American Journal of Epidemiology

Vol. 189, No. 5
DOI: 10.1093/aje/kwz225
Advance Access publication:
October 15, 2019

Practice of Epidemiology

Validation of an Optimized Algorithm for Identifying Persons Living With Diagnosed HIV From New York State Medicaid Data, 2006–2014

Sarah E. Macinski*, Jayleen K. L. Gunn, Mona Goyal, Charles Neighbors, Rajeev Yerneni, and Bridget J. Anderson

* Correspondence to Sarah E. Macinski, Bureau of HIV/AIDS Epidemiology, AIDS Institute, New York State Department of Health, Empire State Plaza, Corning Tower, Room 717, Albany, NY 12237-0627 (e-mail: sarah.macinski@health.ny.gov).

Initially submitted February 27, 2019; accepted for publication September 23, 2019.

Algorithms are regularly used to identify persons living with diagnosed human immunodeficiency virus (HIV) (PLWDH) in Medicaid data. To our knowledge, there are no published reports of an HIV algorithm from Medicaid claims codes that have been compared with an HIV surveillance system to assess its sensitivity, specificity, positive predictive value, and negative predictive value in identifying PLWDH. Therefore, our aims in this study were to 1) develop an algorithm that could identify PLWDH in New York State Medicaid data from 2006–2014 and 2) validate this algorithm using the New York State HIV surveillance system. Classification and regression tree analysis identified 16 nodes that we combined to create a case-finding algorithm with 5 criteria. This algorithm identified 86,930 presumed PLWDH, 88.0% of which were verified by matching to the surveillance system. The algorithm yielded a sensitivity of 94.5%, a specificity of 94.4%, a positive predictive value of 88.0%, and a negative predictive value of 97.6%. This validated algorithm has the potential to improve the utility of Medicaid data for assessing health outcomes and programmatic interventions.

acquired immunodeficiency syndrome; algorithms; classification and regression trees; HIV; Medicaid; validation studies

Abbreviations: AIDS, acquired immunodeficiency syndrome; APR-DRG, All Patient Refined Diagnosis Related Groups; CART, classification and regression tree; CDC, Centers for Disease Control and Prevention; CFA, case-finding algorithm; HH, Health Home; HIV, human immunodeficiency virus; ICD-9-CM, *International Classification of Diseases, Ninth Revision, Clinical Modification*; NPV, negative predictive value; NYS, New York State; NYSDOH, New York State Department of Health; PLWDH, persons living with diagnosed HIV; PPV, positive predictive value.

The federal Medicaid program is the largest health insurer of people who have been diagnosed with human immunodeficiency virus (HIV) infection (40%) (1, 2). Consequently, Medicaid data are a robust source of health-care information and are often used to study multiple facets of HIV disease in the United States, including health-care expenses (3–5), health disparities (6), trends in hospitalization (7, 8), and quality of HIV care (9–11). The size of the Medicaid data set allows for high statistical power and subgroup analysis, as well as the study of low-prevalence events, such as HIV diagnosis, with high statistical power (12–14).

While there are many advantages of using Medicaid data to study HIV care outcomes, data are collected for billing rather than clinical or research purposes (14–17). Medicaid researchers have indicated problems with accuracy and reliability when using Medicaid data (16, 18–25), as claims codes often vary by procedure, type of medical encounter, and clinical condition (14, 22–24, 26). A variety of algorithms have been developed to identify likely persons living with diagnosed HIV (PLWDH) within the Medicaid population (4, 7, 9–11, 13, 22, 23, 27–29). The accuracy and reliability of using diagnosis information in Medicaid data are higher for inpatient claims than for outpatient claims (13), but the number of hospitalizations for PLWDH are decreasing (8, 30). Because most Medicaid claims for PLWDH are for outpatient services (13) and not for diagnostic purposes, new algorithms need be sensitive to a variety of procedure and service codes, in addition to diagnosis codes.

Optimal algorithms should share attributes of the Centers for Disease Control and Prevention (CDC)-defined "optimal surveillance system" (31), such as simplicity, flexibility, representativeness, sensitivity, and a high positive predictive value (PPV) (32). Despite the regular use of algorithms to identify PLWDH from Medicaid data (7, 9–11, 13, 22, 27), to our knowledge there are no published reports of an HIV algorithm derived from Medicaid data that has been validated by matching the results to a population-based HIV surveillance system (the gold standard) to test the algorithm's sensitivity, specificity, PPV, and negative predictive value (NPV). The New York State (NYS) Medicaid program utilizes its own internally developed algorithm to identify HIV-positive persons. That algorithm has not been validated against the NYS HIV surveillance system and has not been published.

NYS has approximately 113,000 PLWDH (33) and more PLWDH on Medicaid than any other state (34). A validated algorithm that accurately identifies PLWDH in NYS Medicaid data would enable researchers and policy-makers to better understand the impact of expanded Medicaid programs for PLWDH (34). Thus, our aims in this study were 2-fold: 1) to develop an algorithm for identifying PLWDH using NYS Medicaid data and 2) to validate the algorithm using the population-based NYS HIV surveillance system.

METHODS

Case identification and validation

Name-based reports for all HIV-related laboratory test results (e.g., CD4 cell counts and percentages, viral load tests, genotype resistance testing) are received by the New York State Department of Health's (NYSDOH) HIV surveillance system for persons who reside or receive HIV-related care in NYS. This information is supplemented with data from the medical provider and partner report forms. In accordance with NYS public health law, clinicians are required to submit a report for all new or previously unreported diagnoses of HIV infection/acquired immunodeficiency syndrome (AIDS) within 14 days of diagnosis. The surveillance system routinely undergoes quality control and verification processes to ensure quality and completeness of the data. For this analysis, PLWDH were defined as persons with confirmed cases of HIV infection per the revised 2014 CDC definition (35). Only confirmed HIV diagnoses (regardless of disease stage) that met the CDC case definition were included in these analyses (Figure 1).

Medicaid encounter and claims data for 2006–2014 were obtained from the NYS Medicaid Data Mart for persons aged 18–64 years who met eligibility criteria for the NYS Medicaid Health Home (HH) Program. To qualify for the HH Program in NYS, individuals must have 2 or more medically documented chronic conditions included in the "major" categories of the 3M (3M Company, St. Paul, Minnesota) Clinical Risk Groups (e.g., alcohol and substance use disorder, mental health, cardiovascular disease, metabolic disease, or respiratory disease) or have a single qualifying diagnosis such as HIV, AIDS, and/or a serious mental illness (36) (Figure 1).

Because of the resource-heavy nature of matching large data sets, it was infeasible to compare all New Yorkers enrolled in Medicaid against the NYS HIV surveillance system; rather, for this study, an inclusive cohort of individuals—designated "potential PLWDH"—were identified from Medicaid-enrolled HH-eligible persons (n = 269,257). Published literature (3–5, 7, 9–11, 13, 22, 27, 28, 37–41) and an internal NYSDOH–AIDS Institute algorithm (42) were used to establish criteria for identification of potential PLWDH in Medicaid data. Criteria for identification of PLWDH in the HH-eligible Medicaid data included: 1) a diagnosis code for HIV (3–5, 7, 9–11, 13, 22, 27, 28, 37–40, 42); 2) rate codes for AIDS hospice care (4, 38, 42), HIV-related outpatient services (4, 37), or an HIV Special Needs Plan (4, 38, 42); 3) a National Drug Code for an HIV antiretroviral medication (4, 13, 37, 38, 40, 43, 44); 4) a procedure code for HIV viral quantification, phenotyping, or genotyping (4, 9, 38, 40, 43, 44); or 5) 3 claims with a 3M Clinical Risk Group between 102 and 105 (45). Individuals were also identified as potential PLWDH if they had a combination of claims with codes for HIV-related outpatient services (4, 5, 13, 37, 38, 42, 44), provider specialties (4, 37, 42), or a diagnosis of an opportunistic infection (4, 42, 44). (See Web Table 1, available at https://academic.oup.com/aje, for a detailed description of the inclusion criteria.)

A potential PLWDH was categorized as a true case if the record could be matched to an existing record in the NYS HIV surveillance system, which was considered the gold standard for this analysis. Potential PLWDH were probabilistically matched to the surveillance system using combinations of name, sex, race/ethnicity, address, date of birth, health-care provider, and identification numbers from Medicaid and the NYS Department of Corrections and Community Supervision. Matching was conducted using IBM InfoSphere QualityStage software (IBM, Armonk, New York). Through probabilistic matching and manual review, 80,899 persons from NYS Medicaid data were confirmed as true PLWDH (Table 1); the software identified 63,047 of these cases as exact matches, and 17,852 were identified through manual review of uncertain matches by trained surveillance staff.

Demographic information was obtained from the Medicaid eligibility data. Age was defined as an individual's age at which he/she was first identified for inclusion in this study. Race/ethnicity was defined as non-Hispanic white, non-Hispanic black, non-Hispanic Asian or Pacific Islander, non-Hispanic American Indian or Alaska Native, Hispanic, or other. Persons without a race or ethnicity listed were recoded as having an unknown race/ethnicity. Records were obtained for each month in which an individual was eligible for Medicaid. Because a person's eligibility status can fluctuate, the Medicaid enrollees may have had data in some years and not in others. Therefore, data were evaluated by year.

Statistical evaluation of algorithm performance

Classification and regression tree (CART) analysis was used to predict which of the potential PLWDH had diagnosed HIV ("presumptive PLWDH"). CART statistically

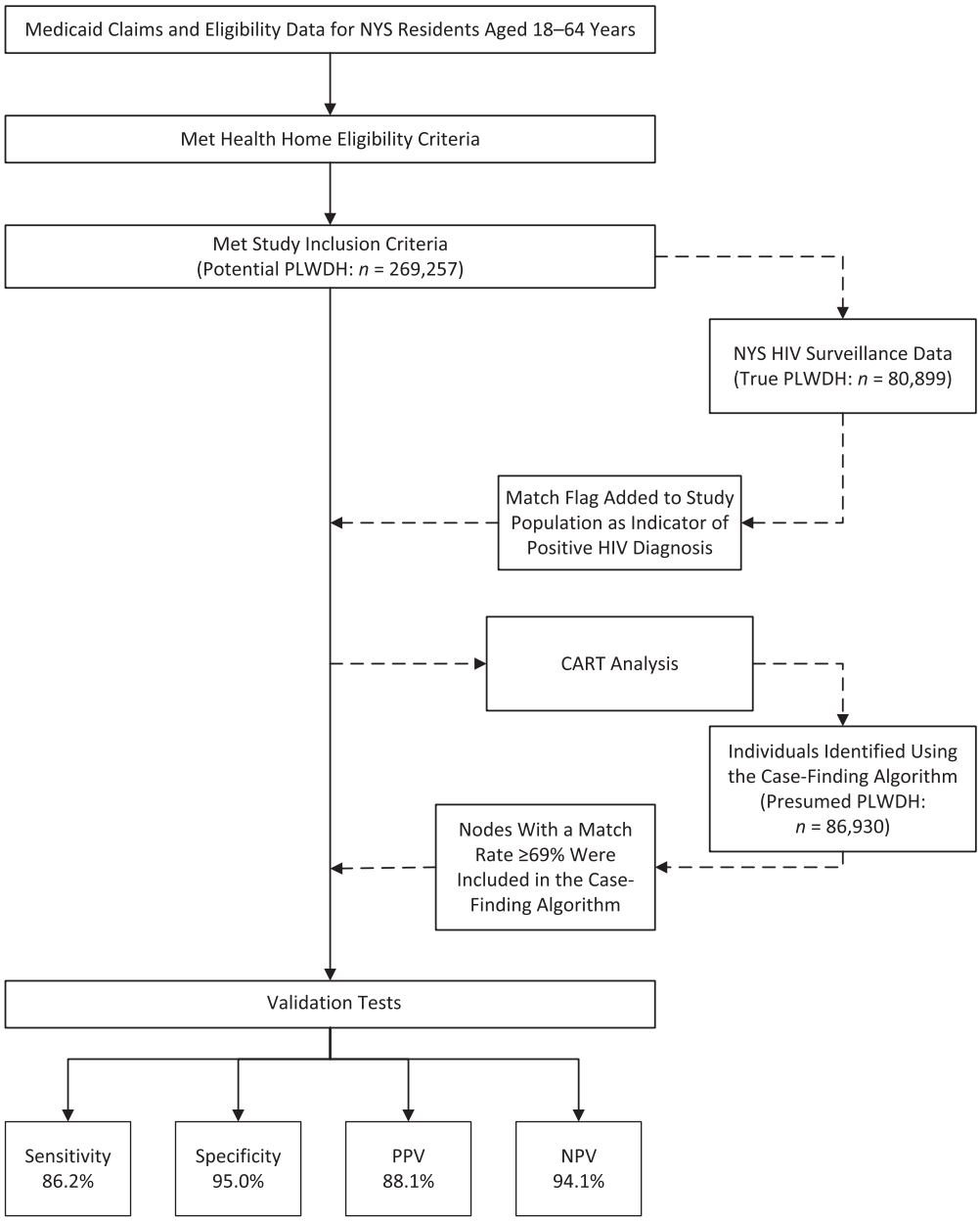

Figure 1. Identification and evaluation of presumptive persons living with diagnosed HIV (PLWDH) from New York State (NYS) Medicaid data, 2006–2014. CART, classification and regression tree; HIV, human immunodeficiency virus; NPV, negative predictive value; PPV, positive predictive value.

determines which factors in a model have the greatest explanatory power (46). The data from 2013 were chosen to develop the case-finding algorithm (CFA) through CART analysis because the 2013 data allowed for both retrospective (2006–2012) and prospective (2014) cross-validation through unique CART analyses. Because of a natural break in the data, nodes that had a match rate of 69% or higher were retained for the CFA (Web Figure 1). The sensitivity, specificity, PPV, and NPV of the CFA were evaluated among the potential PLWDH by comparing the presumptive PLWDH with the gold standard NYS HIV surveillance system. The flexibility and representativeness of the CFA were assessed by stratifying the population by claim year, sex, race/ethnicity, and age.

Ethical considerations

This study was designated as "Exemption 5—Public Benefit or Service Programs" by the NYSDOH Institutional Review Board. Claims data for 2006–2014 were used

Table 1. Demographic Characteristics of Potential and Presumptive Persons Living With Diagnosed HIV in Medicaid Data, New York State, 2006–2014

Characteristic	No. of Potential PLWDH	No. of Presumptive PLWDH
Total	269,257	86,930
Year[a]		
2006	63,650	41,136
2007	59,145	40,750
2008	57,475	41,645
2009	61,421	42,857
2010	64,911	44,443
2011	68,344	45,144
2012	72,017	45,446
2013	100,892	47,067
2014	160,725	49,177
Sex[b]		
Female	134,759	32,556
Male	134,495	54,374
Age group, years		
Missing data	5,920	4,628
≤19	4,513	1,732
20–24	19,139	4,311
25–29	23,466	5,636
30–39	50,518	16,293
40–49	71,452	30,012
50–59	62,994	19,777
≥60	31,255	4,541
Race/ethnicity		
Non-Hispanic white	52,884	11,705
Non-Hispanic black	78,443	38,493
Non-Hispanic Asian/ Pacific Islander	17,037	981
Non-Hispanic American Indian/Alaska Native	409	75
Hispanic	76,585	27,078
Other	13,840	4,376
Unknown	30,059	4,222

Abbreviations: HIV, human immunodeficiency virus; PLWDH, persons living with diagnosed HIV.

[a] Individuals were often identified in multiple years, so the sum for all years may be greater than the total number of individuals.

[b] Sex was unknown for 3 persons.

pursuant to Data Exchange and Application and Data Use Agreements between the NYSDOH Bureau of HIV/AIDS Epidemiology, the NYSDOH Office of Health Insurance Programs, and the National Center on Addiction and Substance Abuse.

RESULTS

Demographic characteristics of potential PLWDH

The potential PLWDH included 269,257 unduplicated individuals meeting the inclusion criteria (Table 1). There were equal numbers of males and females (134,495 and 134,759, respectively; sex was unknown for 3 persons). More than half of the potential PLWDH were non-Hispanic black (29.1%) or Hispanic (28.4%). Persons aged 40–49 years constituted the largest category of the study population (26.5%). The number of people identified for inclusion increased over the study years. Between 2006 and 2012, the annual number of persons identified for inclusion ranged between 57,475 (2008) and 72,017 (2012). Coinciding with the implementation of the NYS HH Program in 2013, the number of potential PLWDH that met inclusion criteria increased to 100,892 in 2013 and 160,725 in 2014.

Use of CART analysis to identify presumptive PLWDH

The 2013 CART model (Web Figure 1) identified 21 unique nodes, each representing a combination of selection criteria. Sixteen of these nodes indicated presumptive PLWDH and were combined to create the CFA (Table 2). The CFA identified "presumptive PLWDH" if a person met the criteria in one of the following categories: 1) 2 outpatient claims with an *International Classification of Diseases, Ninth Revision, Clinical Modification* (ICD-9-CM) (38) diagnosis code or an All Patients Refined Diagnosis Related Groups (APR-DRG) (42) code indicating an HIV diagnosis; 2) 1 inpatient claim with an ICD-9-CM diagnosis code or an APR-DRG code indicating an HIV diagnosis; 3) 1 claim for an HIV Special Needs Plan rate code; 4) 1 claim with a procedure code for viral quantification, phenotyping, or genotyping combined with 1 National Drug Code for an antiretroviral medication and an ICD-9-CM diagnosis code for an opportunistic infection; or 5) 1 claim for an HIV antiretroviral medication plus 1 claim with a rate code for outpatient services, such as an HIV clinic, an HIV medical evaluation visit, or HIV HH case management. By applying these 5 criteria to the whole sample, we identified 86,930 presumptive PLWDH in the Medicaid claims data. As Table 2 demonstrates, 82.6% of these presumptive PLWDH had at least 2 outpatient claims with a code indicating an HIV diagnosis.

Demographic characteristics of presumptive PLWDH identified by the CFA

The CFA identified more males than females (62.6% and 37.5%, respectively) (Table 1). Presumed PLWDH varied by race/ethnicity (non-Hispanic black, 44.3%; Hispanic, 31.2%; non-Hispanic white, 13.5%; non-Hispanic Asian/Pacific Islander, 1.1%; non-Hispanic American Indian/Alaska Native, 0.1%; other, 5.0%). Persons aged 40–49 years were the largest age group identified, constituting 34.5% of the presumptive PLWDH. On average, the CFA identified 44,184 presumptive PLWDH each year between 2006 and 2014, with a slight increase in the yearly number of presumptive PLWDH identified between 2011 and 2014.

Table 2. Components of a Case-Finding Algorithm Used to Identify 86,930 Presumptive Persons Living With Diagnosed HIV in Medicaid Data, New York State, 2006–2014

Component	Total[a] Component Contribution		Unique[b] Component Contribution	
	No. of Persons	%	No. of Persons	%
2 outpatient claims with an ICD-9-CM diagnosis code or an APR-DRG code indicating an HIV diagnosis	71,776	82.6	33,370	38.4
1 inpatient claim with an ICD-9-CM diagnosis code or an APR-DRG code indicating an HIV diagnosis	27,502	31.6	9,779	11.3
1 claim for an HIV Special Needs Plan rate code	19,755	22.7	1,772	2.0
1 claim with a procedure code for viral quantification, phenotyping, or genotyping combined with 1 National Drug Code for an antiretroviral medication and an ICD-9-CM diagnosis code for an opportunistic infection	14,086	16.2	416	0.5
1 claim for an HIV antiretroviral medication plus 1 claim with a rate code for outpatient services	16,857	19.4	1,210	1.4

Abbreviations: APR-DRG, All Patient Refined Diagnosis Related Groups; HIV, human immunodeficiency virus; ICD-9-CM, *International Classification of Diseases, Ninth Revision, Clinical Modification*; PLWDH, persons living with diagnosed HIV.

[a] 46.5% of presumptive PLWDH were captured by more than 1 component of the case-finding algorithm, so the column percentages may not add to 100.

[b] 53.6% of presumptive PLWDH were captured by only 1 component of the case-finding algorithm, so the column percentages may not add to 100.

Comparison of potential PLWDH and presumptive PLWDH

Demographic characteristics varied between the potential PLWDH in the Medicaid data and the presumptive PLWDH identified from the CART analysis (Table 1). The proportion of females dropped from 50.0% to 37.5%. The proportion increased for persons aged 40–49 years (from 26.5% to 34.5%) and decreased for those aged 20–24 years (from 7.1% to 5.0%), 25–29 years (from 8.7% to 6.5%), and ≥60 years (from 11.6% to 5.2%). The proportions of non-Hispanic black (from 29.1% to 44.3%) and Hispanic (from 28.4% to 31.1%) presumptive PLWDH increased. The proportions of non-Hispanic whites (from 19.6% to 13.5%), non-Hispanic Asians/Pacific Islanders (from 6.3% (from 11.2% to 4.9%) decreased.

Roughly one-third of potential PLWDH and 88.0% of presumptive PLWDH were true cases (Table 3). A higher percentage of true cases were identified among the presumptive PLWDH for each of the demographic categories assessed.

Measures of validity for the CFA

The CFA had a sensitivity of 94.5%, a specificity of 94.4%, a PPV of 88.0%, and an NPV of 97.6% (Table 4). The measures of validity were similar between male and female presumptive PLWDH. Sensitivity was 94.0% for males and 95.5% for females, and specificity was 92.8% for males and 95.8% for females. The PPV was 88.9% for males and 86.4% for females, while the NPV was 96.1% for males and 98.7% for females.

The measures of validity for presumptive PLWDH were similar when they were compared across racial/ethnic categories (Table 4). Sensitivity ranged from 92.3% among non-Hispanic white presumptive PLWDH to 96.0% among Hispanic presumptive PLWDH. Specificity ranged from 91.7% among non-Hispanic black presumptive PLWDH to 98.9% among non-Hispanic American Indian/Alaska Native presumptive PLWDH. The PPV ranged from 76.6% among non-Hispanic Asian/Pacific Islander presumptive PLWDH to 94.7% among non-Hispanic American Indian/Alaska Native presumptive PLWDH. The NPV ranged from 96.1% among non-Hispanic black presumptive PLWDH to 99.6% among non-Hispanic Asian/Pacific Islander presumptive PLWDH.

The measures of validity were similar across age groups (Table 4). Sensitivity ranged from 91.7% among presumptive PLWDH aged ≥60 years to 95.5% for presumptive PLWDH aged 18–19 years. Specificity ranged from 93.0% among presumptive PLWDH aged 40–49 years to 96.4% among presumptive PLWDH aged 20–24 years. The PPV ranged from 73.6% among presumptive PLWDH aged ≥60 years to 92.2% among presumptive PLWDH aged 18–19 years. The NPV varied from 96.7% among presumptive PLWDH aged 40–49 years to 98.9% among presumptive PLWDH aged ≥60 years.

The measures of validity improved over the study years (Table 4). Sensitivity ranged from 81.2% in 2006 to 90.5%

Table 3. Numbers of Potential and Presumptive Persons Living With Diagnosed HIV in New York State Medicaid Data Matched to New York State HIV Surveillance Data, 2006–2014

Characteristic	Potential PLWDH			Presumptive PLWDH		
	Total No. of Persons	True Cases		Total No. of Persons	True Cases	
		No.	%		No.	%
Total	269,257	80,899	30.1	86,930	76,467	88.0
Year[a]						
2006	63,650	48,951	76.9	41,136	39,751	96.6
2007	59,145	44,793	75.7	40,750	39,739	97.5
2008	57,475	45,301	78.8	41,645	40,552	97.4
2009	61,421	46,361	75.5	42,857	41,628	97.1
2010	64,911	47,577	73.3	44,443	43,049	96.9
2011	68,344	48,601	71.1	45,144	43,968	97.4
2012	72,017	49,192	68.3	45,446	44,036	96.9
2013	100,892	49,739	49.3	47,067	44,199	93.9
2014	160,725	55,029	34.2	49,177	44,800	91.1
Sex[b]						
Female	134,759	29,445	21.9	32,556	28,115	86.4
Male	134,495	51,454	38.3	54,374	48,352	88.9
Age group, years						
Missing data	5,920	4,541	76.7	4,628	4,374	94.5
≤19	4,513	1,672	37.1	1,732	1,597	92.2
20–24	19,139	4,045	21.1	4,311	3,773	87.5
25–29	23,466	5,167	22.0	5,636	4,859	86.2
30–39	50,518	15,321	30.3	16,293	14,483	88.9
40–49	71,452	28,384	39.7	30,012	27,005	90.0
50–59	62,994	18,124	28.8	19,777	17,035	86.1
≥60	31,255	3,645	11.7	4,541	3,341	73.6
Race/ethnicity						
Non-Hispanic white	52,884	10,649	20.1	11,705	10,130	86.5
Non-Hispanic black	78,443	34,752	44.3	38,493	34,693	90.1
Non-Hispanic Asian/ Pacific Islander	17,037	779	4.6	981	751	76.6
Non-Hispanic American Indian/ Alaska Native	409	75	18.3	75	71	94.7
Hispanic	76,585	23,630	30.9	27,078	23,667	87.4
Other	13,840	3,687	26.6	4,376	3,631	83.0
Unknown	30,059	7,327	24.4	4,222	3,515	83.3

Abbreviations: HIV, human immunodeficiency virus; PLWDH, persons living with diagnosed HIV.
[a] Individuals were often identified in multiple years, so the column percentages may not add to 100.
[b] Sex was unknown for 3 persons.

in 2010 and 2011. Specificity increased, from 90.6% in 2006 to 95.9% in 2014. The PPV ranged from 91.1% in 2014 to 97.5% in 2007. The NPV increased from 59.1% in 2006 to 90.8% in 2014.

DISCUSSION

The CFA developed through CART analysis successfully identified PLWDH from the NYS Medicaid data. The CFA was designed with the CDC's 2014 updated guidelines for evaluating public health surveillance systems and validated using multiple measures of accuracy (sensitivity, specificity, PPV, and NPV). All measures of validity demonstrated that the CFA had a high likelihood of accurately capturing HIV status among persons listed in the Medicaid data.

German et al. (32) defined a flexible system as one that can adapt easily to changing information and stated that flexibility is best evaluated through retrospective analysis.

Table 4. Validity Statistics for a Case-Finding Algorithm Created to Identify Presumptive Persons Living With Diagnosed HIV in Medicaid Data, New York State, 2006–2014

Characteristic	Total No. of PLWDH	Match to Surveillance Data			Measure of Validity			
		No. of Persons	%	95% CI	Sensitivity, %	Specificity, %	PPV, %	NPV, %
Total	86,930	76,467	88.0	87.8, 88.2	94.5	94.4	88.0	97.6
Year[a]								
2006	41,136	39,751	96.6	96.5, 96.8	81.2	90.6	96.6	59.1
2007	40,750	39,739	97.5	97.4, 97.7	88.7	93.0	97.5	72.5
2008	41,645	40,552	97.4	91.2, 97.5	89.5	91.0	97.4	70.0
2009	42,857	41,628	97.1	97.0, 97.3	89.8	92.0	97.1	74.9
2010	44,443	43,049	96.9	96.7, 97.0	90.5	92.0	96.9	77.9
2011	45,144	43,968	97.4	97.3, 97.5	90.5	94.0	97.4	80.0
2012	45,446	44,036	96.9	96.7, 97.1	89.5	93.8	96.9	80.6
2013	47,067	44,199	93.9	93.7, 94.1	88.9	94.4	93.9	89.7
2014	49,177	44,800	91.1	90.9, 91.4	81.4	95.9	91.1	90.8
Sex[b]								
Female	32,556	28,115	86.4	86.0, 86.7	95.5	95.8	86.4	98.7
Male	54,374	48,352	88.9	88.7, 89.2	94.0	92.8	88.9	96.1
Age group, years								
Missing data	4,628	4,374	94.5	93.9, 95.2	96.3	81.6	94.5	87.1
≤19	1,732	1,597	92.2	90.9, 93.5	95.5	95.2	92.2	97.3
20–24	4,311	3,773	87.5	86.5, 88.5	93.3	96.4	87.5	98.2
25–29	5,636	4,859	86.2	85.3, 87.1	94.0	95.8	86.2	98.3
30–39	16,293	14,483	88.9	88.4, 89.4	94.5	94.9	88.9	97.6
40–49	30,012	27,005	90.0	89.6, 90.3	95.1	93.0	90.0	96.7
50–59	19,777	17,035	86.1	85.7, 86.6	94.0	93.9	86.1	97.5
≥60	4,541	3,341	73.6	72.3, 74.9	91.7	95.7	73.6	98.9
Race/ethnicity								
Non-Hispanic white	11,705	10,130	86.5	85.9, 87.2	92.3	96.4	86.5	98.0
Non-Hispanic black	38,493	34,693	90.1	89.8, 90.4	95.3	91.7	90.1	96.1
Non-Hispanic Asian/ Pacific Islander	981	751	76.6	73.9, 79.2	92.5	98.7	76.6	99.6
Non-Hispanic American Indian/ Alaska Native	75	71	94.7	89.6, 99.8	93.4	98.9	94.7	98.6
Hispanic	27,078	23,667	87.4	87.0, 87.8	96.0	93.9	87.4	98.2
Other	4,376	3,631	83.0	81.9, 84.1	93.7	93.1	83.0	97.6
Unknown	4,212	3,515	83.5	82.3, 84.6	85.6	95.3	83.5	95.9

Abbreviations: CI, confidence interval; HIV, human immunodeficiency virus; NPV, negative predictive value; PLWDH, persons living with diagnosed HIV; PPV, positive predictive value.

[a] Individuals were often identified in multiple years, so the column percentages may not add to 100.

[b] Sex was unknown for 3 persons.

The CFA was created using data from 2013 and retrospectively applied to data from 2006–2012, and then prospectively applied to 2014 data. This time period included the implementation of the Affordable Care Act (47), expanded Medicaid service, and the initiation of the HH Program in NYS (48). The CFA adapted well to these changes; each year, the sensitivity was above 80% and the PPV was above 90%. While the CFA presented in this paper is not as simple as the CFAs used by other researchers (9–11, 13, 22), the CFA presented here was easy to use, easy to understand, and replicable by year. Researchers should be able to use this algorithm to identify PLWDH in their states' Medicaid data.

The NPV indicated that 98% of PLWDH identified by the CFA as HIV-negative were truly negative (not diagnosed

with HIV) based on comparison with the gold standard NYS HIV surveillance system. Similarly, the PPV indicated that 88% of the PLWDH identified by the CFA were truly positive. Additionally, the high specificity (94%) observed with the CFA is important in an algorithm, because specificity and PPV are closely related (49) and a decrease in specificity leads to both a decrease in PPV and an increase in the number of people falsely identified by the algorithm. The overall sensitivity was 95%, indicating that the CFA identified nearly all of the true cases from the potential PLWDH.

These measures of validity varied but remained high for each of the racial/ethnic, sex, and age groups examined. This demonstrated that the CFA works well for different groups and is representative for all of the subgroups tested (Table 4). A representative system better characterizes the epidemiologic features of events and interventions (32).

Strengths and weaknesses

The CFA performed well with regard to important characteristics of a public health surveillance system, including specificity, sensitivity, PPV, and NPV, demonstrating patterns of flexibility and representativeness across several years. However, limitations of this study should be noted. In 2017, Ngugi et al. (50) used content analysis methods to systematically evaluate the data quality of the CDC's HIV surveillance data reports and identified problems in several areas, including potential incomplete and incorrect gaps in the HIV surveillance data. Because of NYS's stringent name-based reporting laws, which require all HIV-related laboratory test results (e.g., CD4 cell counts and percentages, viral load, genotypic resistance) for persons who reside or receive HIV-related care in the state to be reported to the NYSDOH, it is thought that the effect of underreporting to the NYS HIV surveillance system is minimal. However, there is potential that PLWDH may not be reported in NYS if 1) they received HIV testing or care under a research protocol approved by an institutional review board, 2) they received anonymous testing, 3) testing and diagnosis were performed in the federal health-care system (i.e., the Veterans Affairs system, the military, or federal prisons), or 4) they received testing from an out-of-state laboratory that failed to report this information to the NYSDOH (51). Even with these limitations, utilizing the HIV surveillance system as the gold standard for the development and evaluation of the CFA is thought to have provided the most accurate picture of PLWDH in NYS.

A second limitation of the CFA is that the analysis did not include all Medicaid data. Therefore, some potential PLWDH who had Medicaid coverage may not have been included in the analysis. However, it is believed that the number of missing cases was minimal, as the initial selection of potential PLWDH was designed to be as inclusive as possible.

In NYS, there were over 6 million people on Medicaid in 2013 (1), of whom fewer than 2% were diagnosed with HIV (34). Increasing the number of true negatives by millions through screening of the entire Medicaid database would increase the NPV and specificity to a level where the estimates would be close to 100%. This would be not only costly but unfeasible in a short period of time. In addition, the annual number of potential PLWDH in this algorithm almost tripled between 2006 and 2014, with a sharp increase in potential PLWDH occurring in 2013 and 2014. This correlates with the increasing number of enrollees in the NYS HH Program and was validated by examining the numbers of HH outreach rate codes by year. This change increased the numbers of true negatives in the analysis but had minimal impact on the number of presumptive PLWDH, adding confidence to the assumption that running the CFA using the entire Medicaid data set would have a limited impact.

Thirdly, the NYS HIV surveillance system and the Medicaid data system collect demographic information differently, making it difficult to compare demographic distributions of the CFA with those of the HIV surveillance system. An example is that the HIV surveillance system collects information on both sex at birth and current sex; however, Medicaid data do not differentiate between these 2 variables. With regard to demographic characteristics that provided a direct comparison, the CFA and the HIV surveillance system yielded similar proportions (data not shown (33)). This adds to the confidence that the CFA was not biased on the basis of individual demographic information.

Fourthly, a limit to the generalizability of the CFA is that it uses NYS-specific Medicaid rate codes. However, the impact of these codes is minimal, as 46.5% of the presumptive PLWDH identified by the CFA had claims in 2 or more of the categories, and 91.8% of the presumptive PLWDH identified using rate codes had claims in 2 or more of the categories. It is also important to note that algorithms developed for internal programmatic purposes may need to consider the inclusion of program-specific billing codes.

Lastly, the CFA used ICD-9-CM diagnosis codes from the Medicaid billing data. As of October 1, 2015, the ICD-9-CM was replaced by the *International Classification of Diseases, Tenth Revision, Clinical Modification* (52). To properly apply this CFA to Medicaid billing data after September 2015, the ICD-9-CM diagnosis codes would need a cross-walk comparison with *International Classification of Diseases, Tenth Revision, Clinical Modification* codes.

Implications

Utilizing billing data is an unobtrusive method of analysis; therefore, the privacy risk posed to PLWDH whose information is being used is minimal. The successful derivation and application of the CFA have promising implications for future HIV analyses. Utilizing the CFA could help improve both clinical research and program evaluation, leading to better care for PLWDH. Applied to the whole Medicaid data set, the CFA may have the ability to look for PLWDH who have not been reported to the NYSDOH HIV surveillance system. Identifying these potential gaps in reporting would allow PLWDH to be better surveilled, further bolstering the HIV surveillance system.

Conclusion

Medicaid is a robust data system, containing treatment information not captured in the NYS HIV surveillance

system, such as claims for health-care provider visits, tests, and antiretroviral medications. Researchers have employed other algorithms to answer questions using Medicaid data, but those algorithms have not been evaluated with respect to their sensitivity, specificity, PPV, and NPV by comparison with the gold standard, an HIV surveillance system. The CFA described in this paper was evaluated against this gold standard and performed well. It was deemed simple, flexible, and representative and had high sensitivity and PPV. It successfully identified PLWDH in the Medicaid data, improving the utility of those data for assessing health outcomes and programmatic interventions such as HHs. Findings from this paper provide future researchers with a tool for identifying PLWDH through Medicaid data, potentially helping to increase linkage with HIV care and retention in care among PLWDH and decreasing future transmission.

ACKNOWLEDGMENTS

Author affiliations: Bureau of HIV/AIDS Epidemiology, AIDS Institute, New York State Department of Health, Albany, New York (Sarah E. Macinski, Jayleen K. L. Gunn, Bridget J. Anderson); and National Center for Alcohol and Substance Abuse, New York, New York (Mona Goyal, Charles Neighbors, Rajeev Yerneni).

This project was partially funded by the National Institute on Drug Abuse (grant 1R01DA038193) and the Centers for Disease Control and Prevention (grant PS18-1802).

We thank Dr. Franklin Laufer, Wendy Levey, and Nicole Levesque of the Office of Medicaid Policy and Programs, AIDS Institute, New York State Department of Health, for technical support regarding the development and implementation of a case-finding algorithm. We also thank Brenda Moncur, Joyce Chicoine, and Kathy Harper of the New York State Department of Health's Bureau of HIV/AIDS Epidemiology for their assistance in matching Medicaid data with HIV surveillance data. We additionally thank Dr. Daniel Gordon for his assistance at the inception of this study. Lastly, we thank Deepa Rajulu and the surveillance staff of the Bureau of HIV/AIDS Epidemiology for their assistance with clerical review of the match results.

Conflict of interest: none declared.

REFERENCES

1. Henry J. Kaiser Family Foundation. *Medicaid and HIV*. San Francisco, CA: Henry J. Kaiser Family Foundation; 2019. http://www.kff.org/hivaids/fact-sheet/medicaid-and-hiv/. Accessed August 28, 2017.
2. Kates J, Garfield R, Young K, et al. *Assessing the Impact of the Affordable Care Act on Health Insurance Coverage of People with HIV*. San Francisco, CA: Henry J. Kaiser Family Foundation; 2014. https://www.kff.org/hivaids/issue-brief/assessing-the-impact-of-the-affordable-care-act-on-health-insurance-coverage-of-people-with-hiv/. Accessed January 18, 2014.
3. Rothbard AB, Metraux S, Blank MB. Cost of care for Medicaid recipients with serious mental illness and HIV infection or AIDS. *Psychiatr Serv*. 2003;54(9):1240–1246.
4. Chesnut TJ, Laufer FN, Carrascal AF, et al. An expenditure analysis of high-cost Medicaid recipients with HIV disease in New York State. *J Health Care Poor Underserved*. 2011; 22(1):330–345.
5. Fleishman JA, Monroe AK, Voss CC, et al. Expenditures for persons living with HIV enrolled in Medicaid, 2006–2010. *J Acquir Immune Defic Syndr*. 2016;72(4):408–415.
6. Zhang S, McGoy SL, Dawes D, et al. The potential for elimination of racial-ethnic disparities in HIV treatment initiation in the Medicaid population among 14 southern states. *PloS One*. 2014;9(4):e96148.
7. Berry S, Fleishman J, Moore RD, et al. Thirty-day hospital readmissions for adults with and without HIV infection. *HIV Med*. 2016;17(3):166–177.
8. Berry SA, Fleishman JA, Moore RD, et al. Trends in reasons for hospitalization in a multisite United States cohort of persons living with HIV, 2001–2008. *J Acquir Immune Defic Syndr*. 2012;59(4):368–375.
9. Byrd KK, Furtado M, Bush T, et al. Evaluating patterns in retention, continuation, gaps, and re-engagement in HIV care in a Medicaid-insured population, 2006–2012, United States. *AIDS Care*. 2015;27(11):1387–1395.
10. Davis AC, Watson G, Pourat N, et al. Disparities in CD4+ T-lymphocyte monitoring among human immunodeficiency virus-positive Medicaid beneficiaries: evidence of differential treatment at the point of care. *Open Forum Infect Dis*. 2014; 1(2):ofu042.
11. Landovitz RJ, Desmond KA, Gildner JL, et al. Quality of care for HIV/AIDS and for primary prevention by HIV specialists and nonspecialists. *AIDS Patient Care STDS*. 2016;30(9):395–408.
12. Schneeweiss S, Avorn J. A review of uses of health care utilization databases for epidemiologic research on therapeutics. *J Clin Epidemiol*. 2005;58(4): 323–337.
13. 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(3):567–574.
14. Crystal S, Akincigil A, Bilder S, et al. Studying prescription drug use and outcomes with Medicaid claims data: strengths, limitations, and strategies. *Med Care*. 2007;45(10 suppl 2): S58–S65.
15. Iezzoni LI. Assessing quality using administrative data. *Ann Intern Med*. 1997;127(8):666–674.
16. Gandhi S, Salmon JW, Kong SX, et al. Administrative databases and outcomes assessment: an overview of issues and potential utility. *J Manag Care Spec Pharm*. 1999;5(3): 215–222.
17. Seiber EE. Physician code creep: evidence in Medicaid and state employee health insurance billing. *Health Care Financ Rev*. 2007;28(4):83–93.
18. Solberg LI, Engebretson KI, Sperl-Hillen JM, et al. Are claims data accurate enough to identify patients for performance measures or quality improvement? The case of diabetes, heart disease, and depression. *Am J Med Qual*. 2006;21(4):238–245.
19. Tang PC, Ralston M, Arrigotti MF, et al. Comparison of methodologies for calculating quality measures based on administrative data versus clinical data from an electronic health record system: implications for performance measures. *J Am Med Inform Assoc*. 2007;14(1):10–15.

20. Angier H, Gold R, Gallia C, et al. Variation in outcomes of quality measurement by data source. *Pediatrics*. 2014;133(6): e1676–e1682.
21. Rector TS, Wickstrom SL, Shah M, et al. Specificity and sensitivity of claims-based algorithms for identifying members of Medicare+Choice health plans that have chronic medical conditions. *Health Serv Res*. 2004;39(6): 1839–1858.
22. Walkup JT, Wei W, Sambamoorthi U, et al. Sensitivity of an AIDS case-finding algorithm: who are we missing? *Med Care*. 2004;42(8):756–763.
23. Leibowitz AA, Desmond K. Do only 21% of HIV-positive Medicaid enrollees link to treatment? Challenges in interpreting Medicaid claims data [letter]. *Sex Transm Dis*. 2013;40(7):582.
24. Peabody JW, Luck J, Jain S, et al. Assessing the accuracy of administrative data in health information systems. *Med Care*. 2004;42(11):1066–1072.
25. Bright RA, Avorn J, Everitt DE. Medicaid data as a resource for epidemiologic studies: strengths and limitations. *J Clin Epidemiol*. 1989;42(10):937–945.
26. Bailey SR, Heintzman JD, Marino M, et al. Measuring preventive care delivery: comparing rates across three data sources. *Am J Prev Med*. 2016;51(5):752–761.
27. Blank MB, Mandell DS, Aiken L, et al. Co-occurrence of HIV and serious mental illness among Medicaid recipients. *Psychiatr Serv*. 2002;53(7):868–873.
28. Prince JD, Walkup J, Akincigil A, et al. Serious mental illness and risk of new HIV/AIDS diagnoses: an analysis of Medicaid beneficiaries in eight states. *Psychiatr Serv*. 2012; 63(10):1032–1038.
29. Keyes M, Andrews R, Mason M-L. A methodology for building an AIDS research file using Medicaid claims and administrative data bases. *J Acquir Immune Defic Syndr*. 1991;4(10):1015–1024.
30. Buchacz K, Baker RK, Moorman AC, et al. Rates of hospitalizations and associated diagnoses in a large multisite cohort of HIV patients in the United States, 1994–2005. *AIDS*. 2008;22(11):1345–1354.
31. Zhong VW, Pfaff ER, Beavers DP, et al. Use of administrative and electronic health record data for development of automated algorithms for childhood diabetes case ascertainment and type classification: the SEARCH for Diabetes in Youth Study. *Pediatr Diabetes*. 2014;15(8): 573–584.
32. German RR, Lee L, Horan J, et al. Updated guidelines for evaluating public health surveillance systems. *MMWR Recomm Rep*. 2001;50(RR-13):1–35.
33. Bureau of HIV/AIDS Epidemiology, AIDS Institute, New York State Department of Health. *New York State HIV/AIDS Annual Surveillance Report for Cases Diagnosed Through December 2016*. Albany, NY: New York State Department of Health; 2017. https://www.health.ny.gov/diseases/aids/general/statistics/annual/2016/2016_annual_surveillance_report.pdf. Accessed February 21, 2018.
34. Henry J. Kaiser Family Foundation. *Medicaid Enrollment and Spending on HIV/AIDS*. San Francisco, CA: Henry J. Kaiser Family Foundation; 2017. http://kff.org/hivaids/state-indicator/enrollment-spending-on-hiv/. Accessed May 24, 2017.
35. Centers for Disease Control and Prevention. Revised surveillance case definition for HIV infection—United States, 2014. *MMWR Recomm Rep*. 2014;63(RR-3):1–10.
36. New York State Department of Health. *Health Home Eligibility Policy*. Albany, NY: New York State Department of Health; 2014. https://www.health.ny.gov/health_care/medicaid/program/medicaid_health_homes/docs/09-23-2014_hh_eligibility_policy.pdf. Accessed May 5, 2017.
37. Antoniou T, Zagorski B, Loutfy MR, et al. Validation of case-finding algorithms derived from administrative data for identifying adults living with human immunodeficiency virus infection. *PLoS One*. 2011;6(6):e21748.
38. New York State Department of Health. *Quality Assurance Reporting Requirements: Technical Specifications Manual (2015 QARR/HEDIS® 2015)*. Albany, NY: New York State Department of Health; 2015.
39. Walkup J, Akincigil A, Hoover DR, et al. Use of Medicaid data to explore community characteristics associated with HIV prevalence among beneficiaries with schizophrenia. *Public Health Rep*. 2011;126(3 suppl):89–101.
40. Mrus JM, Shireman TI, Moomaw C, et al. Development of an HIV research database using Medicaid claims data. *AIDS Public Policy J*. 2001;16(3-4):48–54.
41. HIV/AIDS Bureau, Health Resources and Services Administration, US Department of Health and Human Services. *Improving Health Outcomes Through Data Utilization: 2016 Ryan White HIV/AIDS Program Highlights*. Washington, DC: US Department of Health and Human Services; 2016. https://hab.hrsa.gov/sites/default/files/hab/data/biennialreports/2016_HRSA_Biennial_Report.pdf. Accessed August 28, 2017.
42. Office of Medicaid Policy and Programs, AIDS Institute, New York State Department of Health. *AIDS Institute Screening Criteria Used to Identify Records for Match With HIV Registry*. Albany, NY: New York State Department of Health; 2014.
43. HIV/AIDS Bureau, Health Resources and Services Administration, US Department of Health and Human Services. *HIV/AIDS Bureau Performance Measures*. Washington, DC: US Department of Health and Human Services; 2015. https://hab.hrsa.gov/sites/default/files/hab/clinical-quality-management/coremeasures.pdf. Accessed February 2, 2017.
44. National Quality Center. *Guideline-Based Quality Indicators for HIV Care*. New York, NY: National Quality Center; 2008. http://nationalqualitycenter.org/resources/guideline-based-quality-indicators-for-hiv-care-pdf. Accessed February 2, 2017.
45. Averill RF, Goldfield NI, Eisenhandler J, et al. *Development and Evaluation of Clinical Risk Groups (CRGs)*. Wallingford, CT: 3M Health Information Systems; 1999. https://www.semanticscholar.org/paper/Development-and-Evaluation-of-Clinical-Risk-Groups-Averill-Goldfield/505edb4f2cf344b206aadc4a351de4ac956c9a60. Accessed May 7, 2017.
46. Morgan J. *Classification and Regression Tree Analysis*. (Technical Report no. 1). Boston, MA: School of Public Health, Boston University; 2014. http://www.bu.edu/sph/files/2014/05/MorganCART.pdf. Accessed August 14, 2017.
47. 111th US Congress. *The Patient Protection and Affordable Care Act*. 42 USC §18001. 2010.
48. New York State Department of Health. *Health Homes Provider Manual: Billing Policy and Guidance*. Albany, NY: New York State Department of Health; 2013. https://www.health.ny.gov/health_care/medicaid/program/medicaid_health_homes/docs/hh_prov_manual.pdf. Accessed June 29, 2017.
49. Fenton JJ, Onega T, Zhu W, et al. Validation of a Medicare claims-based algorithm for identifying breast cancers detected at screening mammography. *Med Care*. 2016;54(3):e15–e22.

50. Ngugi BK, Harrington B, Porcher EN, et al. Data quality shortcomings with the US HIV/AIDS surveillance system. *Health Informatics J*. 2019;25(2): 304–314.
51. New York State Department of Health. HIV reporting and partner notification questions and answers. https://www.health.ny.gov/diseases/aids/providers/regulations/reporting_and_notification/question_answer.htm#onethru16. Accessed November 8, 2017.
52. National Center for Health Statistics, Centers for Disease Control and Prevention. *International Classification of Diseases* (ICD-10-CM/PCS) transition—background. 2015. https://www.cdc.gov/nchs/icd10cm_pcs_background.htm. Accessed November 7, 2019.