JAMIA Open, 5(2), 2022, 1–6
https://doi.org/10.1093/jamiaopen/ooac033
Brief Communications

Brief Communications

Comparison of algorithms for identifying people with HIV from electronic medical records in a large, multi-site database

Jessica P. Ridgway, Joseph A. Mason, Eleanor E. Friedman, Samantha Devlin, Junlan Zhou, David Meltzer, and John Schneider

Department of Medicine, University of Chicago, Chicago, Illinois, USA

Corresponding Author: Jessica P. Ridgway, MD, MS, Department of Medicine, University of Chicago, 5841 S Maryland Ave, MC 5065, Chicago, IL 60637, USA; jessica.ridgway@uchospitals.edu

Received 9 November 2021; Revised 31 January 2022; Editorial Decision 26 April 2022; Accepted 9 May 2022

ABSTRACT

Objective: As electronic medical record (EMR) data are increasingly used in HIV clinical and epidemiologic research, accurately identifying people with HIV (PWH) from EMR data is paramount. We sought to evaluate EMR data types and compare EMR algorithms for identifying PWH in a multicenter EMR database.

Materials and Methods: We collected EMR data from 7 healthcare systems in the Chicago Area Patient-Centered Outcomes Research Network (CAPriCORN) including diagnosis codes, anti-retroviral therapy (ART), and laboratory test results.

Results: In total, 13 935 patients had a positive laboratory test for HIV; 33 412 patients had a diagnosis code for HIV; and 17 725 patients were on ART. Only 8576 patients had evidence of HIV-positive status for all 3 data types (laboratory results, diagnosis code, and ART). A previously validated combination algorithm identified 22 411 patients as PWH.

Conclusion: EMR algorithms that combine laboratory results, administrative data, and ART can be applied to multicenter EMR data to identify PWH.

Key words: HIV, clinical phenotyping, EMR, clinical informatics, diagnostic algorithm

LAY SUMMARY

Electronic medical record (EMR) data are increasingly utilized for HIV-related research. Therefore, it is important to accurately identify people who are HIV-positive from data present in EMRs. We evaluated different types of EMR data and compared EMR algorithms for identifying people with HIV (PWH) in a multicenter EMR database. Our data source was the Chicago Area Patient-Centered Outcomes Research Network (CAPriCORN), which contains EMR data from diverse healthcare systems in Chicago. We collected different EMR data types from CAPriCORN, including diagnosis codes, HIV medication data, and laboratory test results, to determine which data types were most helpful for determining if patients were HIV-positive. In the database, 13 935 patients had a positive laboratory test for HIV; 33 412 patients had a diagnosis code for HIV; and 17 725 patients were prescribed HIV-specific medication. Only 8576 patients were identified as HIV-positive in all 3 data types (laboratory results, diagnosis code, and medications). We applied an algorithm that utilized combinations of different data types, and it identified 22 411 patients as PWH. In conclusion, we found that EMR algorithms that combine laboratory results, diagnosis codes, and medications can be applied to multicenter EMR data to identify PWH.

INTRODUCTION

In the era of widespread electronic medical records (EMRs), information from EMR is increasingly used for clinical research in a variety of disciplines, including HIV clinical research.[1–3] In addition, public health agencies utilize HIV-related EMR data for epidemiologic purposes, such as tracking HIV care outcomes.[4–6] To ensure the validity of these analyses, accurately identifying people with HIV (PWH) from EMR data is critical.

Various EMR data sources can be utilized to identify PWH, including administrative diagnosis codes, laboratory test results, and prescriptions for HIV-specific medications. Relying on just one of these EMR data sources to identify PWH could result in misclassification. For example, an erroneous diagnosis code for HIV[7,8] or a prescription of anti-retroviral therapy (ART) for post-exposure prophylaxis rather than treatment of HIV could inaccurately identify a patient as HIV-positive when in fact they are HIV-negative. Conversely, incomplete medical records regarding HIV test results or prescriptions for ART could fail to identify PWH, for example, if a positive HIV test result occurred in a different health system without a shared EMR system. Computable phenotype algorithms that combine multiple EMR data sources can be used to more accurately identify PWH in an EMR system.[9,10]

Paul et al[10] developed and validated 2 EMR-based algorithms to identify PWH. Their first algorithm used laboratory and medication data and had a sensitivity of 78% and specificity of 99% for identifying PWH. Their second algorithm used International Classification of Diseases-9 (ICD-9) codes, medication, and laboratory testing data and had a sensitivity of 78% and specificity of 100%.[10] These algorithms were developed using data from a single medical center and have not been applied to identify PWH in large multicenter databases. The objective of this study was to utilize these algorithms in a large multicenter EMR database and investigate the utility of various electronic data types for identifying PWH.

METHODS

We collected de-identified data from 7 healthcare systems in the Chicago Area Patient-Centered Outcomes Research Network (CAPriCORN).[11] CAPriCORN is a clinical research network with linked data from diverse health care system EMRs in Chicago, including academic medical centers, community hospitals, and clinics.[11] Inclusion in the dataset consisted of all patients in the CAPriCORN database with either a diagnosis code (ICD-9 or ICD-10) for HIV or an HIV viral load test result between January 1, 2011 and September 5, 2019. Patients were de-duplicated via a hashing/matching process that has been previously described.[11]

For each patient in the dataset, we collected EMR data that could be used to determine if a patient was HIV-positive. These included ICD-9 and ICD-10 codes, prescriptions for HIV-specific medications (ie, ART), and laboratory test results (eg, HIV antibody, HIV antigen, HIV viral load). See Table 1 for diagnosis codes for HIV, laboratory test results considered positive for HIV, and HIV-specific medications. Of note, patients with prescriptions for medications used to treat Hepatitis B or pre-exposure prophylaxis for HIV prevention (lamivudine, emtricitabine, tenofovir disoproxil fumarate alone or in combination with emtricitabine, tenofovir alafenamide alone or in combination with emtricitabine) in the absence of other ART medications were not considered to have an 'HIV specific medication' as these medications are used for indications other than treatment of HIV. However, if these medications were prescribed in addition to other ART medications, they were included as HIV-specific medications.

We measured how many participants had a diagnosis code for HIV, positive laboratory test results for HIV, and/or were prescribed HIV-specific medications. We also applied the 2 algorithms previously developed by Paul et al[10] for identifying PWH from EMR data, as described in Figure 1. We explored combining these 2 algorithms to identify patients as HIV-positive if they met any of the criteria in either algorithm, which we labeled Algorithm 3. These 3 algorithms identify patients with positive laboratory tests for HIV and prescriptions for ART, but could potentially miss patients with well-controlled HIV who have been diagnosed at and receive their ART medication at an outside health system. Therefore, we created another algorithm (Algorithm 4) that added additional criteria to identify such patients (ie, diagnosis code for HIV and multiple HIV viral load tests performed). We compared the number and percentage of patients identified by each of these 4 algorithms. This study was approved by the Chicago Area Institutional Review Board.

RESULTS

The study cohort contained EMR data for 45 756 patients in the CAPriCORN research network database. The cohort contained 33 412 (73.0%) patients with a diagnosis code for HIV and 26 452 (57.8%) patients with at least 1 HIV viral load test result. The study cohort included 13 935 (30.5%) patients with a positive laboratory test for HIV (ie, confirmatory HIV antibody, p24 antigen, or HIV viral load >20 copies/mL) and 17 725 (38.7%) patients who were prescribed an HIV-specific medication (Table 2). Figure 2 shows the overlap among patients with a diagnosis code for HIV, those prescribed HIV-specific medication, and those with a positive HIV laboratory test. Only 8576 patients had evidence of HIV-positive status for all 3 data types.

Considering patients who fit multiple criteria for HIV-positive status, there were 16 846 (36.8%) patients with a diagnosis code for HIV who were also prescribed an HIV-specific medication. The cohort had 13 091 (28.6%) patients with a diagnosis code for HIV in addition to a positive HIV laboratory test. There were 13 167 (28.8%) patients with at least 1 HIV viral load test performed and a prescription for HIV-specific medication (Table 2). There were 16 554 (36.2%) patients with a diagnosis code for HIV and had at least 2 HIV viral load tests performed.

Algorithm 1 (positive HIV laboratory test, or at least one HIV viral load test and prescribed HIV-specific medication) identified 18 622 (40.7%) patients. Algorithm 2 (diagnosis code for HIV and a positive HIV laboratory test, or diagnosis code for HIV and prescribed HIV-specific medication) identified 21 361 (46.7%) patients. Algorithm 3 (a combination of Algorithms 1 and 2) identified 22 411 (49.0%) patients. Algorithm 4 identified 24 239 (53.0%) patients, an increase of 1828 patients over Algorithm 3. Figure 3 shows the overlap among patients identified by these 4 algorithms.

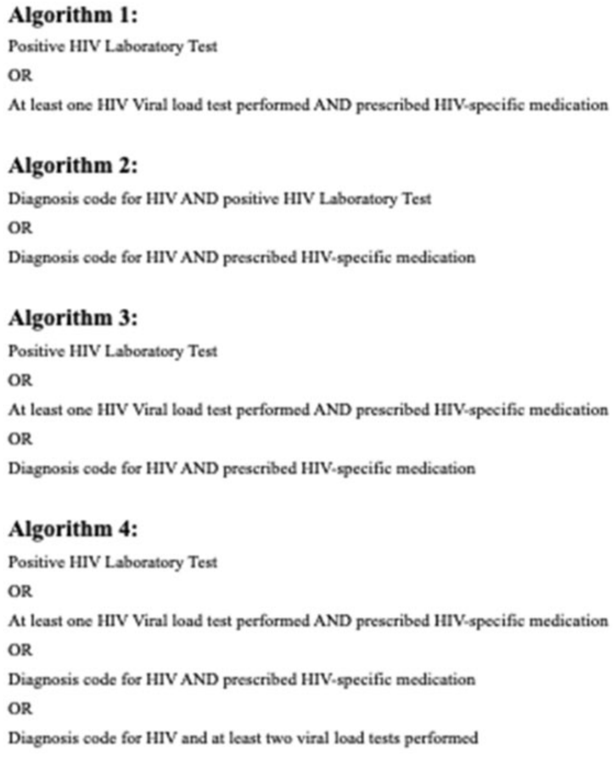
DISCUSSION

In a large multicenter EMR database, we investigated the use of different EMR data types for identifying patients with HIV. We found that only a minority of patients in the study cohort (18.7%) had all

Table 1. Criteria for determining HIV-positive status based on laboratory results, medications, and diagnosis codes, 2011-2019, CAPriCORN, Chicago, IL

	Category	Criteria
Positive laboratory tests for HIV	Confirmatory HIV antibody	Positive Western Blot
		Positive indirect fluorescent antibody
	Antigen	Positive p24 antigen
	NAAT/PCR	HIV viral load >20 copies/mL
HIV-specific medications	Nucleoside reverse transcriptase inhibitors	Abacavir (ABC)
		Didanosine (DDI)
		Emtricitabine (FTC)
		Lamivudine (3TC)
		Stavudine (D4T)
		Tenofovir alafenamide (TAF)
		Tenofovir disoproxil fumarate (TDF)
		Zidovudine (AZT)
	Non-nucleoside reverse transcriptase inhibitors	Delavirdine
		Doravirine
		Efavirenz
		Etravirine
		Nevirapine
	Protease inhibitors	Rilpivirine
		Atazanavir
		Darunavir
		Fosamprenavir
		Indinivir
		Lopinavir
		Nelfinavir
		Ritonavir
		Saquinavir
		Tipranivir
	Integrase inhibitors	Bictegravir
		Dolutegravir
		Elvitegravir
		Raltegravir
	Fusion inhibitor	Enfurvitide
	Entry inhibitor	Maraviroc
	Booster	Cobicistat
	Combination pills	ABC/3TC/dolutegravir
		ABC/3TC
		ABC/3TC/AZT
		AZT/3TC
		Bictegravir/TAF/FTC
		3TC/dolutegravir
		Dolutegravir/rilpivirine
		TAF/FTC
		TAF/FTC/darunavir/cobicistat
		TAF/FTC/elvitegravir/cobicistat
		TAF/FTC/rilpivirine
		TDF/FTC/elvitegravir/cobicistat
		TDF/FTC
		TDF/3TC
		TDF/FTC/efavirenz
		TDF/FTC/rilpivirine
		TDF/3TC/efavirenz
		TDF/3TC/doravirine
Diagnosis codes for HIV	ICD9 codes	042
		042.0
		042.1
		042.2
		042.9
		043.0
		043.1
		043.2
		043.3
		043.9
		044
		044.0
		044.9
		079.53
		795.71
		795.78
		V08
	ICD10 codes	B20
		B21
		B22
		B23
		B24
		R75
		Z21

ICD: International Classification of Diseases; NAAT: nucleic acid amplification test; PCR: polymerase chain reaction.



Algorithm 1:

Positive HIV Laboratory Test

OR

At least one HIV Viral load test performed AND prescribed HIV-specific medication

Algorithm 2:

Diagnosis code for HIV AND positive HIV Laboratory Test

OR

Diagnosis code for HIV AND prescribed HIV-specific medication

Algorithm 3:

Positive HIV Laboratory Test

OR

At least one HIV Viral load test performed AND prescribed HIV-specific medication

OR

Diagnosis code for HIV AND prescribed HIV-specific medication

Algorithm 4:

Positive HIV Laboratory Test

OR

At least one HIV Viral load test performed AND prescribed HIV-specific medication

OR

Diagnosis code for HIV AND prescribed HIV-specific medication

OR

Diagnosis code for HIV and at least two viral load tests performed

Figure 1. Algorithms for identifying people with HIV from electronic medical record data.

3 data types (HIV-specific medication, positive HIV laboratory test, and HIV diagnosis code). This finding suggests that relying on any one data type may lead to under-identification of PWH when using multi-site EMR data for clinical HIV research.

Many prior HIV-related clinical research studies utilizing EMR data have relied on positive HIV test results to identify PWH.[3] Positive HIV test results are highly specific, but relying on test results alone to identify PWH may lack sensitivity. Patients may have their

Table 2. Numbers of patients meeting various criteria for HIV-positive status

Algorithm	Population (*n*)	Percentage
Total patients with at least 1 HIV viral load test performed or diagnosis code for HIV	45 756	100%
Positive laboratory test for HIV (ie, confirmatory HIV antibody, p24 antigen, or HIV viral load >20 copies/mL)	13 935	30.5%
At least 1 encounter with a diagnosis code for HIV	33 412	73.0%
Patients prescribed HIV-specific medication	17 725	38.7%
Patients with at least 1 HIV viral load test performed	26 452	57.8%
Patients with at least 2 HIV viral load tests performed	17 218	37.6%
Diagnosis code for HIV and prescribed HIV-specific medication	16 846	36.8%
Diagnosis code for HIV and at least 2 HIV viral load tests performed	16 554	36.2%
Diagnosis code for HIV and a positive laboratory test for HIV	13 091	28.6%
Patients with at least 1 HIV viral load test performed and prescribed HIV-specific medication	13 167	28.8%
Algorithm 1[a]	18 622	40.7%
Algorithm 2[b]	21 361	46.7%
Algorithm 3[c]	22 411	49.0%
Algorithm 4[d]	24 239	53.0%

[a] Positive HIV Laboratory Test OR (At least one HIV Viral load test performed AND prescribed HIV-specific medication).

[b] (Diagnosis code for HIV AND positive HIV Laboratory Test) OR (Diagnosis code for HIV AND prescribed HIV-specific medication).

[c] Positive HIV Laboratory Test OR (At least one HIV Viral load test performed AND prescribed HIV-specific medication) OR (Diagnosis code for HIV AND prescribed HIV-specific medication).

[d] Positive HIV Laboratory Test OR (At least one HIV Viral load test performed AND prescribed HIV-specific medication) OR (Diagnosis code for HIV AND prescribed HIV-specific medication) OR (Diagnosis code for HIV and at least two viral load tests performed).

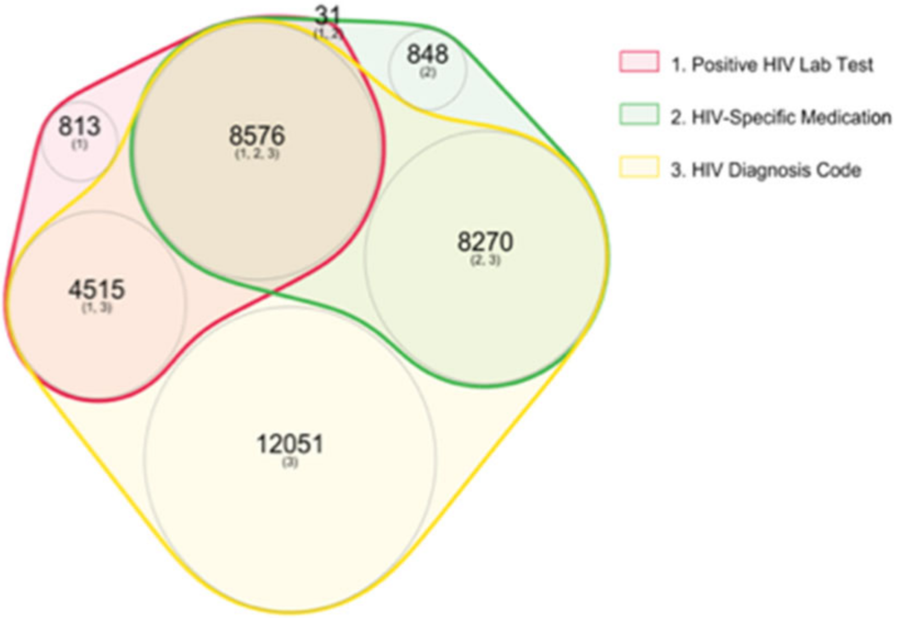
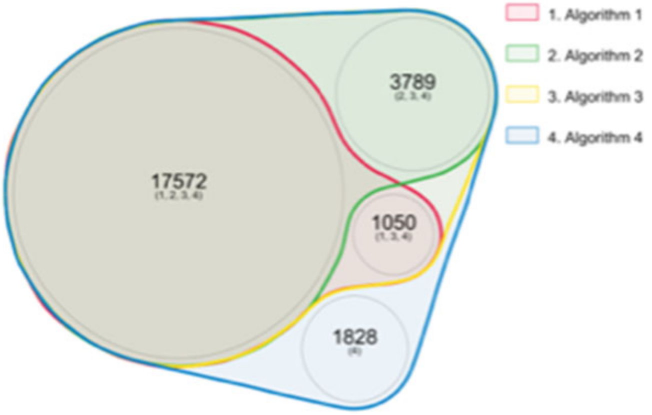


Figure 2. Venn diagram among patients with an HIV diagnosis code, patients prescribed HIV-specific medication, and patients with a positive HIV laboratory test.

initial positive HIV antibody test result in a healthcare system differing from their current place of care. Including detectable HIV viral load test results can identify more patients with HIV than a positive antibody test alone, but PWH who are adherent to ART often have persistently undetectable viral load results. Data sharing of HIV laboratory results among healthcare systems could improve the sensitivity of lab results for identifying PWH. In addition, healthcare systems are required to report HIV laboratory test results for PWH to public health departments. Data sharing between healthcare systems and public health departments could further enhance the sensitivity of EMR lab results for identifying PWH. Indeed, some healthcare systems have utilized such data sharing through the Data to Care program.[4,12]

Other studies have used diagnosis codes to identify PWH from EMR data,[13–15] but diagnosis codes may lack specificity. For example, a patient could have an inaccurate code applied when an HIV-negative patient has an encounter for HIV screening or HIV prevention counseling. In addition, the algorithm developed by Paul et al. included diagnosis codes that indicate non-specific serologic evidence of HIV (e.g., ICD-9 code 795.71) which are sometimes used to connote inconclusive HIV test results and do not necessarily indicate HIV-positive status. Other studies have ex-



Algorithm 1 Positive HIV Laboratory Test OR (At least one HIV Viral load test performed AND prescribed HIV-specific medication)

Algorithm 2. (Diagnosis code for HIV AND positive HIV Laboratory Test) OR (Diagnosis code for HIV AND prescribed HIV-specific medication)

Algorithm 3. Positive HIV Laboratory Test OR (At least one HIV Viral load test performed AND prescribed HIV-specific medication) OR (Diagnosis code for HIV AND prescribed HIV-specific medication)

Algorithm 4. Positive HIV Laboratory Test OR (At least one HIV Viral load test performed AND prescribed HIV-specific medication) OR (Diagnosis code for HIV AND prescribed HIV-specific medication) OR (Diagnosis code for HIV and at least two viral load tests performed)

Figure 3. Venn diagram among patients identified as HIV-positive by 4 different algorithms.

cluded these nonspecific codes for identifying people who are HIV-positive.[16] Because we applied the algorithms developed by Paul in our study, we chose to include these codes despite possible lack of specificity. We performed a sensitivity analysis excluding ICD-9 code 795.71 and ICD-10 code R75, but these codes accounted for <1% of HIV-associated diagnostic codes, and results were very similar.

The addition of other criteria in combination with HIV diagnosis codes, such as prescription of HIV-specific medications, may more accurately identify PWH. In our study, we found that algorithms that combine multiple data types from the EMR to identify PWH, such as the algorithms developed by Paul et al, have improved accuracy and identify more PWH beyond those using just one data type alone. Indeed, Algorithm 4 identified 10 304 more patients as HIV-positive than we would have identified if we had only relied on a positive HIV test result.

When applying the previously validated algorithms from Paul et al to a large multicenter cohort in Chicago, we found several differences in results compared to Paul's single-center study. For Paul, there was significant overlap in patients identified by Algorithms 1 and 2. 91% (970/1063) of patients identified by Algorithm 3 (the combination of Algorithms 1 and 2) were also identified by Algorithms 1 and 2 alone. In our study, only 78.4% (17 572/22 411) of PWH identified by Algorithm 3 were also identified by both Algorithms 1 and 2. The single site in which Paul's study took place may have had more complete EMR information for each patient, allowing for greater consistency between algorithms. In our multicenter study, the lack of overlap of these 2 algorithms could be due to more fragmented care for our patients or incomplete EMR data within the database.

Our study has several limitations. We did not validate the algorithms in our study using manual chart review to determine test sensitivity or specificity because our deidentified database did not include text of clinical notes and was not able to be linked back to medical records for manual chart review. However, we utilized several previously validated algorithms. In addition, while we excluded ART regimens used for PrEP, it is possible that some of the HIV-specific medications we identified were prescribed for HIV-negative patients for post-exposure prophylaxis. To exclude prescriptions for post-exposure prophylaxis, we explored limiting the algorithms to only ART prescriptions with at least one confirmed refill within 6 months. However, 30% of prescriptions in our database were missing refill data, and so we chose to include ART prescriptions for any length of time. Utilizing data from a multisite EMR database could have resulted in discrepancies due to differing internal procedures (eg, ordering laboratory tests, billing, documenting diagnoses in the EMR, etc.). However, using data from multiple sites allows for greater generalizability to other health systems.

CONCLUSION

In conclusion, EMR algorithms that combine laboratory results, administrative data, and ART prescriptions detected more patients with HIV in a large multisite EMR database than use of HIV laboratory test results alone. The use of EMR algorithms across multiple EMR systems within different settings can lead to rapid case detection of PWH and cross-institutional collaboration to facilitate HIV clinical research and epidemiologic studies.

FUNDING

This work was supported by the NIH (K23 MH121190).

AUTHOR CONTRIBUTIONS

JPR and JS conceived of the study. JAM, EEF, and JZ performed data analysis. JPR, JS, JAM, EEF, SD, DM, and JS interpreted results of the study. JPR drafted the manuscript and all other authors provided critical revisions.

CONFLICT OF INTEREST STATEMENT

JPR has received fees for consulting for Gilead Sciences.

DATA AVAILABILITY STATEMENT

The data underlying this article were provided by the Chicago Area Patient-Centered Outcomes Research Network (CAPriCORN). Data will be shared on request to the corresponding author with permission of the CAPriCORN Steering Committee.

REFERENCES

1. Floris-Moore M, Edmonds A, Napravnik S, Adimora AA. Computerized adjudication of coronary heart disease events using the electronic medical record in HIV clinical research: possibilities and challenges ahead. *AIDS Res Hum Retroviruses* 2020; 36 (4): 306–13.
2. Ridgway JP, Lee A, Devlin S, Kerman J, Mayampurath A. Machine learning and clinical informatics for improving HIV care continuum outcomes. *Curr HIV/AIDS Rep* 2021; 18 (3): 229–36.
3. Erly S, Roberts DA, Kerani R, *et al.* Assessing HIV care outcomes among African-born people living with HIV in Seattle: an analysis of the University of Washington Electronic Medical Record. *J Immigr Minor Health* 2021; 23 (6): 1136–44.
4. Ridgway JP, Almirol E, Schmitt J, Wesley-Madgett L, Pitrak D. A clinical informatics approach to reengagement in HIV care in the emergency department. *J Public Health Manag Pract* 2019; 25 (3): 270–3.
5. Dean BB, Hart RL, Buchacz K, *et al.*; HOPS Investigators. HIV laboratory monitoring reliably identifies persons engaged in care. *J Acquir Immune Defic Syndr* 2015; 68 (2): 133–9.
6. Arey AL, Cassidy-Stewart H, Kurowski PL, Hitt JC, Flynn CP. Evaluating HIV surveillance completeness along the continuum of care: supplementing surveillance with health center data to increase HIV data to care efficiency. *J Acquir Immune Defic Syndr* 2019; 82 (Suppl 1): S26–S32.
7. van Walraven C, Bennett C, Forster AJ. Administrative database research infrequently used validated diagnostic or procedural codes. *J Clin Epidemiol* 2011; 64 (10): 1054–9.
8. 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): 1620–39.
9. 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 (7): 626–33.
10. 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 (2): 150–7.
11. Kho AN, Hynes DM, Goel S, *et al.*; CAPriCORN Team. CAPriCORN: Chicago area patient-centered outcomes research network. *J Am Med Inform Assoc* 2014; 21 (4): 607–11.
12. Sweeney P, DiNenno EA, Flores SA, *et al.* HIV data to care-using public health data to improve HIV care and prevention. *J Acquir Immune Defic Syndr* 2019; 82 (Suppl 1): S1–S5.
13. Levison J, Triant V, Losina E, Keefe K, Freedberg K, Regan S. Development and validation of a computer-based algorithm to identify foreign-born patients with HIV infection from the electronic medical record. *Appl Clin Inform* 2014; 5 (2): 557–70.
14. Friedman EE, Devlin SA, McNulty MC, Ridgway JP. SARS-CoV-2 percent positivity and risk factors among people with HIV at an urban academic medical center. *PLoS One* 2021; 16 (7): e0254994.
15. Sigel K, Swartz T, Golden E, *et al.* Coronavirus 2019 and people living with human immunodeficiency virus: outcomes for hospitalized patients in New York City. *Clin Infect Dis* 2020; 71 (11): 2933–8.
16. 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): 1318–25.