HHS Public Access

Author manuscript

Methods Inf Med. Author manuscript; available in PMC 2022 September 30.

Published in final edited form as:

Methods Inf Med. 2021 September ; 60(3-04): 84–94. doi:10.1055/s-0041-1735619.

Optimizing Identification of People Living with HIV from Electronic Medical Records: Computable Phenotype Development and Validation

Yiyang Liu[1], **Khairul A. Siddiqi**[2], **Robert L. Cook**[1], **Jiang Bian**[2], **Patrick J. Squires**[3], **Elizabeth A. Shenkman**[2], **Mattia Prosperi**[1], **Dushyantha T. Jayaweera**[4]

[1]Department of Epidemiology, College of Public Health and Health Professions and College of Medicine, University of Florida, Gainesville, Florida, United States

[2]Department of Health Outcomes and Biomedical Informatics, College of Medicine, University of Florida, Gainesville, Florida, United States

[3]Department of Pharmaceutical Outcomes and Policy, College of Pharmacy, University of Florida, Gainesville, Florida, United States

[4]Department of Medicine, Miller School of Medicine, University of Miami, Miami, Florida, United States

Abstract

Background—Electronic health record (EHR)-based computable phenotype algorithms allow researchers to efficiently identify a large virtual cohort of Human Immunodeficiency Virus (HIV) patients. Built upon existing algorithms, we refined, improved, and validated an HIV phenotype algorithm using data from the OneFlorida Data Trust, a repository of linked claims data and EHRs from its clinical partners, which provide care to over 15 million patients across all 67 counties in Florida.

Methods—Our computable phenotype examined information from multiple EHR domains, including clinical encounters with diagnoses, prescription medications, and laboratory tests. To identify an HIV case, the algorithm requires the patient to have at least one diagnostic code for HIV and meet one of the following criteria: have 1+ positive HIV laboratory, have been prescribed with HIV medications, or have 3+ visits with HIV diagnostic codes. The computable phenotype was validated against a subset of clinical notes.

Results—Among the 15+ million patients from OneFlorida, we identified 61,313 patients with confirmed HIV diagnosis. Among them, 8.05% met all four inclusion criteria, 69.7% met the 3+ HIV encounters criteria in addition to having HIV diagnostic code, and 8.1% met all criteria

Address for correspondence Yiyang Liu, PhD, MPH, Department of Epidemiology, College of Public Health and Health Professions & College of Medicine, University of Florida, 2004 Mowry Road, PO Box 100231, Gainesville 32610-0231, Florida, United States, (yliu26@ufl.edu).

Conflict of Interest

The authors do not report any conflicts of interests. The content is solely the responsibility of the authors and does not necessarily represent the official views of the Patient-Centered Outcomes Research Institute (PCORI), its Board of Governors or Methodology, the OneFlorida Clinical Research Consortium, the University of Florida's Clinical and Translational Science Institute, the Florida Department of Health, or the National Institutes of Health.

except for having positive laboratories. Our algorithm achieved higher sensitivity (98.9%) and comparable specificity (97.6%) relative to existing algorithms (77–83% sensitivity, 86–100% specificity). The mean age of the sample was 42.7 years, 58% male, and about half were Black African American. Patients' average follow-up period (the time between the first and last encounter in the EHRs) was approximately 4.6 years. The median number of all encounters and HIV-related encounters were 79 and 21, respectively.

Conclusion—By leveraging EHR data from multiple clinical partners and domains, with a considerably diverse population, our algorithm allows more flexible criteria for identifying patients with incomplete laboratory test results and medication prescribing history compared with prior studies.

Keywords

HIV; computable phenotype; electronic health records; virtual cohort; diagnosis

Introduction

In the United States (U.S.), the goals of the National Human Immunodeficiency Virus and Acquired Immunodeficiency Syndrome (HIV/AIDS) Strategy 2020 are to diagnose 90% of people with HIV (PWH), link 90% of diagnosed people to care and achieve effective viral suppression for 90% of diagnosed PLW.[1] Additionally, the Ending of HIV Epidemic: A Plan for America (EHE) initiative aims to reduce HIV incidence by at least 90% by 2030 through focusing on four pillars of diagnoses, treat, prevent, and respond.[2] The latest estimates indicated that almost two-thirds of PWH (65%) received any care, and only 56% were virally suppressed in 2018.[3] In addition to timely diagnosis, it is crucial to identify PWH who are not retained in care and/or not suppressed to design health intervention strategies at the individual and population level.

The widespread adoption of electronic health record (EHR) systems and the creation of clinical research networks with extensive collections of EHR data made it possible to develop and validate computable phenotype (CP) algorithms that search and classify patients with specific phenotypes (e.g., living with HIV) using clinical documentation, billing, laboratory data, and other EHR-related data sources. EHR-based CP algorithms to detect HIV cases can help researchers efficiently identify a large virtual cohort of PWH with longitudinal trajectories of engagement through routine care both before and after their HIV diagnosis. Further, secondary analyses of data from the identified EHR cohort can provide real-world evidence[4] to assess gaps in the HIV care continuum, comorbidity burdens, patterns in utilization of services, and quality of care.

HIV/AIDS case detection algorithms have been developed in previous literature using Medicare and/or Medicaid claims data[5–9] or Veterans Health Administration (VHA) data.[10,11] These algorithms heavily relied on diagnostic codings, such as the use of International Classification of Disease (ICD) codes. However, the diagnosis coding is designed for administrative/billing purposes instead of clinical or research use.[12] A single record of a disease diagnostic code may not accurately reflect the actual disease that the patient had.[13] To overcome this shortcoming, some algorithms classify HIV cases only when

patients have multiple claims/encounters with a corresponding HIV diagnosis (e.g., at least one inpatient or two outpatient claims).[5,9,11]

EHR-based algorithms for HIV/AIDS have improved upon previous claims-based algorithms mainly by addressing the inherent methodologic limitations of claims data via incorporating the additional rich information available in EHRs. The algorithm developed by Felsen et al[14] was based on the EHR from one medical center in New York City, and it incorporated information on HIV-related ICD (ver. 9) diagnostic codes, encounters, and laboratories (including western blot, antibody test, viral load, and CD4 count). Paul et al[15] designed and validated two algorithms to identify HIV patients using the EHR from the Duke University Health System. One of their algorithms was laboratory-based and detected an HIV case if the patient met one of the following two criteria: (1) having a positive HIV antibody test confirmed with a Western blot, nucleic acid test, or another positive HIV antigen test, (2) having been prescribed HIV-specific medications. Their second algorithm was ICD-9-based and classified patients as HIV cases when any HIV-related ICD-9 code was accompanied by confirmatory medication or laboratory results consistent with HIV. This second algorithm was designed to better deal with the real-world data with incomplete medical records while still relying heavily on laboratory testing results and prescription history for case detection. However, it should be acknowledged that both algorithms still possess limitations. Felsen et al's algorithm did not incorporate HIV medication information, while the algorithms developed by Paul et al did not consider encounter patterns. Moreover, both studies were based on ICD-9 codes and developed based on EHR from one health care system, and therefore are not very generalizable.

Florida is the topmost contributor to the United States' national cases of HIV, with an estimated total of 116,689 PWH in 2019 and more than 4,500 new HIV infections every year.[3,16] Leveraging local EHR data can help better identify gaps in the HIV care continuum to develop the targeted intervention. As one of the nine clinical data research networks (CDRNs) of the National Patient-Centered Clinical Research Network (PCORnet), the OneFlorida Clinical Research Consortium operates a statewide clinical data warehouse. It partners with 11 clinical health care systems across Florida, and all EHRs were mapped to the PCORnet common data model.[17] Since patient-generated EHR data are generally customized and inconsistent across health care systems,[18,19] calibrating the EHR-based algorithm using OneFlorida EHR data are essential to detect and characterize the HIV cohort in Florida adequately. This study takes a comprehensive approach to construct and validate an algorithm that will improve previous published EHR-based phenotype algorithms by leveraging data from more domains of the PCORnet common data model. Our algorithm allows more flexible inclusion and exclusion criteria for patients with missing or incomplete laboratory test results and medication prescribing history. Additionally, this algorithm is compatible with newly adopted ICD-10 and Systemized Nomenclature of Medicine (SNOMED) diagnostic codes and can be replicable to other PCORnet Network Partners.

Materials and Methods

Study Population and Ethics Statement

We abide by the Declaration of Helsinki, and the study protocol was approved by the University of Florida's (UF) Institutional Review Board (IRB). All Protected health information was de-identified using the Health Information Portability and Accountability Act (HIPAA) Safe Harbor method.[20]

We extracted data from the OneFlorida Data Trust, a repository of claims data and EHRs from its clinical partners following the national PCORnet Common Data Model (CDM).[17,21] OneFlorida partners provide health care to over 15 million patients (>60% of Floridians) across all 67 counties in Florida (Fig. 1). Data from OneFlorida goes through rigorous quality checks at its data coordinating center at UF, including a process to de-duplicate the same patients from multiple health care using a privacy-preserving record linkage method.[22]

We analyzed OneFlorida data between 2012 and 2020, including the following PCORnet common data model tables: DEMOGRAPHIC, ENCOUNTER, DIAGNOSIS, PROCEDURES, LAB_RESULT_CM, CONDITION, and PRESCRIBING.

Computable Phenotype Algorithm

Our CP algorithm was developed based on the strengths of previously validated algorithms[15,23] and expanded to incorporate information from different domains. The algorithm was iteratively refined through several rounds of revisions based on findings and expert feedback. The final algorithm involves four steps, as visualized in Fig. 2.

Step 1 (Diagnosis and Condition): Our algorithm first screened both diagnosis and condition tables to identify all patients with at least one ICD-9, ICD-10, or SNOMED code for HIV (Supplementary Table S1, available in the online version only). When selecting these codes, only codes corresponding to a confirmed HIV positive status were included. Codes that are HIV related but not corresponding to confirmed HIV status were excluded, for example, codes used for HIV counseling (ICD-9 code V65.44, IC-10 Z71.7), inconclusive HIV laboratory evidence (ICD-9 code 795.71, ICD-10 code R75), and exposure to HIV (ICD-9 code V01.79, ICD-10 code Z20.6).

Step 2 (Laboratory): Among people who had at least one confirmatory HIV diagnostic code, laboratory results were screened (the corresponding criteria are summarized in Supplementary Table S2, available in the online version only). All HIV-related laboratory Logical Observation Identifiers Names and Codes (LOINC), raw and cleaned laboratory results, and raw laboratory names were manually reviewed to ensure (1) the algorithm catches all HIV laboratory tests, (2) missing data and invalid results are coded correctly, and (3) appropriate thresholds are selected that can be uniformly applied to different tests. CD4 count and CD4-CD8 ratio laboratories are typically only prescribed to patients with HIV diagnosis to monitor response to HIV medications and guide treatment choices.[14,24] Therefore, in our algorithm, regardless of the laboratory results, the presence of CD4 laboratory prescribing history along with an HIV diagnostic code confirmed an HIV diagnosis. Similarly, HIV genotyping and phenotyping tests are only performed for patients

with known HIV diagnoses to assess drug resistance.[15,25] The presence of these laboratory procedures confirmed an HIV diagnosis. HIV antibody and antigen tests are commonly used for HIV screening and diagnosis. For these tests, only "reactive" results were considered to be positive and considered to confirm an HIV diagnosis, whereas "nonreactive" and "indeterminate" results were treated as negative and cannot be used to confirm a diagnosis. Viral load measurements can both be used to monitor disease progression among HIV patients and used to diagnose new acute HIV infection, and cutoff points were applied to categorize the results into positive or negative for case detection. Viral load ≥20 copies/mL or log viral load ≥1.3 were considered positive and viral load <20 copies/mL or log viral load <1.3 were considered negative and did not confirm a diagnosis.

Step 3 (Prescribing): Patients with all negative HIV laboratory results or with no records for HIV laboratories were further screened for HIV medications from the Prescribing table. Supplementary Table S3 (available in the online version only) lists the generic and brand names for all FDA-approved HIV medications for adults.[26] In the PCORnet CDM, medications are coded in raw medication names and the RxNorm Concept Unique Identifier (CUI). Through the manual review of data, we observed that the raw medication name has better data quality than RxNorm CUI (less missing and fewer misspellings). Therefore, our algorithm applied a text screen to raw medication names to identify all HIV medications. Patients with at least one prescription of HIV medications listed in Supplementary Table S3 (available in the online version only) were considered a confirmed case. This step will capture patients with an undetectable viral load as the result of effective antiretroviral therapy and who were not confirmed as a case in step 2. People who had been prescribed with Emtricitabine/Tenofovir alone or with Dolutegravir or Raltegravir but had no other HIV medication prescriptions were classified as potential pre-exposure prophylaxis (PrEP) or post-exposure prophylaxis users. We would assess the appropriateness of excluding them in the final algorithm through the validation and chart review process.

Step 4 (Encounter): In the last step, based on the Encounter table, patients who have three or more encounters (regardless of inpatient or outpatient) with an HIV diagnostic code (ICD-9 or ICD-10) listed in Supplementary Table S1 (available in the online version only) were considered to have confirmed HIV diagnosis. Considering diagnostic codes for HIV may be inappropriately used for HIV screening, the threshold of three visits was set to exclude patients who came in for one visit to do an HIV screening and came back to another visit to review the results. However, this inclusion criterion may introduce some false negatives for patients who have HIV but have only one or two HIV-related encounters. This may happen when the primary health care system that provides HIV care to the patients was not partnered with OneFlorida, or the patient had low retention in care. A sensitivity analysis was performed by changing the threshold number of encounters from three to two.

To summarize, in our CP algorithm, to be considered as an HIV case, patients needed to have at least one ICD-9, 10, or SNOMED code for HIV diagnosis or condition and met one of the following three inclusion criteria: (1) had at least one positive HIV laboratory, (2) had at least once been prescribed with HIV medications, or (3) had at least three HIV-related encounters.

Validation through Manual Chart Review

Manual review of the electronic records and clinical notes is considered the gold standard for validating the EHR-based CP algorithms.[27] In the validation process, we extracted clinical notes from one health provider in the OneFlorida network—the UF Health—and manually reviewed them to determine if the presence of confirmative evidence supports the patient's HIV status.

A total of 16 mutually exclusive combinations were created corresponding to the different grouping of inclusion and exclusion criteria and cutoff point used among people who had at least one HIV-related element (like an HIV screening procedure) documented in the structured EHR. To estimate the overall performance of the CP, we used a stratified random sampling method to select 150 patients then compared the CP classification with a manual chart review classification. Because some combinations have a large number of patients (44,072 with the combination of "has Dx and Enc, no lab and Rx," 62,428 with the combination of "has Dx only," 423,277 with the combination of "has HIV screening procedure"), we decided to downsample for these combinations to ensure a balanced sample. Twenty patients were selected from the "has Dx and Enc, no lab and Rx" combination; another 20 patients from the "has Dx only" or "has HIV screening procedure" combinations; 110 patients from the rest 13 combinations.

Additionally, to identify potential ways to refine our CP and ensure it is validated in each inclusion combination, we used a proportional sampling strategy to generate an additional focused validation sample. In this sample, at least two patients were selected for each of the 16 combinations. This focused validation sample will allow us to assess the appropriateness of inclusion criteria. After the chart review of the clinical notes, sensitivity, specificity, positive predictive value (PPV), and negative predictive value were calculated. Reasons for discordance between chart review and algorithm classification were examined to refine the algorithm if applicable. Algorithm coding and data analyses were conducted in SAS 9.4 (SAS Institute, Cary, North Carolina, United States).

Results

Around 0.6 billion encounters, 1.2 billion diagnoses, 0.3 billion prescriptions, and 0.5 billion claims were documented in OneFlorida among the 15+ million patients. Approximately 5 million patients had at least one care visit to OneFlorida partner each year. The majority (54.8%) of the sample were females, 27.6% were Hispanic/Latinx, and 22.0% were Black African American. The mean age was 30 years.

Algorithms

As shown in the flow chart in Fig. 2, among all patients seen in OneFlorida between 2012 and 2020, 124,293 met the must-have inclusion criteria in step 1 (at least one ICD-9, 10, or SNOMED code for HIV diagnosis or HIV-related condition). Among them, 11,660 had the HIV diagnosis confirmed by a positive HIV laboratory result. Of people with no laboratory testing records or HIV test results being negative, 6,933 patients had their HIV diagnosis confirmed by at least one HIV medication prescription. Lastly, 42,720 patients were entered

into the cohort in step 4 by having three or more HIV-related encounters in the study period. In total, our algorithm identified 61,313 patients with confirmed HIV diagnosis. The combination patterns of inclusion criteria met by the entire study sample are summarized in Table 1. In the identified HIV cohort, 9.72% met all four inclusion criteria, 69.68% only met the criteria for having HIV diagnostic codes and three or more HIV encounters, 8.05% met all criteria except for having positive laboratories, 6.85% met all criteria except for having HIV medication, 3.25% met having HIV diagnostic codes and medications criteria only, 1.99% met having diagnostic codes and HIV laboratories criteria only, and 0.46% met all criteria except for having three or more encounters.

Algorithm Refinement and Validation against Clinical Notes

The percentages of people correctly classified by our CP among the 16 validation groups are listed in Table 2. In addition to different combinations of the inclusion criteria, we created two targeted validation groups (groups 1 and 2) to assess the appropriateness of some of the criteria used. Group 1 consisted of people who would not have been included as HIV cases if we exclude people who had only been prescribed medications with PrEP indication. Among this group, according to the chart review, six out of eight people had HIV diagnoses. Therefore, in our final algorithm, we did not exclude potential PrEP users. Another validation group represents people who would not be included as cases if we used the ≥50 copies/mL as the cutoff point for case detection viral load laboratories instead of using ≥20 copies/mL. Of the patients selected for this group, one is HIV positive, and another is HIV negative according to the chart review. Given the influence of the choice of the cutoff is likely to be minimal (only 10 people affected), we decided to keep ≥20 copies/mL, the default threshold in the dataset, as the viral load cutoff in the final algorithm.

Comparing the classification of the finalized CP to the gold standard of chart review, the performance of the algorithm is listed in Table 3. After adjusting for the sample weight, the sensitivity (recall) and specificity of our algorithm were found to be 98.5% (95% CI 92.2–100%) and 97.6% (95% CI 95.9–100%), respectively. Additionally, the positive predictive value (precision) and negative predictive value were found to be 80.9% (95% CI 62.0–99.8%) and 99.8% (95% CI 99.2–100%). An additional 15,761 patients were classified as having HIV in a sensitivity analysis when the threshold of encounter changed from three to two. However, the algorithm precision reduced to 45.33% with <1% gain in recall (Supplementary Table S4, available in the online version only).

Characteristics of the OneFlorida HIV Cohort Identified by Our Algorithm

The population characteristics of the OneFlorida HIV cohort identified by our algorithm are summarized in Table 4. The majority of the sample (58.15%) was male, 49.74% of subjects were Black African American, 29.74% were White, and 16.56% were Hispanic/Latinx. The mean age was 42.66 years (SD = 13.39). The median number of all encounters and HIV-related encounters were 79 (interquartile range [IQR] 172) and 21 (IQR 47), respectively. The average follow-up period (estimated as the duration between first and last encounter in the EHR) for the virtual HIV cohort was 1,682.43 days (approximately 4.6 years, SD 987.16 days). Over 40% of the sample entered OneFlorida in 2012, when the network was first

established. Over 12% of the sample entered in 2013; around 8% entered between 2014 and 2017 each year; 6.09%, 4.59%, and 0.17% entered in 2018, 2019, and 2020, respectively.

Discussion

In this study, we developed and validated a CP algorithm for identifying patients with confirmed HIV diagnoses in EHRs, leveraging information from multiple health care providers and EHR domains in the PCORnet common data model. Our study extended the previously published algorithms in several important ways. First, we expanded the diagnostic codes to include ICD-10-CM and SNOMED codes in addition to ICD-9-CM-codes. Second, our algorithm allowed more flexible inclusion and exclusion criteria by leveraging information from multiple EHR domains and less reliance on complete screening test laboratory results relative to the algorithm designed by Paul et al.[15] Third, in addition to considering HIV screen tests, our phenotype considered laboratory procedures commonly prescribed for patients with known HIV status (such as HIV genotyping and CD4 tests). Lastly, our algorithm was developed using EHRs that follow the PCORnet common data model and can be easily replicable to other PCORnet Network health systems with local recalibration.

Our final algorithm achieved higher sensitivity (98.5%) compared with previous algorithms (77–78% in Paul et al,[15] 83% in Goetz et al[23]) with comparable specificity (97.6% in our algorithm, vs. 99–100% in Paul et al,[15] 86% in Goetz et al[23]). The main reason for false negatives was because HIV was documented as past medical history in clinical notes, but no HIV diagnostic codes were used in the structured EHR. Misclassification of the false positives was primarily due to incomplete HIV laboratories and prescribing records. The use of two HIV-related encounter criteria may incorrectly classify a small proportion of people without an HIV diagnosis. For example, a patient may have had two HIV-related ICD codes documented in EHR for routine screening of HIV.

The high performance of EHR-based CPs offers rapid identification of patients from EHRs of different health care systems, which help the development of multicenter patient cohorts for research, clinical care, and public health initiatives.[15] The HIV cases (61,313 patients) identified by our algorithm reflect the number of PWH in Florida who received some care in OneFlorida partners. In 2019, around 80% of 116,689 PWH received an HIV care,[16] and OneFlorida partners provided care for 15 million patients in past years, approximately 70% of the 2019 population.[28] The age and race composition in our HIV cohort were similar to the characteristics of PWH estimated by the State's Health Department, yet males (58.2% in our cohort vs. 72.7% in the state's estimate), and Hispanics (16.6% in our cohort vs. 27.4% in the state's estimate) were less represented in our cohort.[29] Patients included in the OneFlorida dataset generally have access to a major health system network in Florida. PWH who are out of care or seek care outside of the OneFlorida network, like exclusively from county health departments, may not be representative of our cohort. Females are generally more frequent users of the health care system than males[30–33]; this may partially explain the overrepresentation of females relative to the state's estimate. Moreover, research has documented lower access to care among the Hispanic/Latinx population due to lack of insurance, language barriers, and low socioeconomic status.[34–36] Additionally, 9.7% of

ethnicity was unknown in our HIV cohort, which may also underestimate the true proportion of Hispanic/Latinx.

Unlike previous algorithms built solely upon the EHR of one hospital system or single-payer claims data, OneFlorida combines both all-payer claims data and hospital EHRs to capture all potential encounters of the patients, which intensify the identification of a virtual cohort with any conditions. Building on the strength of the OneFlorida network, the HIV virtual cohort identified through our CP can provide real-world evidence that helps better understand the disease burden and patterns in access to care among PWH in Florida and inform targeted intervention to improve their health outcomes. Potential future research like geo-mapping of the identified patients would be possible to examine the disparity burden by county, urbanicity, or other geolocation groups. Additionally, the rich longitudinal EHR data could also be mined to assess comorbidity and mortality burdens, gaps in the HIV care continuum, patterns in utilization of services, and quality of care among HIV patients. Another major strength of our virtual HIV cohort in OneFlorida EHR is that it captures patients' interaction with the health care system both before and after their HIV diagnosis, whereas the other patient cohorts identified from the Ryan White HIV/AIDS Program and state HIV surveillance program, like the Enhanced HIV/AIDS Reporting System, only focused on health services after their HIV diagnosis. Examining health care utilization patterns before HIV diagnosis can help researchers to identify missed opportunities for future HIV prevention efforts. Furthermore, instead of focusing solely on HIV-related outcomes, our EHR-based virtual cohort can be used to examine the full spectrum of comorbidities along with their related engagement in care, laboratory results, and treatment. PWH often experiences comorbidities. Additionally, there is an increasing recognition that HIV may not always be a patient's top priority or chief complaint while interacting with the health care system.[37,38] Secondary data analysis of the identified virtual HIV cohort could provide real-world evidence on their comorbidities and treatment burdens and thus provide insights into future patient-centered HIV care.

Our identified virtual cohort faces some unique challenges in assessing gaps in the HIV care continuum. Many patients in the virtual cohort do not have complete HIV laboratory results. One reason could be that OneFlorida incorporates Medicaid claim, which does not contain laboratory results. These missing laboratory results might limit the data in its potential for monitoring viral suppression over time. Another challenge is that we cannot distinguish individuals who have all of their health care documented within the OneFlorida EHR versus those who do not. As a result, we may underestimate the proportion of patients who were engaged in care as some of them may be misclassified as "not engaged in care" because they seek health care both in and out of the OneFlorida partners.

Our work has some limitations. First, not all sites and partners within the OneFlorida network have complete PCORnet common model elements, which limits the generalizability of our identified HIV cohort and the performance of our CP algorithm. Another limitation involves our validation process as it was limited to one health care system. Ideally, a random validation sample would be selected from the OneFlorida Clinical Research Consortium to extract clinical notes to perform chart reviews. However, we were not able to access medical notes from each of the partner health care systems. Additionally, the validation sample was

drawn based on HIV-related inclusion clusters. This may inflate the prevalence of HIV among the validation sample and lead to a slight overestimation of PPV and underestimation of negative predictive value.

Conclusion

In summary, the CP we developed achieved high sensitivity and specificity in identifying patients with confirmed HIV diagnosis using EHR data. Relative to the published algorithms, our method relies less on complete laboratory results and prescribing data by leveraging the multi-provider, multi-domain EHR. Our algorithm was compatible with the PCORNet Common Data Model and can be translatable to other health care systems using the same Common Data Model. The CP will enable future researchers to identify a large cohort of PWH efficiently. The identified HIV cohort has the potential to improve research and programs in the area of HIV prevention and care and other important health outcomes of PWH in Florida.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

The authors acknowledge the OneFlorida Clinical Research Consortium, the University of Florida Integrated Data Repository (IDR), and the UF Health Office of the Chief Data Officer for providing the analytic data set for this project.

Funding

This work was supported by the National Institute of Allergy and Infectious Diseases (NIAID) under Award Number R01AI145552 (Co-PIs: Salemi, Prosperi) and a pilot grant from the Center for AIDS Research (CFAR) (PI: Jayaweera) from the National Institute of Allergy and Infectious Diseases (NIAID) under Award Number 5P30AI073961-13 (PI: Pahwa). The work was also, in part, funded by CDC U18DP006512 and NCI R01CA246418. Additionally, the research reported in this publication was supported by the National Center for Advancing Translational Sciences of the National Institutes of Health under University of Florida Clinical and Translational Science Awards UL1TR000064 and UL1TR001427. The OneFlorida Clinical Research Consortium was funded by the Patient-Centered Outcomes Research Institute number CDRN-1501-26692 and RICRN-2020-005; in part by the OneFlorida Cancer Control Alliance, funded by the Florida Department of Health's James and Esther King Biomedical Research Program #4KB16.

References

1. Office of National AIDS Policy. National HIV/AIDS Strategy for the United States: Updated to 2020. 2015; Accessed May 5, 2021 at: https://files.hiv.gov/s3fs-public/nhas-update.pdf
2. Fauci AS, Redfield RR, Sigounas G, Weahkee MD, Giroir BP. Ending the HIV epidemic: a plan for the United States. JAMA 2019;321 (09):844–845 [PubMed: 30730529]
3. Centers for Disease Control and Prevention. HIV in the United States by Region. 2021; Accessed May 5, 2021 at: https://www.cdc.gov/hiv/statistics/overview/geographicdistribution.html
4. Food and Drug Administration. Framework for FDA's Real-World Evidence Program. 2018. Accessed September 20, 2021 at: https://www.fda.gov/media/120060/download
5. Fasciano NJ, Cherlow AL, Turner BJ, Thornton CV. Profile of medicare beneficiaries with AIDS: application of an AIDS case finding algorithm. Health Care Financ Rev 1998;19(03): 1–20
6. Thornton C, Fasciano N, Turner BJ, Cherlow A, Bencio DS. Methods for Identifying AIDS Cases in Medicare and Medicaid Claims Data. Health Care Financ Admin Princeton, NJ: Mathematica Policy Research; 1997

7. Keyes M, Andrews R, Mason ML. A methodology for building an AIDS research file using Medicaid claims and administrative data bases. J Acquir Immune Defic Syndr (1988) 1991;4(10): 1015–1024 [PubMed: 1832459]
8. Leibowitz AA, Desmond K. Identifying a sample of HIV-positive beneficiaries from Medicaid claims data and estimating their treatment costs. Am J Public Health 2015;105(03):567–574 [PubMed: 25602870]
9. Walkup JT, Wei W, Sambamoorthi U, Crystal S. Sensitivity of an AIDS case-finding algorithm: who are we missing? Med Care 2004;42(08):756–763 [PubMed: 15258477]
10. McGinnis KA, Fine MJ, Sharma RK, et al. ; Veterans Aging Cohort 3-Site Study (VACS 3) Understanding racial disparities in HIV using data from the veterans aging cohort 3-site study and VA administrative data. Am J Public Health 2003;93(10):1728–1733 [PubMed: 14534229]
11. Fultz SL, Skanderson M, Mole LA, et al. Development and verification of a "virtual" cohort using the National VA Health Information System. Med Care 2006;44(8, suppl 2):S25–S30 [PubMed: 16849965]
12. O'Malley KJ, Cook KF, Price MD, Wildes KR, Hurdle JF, Ashton CM. Measuring diagnoses: ICD code accuracy. Health Serv Res 2005;40 (5 Pt 2):1620–1639 [PubMed: 16178999]
13. Peabody JW, Luck J, Jain S, Bertenthal D, Glassman P. Assessing the accuracy of administrative data in health information systems. Med Care 2004;42(11):1066–1072 [PubMed: 15586833]
14. Felsen UR, Bellin EY, Cunningham CO, Zingman BS. Development of an electronic medical record-based algorithm to identify patients with unknown HIV status. AIDS Care 2014;26(10): 1318–1325 [PubMed: 24779521]
15. Paul DW, Neely NB, Clement M, et al. Development and validation of an electronic medical record (EMR)-based computed phenotype of HIV-1 infection. J Am Med Inform Assoc 2018;25(02): 150–157 [PubMed: 28645207]
16. Florida Department of Health. HIV Data Center: HIV Care Data. 2020; Accessed October 13, 2020 at: http://www.floridahealth.gov/diseases-and-conditions/aids/surveillance/index.html
17. Shenkman E, Hurt M, Hogan W, et al. OneFlorida Clinical Research consortium: linking a clinical and translational science institute with a community-based distributive Medical Education Model. Acad Med 2018;93(03):451–455 [PubMed: 29045273]
18. Botsis T, Hartvigsen G, Chen F, Weng C. Secondary use of EHR: data quality issues and informatics opportunities. Summit On Translat Bioinforma 2010;2010:1–5
19. Ehrenstein V, Kharrazi H, Lehmann H, Taylor CO. Obtaining Data from Electronic Health Records. In: Gliklich RE, Leavy MB, Dreyer NA, eds. Tools and Technologies for Registry Interoperability, Registries for Evaluating Patient Outcomes: A User's Guide, 3rd ed., Addendum 2. Chapter 4 Obtaining Data From Electronic Health Records. Rockville, MD: Agency for Healthcare Research and Quality (US); 2019
20. Benitez K, Malin B. Evaluating re-identification risks with respect to the HIPAA privacy rule. J Am Med Inform Assoc 2010;17(02): 169–177 [PubMed: 20190059]
21. Fleurence RL, Curtis LH, Califf RM, Platt R, Selby JV, Brown JS. Launching PCORnet, a national patient-centered clinical research network. J Am Med Inform Assoc 2014;21(04):578–582 [PubMed: 24821743]
22. Bian J, Loiacono A, Sura A, et al. Implementing a hash-based privacy-preserving record linkage tool in the OneFlorida clinical research network. JAMIA Open 2019;2(04):562–569 [PubMed: 32025654]
23. 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 2014;30(07):626–633 [PubMed: 24564256]
24. Duro R, Rocha-Pereira N, Figueiredo C, et al. Routine CD4 monitoring in HIV patients with viral suppression: is it really necessary? A Portuguese cohort. J Microbiol Immunol Infect 2018;51(05):593–597 [PubMed: 28712820]
25. Ambrosioni J, Mosquera M, Miró JM. Baseline Genotype Testing to Assess Drug Resistance Before Beginning HIV Treatment. JAMA 2018;320(20):2153–2154 [PubMed: 30480719]
26. Food and Drug Administration. HIV Treatment Information for Adults. 2020 Accessed September 25, 2020 at: https://www.fda.gov/drugs/hiv-treatment/hiv-treatment-information-adults

27. Peissig PL, Rasmussen LV, Berg RL, et al. Importance of multi-modal approaches to effectively identify cataract cases from electronic health records. J Am Med Inform Assoc 2012;19(02): 225–234 [PubMed: 22319176]

28. United States Census Bureau. Annual Estimates of the Resident Population, QuickFacts Florida. 2021. Accessed May 05, 2021 at: https://www.census.gov/quickfacts/FL

29. AIDSVu. Local Data: Florida. 2021. Accessed May 4, 2021 at: https://aidsvu.org/local-data/united-states/south/florida/

30. Hibbard JH, Pope CR. Gender roles, illness orientation and use of medical services. Soc Sci Med 1983;17(03):129–137 [PubMed: 6836347]

31. Cleary PD, Mechanic D, Greenley JR. Sex differences in medical care utilization: an empirical investigation. J Health Soc Behav 1982;23(02):106–119 [PubMed: 7108177]

32. Vaidya V, Partha G, Karmakar M. Gender differences in utilization of preventive care services in the United States. J Womens Health (Larchmt) 2012;21(02):140–145 [PubMed: 22081983]

33. Shergill Y, Rice D, Smyth C, et al. Characteristics of frequent users of the emergency department with chronic pain. CJEM 2020;22(03):350–358 [PubMed: 32213214]

34. Luque JS, Soulen G, Davila CB, Cartmell K. Access to health care for uninsured Latina immigrants in South Carolina. BMC Health Serv Res 2018;18(01):310 [PubMed: 29716586]

35. Betancourt JR, Carrillo JE, Green AR, Maina A. Barriers to health promotion and disease prevention in the Latino population. Clin Cornerstone 2004;6(03):16–26, discussion 27–29 [PubMed: 15707259]

36. De Jesus M, Xiao C. Cross-border health care utilization among the Hispanic population in the United States: implications for closing the health care access gap. Ethn Health 2013;18(03):297–314 [PubMed: 23043379]

37. Boyd CM, Lucas GM. Patient-centered care for people living with multimorbidity. Curr Opin HIV AIDS 2014;9(04):419–427 [PubMed: 24871089]

38. Harris MF, Dennis S, Pillay M. Multimorbidity: negotiating priorities and making progress. Aust Fam Physician 2013;42(12): 850–854 [PubMed: 24324984]

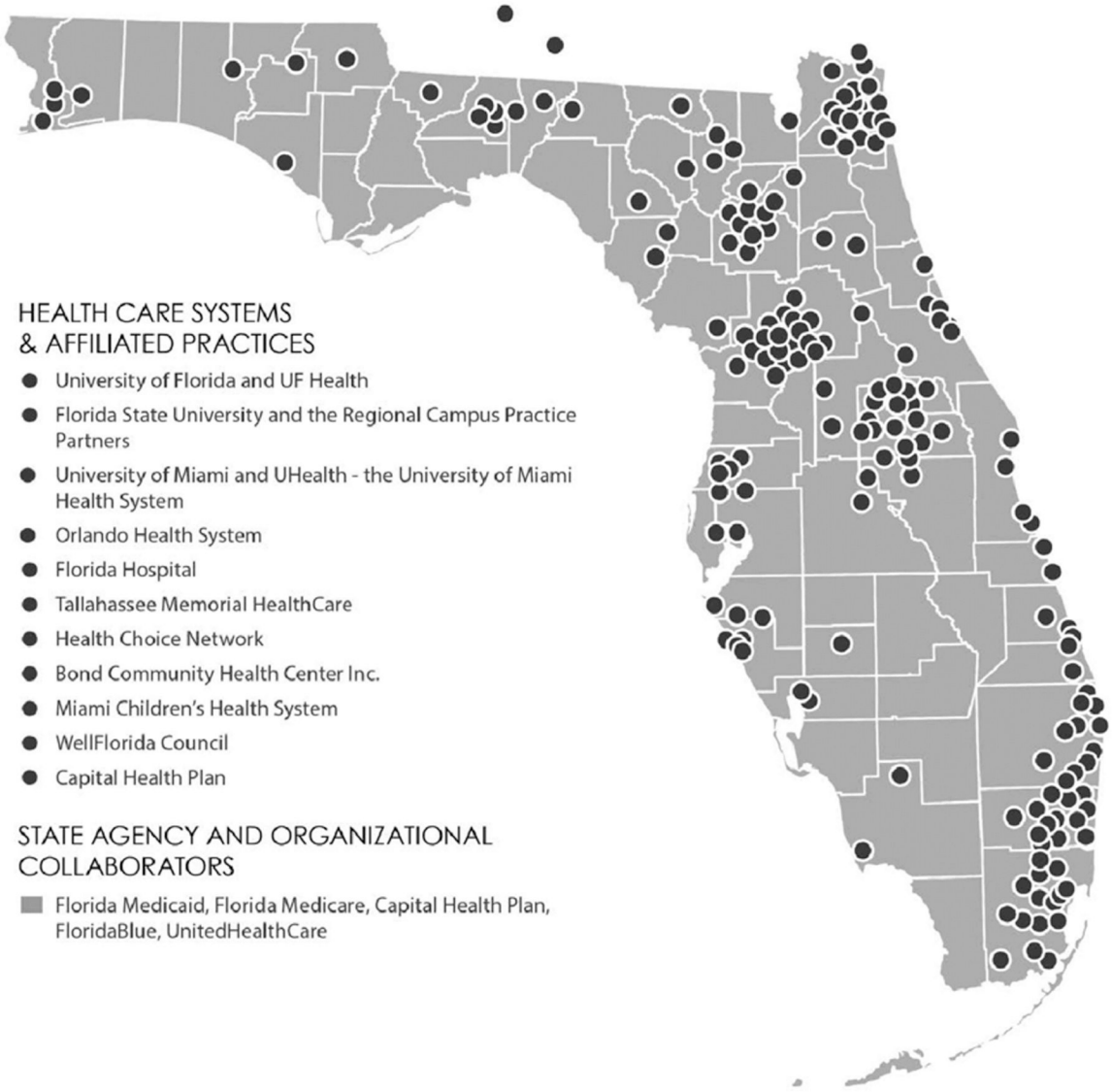

Fig. 1.
Map of OneFlorida partners health care providers throughout Florida.

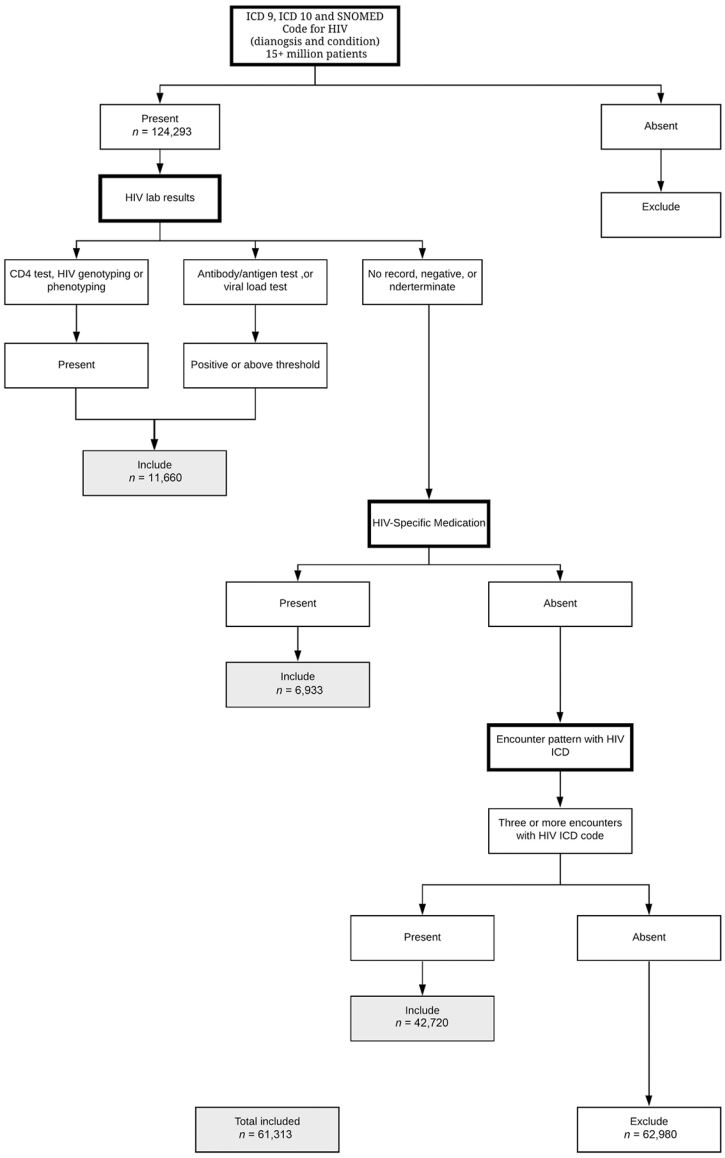

Fig. 2.

Flow chart of confirmed HIV cases identified by the computable phenotype algorithms.

Table 1

Patterns of the inclusion criteria combinations

Inclusion criteria				Algorithm classification	Frequency	Percent among the entire sample	Percent among the identified cohort
HIV diagnostic codes	Positive laboratories	HIV medication	Three or more HIV encounters				
				Excluded	2,399,275	94.79	NA
		√		Excluded	2,659	0.11	NA
				Excluded	4,745	0.19	NA
	√	√		Excluded	75	0.00	NA
√				Excluded	62,980	2.49	NA
√			√	Included	42,720	1.69	69.68
√		√		Included	1,995	0.08	3.25
√		√	√	Included	4,938	0.20	8.05
√	√			Included	1,221	0.05	1.99
√	√		√	Included	4,197	0.17	6.85
√	√	√		Included	285	0.01	0.46
√	√		√	Included	5,957	0.24	9.72

Table 2

The selection of the validation groups and corresponding proportion of people correctly classified by our computable phenotype

	N **(mutually exclusive)**	**Algorithm classification**	**Stratified sample group**	**Stratified sample group size**	**Focus sample group**	**Focus sample group size**	**Proportion of correct classify (target sample)**	**Proportion of correct classify (all sample)**
Pt not identified if exclude PrEP only	83	Included	1	Group1 (Randomly select 110)	1	5	4/5	6/8
Pt not identified if viral load using 50 cutoff	13				2	2	1/2	1/2
Pt has Dx, laboratory, Rx, and Enc	5,957				3	3	3/3	33/33
Pt has Dx, laboratory, Rx, NO enc	285				4	3	3/3	4/4
Pt has Dx, laboratory, enc, NO Rx	4,197				5	3	2/3	5/6
Pt has Dx, Rx, enc, NO laboratory	4,938				6	3	3/3	35/35
Pt has Dx and Rx, NO laboratory and enc	1,912				7	3	3/3	11/13
Pt has Dx and laboratory, NO Rx and enc	1,208				8	3	1/3	2/4
Pt has DX and Enc, NO laboratory and Rx	42,720		2	Group2 (Randomly select 20)	9	3	3/3	16/21
Pt has laboratory and Rx, no Dx	75	Excluded	1	Group1 (Randomly select 110)	10	5	5/5	6/6
Pt has HIV Geno procedure	72				11	5	5/5	3/3
Pt has HIV viral load procedure (CPT 87536)	4,991				12	2	2/2	11/12
Pt has Rx only, no Dx	2,565				13	2	2/2	16/17
Pt has laboratory only, no DX	4,594				14	3	3/3	3/3
Pt has Dx only	61,921		3	Group3 (Randomly select 20)	15	3	3/3	4/4
Pt has HIV screening procedure	422,831		3		16	2	2/2	19/19

Table 3

Crude and weighted algorithm performance

	Crude (95% CI)	**Weighted (95% CI)**
Sensitivity (recall)	96.81 (90.96, 99.34)	98.54 (92.17, 100)
Specificity	84.31 (71.41,92.98)	97.55 (94.88, 100)
PPV (precision)	91.92 (86.55, 97.29)	80.9 (61.97, 99.82)
NPV	93.48 (86.34, 100)	99.84 (99.15, 100)

Abbreviations: NPV, negative predictive value; PPV, positive predictive value.

Table 4

Sample characteristics of the OneFlorida HIV cohort as identified by our algorithm (n = 61,313)

Characteristics	Frequency (n)/mean/median	Percent (%)/SD, IQR
Sex		
Female	25,635	41.81
Male	35,654	58.15
Unknown	24	0.04
Race		
White	18,234	29.74
Black or African American	30,265	49.36
American Indian or Alaska Native	107	0.17
Asian	264	0.43
Native Hawaiian or Other Pacific Islander	19	0.03
Multiple race	231	0.38
Unknown	4,867	7.94
Other	7,326	11.95
Hispanic/Latinx		
Yes	10,155	16.56
No	45,230	73.77
Unknown	5,928	9.67
Mean age	42.66 (mean)	13.39 (SD)
Age group		
<18	1,891	3.25
18–29	9,094	15.61
30–44	17,056	29.28
45–64	27,870	47.85
65+	2,334	4.01
Median numberof HIV related encounter	21 (median)	47 (IQR)
Median total encounter	79 (median)	172 (IQR)
Mean length of follow up (number of days from first to last encounter)	1,682.43 (mean)	987.16 (SD)
Year entered the cohort		
2012	27,065	44.18
2013	7,675	12.53
2014	5,293	8.64
2015	4,826	7.88
2016	5,259	8.59
2017	4,490	7.33
2018	3,733	6.09
2019	2,810	4.59
2020	104	0.17

Abbreviations: IQR, interquartile range; SD, standard deviation.