Original Paper

Identifying People Living With or Those at Risk for HIV in a Nationally Sampled Electronic Health Record Repository Called the National Clinical Cohort Collaborative: Computational Phenotyping Study

Eric Hurwitz[1*], PhD; Cara D Varley[2*], MPH, MD; A Jerrod Anzolone[3], PhD; Vithal Madhira[4], MS; Amy L Olex[5], PhD; Jing Sun[6], MPH, MD, PhD; Dimple Vaidya[7], MS; Nada Fadul[8], MD; Jessica Y Islam[9,10], MPH, PhD; Lesley E Jackson[11], MD; Kenneth J Wilkins[12], PhD; Zachary Butzin-Dozier[13], MPH, PhD; Dongmei Li[14], PhD; Sandra E Safo[15], PhD; Julie A McMurry[1], MPH; Pooja Maheria[11], MS; Tommy Williams[11]; Shukri A Hassan[7], BS; Melissa A Haendel[1], PhD; Rena C Patel[11], MPH, MPhil, MD; The National Clinical Cohort Collaborative (N3C) Consortium[16*]

[1]Department of Genetics, University of North Carolina at Chapel Hill, Chapel Hill, NC, United States

[2]School of Medicine, Oregon Health & Science University, Portland, OR, United States

[3]Department of Biostatistics, University of Nebraska Medical Center, Omaha, NE, United States

[4]Palila Software, Reno, NV, United States

[5]Wright Center for Clinical and Translational Research, Virginia Commonwealth University, Richmond, VA, United States

[6]Department of Epidemiology, Johns Hopkins Bloomberg School of Public Health, Baltimore, MD, United States

[7]Department of Medicine, University of Washington, Seattle, WA, United States

[8]Department of Internal Medicine, University of Nebraska Medical Center, Omaha, NE, United States

[9]Department of Cancer Epidemiology, H. Lee Moffitt Cancer Center and Research Institute, Tampa, FL, United States

[10]Department of Oncologic Sciences, University of South Florida, Tampa, FL, United States

[11]Department of Medicine, University of Alabama at Birmingham, Birmingham, AL, United States

[12]Biostatistics Program Office of the Director, National Institute of Diabetes and Digestive and Kidney Diseases, National Institutes of Health, Bethesda, MD, United States

[13]Division of Biostatistics, University of California Berkeley School of Public Health, Berkeley, CA, United States

[14]Department of Clinical and Translational Research, University of Rochester Medical Center, Rochester, NY, United States

[15]Department of Biostatistics, School of Public Health, University of Minnesota, Minneapolis, MN, United States

[16]See Acknowledgments

[*]these authors contributed equally

Corresponding Author:
Eric Hurwitz, PhD
Department of Genetics
University of North Carolina at Chapel Hill
120 Mason Farm Road
Chapel Hill, NC 27514
United States
Phone: 1 919 843 8724
Email: eric_hurwitz@med.unc.edu

Abstract

Background: Electronic health records (EHRs) provide valuable insights to address clinical and epidemiological research concerning HIV, including the disproportionate impact of the COVID-19 pandemic on people living with HIV. To identify this population, most studies using EHR or claims databases start with diagnostic codes, which can result in misclassification without further refinement using drug or laboratory data. Furthermore, given that antiretrovirals now have indications for both HIV and COVID-19 (ie, ritonavir in nirmatrelvir/ritonavir), new phenotyping methods are needed to better capture people living with HIV. Therefore, we created a generalizable and innovative method to robustly identify people living with HIV, preexposure prophylaxis

(PrEP) users, postexposure prophylaxis (PEP) users, and people not living with HIV using granular clinical data after the emergence of COVID-19.

Objective: The primary aim of this study was to use computational phenotyping in EHR data to identify people living with HIV (cohort 1), PrEP users (cohort 2), PEP users (cohort 3), or "none of the above" (people not living with HIV; cohort 4) and describe COVID-19–related characteristics among these cohorts.

Methods: We used diagnostic and laboratory measurements and drug concepts in the National Clinical Cohort Collaborative to create a computational phenotype for the 4 cohorts with confidence levels. For robustness, we conducted a randomly sampled, blinded clinician annotation to assess precision. We calculated the distribution of demographics, comorbidities, and COVID-19 variables among the 4 cohorts.

Results: We identified 132,664 people living with HIV with a high level of confidence, 36,088 PrEP users, 4120 PEP users, and 20,639,675 people not living with HIV. Most people living with HIV were identified by a combination of medical conditions, laboratory measurements, and drug exposures (74,809/132,664, 56.4%), followed by laboratory measurements and drug exposures (15,241/132,664, 11.5%) and then by medical conditions and drug exposures (14,595/132,664, 11%). A higher proportion of people living with HIV experienced COVID-19–related hospitalization (4650,132,664, 3.5%) or mortality (828/132,664, 0.6%) and all-cause mortality (2083/132,664, 1.6%) compared to other cohorts.

Conclusions: Using an extensive phenotyping algorithm leveraging granular data in an EHR repository, we have identified people living with HIV, people not living with HIV, PrEP users, and PEP users. Our findings offer transferable lessons to optimize future EHR phenotyping for these cohorts.

(JMIR Med Inform 2025;13:e68143) doi: 10.2196/68143

KEYWORDS

HIV; electronic health records phenotype; epidemiologic methods; COVID-19; preexposure prophylaxis; postexposure prophylaxis

Introduction

Background

In the United States, there are an estimated 1.2 million people living with HIV, with 32,100 new cases diagnosed in 2021 [1]. The COVID-19 pandemic has disproportionately impacted people living with HIV, who have a 4-fold higher likelihood of contracting COVID-19 and experiencing severe outcomes, such as increased disease severity, higher hospitalization rates, and elevated mortality [2-4]. Initially, the COVID-19 pandemic introduced challenges to optimal HIV management, including a substantial reduction in HIV testing, resulting in underdiagnosis and barriers to necessary treatment [5-8]. Beyond its effects on people living with HIV, the COVID-19 pandemic also disrupted preexposure prophylaxis (PrEP) use and availability, potentially increasing the vulnerability of those at risk of acquiring HIV [9]. For example, PrEP users experienced a 6% to 11.5% decline in medication coverage early in the pandemic [5,9,10]. Disparities related to social determinants of health that have historically affected minoritized (ie, racial, ethnic, gender, rural, and socioeconomic) communities with HIV became more apparent since the onset of the COVID-19 pandemic [11-15]. The pronounced impact of COVID-19 on people living with HIV and those at risk for HIV underscores the urgent need for focused research within this population for better preparedness for the next public health crisis that may affect an already vulnerable population.

Electronic health records (EHRs) provide valuable insights to address specific clinical and epidemiological research concerning HIV, including the disproportionate impact of the COVID-19 pandemic on this population [2,13-15]. The National Clinical Cohort Collaborative (N3C) is a nationally sampled EHR repository in the United States with granular, individual-level clinical data from >98 data partner sites and houses >32 billion rows of data for >22 million individuals [16]. These rich EHR data have allowed investigation of questions related to the COVID-19 pandemic and HIV, including identification of an elevated risk of poor COVID-19 outcomes in people living with HIV with a low cluster of differentiation 4 (CD4) count (<200 cells/µL) and racial disparities with COVID-19 positivity among people living with HIV, in addition to both individual and area-level social determinants of health with COVID-19 outcomes [2,13-15]. The N3C Enclave not only provides a large sample size for answering research questions but also contains granular individual-level information, such as social determinants of health [17], visit and prescription frequencies, and geographic information to better identify health disparities during the COVID-19 pandemic [18,19].

Nonetheless, identifying potential people living with HIV, PrEP users, or postexposure prophylaxis (PEP) users in any EHR dataset necessitates indirectly inferring data from source information. Most studies using EHR or claims databases start with diagnostic codes, which can result in misclassification without further refinement using drug or laboratory data [20-22]. While these previous studies have performed computational phenotyping of HIV on smaller cohorts, with data or validation often limited to a single institution or health system, one study [23], conducted with All of Us data, has included additional complexity such as self-reporting of medical conditions, which is not always feasible with large datasets [22-26]. While each approach has strengths, they also suffer from limitations, such as (1) variation in the availability of EHR data among individuals that potentially leads to differential misclassification, (2) use of nonstandardized codes specific to a single health system preventing interoperability across health systems, (3) inability to distinguish between individuals with high- or low-confidence designations, (4) additional misclassification

with increased uptake of PrEP or PEP, and (5) the use of antiretrovirals that now have indications for both HIV and COVID-19 treatment (ie, ritonavir in nirmatrelvir/ritonavir) resulting in misclassification.

Objectives

Our objective was to use an innovative approach after the emergence of COVID-19 to robustly identify people living with HIV, PrEP users, PEP users, and people not living with HIV using granular clinical data in N3C, augmented by clinician annotation without chart abstraction of source data. We used the standardized Observational Medical Outcomes Partnership (OMOP) Common Data Model to create 4 cohorts or "phenotypes," including people living with HIV (cohort 1), PrEP users (cohort 2), PEP users (cohort 3), and people not living with HIV (cohort 4), based on medical conditions, laboratory measurements, and drug data available in N3C. We conducted a clinician-curated annotation to reduce and describe misclassification. In addition, we introduced confidence levels to classify individuals into categories of low, medium, or high confidence in their status as people living with HIV, allowing researchers to tailor the cohorts to their specific research needs. These cohorts will serve as a framework for future research to address questions related to HIV within N3C, and our methods inform more generalizable approaches to phenotyping HIV in EHR data.

Methods

Study Population

The N3C Enclave contains EHR data from clinical sites and represents the largest limited dataset of COVID-19 cases and controls in the United States. N3C contains harmonized EHR data for individuals who tested positive and negative for COVID-19 from January 1, 2020, to present through routine weekly updates to incorporate near real-time clinical encounters and procedures, in addition to historical data dating back to January 1, 2018. Data include individuals who tested positive for COVID-19 matched with 2 controls who tested negative for COVID-19 based on a maximum of 4 sociodemographic variables (age, sex, race, and ethnicity) whenever available by data partner site. Our cohorts of people living with HIV, PrEP users, PEP users, and people not living with HIV were defined based on the N3C Enclave data release as of November 2, 2023 (version 148; from 93 data partner sites) using a limited dataset.

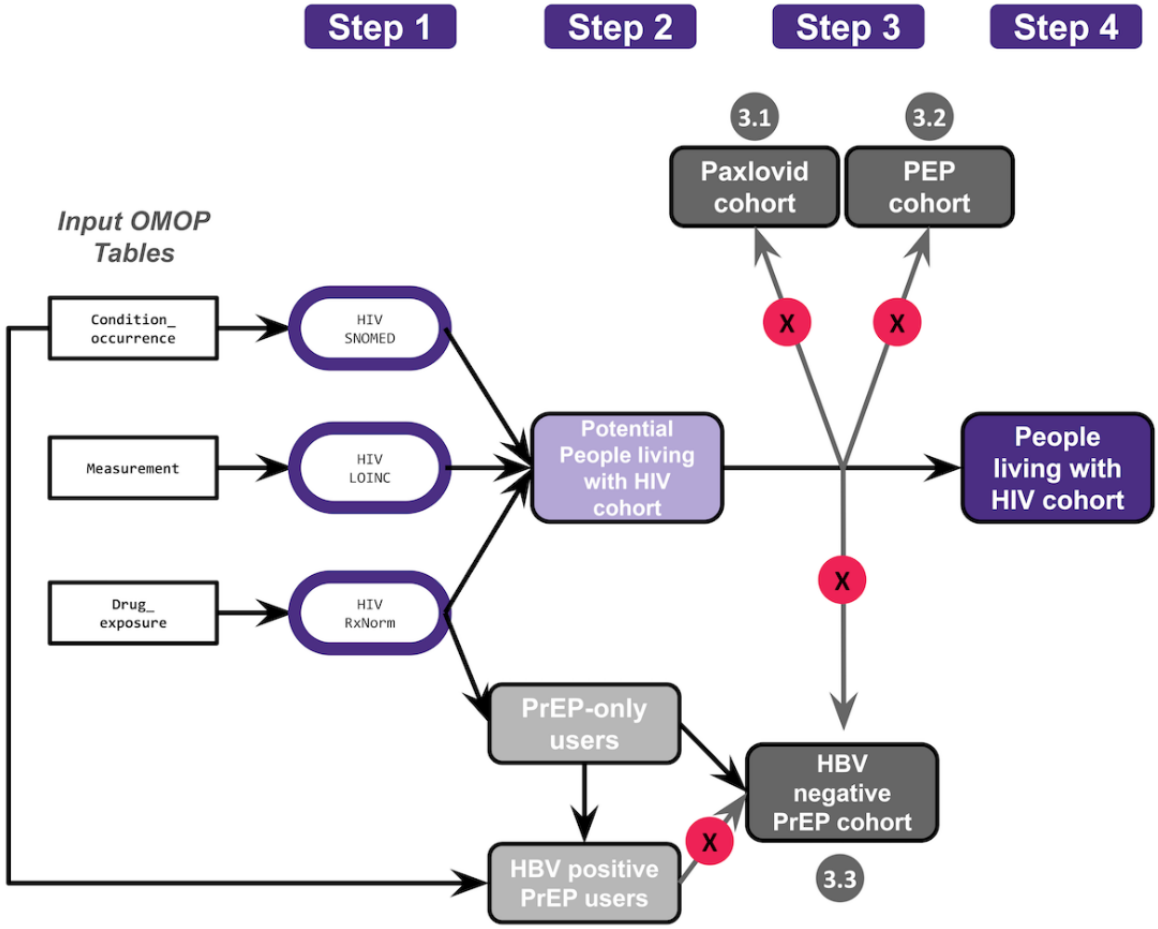
Procedures

HIV Phenotyping Overview Scheme

Our cohorts were identified in the N3C Enclave through the use of OMOP concepts, where concept IDs, concept set names, and our phenotyping pipeline can be searched in our public-facing GitHub repository for transferable methods in and out of N3C [27]. These standardized concepts correspond to three major sources of data: (1) Systematized Medical Nomenclature for Medicine Clinical Terminology (SNOMED CT) codes and *International Classification of Diseases, Tenth Revision* (*ICD-10*) mapped to SNOMED CT codes for HIV diagnosis from the HIV-related condition occurrence table; (2) Logical Observation Identifiers Names and Codes for HIV measurements from the measurement table; and (3) HIV medication exposures from the drug exposure table in RxNorm [28]. The selection of OMOP concepts across these 3 domains was carried out using the Observational Health Data Sciences and Informatics Atlas tool [29,30], in collaboration with HIV clinicians, to identify and include all relevant HIV-related OMOP concept IDs.

Conceptually, our approach identified possible people living with HIV, followed by subtyping PrEP and PEP users and removing people using HIV drugs for only chronic hepatitis B virus (HBV) or COVID-19 infection (Figure 1). From the potential cohort of people living with HIV, using the table of individuals prescribed HIV-related drugs, we phenotyped PrEP users using concept IDs in the *(immunosuppressed or compromised) PrEP concept set* and excluding individuals with an HIV diagnostic code or positive laboratory measurement (eg, antibody and antigen, viral load [VL] ≥50 copies/mL, or CD4 ≤200 cells/μL) [27]. The final cohort of PrEP users was created by excluding individuals with HBV infection to avoid potential misclassification (as tenofovir disoproxil fumarate [TDF] alone is often used for HBV treatment and avoided for PrEP to reduce the risk of acute hepatitis flare upon discontinuation of TDF for PrEP). From the potential cohort of people living with HIV, identified earlier by drug exposures (after excluding PrEP users), PEP users were identified using the concepts in the *PEP concept set* [27]. Individuals with ritonavir-only exposure (ie, those with no other antiretroviral exposure for HIV treatment, no medical conditions, and no laboratory measurements), were categorized as "uncertain classification," given the overlap of ritonavir use for COVID-19 treatment in nirmatrelvir/ritonavir using concepts listed in *Ritonavir concept set* and *Paxlovid concept set* [27]. In addition to our "ritonavir only" group, we identified those with uncertain classification status for whom, despite our best efforts, we could not adjudicate with precision which cohort they belonged to and therefore excluded these individuals from all cohorts. For instance, some individuals had repeat HIV VL testing without any results or other HIV-related laboratory measurements, drug exposures, or medical conditions, where the repeat testing could represent monitoring not only HIV treatment but also HIV screening. We considered this uncertain classification to be of low confidence, with high concern for misclassification. Given the risk of additional misclassification that assigning such individuals to one of the 4 specific cohorts would cause, we chose to remove them from the 4 specific cohorts altogether. Thus, the final cohort of people living with HIV was created by including individuals positive for HIV by medical conditions, laboratory measurements, or drug exposures and excluding any individuals listed in the final PrEP, PEP, and "ritonavir-only" cohorts (Figure 1).

Figure 1. A schematic detailing the process of creating our people living with HIV cohort. Individuals were included in the HIV cohort by a combination of medical conditions, laboratory measurements, or drug exposures (step 1) to create a potential cohort of people living with HIV (step 2). Individuals taking nirmatrelvir/ritonavir (step 3.1), postexposure prophylaxis (PEP) users (step 3.2), and pre-exposure prophylaxis (PrEP) users (step 3.3) were excluded from the intermediate cohort (step 3), which is depicted by "X" in the red circle, to create a final cohort of people living with HIV (step 4). HBV: hepatitis B virus; LOINC: Logical Observation Identifiers Names and Codes; OMOP: Observational Medical Outcomes Partnership; SNOMED: Systematized Medical Nomenclature for Medicine.



Data Preprocessing

Laboratory measurements, such as absolute CD4 cell count and percentage and VL measures, are crucial to research among people living with HIV. To extract meaningful data related to CD4 and VL measurements, we first identified concept IDs for CD4 and VL measurements seen in the *CD4 Cell Count (Absolute)* and the *HIV Viral Load (version 3) concept sets* [27]. CD4 and VL phenotyping were performed on all individuals in N3C using the measurement table.

Applying Levels of Confidence to the Cohort of People Living With HIV

To determine, with varying levels of confidence, who within the cohort were people living with HIV, we assigned a confidence level between 1 and 3 (with 1 indicating the highest confidence) per individual to reflect the degree of certainty in their status as someone living with HIV. Briefly, confidence level 1 included individuals with positive HIV laboratory results; confidence level 2 comprised individuals identified through a combination of medical conditions, the presence of laboratory measurements, and HIV medication exposures but without available laboratory results; and confidence level 3 consisted of individuals identified solely based on HIV drug exposures or only 1 record of an HIV diagnosis. More details are provided in the applying levels of confidence to the cohort of people living with HIV section in Multimedia Appendix 1 [16,31-33].

PrEP Phenotyping

In order to identify PrEP users, we first identified those using PrEP drugs (emtricitabine and tenofovir) from the possible HIV cohort who were identified by drug exposures only (ie, excluding individuals with an HIV diagnostic code or positive laboratory measurement, such as antibody and antigen, VL ≥50 copies/mL, or CD4 ≤200 cells/μL). Due to low numbers, cabotegravir for PrEP was not included. Among these, as noted earlier, individuals with likely chronic HBV infection were excluded based on concept IDs listed in the *HBV concept set* [27]. Further details about PrEP phenotyping and assigning confidence levels for PrEP users are provided in Multimedia Appendix 1.

PEP Phenotyping

PEP phenotyping was performed using the potential cohort of people living with HIV identified by drug exposures after excluding those in the PrEP cohort. We then filtered individuals based on the following criteria: (1) had a record of emtricitabine/TDF and dolutegravir or emtricitabine/TDF and

raltegravir use on the same date, (2) had a supply of <30 days (or the supply information was null), and (3) only had one row of data (indicating they were not receiving an HIV treatment regimen repeatedly, which would otherwise be more indicative of an HIV-infection–related treatment). Given that our phenotyping approach was dependent on accounting for all data available over time, it was not possible to distinguish, with confidence, PEP users who later acquired HIV. Concepts used in this workflow are available in the *PEP concept set* [27]. Given the overall small size of the PEP cohort, we chose not to pursue confidence level typing of this group.

Ritonavir-Only Phenotyping

We recognized that several individuals may have been prescribed ritonavir as part of nirmatrelvir/ritonavir (ie, Paxlovid) or another off-label regimen as a COVID-19 therapeutic. Therefore, those entering our cohort of people living with HIV by drug exposures only and who had only been prescribed ritonavir (without any other antiretrovirals used for HIV treatment documented) after January 1, 2020, were categorized separately for further evaluation using concepts described earlier [27]. We then identified those taking ritonavir as part of nirmatrelvir/ritonavir via concepts and drug source values. Individuals who only used ritonavir as part of nirmatrelvir/ritonavir were subsequently excluded. Given that there were no other laboratory tests, medical conditions, or drug exposures consistent with HIV, we suspected this group reflected people not living with HIV; however, due to the uncertainty of the indication for the ritonavir prescription, we categorized them in an uncertain classification.

Clinician Annotation Activity

To ensure the robustness of our cohorts of people living with HIV, PrEP users, and PEP users, clinicians experienced in treating HIV conducted a blinded annotation activity. Briefly, clinician annotation involving 120 randomly selected individuals (n=90, 75% people living with HIV; n=10, 8.3% PrEP users; n=10, 8.3% PEP users; and n=10, 8.3% individuals not living with HIV) was performed. Three clinicians with experience of treating HIV, blinded to the study procedures, reviewed row-level HIV EHR data (including medical conditions, laboratory measurements, and drug exposures) and classified each individual as person living with HIV, PrEP user, PEP user, or person not living with HIV based on the available data for each individual. For each clinician, we generated a confusion matrix and calculated sensitivity, specificity, positive predictive value, negative predictive value, precision, recall, and F_1-score. Multimedia Appendix 1 provides further details about this process.

Demographics and Comorbidity Definitions

Demographics and comorbidities of individuals were defined using standardized N3C-wide definitions. The set of comorbidities selected included those in the Charlson Comorbidity Index. Concept sets were created for each medical condition listed using the primary conditions listed in the SNOMED CT hierarchy, including all descendants. Specific details for each concept set can be found in the N3C concept set browser and are described in the data dictionary [16,34].

COVID-19 Outcome Definitions

The COVID-19 severity and outcomes for individuals were defined using N3C-wide definitions. In this analysis, COVID-19 positivity was defined by (1) a set of a priori–defined SARS-CoV-2 laboratory tests (that included polymerase chain reaction or antigen positivity but not antibody positivity) or (2) a "strong positive" diagnostic code (this cohort code is available on GitHub) [35,36].

Ethical Considerations

Direct patient consent was not obtained for this repository of deidentified data per N3C policies. The N3C received a waiver of consent from the National Institutes of Health (NIH) Institutional Review Board (IRB), and NIH takes care to ensure the highest privacy and security requirements are met and adhered to for housing and protecting these data in the NIH-managed N3C Enclave. More details can be found in N3C resources [37].

The N3C Enclave is approved through the NIH IRB. Each individual data partner site maintains its own IRB-approved data transfer agreement or joins under a Johns Hopkins University Reliance Protocol (IRB00249128). Each investigator accessing the N3C Enclave receives local IRB approval from their respective institutions. The N3C Data Access Committee approved this project (RP-CA3365).

Results

Identifying People Living With HIV, PrEP Users, and PEP Users in N3C

Using a computational phenotype through a combination of medical conditions, laboratory measurements, and medication data, we identified 152,282 (0.7%) individuals as people living with HIV from a total population of 20,928,656 (Figure 2). Among these cases, many were recognized through the combination of medical conditions, laboratory measurements, and drug exposures (n=74,809, 49.1%), followed by medical conditions alone (n=15,617, 10.3%), and then by laboratory measurements and drug exposures (n=15,241, 10%; Figure 3). When categorizing people living with HIV based on confidence levels, we found that most individuals fell into confidence level 1 (n=108,068, 71%), followed by level 2 (n=24,596, 16.2%), and then level 3 (n=19,618, 12.9%; Table 1). Our phenotyping process also identified 36,088 PrEP users (after excluding 303 individuals for HBV monoinfection), 4120 PEP users (the total number of PEP users is presented as an estimate due to N3C governance rules to obfuscate counts <20), and 96,491 individuals with uncertain classification (Table 2). In terms of our confidence levels for each PrEP user, most individuals were categorized as level 1 (20,742/36,088, 57.5%), followed by level 3 (11,560/36,088, 32%) and then level 2 (3786/36,088, 10.5%; Table 3). We did observe a small number of individuals from the people living with HIV cohort (24/152,282, 0%) with possible PrEP exposure before a diagnosis of HIV.

Figure 2. Flow diagram describing the inclusion and exclusion criteria of the HIV cohort. *Total number of postexposure prophylaxis (PEP) users is presented as an estimate due to National Clinical Cohort Collaborative (N3C) governance rules to obfuscate counts <20. HBV: hepatitis B virus; PrEP: pre-exposure prophylaxis.

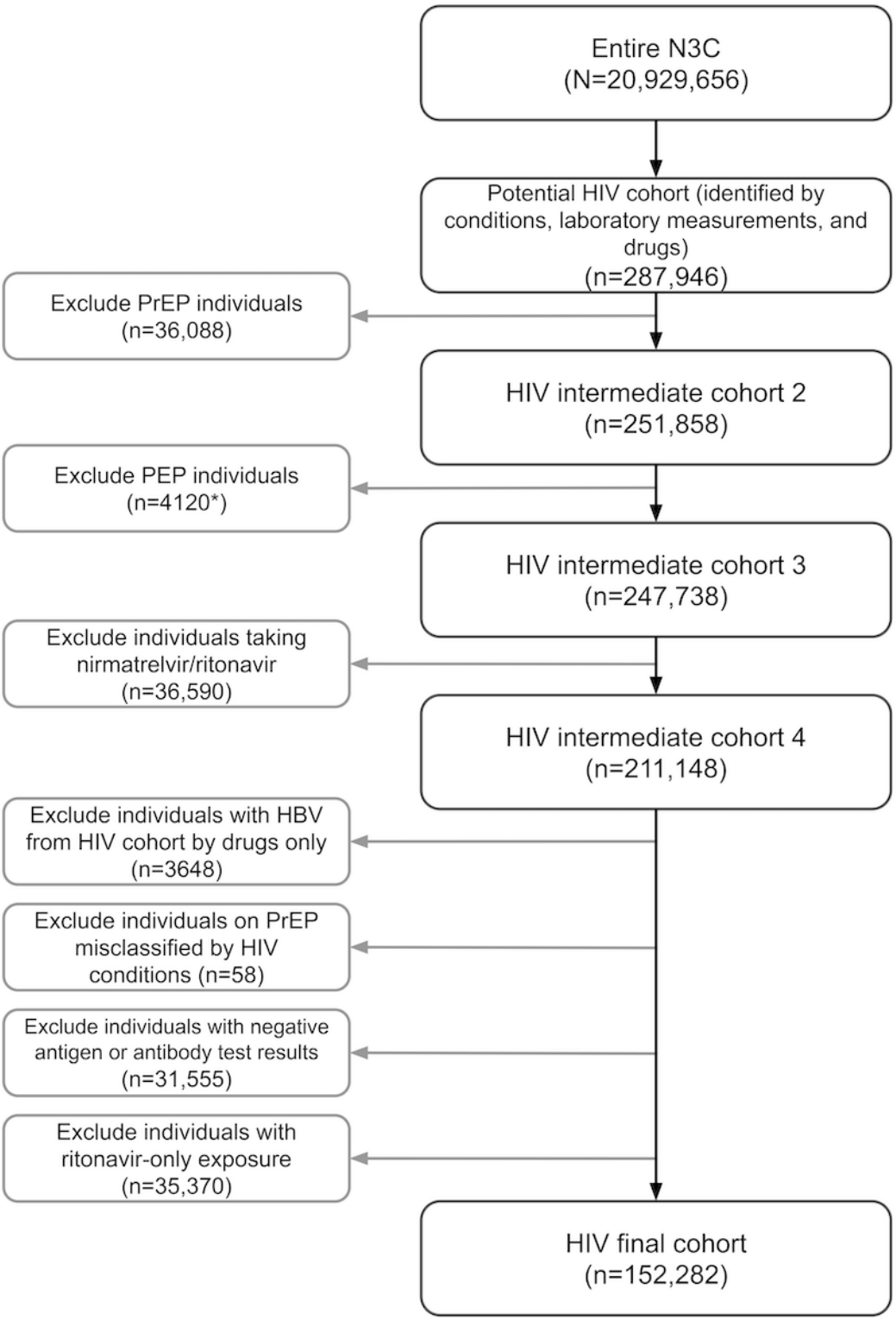


Figure 3. A Venn diagram displaying the number of individuals in the HIV cohort identified by medical conditions, laboratory measurements, and drug exposures (n=152,282). Top numbers represents the count (and percentage) of individuals in specific categories (eg, 14,595 individuals identified by the combination of medications and medical conditions), while bottom numbers display the number of individuals broadly identified by medical conditions, laboratory measurements, and medications (eg, 116,155 individuals in the HIV cohort have an HIV-related condition).

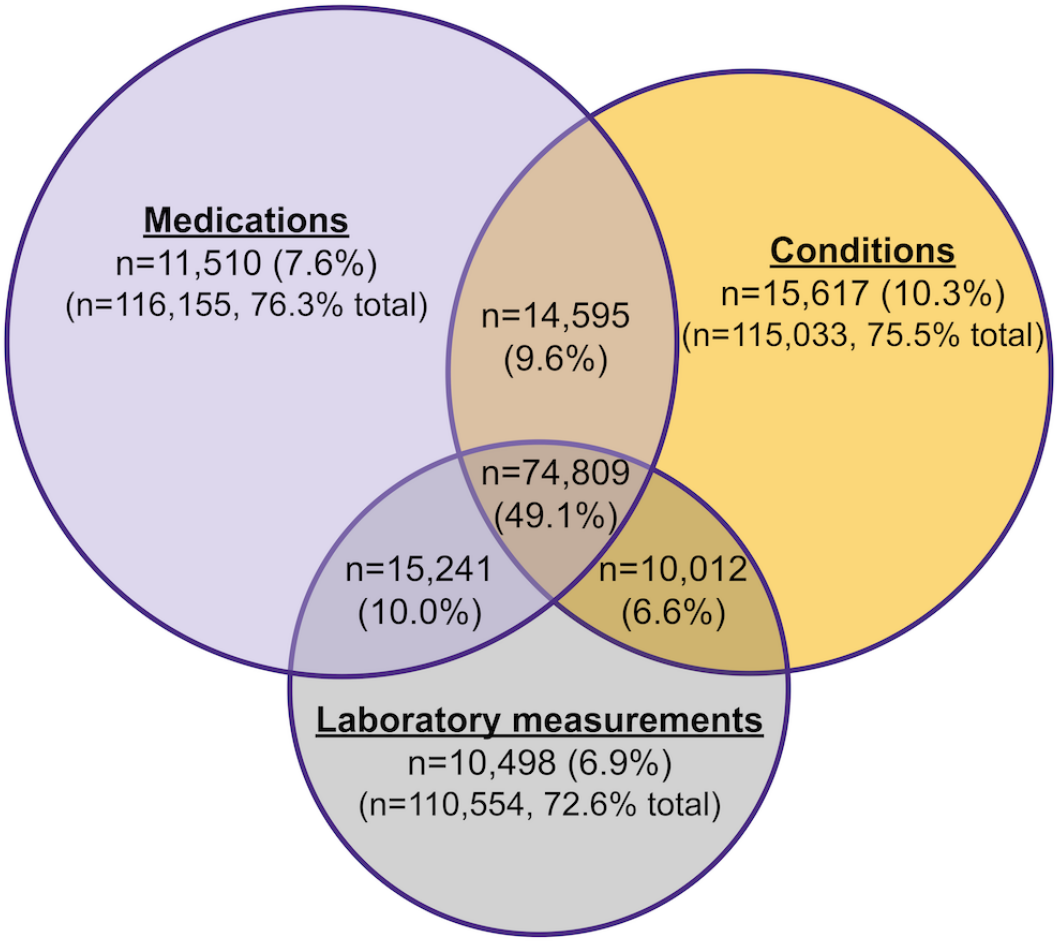


Table 1. Counts and percentages of people living with HIV in each confidence level by criterion (n=152,282).

Confidence level	Count, n (%)
Level 1 total	108,068 (71)
Criterion 1	41,923 (27.5)
Criterion 2	28,512 (18.7)
Criterion 3	1379 (0.9)
Criterion 4	5606 (3.7)
Criterion 5	18,499 (12.1)
Criterion 6	12,495 (8.2)
Level 2 total	24,596 (16.2)
Criterion 7	264 (0.2)
Criterion 8	13,061 (8.6)
Criterion 9	7509 (4.9)
Level 3 total	19,618 (12.9)
Criterion 10	7078 (4.6)
Criterion 11	4432 (2.9)
Criterion 12	8108 (5.3)

Table 2. Counts and percentages of individuals in each cohort (N=20,928,656)[a].

Cohort	Count, n (%)
People living with HIV (confidence levels 1-3)	152,282 (0.7)
People not living with HIV	20,639,675 (98.6)
PrEP[b] users	36,088 (0.2)
PEP[c] users	4120 (0)
Uncertain classification	96,491 (0.5)

[a]Count for the entire National Clinical Cohort Collaborative.

[b]PrEP: preexposure prophylaxis.

[c]PEP: postexposure prophylaxis.

Table 3. Counts and percentages of preexposure prophylaxis (PrEP) users in each confidence level (n=36,088).

Confidence level	Count, n (%)
Level 1	20,742 (57.5)
Level 2	3786 (10.5)
Level 3	11,560 (32)

Verification of Cohort of People Living With HIV, PrEP Users, and PEP Users via a Clinician Annotation Activity

To enhance the reliability of our methods, we conducted a clinician annotation process aimed at assessing the accuracy of our phenotyping methods for people living with HIV, PrEP users, and PEP users. Our results demonstrated excellent sensitivity (range 0.90-1), specificity (range 0.97-1), positive predictive value (range 0.77-1), negative predictive value (range 0.88-1), precision (range 0.77-1), recall (range 0.90-1), and F_1-score (range 0.87-1; Multimedia Appendix 2). Similarly, the κ values exhibited high levels of interrater reliability (range 0.91-0.98; Multimedia Appendix 2).

Characteristics of People Living With HIV, PrEP Users, and PEP Users in N3C

To obtain a more comprehensive understanding of the individuals within our various cohorts, we described their sociodemographic characteristics. Within the cohort of people living with HIV in the highest 2 confidence levels (132,664/152,282, 87.1% of all potential people living with HIV identified), the highest proportion of non-Hispanic Black individuals is in the people living with HIV cohort (49,769/132,664, 37.5%), followed by PEP users (784/4120, 19%), people not living with HIV (2,909,443/20,639,675, 14.1%), and finally among PrEP users (4534/36,088, 12.6%; Table 4). Notably, PrEP users contained the highest proportion of Hispanic or Latinx individuals (5829/36,088, 16.2%) of any race and ethnicity compared to other cohorts (people living with HIV: 20,925/132,664, 15.8%; PEP users: 403/4120, 9.8%; people not living with HIV: 2,628,308/20,639,675, 12.7%; Table 4). Male individuals constituted a higher proportion of people living with HIV (90,496/132,664, 68.2%) and PrEP users (31,722/36,088, 87.9%), which was notably higher than the proportion of male individuals among PEP cohorts (1399/4120, 34%) and people not living with HIV (9,044,853/20,639,675, 43.8%; Table 4). The median ages of people living with HIV, PrEP users, PEP users, and people not living with HIV were 51 (IQR 38-61), 36 (IQR 30-46), 35 (IQR 28-44), and 44 (IQR 26-63) years, respectively, demonstrating that PrEP and PEP users were generally younger than people living with HIV and people living without HIV (Table 4). We also examined comorbidities among people living with HIV compared to other cohorts and found that people living with HIV experienced chronic lung disease, liver disease, hypertension, kidney disease, and diabetes more often compared to PrEP users, PEP users, and people not living with HIV (Table 4).

Table 4. Descriptive statistics of people living with HIV, preexposure prophylaxis (PrEP) users, postexposure prophylaxis (PEP) users, and people not living with HIV cohorts.

Characteristics	People living with HIV (n=132,664; confidence level 1-2)	People living with HIV (n=152,282; confidence level 1-3)	PrEP users (n=36,088)	PEP users (n=4120)[a]	People not living with HIV (n=20,639,675)
Race or ethnicity, n (%)					
Hispanic or Latinx, any race	20,925 (15.8)	23,346 (15.3)	5829 (16.2)	403 (9.8)	2,628,308 (12.7)
Non-Hispanic American Indian or Alaska Native	615 (0.5)	684 (0.4)	163 (0.5)	<20 (0.5)	85,524 (0.4)
Non-Hispanic Asian	2753 (2.1)	3517 (2.3)	1726 (4.8)	222 (5.4)	729,491 (3.5)
Non-Hispanic Black or African American	49,769 (37.5)	55,144 (36.2)	4534 (12.6)	784 (19)	2,909,443 (14.1)
Non-Hispanic combined[b]	903 (0.7)	1150 (0.8)	505 (1.4)	40 (1)	281,450 (1.4)
Non-Hispanic Native Hawaiian or Pacific Islander	247 (0.2)	300 (0.2)	88 (0.2)	<20 (0.5)	67,808 (0.3)
Non-Hispanic White	48,604 (36.6)	57,479 (37.7)	19,663 (54.5)	2119 (51.4)	11,966,553 (58)
Unknown	8848 (6.7)	10,662 (7)	3580 (9.9)	512 (12.4)	1,971,098 (9.6)
Age (y), median (IQR)	51 (38-61)	51 (37-61)	36 (30-46)	35 (28-44)	44 (26-63)
Sex, n (%)					
Female	42,093 (31.7)	49,846 (32.7)	4253 (11.8)	2693 (65.4)	11,579,223 (56.1)
Male	90,496 (68.2)	102,332 (67.2)	31,722 (87.9)	1399 (34)	9,044,853 (43.8)
Combined or unknown[c]	75 (0.1)	104 (0.1)	113 (0.3)	<20 (0.5)	15,599 (0.1)
Comorbidity, n (%)					
Myocardial infarction	9265 (7)	10,349 (6.8)	482 (1.3)	57 (1.4)	795,829 (3.9)
Congestive heart failure	10,926 (8.2)	12,106 (7.9)	406 (1.1)	42 (1)	998,313 (4.8)
Peripheral vascular disease	8002 (6)	8971 (5.9)	352 (1)	52 (1.3)	787,932 (3.8)
Cerebrovascular disease	9082 (6.8)	10,123 (6.6)	492 (1.4)	76 (1.9)	977,070 (4.7)
Dementia	3177 (2.4)	3638 (2.4)	225 (0.6)	63 (1.5)	475,742 (2.3)
Chronic lung disease	40,080 (30.2)	44,328 (29.1)	5617 (15.6)	716 (17.5)	3,562,138 (17.3)
Rheumatologic disease	10,245 (7.7)	11,502 (7.6)	1280 (3.5)	139 (3.4)	1,097,753 (5.3)
Peptic ulcer	4445 (3.4)	4919 (3.2)	466 (1.3)	61 (1.5)	335,672 (1.6)
Liver disease	32,690 (24.6)	35,536 (23.3)	2965 (8.2)	320 (7.8)	1,314,849 (6.4)
Diabetes	28,307 (21.3)	31,731 (20.8)	2631 (7.3)	321 (7.8)	2,750,693 (13.3)
Hemiplegia or paraplegia	3751 (2.8)	4124 (2.7)	121 (0.3)	32 (0.8)	281,201 (1.4)
Kidney disease	26,015 (19.6)	28,288 (18.6)	1142 (3.2)	157 (3.8)	1,681,615 (8.1)
Cancer	20,035 (15.1)	22,447 (14.7)	1399 (3.9)	143 (3.5)	1,866,452 (9)
Hypertension	63,025 (47.5)	69,487 (45.6)	7464 (20.7)	691 (16.9)	5,904,993 (28.6)
Tobacco smoking	34,189 (25.8)	37,120 (24.4)	5830 (16.2)	550 (13.4)	1,844,538 (8.9)
BMI, n (%)					
Underweight	2280 (1.7)	2931 (1.9)	378 (1)	90 (2.2)	1,156,445 (5.6)
Healthy weight	20,789 (15.7)	23,743 (15.6)	6684 (18.5)	710 (17.3)	2,837,262 (13.7)
Overweight	35,471 (26.7)	39,718 (26.1)	10,600 (29.4)	864 (21.1)	3,565,102 (17.3)
Obese	52,872 (39.9)	59,238 (38.9)	12,240 (33.9)	1369 (33.4)	6,260,259 (30.3)
Missing	21,252 (16)	26,652 (17.5)	6186 (17.1)	1066 (26)	6,820,607 (33)

[a]The PEP cohort is estimated to obfuscate any groups with counts <20 to follow National Clinical Cohort Collaborative governance rules. This will also cause percentages not to sum to 100 in this group.

[b]Includes individuals with race and ethnicity variable values that do not include Hispanic, White, Black or African American, Asian, American Indian or Alaska Native, and Native Hawaiian or Other Pacific Islander.

[c]Includes individuals with sex variable values that do not include male or female.

HIV-Related Characteristics Among People Living With HIV in N3C

Of the 132,664 people living with HIV in the highest 2 confidence levels, 86,148 (64.9%) had at least 1 VL measurement, and 69,704 (52.5%) had at least 1 CD4 count recorded. The median number of VL and CD4 count laboratory tests was 5 (IQR 2-9) and 4 (IQR 2-8), respectively, per individual (Table 5). Of those with recorded VL and CD4 count measurements, 76,456 (57.6%) had VL and 69,254 (52.2%) had CD4 count results available for analysis (Table 5). Of those who had results, 41,923 (31.6%) had ≥1 detectable VLs (defined as VL >50 copies/mL) and 51,707 (39%) had ≥1 CD4 count <200 cells/µL (Table 5).

Table 5. HIV-related characteristics of people living with HIV, preexposure prophylaxis (PrEP) user, postexposure prophylaxis (PEP) user, and people not living with HIV cohorts.

Characteristics	People living with HIV (n=132,664; confidence level 1-2)	People living with HIV (n=152,282; confidence level 1-3)	PrEP users (n=36,088)	PEP users (n=4120)[a]	People not living with HIV (n=20,639,675)
Individuals with VL[b] and CD4[c] measurements, n (%)					
≥1 VL measurement	86,148 (64.9)	86,148 (56.6)	3983 (11)	0 (0)	11,390 (0.1)
≥1 CD4 measurement	69,704 (52.5)	69,704 (45.8)	259 (0.7)	0 (0)	36,528 (0.2)
VL measurements for analysis	76,456 (57.6)	76,456 (50.2)	3483 (9.7)	0 (0)	7599 (0)
CD4 measurement for analysis	69,254 (52.2)	69,254 (45.5)	238 (0.7)	0 (0)	36,121 (0.2)
VL and CD4 measurements per person (n), median (IQR)					
VL measurement count	5 (2-9)	5 (2-9)	1 (1-2)	0 (0-0)	1 (1-1)
CD4 measurement count	4 (2-8)	4 (2-8)	1 (1-1)	0 (0-0)	1 (1-3)
Individuals per VL category (copies/mL), n (%)					
VL ≥50	41,923 (31.6)	41,923 (27.5)	0 (0)	0 (0)	0 (0)
VL <50	34,533 (26)	34,533 (22.7)	3483 (9.7)	0 (0)	7599 (0)
VL unknown	9692 (7.3)	9692 (6.4)	500 (1.4)	4120[a] (100)	3792 (0)
VL null	46,516 (35.1)	66,134 (43.4)	32,105 (89)	0 (0)	20,628,284 (99.9)
Individuals per CD4 category (cells/µL), n (%)					
CD4 >200	51,707 (39)	51,707 (34)	238 (0.7)	0 (0)	27,718 (0.1)
CD4 ≤200	17,547 (13.2)	17,547 (11.5)	0 (0)	0 (0)	8403 (0)
CD4 unknown	450 (0.3)	450 (0.3)	21 (0.1)	4120[a] (100)	407 (0)
CD4 null	62,960 (47.5)	82,578 (54.2)	35,829 (99.3)	0 (0)	20,603,147 (99.8)
Individuals with HIV-related conditions and negative laboratory results, n (%)	11,048 (8.3)	11,048 (7.3)	0 (0)	0 (0)	0 (0)

[a]The PEP cohort is estimated to obfuscate any groups with counts <20 to follow National Clinical Cohort Collaborative governance rules.

[b]VL: viral load.

[c]CD4: cluster of differentiation 4.

COVID-19 Outcomes in People Living With HIV, PrEP Users, and PEP Users in N3C

We examined the distribution of COVID-19–related variables within the cohorts of people living with HIV, PrEP users, and PEP users in comparison to people not living with HIV. The number of individuals who were tested for COVID-19, regardless of the results, was 104,903 (79.1%) among 132,664 people living with HIV (confidence levels 1 and 2), 30,799 (85.3%) for 36,008 PrEP users, 3704 (89.9%) for 4120 PEP users, and 17,756,497 (86%) in 20,639,675 people not living with HIV (Table 6). The proportions who tested positive for COVID-19 were similar between the cohorts; there were 34.4% (45,609/132,664) people living with HIV, 33.6% (12,121/36,088) PrEP users, 37% (1525/4120) PEP users, and 38.6% (7,970,336/20,639,675) people not living with HIV who tested positive for COVID-19 (Table 6). When categorizing individuals based on the severity of COVID-19, a higher

proportion of people living with HIV experienced a COVID-19–related hospitalization (4650/132,664, 3.5%), COVID-19–related mortality (828/132,664, 0.6%), and all-cause mortality (2083/132,664, 1.6%) outcomes compared to people not living with HIV who had a lower proportion of COVID-19–related hospitalization (407,894/20,639,675, 2%), COVID-19–related mortality (95,421/20,639,675, 0.5%), and all-cause mortality (192,519/20,639,675, 0.9%; Table 6). With regard to receipt of COVID-19 vaccinations, people living with HIV exhibited a lower proportion of full primary (ie, completing 2 doses when applicable) vaccination documentation (15,387/132,664, 11.6%) compared to PrEP users (5414/36,088, 15%), although it was higher compared to PEP users (441/4120, 10.7%) and similar to people not living with HIV (2,327,522/20,639,675, 11.3%; Table 6). The booster vaccination documentation displayed a lower proportion of people living with HIV receiving any booster vaccination (28,520/132,664, 21.5%) compared to PrEP users (12,547/36,088, 34.8%) but higher than PEP users (559/4120, 13.6%) and people not living with HIV (2,521,638/20,639,675, 12.2%; Table 6).

Table 6. COVID-19–related characteristics of the people living with HIV, preexposure prophylaxis (PrEP) user, postexposure prophylaxis (PEP) user, and people not living with HIV cohorts.

Characteristics	People living with HIV (n=132,664; confidence level 1-2), n (%)	People living with HIV (n=152,282; confidence level 1-3), n (%)	PrEP users (n=36,088), n (%)	PEP users (n=4120)[a], n (%)	People not living with HIV (n=20,639,675), n (%)
COVID-19 testing and outcomes					
COVID-19 screening	104,903 (79.1)	121,130 (79.5)	30,799 (85.3)	3704 (90.4)	17,756,497 (86)
COVID-19 positivity	45,609 (34.4)	52,223 (34.3)	12,121 (33.6)	1525 (37.2)	7,970,336 (38.6)
COVID-19–associated hospitalization	4650 (3.5)	5259 (3.5)	226 (0.6)	51 (1.2)	407,894 (2)
COVID-19–associated mortality	828 (0.6)	1027 (0.7)	<20 (0.1)	<20 (0.5)	95,421 (0.5)
All-cause mortality among individuals with COVID-19	2083 (1.6)	2412 (1.6)	40 (0.1)	<20 (0.5)	192,519 (0.9)
Had ≥1 COVID-19 reinfection	2864 (2.2)	3204 (2.1)	612 (1.7)	118 (2.9)	353,177 (1.7)
COVID-19 vaccination					
No vaccination documented	81,096 (61.1)	95,300 (62.6)	15,855 (43.9)	2922 (71.3)	15,087,131 (73.1)
Partial primary vaccination series documented	7661 (5.8)	8464 (5.6)	2272 (6.3)	177 (4.3)	703,384 (3.4)
Full primary vaccination series documented	15,387 (11.6)	17,301 (11.4)	5414 (15)	441 (10.7)	2,327,522 (11.3)
Booster vaccination documented	28,520 (21.5)	31,217 (20.5)	12,547 (34.8)	559 (13.6)	2,521,638 (12.2)

[a]The PEP cohort is estimated to obfuscate any groups with counts <20 to follow National Clinical Cohort Collaborative governance rules.

Discussion

Principal Findings

Using an extensive phenotyping algorithm that leverages granular data for medical conditions, laboratory measurements, and drug exposures available in an EHR repository, we have identified people living with HIV, PrEP users, and PEP users, augmented by confidence levels to provide flexibility in cohort selection to address potential misclassification. Our algorithm was refined using an iterative process with multiple reviews by clinicians with HIV experience to reduce misclassification. Our work allows for rapid identification of people living with HIV in large datasets originating from multiple health systems, with reduced misclassification among PrEP or PEP users, those with HBV infection only, or those prescribed ritonavir only for COVID-19 treatment. Our computational phenotyping approach allows for greater transferability to other EHR data sources, accounts for nuances after the emergence of COVID-19, and is applicable for optimizing future research. Our approach lays the groundwork for large epidemiological investigations among people living with HIV and those at risk of acquiring HIV [13,14,38-41], noting that analyses to date that use this phenotyping have influenced guidance for COVID-19 vaccination prioritization for people living with HIV and can inform interventions, ranging from addressing clinical to social needs.

We have identified one of the largest cohorts of people living with HIV in the United States, leveraging a nationally sampled EHR repository. The overall distribution of sociodemographic, comorbidity, and COVID-19 findings of our cohorts is consistent with available literature. We identified a higher proportion of diabetes, hypertension, and chronic lung disease in people living with HIV than in people not living with HIV, which reflects the epidemiology of comorbidity burden for HIV in the United

States [1,42]. Similarly, where laboratory results were available, our proportion of people living with HIV with undetectable VL (34,533/132,664, 26%) reflects the HIV care continuum for the United States [1]. In addition, the demographics for PrEP users are also consistent with what has been previously reported in the United States. Similarly, with regard to COVID-19, we saw increased COVID-19–related hospitalizations, mortality, and all-cause mortality in people living with HIV than in people not living with HIV, which has been previously reported in other cohorts of people living with HIV [43-46]. Nonetheless, appreciating the nuanced similarities and differences between our N3C cohorts and others helps articulate places of generalizability to the HIV or PrEP populations in the United States.

One key element of our computational phenotype, with clinician annotation, aimed to minimize potential misclassification. For instance, in the entire N3C, we identified 121,099 individuals with a condition for HIV, but 16,185 (13.4%) of them had tested negative for HIV, suggesting these individuals may be misclassified if we used medical conditions alone. In our experience, it also appeared that medical conditions were occasionally applied inappropriately for those who were being screened for HIV or receiving prescriptions for PrEP or PEP. Many of the published HIV case-finding algorithms, with specificity frequently >90%, were validated before approval of and more expanded access to PrEP or nonoccupational PEP [47,48]. Thus, the potential for misclassification is greater now than ever.

Our computational phenotyping approach parallels other recent work, although arguably, it adds nuances that help reduce potential misclassification. Yang et al [23] conducted a study involving computational phenotyping of people living with HIV and PrEP users within the All of Us research program. They used a hybrid approach, with both EHR and self-reported survey data, to identify people living with HIV. Notably, their findings indicated that the identification of people living with HIV was most prevalent through a combination of drug exposures and medical conditions (n=3324 individuals, constituting 43.4% as opposed to our study with 14,595/152,282, 9.6%) as well as drug exposure alone (n=2191 individuals, accounting for 28.6% as opposed to our study with 11,510/152,282, 7.6%), with 19.9% (43,419/216,971) of the individuals with drug exposure only to ritonavir [23]. Similarly, in the HIV-Phen algorithm, May et al [24] identified misclassification with *ICD-10* codes alone, including roughly 6.5% of individuals who only had a screening test. Within the Veterans Health Administration, using a single *ICD-10* code alone provided a positive predictive value of only 69% [25]. Notably, among our cohort of people not living with HIV, 6065 individuals had inappropriate HIV-related conditions assigned despite their data suggesting they are not living with HIV, and 4159 (68.6%) of them had only a single medical condition; 58 of the PrEP users in our PrEP cohort had a condition for HIV inappropriately assigned. This misclassification was relatively small for the people not living with HIV cohort (6065/20,639,675, 0.03%); however, it would make up 4% of our people living with HIV cohort if they were included. This underscores the fallibility of relying on diagnostic codes alone, and our work can provide estimates of misclassification for researchers who may not have access to additional clinical data (ie, those limited to claims data). To help address the low specificity of medical conditions alone, we used a comprehensive approach evaluating not only positive or negative laboratory results to identify our cohorts but also repeated laboratory measurements over time. For example, an instance of an individual with one row of medical condition data followed by multiple negative HIV antigen or antibody tests is more consistent with an HIV-related condition assigned incorrectly. In contrast, an instance of an individual with one row of medical condition data followed by repeated VL and CD4 laboratory measurements is more consistent with an HIV-related condition being assigned correctly. We estimate that using medical conditions alone in other published algorithms, approximately 13.4% of probable people living with HIV were potentially misclassified according to the methods developed by Yang et al [23]. Adding an algorithm to identify PEP users also allows for reduced misclassification, as these individuals were identified by drug exposures only and would be classified as people living with HIV if not properly removed.

Another novel element of our computational phenotyping is the ability to identify individuals at risk for HIV through both PrEP and PEP exposures. Unfortunately, due to low numbers of people with exposure to long-acting injectables (cabotegravir and lenacapavir) in N3C, these were not included in our phenotyping; however, this work is currently in progress as the use of these agents expands. Among the aforementioned studies, only one [23] conducts computational phenotyping to identify PrEP or PEP users, with the latter being a group of people exposed to HIV drugs but only for a short time and, thus, embodying an element of temporality that is only available in longitudinal datasets. This provides the benefit of reducing potential misclassification between these 3 cohorts with antiretroviral drug exposure for differing reasons and provides additional comparator groups at risk of acquiring HIV for analyses. Nonetheless, readers should consider that our overall approach lacks source validation, as N3C prohibits reidentification of individuals with site data. While our clinician annotation provided measures of high consistency, we are unable to assess accuracy. Thus, we are limited in adjudicating further these reductions in potential misclassification without source validation.

With the emergence of COVID-19 and one of its primary treatments (ie, oral nirmatrelvir/ritonavir), which includes an antiretroviral also used for HIV treatment, using a single antiretroviral alone after COVID-19 can result in misclassification. Out of the entire N3C population, we identified 0.21% (43,149/20,832,144) of the individuals with only ritonavir exposure and 0.03% (6586/20,832,144) with ritonavir exposure and an HIV laboratory measurement. From this point forward, in the post–COVID-19 era, additional steps need to be implemented to reduce misclassification of people not living with HIV as people living with HIV when isolated ritonavir exposures exist. Yang et al [23] identified 2191 individuals (accounting for 28.6% of people living with HIV) by drug exposure alone; 19.9% (43,149/216,971) of our study's drug-only population had ritonavir-only exposures, underscoring the point that, after the emergence of COVID-19, potential

misclassification based on drug alone may be a significant issue [23].

A clinician-annotated approach to computational phenotyping of a cohort involving individuals with HIV is desirable from several points of view, engendering, for example, the use of clinically relevant data and insights. However, such efforts are less reproducible and repeatable over time. Thus, future work in automation, for example, through machine learning, could provide unique insights into how such a cohort can more readily be derived in any EHR dataset. Future research can assess various performance characteristics of clinician-annotated or curated approaches to automation approaches, as we acknowledge that our clinician-annotated approach was time consuming and may face challenges in transferability to other EHR repositories, despite our best efforts in making our phenotyping approach and code as transparent as possible.

Strengths and Limitations

While our work has several strengths, including identifying one of the largest cohorts of people living with HIV in the United States, demonstrating high consistency with our phenotyping using clinician annotation, and identifying PrEP and PEP users, it also faces several challenges. First, as already noted, the lack of source validation or any external validation limits our ability to assess the accuracy of the phenotypes. That being said, given our use of highly specific HIV laboratory data and repeated measures over time, we are extremely confident in our identification of people living with HIV in our confidence levels 1 and 2. Second, our dataset contains a significant amount of noninformative data. For example, there are instances where we can ascertain that an HIV screening test was performed, but the result is not available in an interpretable form, or a CD4 count is available, but the units are missing. This is likely a by-product of data ingestion and harmonization across multiple different health systems; while major strides have been made in EHR interoperability, more work needs to be done to meaningfully use all available data in the various EHRs and improve the quality of data ingestion processes. Third, while the N3C Enclave cohort is inclusive of multiple health systems across the United States, it is still only inclusive of individuals with encounters with the health care system, hence various characteristics, including COVID-19 outcomes, may not be fully generalizable to those affected by HIV with limited or no access to care, who are often the most vulnerable of the vulnerable.

Conclusions

Using an extensive phenotyping algorithm leveraging granular data in an EHR repository for medical conditions, laboratory measurements, and drug exposures, we have identified people living with HIV, PrEP users, and PEP users with high precision in the post–COVID-19 era. Our approach offers flexibility to select cohorts with confidence levels that best fit the needs of the research question under investigation, with regard to potential misclassification. We offer transferable lessons to optimize future EHR phenotyping for these cohorts. Our approach lays the groundwork for large epidemiological investigations among people living with HIV and those at risk for acquiring HIV and can inform interventions addressing clinical and social needs.

Acknowledgments

The authors would like to especially thank patients whose data are used in this analysis and the various National Clinical Cohort Collaborative (N3C) team members, whose work enables this and other research.

ChatGPT (GPT-3.5; OpenAI) and Claude (Anthropic) were used to edit pieces of the manuscript, including language, grammar, and synonyms. All edits by ChatGPT and Claude were reviewed by the authors and were not used for idea or content creation.

The analyses described in this paper were conducted with data or tools accessed through the National Center for Advancing Translational Sciences (NCATS) N3C Data Enclave [49] and N3C Attribution and Publication Policy (version 1.2-2020-08-25b) supported by NCATS U24 TR002306 and Axle Informatics subcontract (NCATS-P00438-B). Individual authors were supported by the following funding sources: National Institute of Mental Health (NIMH; R01131542; principal investigator [PI] RCP). The funding sources or study sponsors had no role in study design; in the collection, analysis, and interpretation of data; in the writing of the report; and in the decision to submit the paper for publication. This research was possible because of the patients whose information is included within the data and the organizations [50] and the scientists who have contributed to the ongoing development of this community resource [51].

The authors gratefully acknowledge the following core contributors to N3C: Adam B Wilcox, Adam M Lee, Alexis Graves, Alfred (Jerrod) Anzalone, Amin Manna, Amit Saha, Andrea Zhou, Andrew E Williams, Andrew Southerland, Andrew T Girvin, Anita Walden, Anjali A Sharathkumar, Benjamin Amor, Benjamin Bates, Brian Hendricks, Brijesh Patel, Caleb Alexander, Carolyn Bramante, Cavin Ward-Caviness, Charisse Madlock-Brown, Christine Suver, Christopher Chute, Christopher Dillon, Chunlei Wu, Clare Schmitt, Cliff Takemoto, Dan Housman, Davera Gabriel, David A Eichmann, Diego Mazzotti, Don Brown, Eilis Boudreau, Elaine Hill, Elizabeth Zampino, Emily Carlson Marti, Emily R Pfaff, Evan French, Farrukh M Koraishy, Federico Mariona, Fred Prior, George Sokos, Greg Martin, Harold Lehmann, Heidi Spratt, Hemalkumar Mehta, Hongfang Liu, Hythem Sidky, JW Awori Hayanga, Jami Pincavitch, Jaylyn Clark, Jeremy Richard Harper, Jin Ge, Joel Gagnier, Joel H Saltz, Joel Saltz, Johanna Loomba, John Buse, Jomol Mathew, Joni L Rutter, Justin Guinney, Justin Starren, Karen Crowley, Katie Rebecca Bradwell, Kellie M Walters, Kenneth R Gersing, Kenrick Dwain Cato, Kimberly Murray, Kristin Kostka, Lavance Northington, Lee Allan Pyles, Leonie Misquitta, Lesley Cottrell, Lili Portilla, Mariam Deacy, Mark M Bissell, Marshall Clark, Mary Emmett, Mary Morrison Saltz, Matvey B Palchuk, Melissa A Haendel, Meredith Adams, Meredith Temple-O'Connor, Michael G. Kurilla, Michele Morris, Nabeel Qureshi, Nasia Safdar, Nicole Garbarini, Noha Sharafeldin, Ofer Sadan, Patricia A Francis, Penny Wung

Burgoon, Peter Robinson, Philip RO Payne, Rafael Fuentes, Randeep Jawa, Rebecca Erwin-Cohen, Richard A Moffitt, Richard L Zhu, Rishi Kamaleswaran, Robert Hurley, Robert T Miller, Saiju Pyarajan, Sam G Michael, Samuel Bozzette, Sandeep Mallipattu, Satyanarayana Vedula, Scott Chapman, Shawn T O'Neil, Soko Setoguchi, Stephanie S Hong, Steve Johnson, Tellen D Bennett, Tiffany Callahan, Umit Topaloglu, Usman Sheikh, Valery Gordon, Vignesh Subbian, Warren A Kibbe, Wenndy Hernandez, Will Beasley, Will Cooper, William Hillegass, and Xiaohan Tanner Zhang.

Details of contributions are available at the N3C Data Enclave website [52].

Additional collaborators were Jasvinder A Singh, Alfred Anzalone, Christopher G Chute, Alyssa Columbus, Mary Emmett, and Evguenia Malaia.

The following include data partner institutions, whose data are released or pending:

Available: Advocate Health Care Network—UL1TR002389: The Institute for Translational Medicine); Aurora Health Care Inc—UL1TR002373: Wisconsin Network For Health Research; Boston University Medical Campus—UL1TR001430: Boston University Clinical and Translational Science Institute; Brown University—U54GM115677: Advance Clinical Translational Research; Carilion Clinic—UL1TR003015: Integrated Translational Health Research Institute of Virginia; Case Western Reserve University—UL1TR002548: The Clinical and Translational Science Collaborative of Cleveland; Charleston Area Medical Center—U54GM104942: West Virginia Clinical and Translational Science Institute; Children's Hospital Colorado—UL1TR002535: Colorado Clinical and Translational Sciences Institute; Columbia University Irving Medical Center—UL1TR001873: Irving Institute for Clinical and Translational Research; Dartmouth College—None (Voluntary) Duke University—UL1TR002553: Duke Clinical and Translational Science Institute; George Washington Children's Research Institute—UL1TR001876: Clinical and Translational Science Institute at Children's National; George Washington University—UL1TR001876: Clinical and Translational Science Institute at Children's National; Harvard Medical School—UL1TR002541: Harvard Catalyst; Indiana University School of Medicine—UL1TR002529: Indiana Clinical and Translational Science Institute; Johns Hopkins University—UL1TR003098: Johns Hopkins Institute for Clinical and Translational Research; Louisiana Public Health Institute—None (Voluntary); Loyola Medicine—Loyola University Medical Center; Loyola University Medical Center—UL1TR002389: The Institute for Translational Medicine; Maine Medical Center—U54GM115516: Northern New England Clinical and Translational Research Network; Mary Hitchcock Memorial Hospital and Dartmouth Hitchcock Clinic—None (Voluntary);Massachusetts General Brigham—UL1TR002541: Harvard Catalyst; Mayo Clinic Rochester—UL1TR002377: Mayo Clinic Center for Clinical and Translational Science; Medical University of South Carolina—UL1TR001450: South Carolina Clinical and Translational Research Institute; MITRE Corporation—None (Voluntary); Montefiore Medical Center—UL1TR002556: Institute for Clinical and Translational Research at Einstein and Montefiore; Nemours—U54GM104941: Delaware Clinical and Translational Research, Accelerating Clinical and Translational Research Program; NorthShore University Health System—UL1TR002389: The Institute for Translational Medicine; Northwestern University at Chicago—UL1TR001422: Northwestern University Clinical and Translational Science Institute; Oregon Community Health Information Network—INV-018455: Bill and Melinda Gates Foundation grant to Sage Bionetworks; Oregon Health and Science University—UL1TR002369: Oregon Clinical and Translational Research Institute; Penn State Health Milton S Hershey Medical Center—UL1TR002014: Penn State Clinical and Translational Science Institute; Rush University Medical Center—UL1TR002389: The Institute for Translational Medicine; Rutgers, The State University of New Jersey—UL1TR003017: New Jersey Alliance for Clinical and Translational Science;Stony Brook University—U24TR002306; The Alliance at the University of Puerto Rico, Medical Sciences Campus—U54GM133807: Hispanic Alliance for Clinical and Translational Research (The Alliance); The Ohio State University—UL1TR002733: Center for Clinical and Translational Science; The State University of New York at Buffalo—UL1TR001412: Clinical and Translational Science Institute; The University of Chicago—UL1TR002389: The Institute for Translational Medicine; The University of Iowa—UL1TR002537: Institute for Clinical and Translational Science; The University of Miami Leonard M. Miller School of Medicine—UL1TR002736: University of Miami Clinical and Translational Science Institute; The University of Michigan at Ann Arbor—UL1TR002240: Michigan Institute for Clinical and Health Research; The University of Texas Health Science Center at Houston—UL1TR003167: Center for Clinical and Translational Sciences; The University of Texas Medical Branch at Galveston—UL1TR001439: The Institute for Translational Sciences; The University of Utah—UL1TR002538: Uhealth Center for Clinical and Translational Science; Tufts Medical Center—UL1TR002544: Tufts Clinical and Translational Science Institute; Tulane University—UL1TR003096: Center for Clinical and Translational Science; The Queens Medical Center—None (Voluntary); University Medical Center New Orleans—U54GM104940: Louisiana Clinical and Translational Science Center; University of Alabama at Birmingham—UL1TR003096: Center for Clinical and Translational Science; University of Arkansas for Medical Sciences—UL1TR003107: UAMS Translational Research Institute; University of Cincinnati—UL1TR001425: Center for Clinical and Translational Science and Training; University of Colorado Denver, Anschutz Medical Campus—UL1TR002535: Colorado Clinical and Translational Sciences Institute; University of Illinois at Chicago—UL1TR002003: University of Illinois Center for Clinical and Translational Science; University of Kansas Medical Center—UL1TR002366: Frontiers: University of Kansas Clinical and Translational Science Institute; University of Kentucky—UL1TR001998: UK Center for Clinical and Translational Science; University of Massachusetts Medical School Worcester—UL1TR001453: The UMass Center for Clinical and Translational Science; University Medical Center of Southern Nevada—None (voluntary); University of Minnesota—UL1TR002494: Clinical and Translational Science Institute; University

of Mississippi Medical Center—U54GM115428: Mississippi Center for Clinical and Translational Research; University of Nebraska Medical Center—U54GM115458: Great Plains IDeA-Clinical and Translational Research; University of North Carolina at Chapel Hill—UL1TR002489: North Carolina Translational and Clinical Science Institute; University of Oklahoma Health Sciences Center—U54GM104938: Oklahoma Clinical and Translational Science Institute; University of Pittsburgh—UL1TR001857: The Clinical and Translational Science Institute; University of Pennsylvania—UL1TR001878: Institute for Translational Medicine and Therapeutics; University of Rochester—UL1TR002001: UR Clinical and Translational Science Institute; University of Southern California—UL1TR001855: The Southern California Clinical and Translational Science Institute; University of Vermont—U54GM115516: Northern New England Clinical and Translational Research Network; University of Virginia—UL1TR003015: Integrated Translational Health Research Institute of Virginia; University of Washington—UL1TR002319: Institute of Translational Health Sciences; University of Wisconsin-Madison—UL1TR002373: UW Institute for Clinical and Translational Research; Vanderbilt University Medical Center—UL1TR002243: Vanderbilt Institute for Clinical and Translational Research; Virginia Commonwealth University—UL1TR002649: C. Kenneth and Dianne Wright Center for Clinical and Translational Research; Wake Forest University Health Sciences—UL1TR001420: Wake Forest Clinical and Translational Science Institute; Washington University in St Louis—UL1TR002345: Institute of Clinical and Translational Sciences; Weill Medical College of Cornell University—UL1TR002384: Weill Cornell Medicine Clinical and Translational Science Center; West Virginia University—U54GM104942: West Virginia Clinical and Translational Science Institute.

Submitted: Icahn School of Medicine at Mount Sinai—UL1TR001433: ConduITS Institute for Translational Sciences; The University of Texas Health Science Center at Tyler—UL1TR003167: Center for Clinical and Translational Sciences; University of California, Davis—UL1TR001860: UCDavis Health Clinical and Translational Science Center; University of California, Irvine—UL1TR001414: The UC Irvine Institute for Clinical and Translational Science; University of California, Los Angeles—UL1TR001881: UCLA Clinical Translational Science Institute; University of California, San Diego—UL1TR001442: Altman Clinical and Translational Research Institute; University of California, San Francisco—UL1TR001872: UCSF Clinical and Translational Science Institute.

Pending: Arkansas Children's Hospital—UL1TR003107: University of Arkansas for Medical Sciences Translational Research Institute; Baylor College of Medicine—None (Voluntary); Children's Hospital of Philadelphia—UL1TR001878: Institute for Translational Medicine and Therapeutics; Cincinnati Children's Hospital Medical Center—UL1TR001425: Center for Clinical and Translational Science and Training; Emory University—UL1TR002378: Georgia Clinical and Translational Science Alliance; HonorHealth—None (Voluntary); Loyola University Chicago—UL1TR002389: The Institute for Translational Medicine; Medical College of Wisconsin—UL1TR001436: Clinical and Translational Science Institute of Southeast Wisconsin; MedStar Health Research Institute—None (Voluntary); Georgetown University—UL1TR001409: The Georgetown-Howard Universities Center for Clinical and Translational Science; MetroHealth—None (Voluntary); Montana State University—U54GM115371: American Indian/Alaska Native CTR; New York University Langone Medical Center—UL1TR001445: Langone Health's Clinical and Translational Science Institute; Ochsner Medical Center—U54GM104940: Louisiana Clinical and Translational Science Center; Regenstrief Institute—UL1TR002529: Indiana Clinical and Translational Science Institute; Sanford Research—None (Voluntary); Stanford University—UL1TR003142: Spectrum: The Stanford Center for Clinical and Translational Research and Education; The Rockefeller University—UL1TR001866: Center for Clinical and Translational Science; The Scripps Research Institute—UL1TR002550: Scripps Research Translational Institute; University of Florida—UL1TR001427: University of Florida Clinical and Translational Science Institute; University of New Mexico Health Sciences Center—UL1TR001449: University of New Mexico Clinical and Translational Science Center; University of Texas Health Science Center at San Antonio—UL1TR002645: Institute for Integration of Medicine and Science; Yale New Haven Hospital—UL1TR001863: Yale Center for Clinical Investigation

The overall work was supported by the National Institutes of Health (NIH; National Institute of Mental Health [NIMH]) R01131542; PI RCP and the National Institute of Diabetes and Digestive and Kidney Diseases Office of the Director; KJW). The funding sources or study sponsors had no role in study design; in the collection, analysis, and interpretation of data; in the writing of the report; and in the decision to submit the paper for publication.

The N3C Publication committee confirmed that this manuscript (MSID:2139.561) is in accordance with N3C data use and attribution policies; however, this content is solely the responsibility of the authors and does not necessarily represent the official views of the NIH or the N3C program.

The N3C data transfer to NCATS is performed under a Johns Hopkins University Reliance Protocol (IRB00249128) or individual site agreements with NIH. The N3C Data Enclave is managed under the authority of the NIH; information can be found at the NCATS website [53].

Data Availability

The National Clinical Cohort Collaborative (N3C) data transfer to the National Center for Advancing Translational Sciences is performed under a Johns Hopkins University Reliance Protocol (IRB00249128) or individual site agreements with the National Institutes of Health (NIH). The N3C Data Enclave is managed under the authority of the NIH; information can be found [54]. N3C Enclave data are protected and can be accessed for COVID-19–related research with an NIH-approved (1) IRB protocol

and (2) institutional Data Use Request (DUR ID: RP-CA3365). A detailed accounting of data protections and access tiers is found at the NCATS website [54]. N3C Enclave and data access instructions can be found at the N3C website [55]; all codes used to produce the analyses in this manuscript are available within the N3C Enclave to users with valid log-in credentials to support reproducibility.

Authors' Contributions

EH and CDV take responsibility for the whole manuscript. RCP, EH, CDV, VM, and ALO were responsible for conceptualization. RCP, CDV, EH, KJW, and AJA were responsible for methodology. RCP, CDV, EH, AJA, VM, and ALO were responsible for phenotype development. RCP, CDV, EH, NF, JS, LEJ, and DV were responsible for phenotype validation. EH was responsible for software validation and data curation. Formal analysis was conducted by EH, KJW, DL, and SES. Investigation was performed by RCP, CDV, and EH. RCP and MAH were responsible for resources. CDV, EH, and RCP were responsible for the original manuscript draft preparation. AJA, VM, ALO, JS, DV, NF, JYI, LEJ, KJW, ZB-D, DL, SES, JAM, PM, TW, SAH, and MAH were responsible for reviewing and editing the manuscript. EH and JAM were responsible for visualization. RCP was responsible for supervision. RCP, SAH, TW, and PM were involved in project administration. RCP and MAH were responsible for funding acquisition.

Conflicts of Interest

None declared.

Multimedia Appendix 1

Supplementary methods for applying confidence levels to people living with HIV, pre-exposure prophylaxis users, the clinician annotation activity, demographic and comorbidity definitions, and COVID-19 outcomes definitions.
[DOC File , 1236 KB-Multimedia Appendix 1]

Multimedia Appendix 2

Performance of our clinician annotation activity in people living with HIV, pre-exposure prophylaxis, postexposure prophylaxis, and people not living with HIV cohorts.
[DOC File , 78 KB-Multimedia Appendix 2]

References

1. Fast facts: HIV in the United States. Centers for Disease Control and Prevention. Apr 22, 2024. URL: https://www.cdc.gov/hiv/data-research/facts-stats/index.html [accessed 2025-05-28]
2. Sun J, Patel RC, Zheng Q, Madhira V, Olex AL, Islam JY, et al. COVID-19 disease severity among people with HIV infection or solid organ transplant in the United States: a nationally-representative, multicenter, observational cohort study. medRxiv. Preprint posted online on July 28, 2021. [FREE Full text] [doi: 10.1101/2021.07.26.21261028] [Medline: 34341798]
3. Kouhpayeh H, Ansari H. HIV infection and increased risk of COVID-19 mortality: a meta-analysis. Eur J Transl Myol. Dec 21, 2021;31(4):10107. [FREE Full text] [doi: 10.4081/ejtm.2021.10107] [Medline: 34962366]
4. Tesoriero JM, Swain CA, Pierce JL, Zamboni L, Wu M, Holtgrave DR, et al. COVID-19 outcomes among persons living with or without diagnosed HIV infection in New York State. JAMA Netw Open. Feb 01, 2021;4(2):e2037069. [FREE Full text] [doi: 10.1001/jamanetworkopen.2020.37069] [Medline: 33533933]
5. Hoover KW, Zhu W, Gant ZC, Delaney KP, Wiener J, Carnes N, et al. HIV services and outcomes during the COVID-19 pandemic - United States, 2019-2021. MMWR Morb Mortal Wkly Rep. Dec 02, 2022;71(48):1505-1510. [FREE Full text] [doi: 10.15585/mmwr.mm7148a1] [Medline: 36454696]
6. Menza TW, Zlot A, Gonzalez-Pena Y, Capizzi J, Bush L, Humphrey S, et al. The ongoing impact of COVID-19 on testing for and diagnoses of HIV and bacterial sexually transmitted infections in Oregon. Sex Transm Dis. Aug 01, 2023;50(8):543-549. [FREE Full text] [doi: 10.1097/OLQ.0000000000001817] [Medline: 37074311]
7. Moitra E, Tao J, Olsen J, Shearer RD, Wood BR, Busch AM, et al. Impact of the COVID-19 pandemic on HIV testing rates across four geographically diverse urban centres in the United States: an observational study. Lancet Reg Health Am. Mar 2022;7:100159. [FREE Full text] [doi: 10.1016/j.lana.2021.100159] [Medline: 34961858]
8. Meyer D, Slone SE, Ogungbe O, Duroseau B, Farley J. Impact of the COVID-19 pandemic on HIV healthcare service engagement, treatment adherence, and viral suppression in the United States: a systematic literature review. AIDS Behav. Jan 2023;27(1):344-357. [FREE Full text] [doi: 10.1007/s10461-022-03771-w] [Medline: 35916951]
9. Tao L, Hojilla J, Shvachko V, Yang J, Carter CC, Das M. Increase in new HIV diagnoses following decrease in use of pre-exposure prophylaxis (PrEP) during the COVID-19 pandemic. Open Forum Infect Dis. 2023;10(Supplement_2):ofad500.1391. [FREE Full text] [doi: 10.1093/ofid/ofad500.1391]

10. Fallahi M, Guadamuz JS, Shooshtari A, Qato DM. Changes in HIV pre-exposure prophylaxis (PrEP) coverage at state and county level during the COVID-19 pandemic in the United States. AIDS Behav. Mar 2024;28(3):799-804. [FREE Full text] [doi: 10.1007/s10461-023-04180-3] [Medline: 37751110]
11. Millett G. New pathogen, same disparities: why COVID-19 and HIV remain prevalent in U.S. communities of colour and implications for ending the HIV epidemic. J Int AIDS Soc. Nov 2020;23(11):e25639. [FREE Full text] [doi: 10.1002/jia2.25639] [Medline: 33222424]
12. Beltran RM, Holloway IW, Hong C, Miyashita A, Cordero L, Wu E, et al. Social determinants of disease: HIV and COVID-19 experiences. Curr HIV/AIDS Rep. Feb 2022;19(1):101-112. [FREE Full text] [doi: 10.1007/s11904-021-00595-6] [Medline: 35107810]
13. Islam J, Madhira V, Sun J, Olex A, Franceschini N, Kirk G, et al. Racial disparities in COVID-19 test positivity among people living with HIV in the United States. Int J STD AIDS. Apr 2022;33(5):462-466. [FREE Full text] [doi: 10.1177/09564624221074468] [Medline: 35306931]
14. Vaidya D, Wilkins KJ, Hurwitz E, Islam JY, Li D, Sun J, et al. Assessing associations between individual-level social determinants of health and COVID-19 hospitalizations: investigating racial/ethnic disparities among people living with human immunodeficiency virus (HIV) in the U.S. National COVID Cohort Collaborative (N3C). J Clin Transl Sci. 2024;8(1):e107. [FREE Full text] [doi: 10.1017/cts.2024.550] [Medline: 39296577]
15. Islam JY, Hurwitz E, Li D, Camacho-Rivera M, Sun J, Safo S, et al. Associations of county-level social determinants of health with COVID-19 related hospitalization among people with HIV: a retrospective analysis of the U.S. National COVID Cohort Collaborative (N3C). AIDS Behav. Oct 2024;28(Suppl 1):136-148. [doi: 10.1007/s10461-024-04466-0] [Medline: 39292319]
16. National COVID cohort collaborative. GitHub. URL: https://github.com/National-COVID-Cohort-Collaborative [accessed 2024-05-23]
17. Phuong J, Hong S, Palchuk MB, Espinoza J, Meeker D, Dorr DA, et al. Advancing interoperability of patient-level social determinants of health data to support COVID-19 research. AMIA Jt Summits Transl Sci Proc. 2022;2022:396-405. [FREE Full text] [Medline: 35854720]
18. Anzalone AJ, Horswell R, Hendricks BM, Chu S, Hillegass WB, Beasley WH, et al. Higher hospitalization and mortality rates among SARS-CoV-2-infected persons in rural America. J Rural Health. Jan 2023;39(1):39-54. [FREE Full text] [doi: 10.1111/jrh.12689] [Medline: 35758856]
19. Yoo YJ, Wilkins KJ, Alakwaa F, Liu F, Torre-Healy LA, Krichevsky S, et al. Geographic and temporal trends in COVID-associated acute kidney injury in the National COVID Cohort Collaborative. Clin J Am Soc Nephrol. Aug 01, 2023;18(8):1006-1018. [FREE Full text] [doi: 10.2215/CJN.0000000000000192] [Medline: 37131278]
20. O'Malley KJ, Cook KF, Price MD, Wildes K, Hurdle JF, Ashton CM. Measuring diagnoses: ICD code accuracy. Health Serv Res. Oct 2005;40(5 Pt 2):1620-1639. [FREE Full text] [doi: 10.1111/j.1475-6773.2005.00444.x] [Medline: 16178999]
21. van Walraven C, Bennett C, Forster AJ. Administrative database research infrequently used validated diagnostic or procedural codes. J Clin Epidemiol. Oct 2011;64(10):1054-1059. [FREE Full text] [doi: 10.1016/j.jclinepi.2011.01.001] [Medline: 21474278]
22. Pocobelli G, Oliver M, Albertson-Junkans L, Gundersen G, Kamineni A. Validation of human immunodeficiency virus diagnosis codes among women enrollees of a U.S. health plan. BMC Health Serv Res. Feb 22, 2024;24(1):234. [FREE Full text] [doi: 10.1186/s12913-024-10685-x] [Medline: 38389066]
23. Yang X, Zhang J, Cai R, Liang C, Olatosi B, Weissman S, et al. Computational phenotyping with the All of Us Research Program: identifying underrepresented people with HIV or at risk of HIV. JAMIA Open. Oct 2023;6(3):ooad071. [FREE Full text] [doi: 10.1093/jamiaopen/ooad071] [Medline: 37614566]
24. May SB, Giordano TP, Gottlieb A. A phenotyping algorithm to identify people with HIV in electronic health record data (HIV-Phen): development and evaluation study. JMIR Form Res. Nov 25, 2021;5(11):e28620. [FREE Full text] [doi: 10.2196/28620] [Medline: 34842532]
25. Fultz SL, Skanderson M, Mole LA, Gandhi N, Bryant K, Crystal S, et al. Development and verification of a "virtual" cohort using the National VA Health Information System. Med Care. Aug 2006;44(8 Suppl 2):S25-S30. [FREE Full text] [doi: 10.1097/01.mlr.0000223670.00890.74] [Medline: 16849965]
26. Paul DW, Neely NB, Clement M, Riley I, Al-Hegelan M, Phelan M, et al. Development and validation of an electronic medical record (EMR)-based computed phenotype of HIV-1 infection. J Am Med Inform Assoc. Feb 01, 2018;25(2):150-157. [FREE Full text] [doi: 10.1093/jamia/ocx061] [Medline: 28645207]
27. National-Clinical-Cohort-Collaborative / CS-ISC. GitHub. URL: https://github.com/National-COVID-Cohort-Collaborative/CS-ISC/tree/main/HIV-Phenotyping [accessed 2024-09-27]
28. RxNorm. National Institutes of Health National Library of Medicine. URL: https://www.nlm.nih.gov/research/umls/rxnorm/index.html [accessed 2025-06-06]
29. OHDSI / Atlas. GitHub. URL: https://github.com/OHDSI/Atlas [accessed 2025-06-06]
30. ATLAS homepage. ATLAS. URL: https://atlas-demo.ohdsi.org/#/home [accessed 2025-06-06]

31. Jones S, Bradwell KR, Chan LE, Olson-Chen C, Tarleton J, Wilkins KJ, et al. Who is pregnant? Defining real-world data-based pregnancy episodes in the National COVID Cohort Collaborative (N3C). medRxiv. Preprint posted online on August 6, 2022. [FREE Full text] [doi: 10.1101/2022.08.04.22278439] [Medline: 35982668]
32. Binswanger IA, Narwaney KJ, Gardner EM, Gabella BA, Calcaterra SL, Glanz JM. Development and evaluation of a standardized research definition for opioid overdose outcomes. Subst Abus. 2019;40(1):71-79. [FREE Full text] [doi: 10.1080/08897077.2018.1546263] [Medline: 30875477]
33. Adult BMI categories. Centers for Disease Control and Prevention. URL: https://www.cdc.gov/bmi/adult-calculator/bmi-categories.html [accessed 2025-06-05]
34. National-Clinical-Cohort-Collaborative / Logic-Liaison-Confirmed-COVID-Positive-Template. GitHub. URL: https://github.com/National-Clinical-Cohort-Collaborative/Logic-Liaison-Confirmed-COVID-Positive-Template/blob/main/pyspark%20version/README%20for%20Logic%20Liaison%20Confirmed%20Covid%20Positive%20Template.pdf [accessed 2024-09-20]
35. Haendel MA, Chute CG, Bennett TD, Eichmann DA, Guinney J, Kibbe WA, et al. The National COVID Cohort Collaborative (N3C): rationale, design, infrastructure, and deployment. J Am Med Inform Assoc. Mar 01, 2021;28(3):427-443. [FREE Full text] [doi: 10.1093/jamia/ocaa196] [Medline: 32805036]
36. Bennett TD, Moffitt RA, Hajagos JG, Amor B, Anand A, Bissell MM, et al. Clinical characterization and prediction of clinical severity of SARS-CoV-2 infection among US adults using data from the US National COVID Cohort Collaborative. JAMA Netw Open. Jul 01, 2021;4(7):e2116901. [FREE Full text] [doi: 10.1001/jamanetworkopen.2021.16901] [Medline: 34255046]
37. N3C COVID enclave forms and resources. National Institutes of Health National Center for Advancing Translational Sciences. URL: https://ncats.nih.gov/research/research-activities/n3c/covid-enclave/resources [accessed 2025-05-12]
38. Nkodo EN, Maheria P, Hurwitz EG, Anzalone J, Li D, Islam J, et al. 1097 – Disparities in COVID-19 therapeutics access among people with and without HIV: an N3C analysis. In: Proceedings of the Conference on Retroviruses and Opportunistic Infections. 2025. Presented at: CROI 2025; March 9-12, 2025; San Francisco, CA. URL: https://www.croiconference.org/abstract/2826-2025/
39. Vaidya D, Wilkins K, Hurwitz EG, Butzin-Dozier Z, Hill E, Maheria P, et al. 1100 – Differential impact of social determinants of health by sex in HIV for COVID-19 hospitalizations. In: Proceedings of the Conference on Retroviruses and Opportunistic Infections. 2025. Presented at: CROI 2025; March 9-12, 2025; San Francisco, CA. URL: https://www.croiconference.org/abstract/666-2025/
40. Yang H, Anzalone J, Hurwitz EG, Nosyk B, Patel RC, Xiao Z. 1120 – Racial/ethnic and regional disparities in HIV testing before, during, and after COVID-19 in the US. In: Proceedings of the Conference on Retroviruses and Opportunistic Infections. 2025. Presented at: CROI 2025; March 9-12, 2025; San Francisco, CA. URL: https://www.croiconference.org/abstract/1683-2025/
41. Kunz M, Rott KW, Hurwitz E, Kunisaki K, Sun J, Wilkins KJ, et al. The intersections of COVID-19, HIV, and race/ethnicity: machine learning methods to identify and model risk factors for severe COVID-19 in a large U.S. national dataset. AIDS Behav. Oct 2024;28(Suppl 1):5-21. [FREE Full text] [doi: 10.1007/s10461-024-04266-6] [Medline: 38326668]
42. Gallant J, Hsue PY, Shreay S, Meyer N. Comorbidities among US patients with prevalent HIV infection-a trend analysis. J Infect Dis. Dec 19, 2017;216(12):1525-1533. [FREE Full text] [doi: 10.1093/infdis/jix518] [Medline: 29253205]
43. Coburn SB, Humes E, Lang R, Stewart C, Hogan BC, Gebo KA, et al. COVID-19 infections post-vaccination by HIV status in the United States. medRxiv. Preprint posted online on December 05, 2021. [FREE Full text] [doi: 10.1101/2021.12.02.21267182] [Medline: 34909791]
44. Coburn SB, Humes E, Lang R, Stewart C, Hogan BC, Gebo KA, et al. Analysis of postvaccination breakthrough COVID-19 infections among adults with HIV in the United States. JAMA Netw Open. Jun 01, 2022;5(6):e2215934. [FREE Full text] [doi: 10.1001/jamanetworkopen.2022.15934] [Medline: 35671054]
45. Jefferson C, Watson E, Certa JM, Gordon KS, Park LS, D'Souza G, et al. Differences in COVID-19 testing and adverse outcomes by race, ethnicity, sex, and health system setting in a large diverse US cohort. PLoS One. 2022;17(11):e0276742. [FREE Full text] [doi: 10.1371/journal.pone.0276742] [Medline: 36417366]
46. Park LS, McGinnis KA, Gordon KS, Justice AC, Leyden W, Silverberg MJ, et al. SARS-CoV-2 testing and positivity among persons with and without HIV in 6 US cohorts. J Acquir Immune Defic Syndr. Jul 01, 2022;90(3):249-255. [FREE Full text] [doi: 10.1097/QAI.0000000000002943] [Medline: 35195574]
47. Antoniou T, Zagorski B, Loutfy MR, Strike C, Glazier RH. Validation of case-finding algorithms derived from administrative data for identifying adults living with human immunodeficiency virus infection. PLoS One. 2011;6(6):e21748. [FREE Full text] [doi: 10.1371/journal.pone.0021748] [Medline: 21738786]
48. Goetz MB, Hoang T, Kan VL, Rimland D, Rodriguez-Barradas M. Development and validation of an algorithm to identify patients newly diagnosed with HIV infection from electronic health records. AIDS Res Hum Retroviruses. Jul 2014;30(7):626-633. [FREE Full text] [doi: 10.1089/aid.2013.0287]
49. National COVID Cohort Collaborative homepage. National COVID Cohort Collaborative. URL: https://covid.cd2h.org [accessed 2025-06-06]

50. Data transfer agreement signatories. National Institutes of Health National Center for Advancing Translational Sciences. URL: https://ncats.nih.gov/n3c/resources/data-contribution/data-transfer-agreement-signatories [accessed 2025-06-06]
51. Haendel MA, Chute CG, Bennett TD, Eichmann DA, Guinney J, Kibbe WA, et al. The National COVID Cohort Collaborative (N3C): rationale, design, infrastructure, and deployment. J Am Med Inform Assoc. Mar 01, 2021;28(3):427-443. [FREE Full text] [doi: 10.1093/jamia/ocaa196] [Medline: 32805036]
52. N3C core contributors. National COVID Cohort Collaborative. URL: https://covid.cd2h.org/contributors/ [accessed 2025-06-06]
53. N3C COVID enclave forms and resources. National Institutes of Health National Center for Advancing Translational Sciences. URL: https://ncats.nih.gov/research/research-activities/n3c/covid-enclave/resources [accessed 2025-06-06]
54. N3C data access forms and resources. National Institutes of Health National Center for Advancing Translational Sciences. URL: https://ncats.nih.gov/n3c/resources/data-access [accessed 2025-06-06]
55. Researcher essentials. National COVID Cohort Collaborative. URL: https://covid.cd2h.org/for-researchers/ [accessed 2025-06-06]

Abbreviations

CD4: cluster of differentiation 4
EHR: electronic health record
HBV: hepatitis B virus
ICD-10: International Classification of Diseases, Tenth Revision
IRB: Institutional Review Board
N3C: National Clinical Cohort Collaborative
NIH: National Institutes of Health
OMOP: Observational Medical Outcomes Partnership
PEP: postexposure prophylaxis
PrEP: preexposure prophylaxis
SNOMED CT: Systemized Medical Nomenclature for Medical Clinical Terminology
TDF: tenofovir disoproxil fumarate
VL: viral load

Edited by A Coristine; submitted 29.10.24; peer-reviewed by Y Liu, P Ebrahimi; comments to author 14.04.25; revised version received 13.05.25; accepted 15.05.25; published 11.07.25

Please cite as:
Hurwitz E, Varley CD, Anzolone AJ, Madhira V, Olex AL, Sun J, Vaidya D, Fadul N, Islam JY, Jackson LE, Wilkins KJ, Butzin-Dozier Z, Li D, Safo SE, McMurry JA, Maheria P, Williams T, Hassan SA, Haendel MA, Patel RC, The National Clinical Cohort Collaborative (N3C) Consortium
Identifying People Living With or Those at Risk for HIV in a Nationally Sampled Electronic Health Record Repository Called the National Clinical Cohort Collaborative: Computational Phenotyping Study
JMIR Med Inform 2025;13:e68143
URL: https://medinform.jmir.org/2025/1/e68143
doi: 10.2196/68143
PMID: