3) To validate the revised or newly developed concept sets by using EHR chart review that is performed independently by two domain experts [16].

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 (PCA), singular value decomposition (SVD), k-nearest neighbor (kNN), least squares, expectation-maximization, and random forest. [20,21] Data retrieval and integration will be implemented using SQL, R, Python, and PySpark, whichever appropriate.

Graphical EHR model. We will develop a customized Graphical model to represent the temporal relations among patients' demographic, clinical, and socio-behavioral 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 modeling 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 (e.g., 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 neighboring EHR data. For

example, symptom onsets could be found in clinical notes (e.g., admission, discharge); trimester and gestational age could be estimated by gestation-related diagnoses (e.g., Z3A: weeks of gestation), procedures (e.g., ultrasound procedures), and clinical notes (e.g., last menstrual period) [22,23]. Second, we will represent temporal information of clinical events by modeling 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 (i.e., edge in a Graph) including the instantaneous impact of an event, and time. These recurrent states will form a recurrent layer to be 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 customize 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 (personalized) as time progresses. We adopt RNN for its unique advantage in capturing temporal dependencies of data [24]. 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 thousand 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 occasions 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 (e.g., 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 [25]. 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 (PDC) to categorize ART adherence levels (e.g., < 50, 50-80, 80-85, 85-90, ≥90%) [26]. We will categorize antiviral medications (ARV) into Integrase inhibitor (INSTI)-based, non-nucleoside reverse transcriptase inhibitor (NNRTI)-based, protease inhibitor (PI)-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 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 [27]. 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 post-acute 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 specialized 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 [24]. 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 [28] but does not take full advantage of temporal dependency (hypothesis 2) and personalized 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 modeling.

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 handpicked by researchers after error analysis of models.

Clinical Decision Support System

Based on the automatic clinical outcomes and disease prognosis prediction model, we will design and implement a clinical decision support (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 (e.g., not hospitalized, hospitalized, post-acute 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.

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
Hospitalized, moderate disease	Hospitalized, no oxygen therapy	4
	Hospitalized, oxygen by mask or nasal prongs	5
Hospitalized, severe disease	Hospitalized, oxygen by NIV or high flow	6
	Intubation and mechanical ventilation, pO_2/FiO_2 ≥150 or SpO_2/FiO_2 ≥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 ECMO	9
Dead	Dead	10

CDS workflow. The proposed CDS is a hybrid of knowledge-based and non-knowledge-based system [29]. 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 (i.e., 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) hospitalized patients for COVID-19 with moderate disease, 3) hospitalized patients for COVID-19 with severe disease, and 4) post-acute phase of COVID-19 (i.e., from beyond 4 weeks after symptom onset) [30]. 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 (i.e., acute, and post-acute phases every 3 months. The prediction will include primary COVID-19 clinical outcomes (***Table 2***) developed by the WHO Working Group on the Clinical Characterisation and Management of COVID-19 [31].

Table 2. 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 • Chest CT scan, chest x-ray
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 (SMM) measures
Mortality • All-cause mortality at hospital discharge
Quality of life • Longer term survival and primary diagnoses for readmission (post-acute phase)

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 specialized in HIV and COVID-19. First, the providers will generate gold-standard judgment 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 upon 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 multi-organ system complications and PASC, referrals, and outreach [32]. Providers' decision-making processes will be programmed using Arden Syntax (v3) or Clinical Quality Language [33] in the knowledge base, which is determined by specific Epic 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 (***Table 2***) 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 quasi-experimental 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 (e.g., same day, 7-, 14-, 30- days), healthcare utilization (e.g., LOS, ER/observation visits, ICU admission). We will use mixed-effect generalized regression models to estimate model effectiveness wherever appropriate.

Usability testing. We will assess CDS usability by adopting the "think aloud" protocol [34]. 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 verbalize their reasoning procedures (e.g., 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 (e.g., 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 (e.g., underlying conditions, ART adherence, comorbidities, socio-behavioral), 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 generalizable 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.

Authors' Contributions

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

Funding Statement

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 Interest Statement

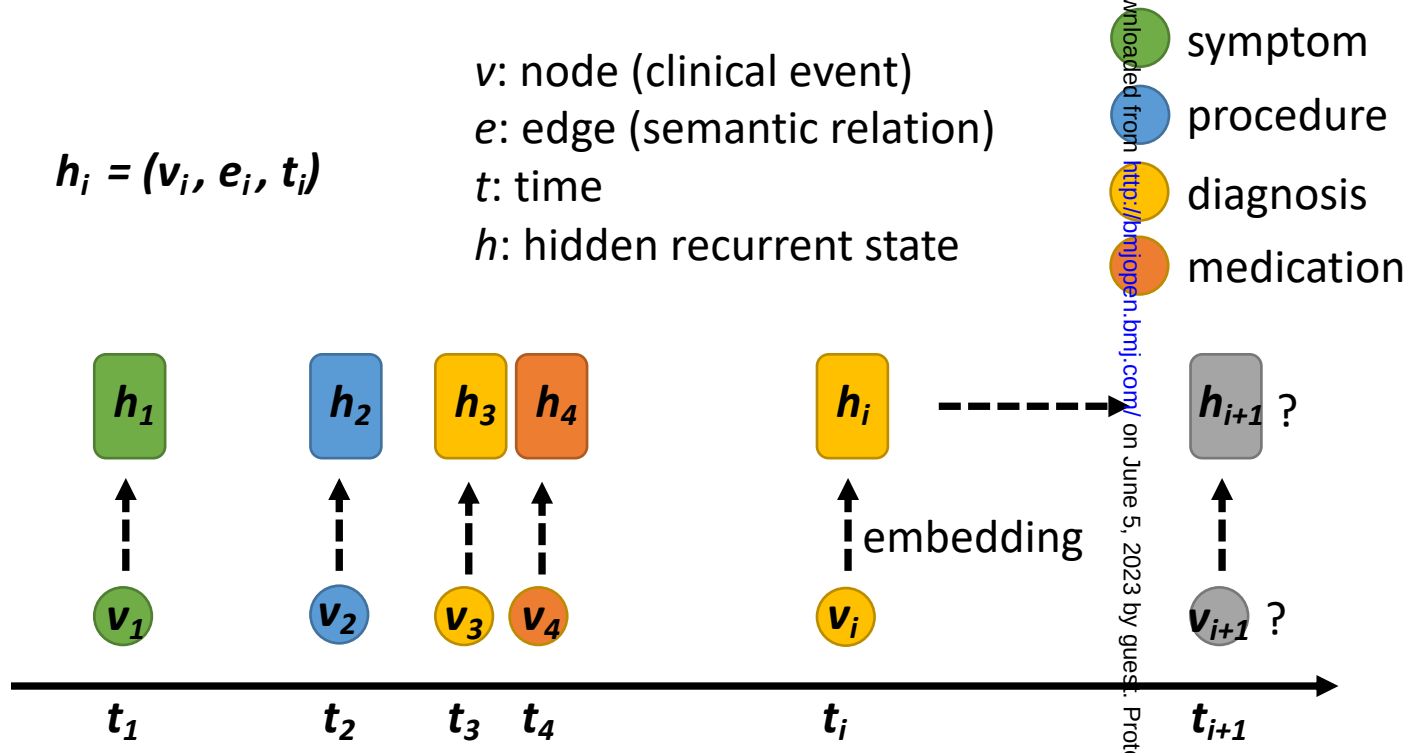
None declared.

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, et al. Characteristics, Comorbidities, and outcomes in a multicenter registry of patients with human immunodeficiency virus and coronavirus disease 2019. *Clin Infect Dis*. Published online 2020.
3. Braunstein SL, Lazar R, Wahnich A, Daskalakis DC, Blackstock OJ. COVID-19 infection among people with HIV in New York City: A population-level analysis of linked surveillance data. *Clinical Infectious Diseases: An Official Publication of the Infectious Diseases Society of America*. Published online 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, Harky A. Coronavirus disease 2019 (COVID-19) outcomes in HIV/AIDS patients: a systematic review. *HIV Med*. 2020;21(9):567-577.
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(2):333-337.
7. Costenaro P, Minotti C, Barbieri E, Giaquinto C, Donà D. SARS-CoV-2 infection in people living with HIV: a systematic review. *Reviews in Medical Virology*. 2021;31(1):1-12.
8. Park LS, Rentsch CT, Sigel K, et al. COVID-19 in the largest US HIV cohort. *AIDS*. 2020;2020:23rd.
9. Haendel MA, Chute CG, Gersing K. The National COVID Cohort Collaborative (N3C): rationale, design, infrastructure, and deployment. *Journal of the American Medical Informatics Association*. Published online 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. Accessed February 20, 2021. https://covid19researchdatabase.org/
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, Olatosi B, Weissman S, Li X. 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*. Published online 2021.
15. Liu C, Wang F, Hu J, Xiong H. Temporal phenotyping from longitudinal electronic health records: A graph based framework. In: *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ; 2015:705-714.
16. Banda JM, Seneviratne M, Hernandez-Boussard T, Shah NH. 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. *Journal of the American Medical Informatics Association*. 2013;20(e2):e226-e231.
18. Weng C, Shah NH, Hripcsak G. Deep phenotyping: Embracing complexity and temporality—Towards scalability, portability, and interoperability. *Journal of Biomedical Informatics*. 2020;105:103433.
19. OHDSI Athena standard vocabularies. Accessed September 1, 2021. https://www.ohdsi.org/analytic-tools/athena-standardized-vocabularies/
20. Luo Y. Evaluating the state of the art in missing data imputation for clinical data. *Briefings in Bioinformatics*. 2022;23(1):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):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. *arXiv preprint arXiv:220502933*. Published online 2022.
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(6):e062294.
24. Goodfellow I, Bengio Y, Courville A, Bengio Y. *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. *Current Hiv/aids Reports*. 2018;15(1):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(3):245.
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. *Morbidity and Mortality Weekly Report*. 2020;69(45):1695.
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(01):120-127.
30. Nalbandian A, Sehgal K, Gupta A, et al. Post-acute COVID-19 syndrome. *Nat Med*. Published online 2021:1-15.
31. Marshall JC, Murthy S, Diaz J, et al. A minimal common outcome measure set for COVID-19 clinical research. *The Lancet Infectious Diseases*. 2020;20(8):e192–e197.
32. Guidelines Working Groups of the NIH Office of AIDS Research Advisory Council. Guidance for COVID-19 and People with HIV.
33. Hripcsak G, Clayton P, Pryor T, Haug P, Wigertz O, der Lei J. The Arden syntax for medical logic modules. In: *Proceedings. Symposium on Computer Applications in Medical Care*. ; 1990:200-204.
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(11):761-772.

Figure 1. EHR model design.



CONSORT 2010 checklist of information to include when reporting a randomised trial*

Section/Topic	Item No	Checklist item	Reported on page No
Title and abstract			
	1a	Identification as a randomised trial in the title	NA
	1b	Structured summary of trial design, methods, results, and conclusions (for specific guidance see CONSORT for abstracts)	2
Introduction			
Background and objectives	2a	Scientific background and explanation of rationale	4-5
	2b	Specific objectives or hypotheses	4-5
Methods			
Trial design	3a	Description of trial design (such as parallel, factorial) including allocation ratio	NA
	3b	Important changes to methods after trial commencement (such as eligibility criteria), with reasons	NA
Participants	4a	Eligibility criteria for participants	5, 8
	4b	Settings and locations where the data were collected	5
Interventions	5	The interventions for each group with sufficient details to allow replication, including how and when they were actually administered	NA
Outcomes	6a	Completely defined pre-specified primary and secondary outcome measures, including how and when they were assessed	11
	6b	Any changes to trial outcomes after the trial commenced, with reasons	NA
Sample size	7a	How sample size was determined	8
	7b	When applicable, explanation of any interim analyses and stopping guidelines	NA
Randomisation:			
Sequence generation	8a	Method used to generate the random allocation sequence	NA
	8b	Type of randomisation; details of any restriction (such as blocking and block size)	NA
Allocation concealment mechanism	9	Mechanism used to implement the random allocation sequence (such as sequentially numbered containers), describing any steps taken to conceal the sequence until interventions were assigned	NA
Implementation	10	Who generated the random allocation sequence, who enrolled participants, and who assigned participants to interventions	NA
Blinding	11a	If done, who was blinded after assignment to interventions (for example, participants, care providers, those	NA

		assessing outcomes) and how	
	11b	If relevant, description of the similarity of interventions	NA
Statistical methods	12a	Statistical methods used to compare groups for primary and secondary outcomes	NA
	12b	Methods for additional analyses, such as subgroup analyses and adjusted analyses	NA
Results			
Participant flow (a diagram is strongly recommended)	13a	For each group, the numbers of participants who were randomly assigned, received intended treatment, and were analysed for the primary outcome	NA
	13b	For each group, losses and exclusions after randomisation, together with reasons	NA
Recruitment	14a	Dates defining the periods of recruitment and follow-up	NA
	14b	Why the trial ended or was stopped	NA
Baseline data	15	A table showing baseline demographic and clinical characteristics for each group	NA
Numbers analysed	16	For each group, number of participants (denominator) included in each analysis and whether the analysis was by original assigned groups	8
Outcomes and estimation	17a	For each primary and secondary outcome, results for each group, and the estimated effect size and its precision (such as 95% confidence interval)	NA
	17b	For binary outcomes, presentation of both absolute and relative effect sizes is recommended	NA
Ancillary analyses	18	Results of any other analyses performed, including subgroup analyses and adjusted analyses, distinguishing pre-specified from exploratory	11
Harms	19	All important harms or unintended effects in each group (for specific guidance see CONSORT for harms)	NA
Discussion			
Limitations	20	Trial limitations, addressing sources of potential bias, imprecision, and, if relevant, multiplicity of analyses	NA
Generalisability	21	Generalisability (external validity, applicability) of the trial findings	9-11
Interpretation	22	Interpretation consistent with results, balancing benefits and harms, and considering other relevant evidence	9-11
Other information			
Registration	23	Registration number and name of trial registry	NA
Protocol	24	Where the full trial protocol can be accessed, if available	NA
Funding	25	Sources of funding and other support (such as supply of drugs), role of funders	12

*We strongly recommend reading this statement in conjunction with the CONSORT 2010 Explanation and Elaboration for important clarifications on all the items. If relevant, we also recommend reading CONSORT extensions for cluster randomised trials, non-inferiority and equivalence trials, non-pharmacological treatments, herbal interventions, and pragmatic trials. Additional extensions are forthcoming: for those and for up to date references relevant to this checklist, see www.consort-statement.org.