Journal of the American Medical Informatics Association, 25(2), 2018, 150–157
doi: 10.1093/jamia/ocx061
Advance Access Publication Date: 22 June 2017
Research and Applications

Research and Applications

Development and validation of an electronic medical record (EMR)-based computed phenotype of HIV-1 infection

Devon W Paul,[1,*] Nigel B Neely,[2,*] Meredith Clement,[2,3] Isaretta Riley,[1] Mashael Al-Hegelan,[1] Matthew Phelan,[2] Monica Kraft,[4] David M Murdoch,[1] Joseph Lucas,[2] John Bartlett,[3] Mehri McKellar,[3] and Loretta G Que[1]

[1]Division of Pulmonary, Allergy, and Critical Care Medicine, Duke University, Durham, NC, USA, [2]Duke Clinical Research Institute, Durham, NC, USA, [3]Division of Infectious Diseases, Duke University, Durham, NC, USA and [4]Department of Internal Medicine, University of Arizona, Tucson, AZ, USA

*Corresponding Author: Devon Paul, Duke Health Systems, Durham, NC 27710, PO Box 3132, USA. E-mail: devon.paul@duke.edu. Phone: +1-919-684-8111

Received 7 September 2016; Revised 27 March 2017; Accepted 18 May 2017

ABSTRACT

Background: Electronic medical record (EMR) computed algorithms allow investigators to screen thousands of patient records to identify specific disease cases. No computed algorithms have been developed to detect all cases of human immunodeficiency virus (HIV) infection using administrative, laboratory, and clinical documentation data outside of the Veterans Health Administration. We developed novel EMR-based algorithms for HIV detection and validated them in a cohort of subjects in the Duke University Health System (DUHS).
Methods: We created 2 novel algorithms to identify HIV-infected subjects. Algorithm 1 used laboratory studies and medications to identify HIV-infected subjects, whereas Algorithm 2 used International Classification of Diseases, Ninth Revision (ICD-9) codes, medications, and laboratory testing. We applied the algorithms to a well-characterized cohort of patients and validated both against the gold standard of physician chart review. We determined sensitivity, specificity, and prevalence of HIV between 2007 and 2011 in patients seen at DUHS.
Results: A total of 172 271 patients were detected with complete data; 1063 patients met algorithm criteria for HIV infection. In all, 970 individuals were identified by both algorithms, 78 by Algorithm 1 alone, and 15 by Algorithm 2 alone. The sensitivity and specificity of each algorithm were 78% and 99%, respectively, for Algorithm 1 and 77% and 100% for Algorithm 2. The estimated prevalence of HIV infection at DUHS between 2007 and 2011 was 0.6%.
Conclusions: EMR-based phenotypes of HIV infection are capable of detecting cases of HIV-infected adults with good sensitivity and specificity. These algorithms have the potential to be adapted to other EMR systems, allowing for the creation of cohorts of patients across EMR systems.

Key words: electronic medical record, diagnostic algorithm, HIV

BACKGROUND AND SIGNIFICANCE

The conversion from paper to electronic medical records (EMRs) by health care systems and networks for clinical documentation is rapidly being done at medical centers and hospitals around the United States. While useful for clinical documentation, this transition to EMRs also yields an unprecedented opportunity to combine and analyze large clinical and ancillary datasets across multiple health care systems for research and quality improvement purposes. The ability to automate this process builds upon traditional retrospective methodology (so-called chart reviews), which involves manually

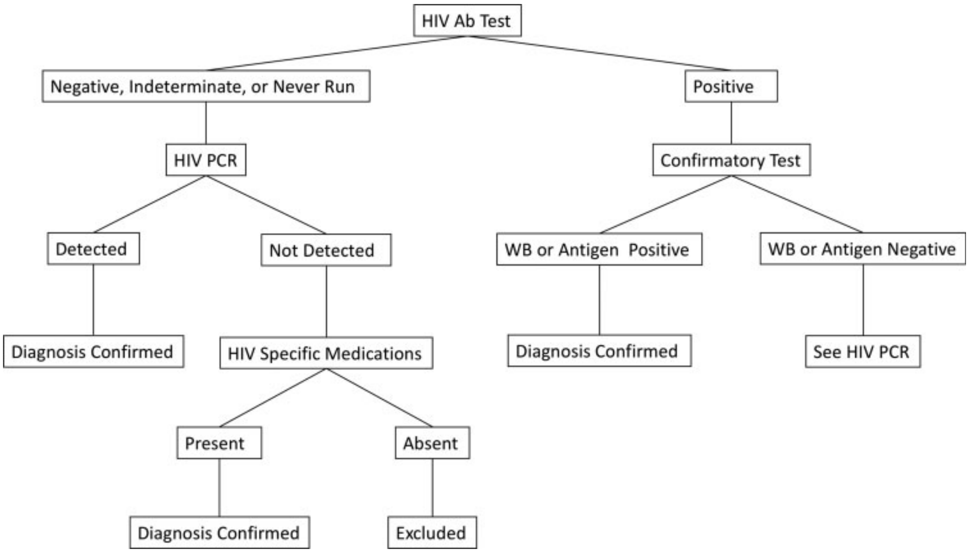

Figure 1. Computable phenotype Algorithm 1. Modified from the most recent Centers for Disease Control algorithm for HIV diagnosis, this algorithm uses lab tests to confirm the presence of HIV infection.

reviewing individual files, often requiring multiple individuals collaborating over long periods of time.[1] Compiling a large sample with a given disease or characteristic could take months and hundreds of person-hours using these traditional methods, limiting the ability of researchers to conduct important clinical studies.[1]

With the advent and adoption of the EMR, researchers are now able to rapidly identify potential disease cases for clinical studies.[2–4] To do this effectively and fully leverage the benefits of an electronic system, computable phenotype algorithms are needed that can search across billing data, laboratory data, and clinical documentation in order to perform case detection. These computable phenotype algorithms can be conceived in a manner to have high sensitivity and specificity for identifying individual subjects' true disease status using methods borrowed from routine clinical care.[2–4]

HIV infection is one disease that lends itself to this algorithm-based case detection, given the reliance on laboratory-based testing.[5] Previous studies that have attempted to identify HIV-infected patients in large datasets have had significant limitations. Many have focused on preselected populations, either Medicare and/or Medicaid enrollees[6–9] or patients receiving care from the Veterans Health Administration.[10,11] Due to their design, these studies lack generalizability outside the government health sector. Their utility is further limited by the use of administrative data alone for case detection, subjecting these studies to the systematic error inherent in this type of data, including missing codes if HIV infection is not the primary issue being addressed in an encounter, or inappropriate use of HIV-related ICD-9 codes to cover diagnostic testing or HIV-prevention counseling.[12] Subsequent algorithms have improved on these methodologic issues, but were targeted to identify specific subgroups of patients, including foreign-born individuals who have known HIV infection[13] and only new cases of HIV infection in a dataset,[8] or determining which patients are not known to be infected with HIV but are at risk of contracting the disease.[14]

In order to maximally utilize the features of the EMR to develop a virtual cohort of HIV-infected patients in a large nongovernmental dataset, we set out to develop novel computable phenotype algorithms for case detection of these patients. We also sought to validate the algorithms, determine their sensitivity and specificity, and use them to determine the prevalence of HIV infection in a large preexisting cohort.

MATERIALS AND METHODS

Study setting

This study was performed within the Duke University Health System in Durham, North Carolina, using an existing, previously described cohort comprising all adults (18+ years old) who had encounters with the health system between 2007 and 2011 and could be identified as Durham County residents at the end of 2011.[15] Within this cohort, a total of 4 408 919 unique patient encounters with an average of 881 784 encounters per year were identified. Data were extracted using the Duke Enterprise Data Unified Content Explorer system, a proprietary, web-based tool designed to interface with the health system's central data repository.[16] Accessible data included clinical, billing, and demographic information for patients across the health system's 3 acute care hospitals as well as ambulatory primary care and specialty clinics. Information in the repository is updated in an ongoing prospective fashion, and includes encounter information preceding the implementation of Epic software as the institution's unified EMR. Data available after the cohort time period were also included. Approval for the study was provided by the Duke Institutional Review Board for Clinical Investigation (Pro00057348).

EMR definition – multiple pathways

In order to maximize the utility and performance characteristics of the algorithms, 2 separate algorithms utilizing complementary approaches to HIV case detection were created. A panel of physicians, including experts in the diagnosis and treatment of HIV infection, provided recommendations for the development of these algorithms and approved their final versions. Algorithm 1 (Figure 1) is a laboratory-based approach modified from the HIV diagnostic algorithm in practice during the period of our cohort enrollment, from 2007 to 2011.[17] For Algorithm 1, the expert panel concluded that a diagnosis of HIV could be reasonably applied if a patient had

1 of 2 additional features. The first was a positive nucleic acid test (HIV RNA or DNA polymerase chain reaction [PCR]), Western blot, or p24 antigen test, based on testing recommendations within our cohort time frame.[17] The second was HIV-specific medications in the patient's prescribed medication list. The addition of HIV-specific medications was felt to represent evidence of contact with a health care provider with expertise in HIV diagnosis and treatment. Recognizing that some medications used to treat HIV infection are also used today for hepatitis B, we looked for medication combinations specific to HIV therapy (Table 1). Monotherapy with lamivudine, emtricitabine, or tenofovir was thus not considered a positive HIV medication unless prescribed in combination with an additional drug. By excluding single drugs only, we recognize that we may have included HIV-negative patients on dual therapy with emtricitabine and tenofovir for pre-exposure prophylaxis (PrEP) and post-exposure prophylaxis (PEP). Additionally, we were unable to exclude patients on triple therapy for PEP as we did not have access to prescription records.

Table 1. HIV-specific medications by generic and trade name with example metadata permutations

Drug Class	Generic Name	Trade Name
Nucleoside reverse transcriptase inhibitor	Abacavir (ABC)	Ziagen
	Didanosine (DDI)	Videx
	Emtricitabine[a,b,c] (FTC)	Emtriva
	Lamivudine[a,c] (3TC)	Epivir
	Stavudine (D4T)	Zerit
	Tenofovir disoproxil fumarate[a,b,c] (TDF)	Viread
	Zidovudine (AZT)[c]	Retrovir
Non-nucleoside reverse transcriptase inhibitor	Efavirenz	Sustiva
	Etravirine	Intelence
	Nevirapine	Viramune
	Rilpivirine[c]	Edurant
Protease inhibitor (PI)	Atazanavir[c]	Reyataz
	Darunavir[c]	Prezista
	Fosampre[c]	Lexiva
	Indinavir	Crixivan
	Nelfinavir	Viracept
	Ritonavir[c]	Norvir
	Saquinavir	Invirase
	Tipranavir	Aptivus
Integrase inhibitor	Dolutegravir[c]	Tivicay
	Elvitegravir	Vitekta
	Raltegravir[c]	Isentress
Fusion inhibitor	Enfuvirtide	Fuzeon
Entry inhibitor	Maraviroc	Selzentry
Booster	Cobicistat	Tybost
Combination pills	Tenofovir/Emtricitabine (TDF/FTC)[a,c]	Truvada
	Tenofovir alafenamide/Emtricitabine (TAF/FTC)[c]	Descovy
	Efavirenz + TDF/FTC	Atripla
	Rilpivirine + TDF/FTC	Complera
	Rilpivirine + Tenofovir alafenamide/FTC	Odefsey
	Elvitegravir + Cobicistat + TDF/FTC	Stribild
	Elvitegravir + Cobicistat + Tenofovir alafenamide/FTC	Genvoya
	Dolutegravir + ABC/3TC	Triumeq
	Zidovudine + Lamivudine (AZT/3TC)[c]	Combivir
	Abacavir + Lamivudine (ABC/3TC)	Epzicom
	ABC + AZT + 3TC	Trizivir
	Lopinavir/Ritonavir[c]	Kaletra

Comprehensive list of the trade names and generic names of HIV-specific medications. Please note that not all of these medications were available during 2007–2011 but were included, as data outside the cohort time period were utilized. Those medications with additional clinical uses including treatment for hepatitis B as well as PEP and/or PrEP are noted as follows: [a]medications also used for hepatitis B virus; [b]medications also used for PrEP; [c]medications also used for PEP.

Algorithm 2 (Figure 2) was designed to detect those patients not identified by Algorithm 1 due to incomplete medical records, because they had fragmented access to care or had diagnostic labs performed outside of the study's health system. Both of these scenarios led to HIV-infected patients not having results for all labs required for Algorithm 1 accessible within the study's health system. Algorithm 2, a complementary algorithm, includes a combination of ICD-9 billing codes, laboratory investigations, and medication lists to create a more "real-world" clinical approach to HIV case detection. Borrowing from traditional retrospective research methodology, Algorithm 2 begins with an ICD-9 code for an HIV-related illness (Table 2), followed by a confirmatory medication list (Table 1) or lab test (HIV antibody, HIV Western blot, or HIV PCR).

Implementation

Once the 2 algorithms were finalized and approved by the physician panel, individual components of each algorithm were converted into programmable, searchable code. Given the evolving nature of data storage in the central repository over time due to changes in HIV testing technology and new medication development, a comprehensive search of the metadata was necessary to identify all permutations of pertinent laboratory results and medications. The Duke Enterprise Data Unified Content Explorer system was used for this search, and a list of all permutations of pertinent laboratory test names, ICD-9 codes, and medication names was compiled (Tables 1 and 2). ICD-9 codes were readily identified and implemented without the need for manual chart review. In contrast, searches for medication names and laboratory test names yielded thousands of results for a small number of pertinent items. The initial metadata search results were manually reviewed by the study team to isolate pertinent test and medication names. By initially reviewing at the metadata level (as opposed to the clinical documentation level), we were able to ensure that each unique lab value and medication label in our data extract was reviewed by a trained clinician in an efficient and systematic manner. In those cases where manual review of the metadata was insufficient to conclude the name of a laboratory test or medication, due either to misspelling during the clinical encounter documentation or formatting issues when transferring from clinical documentation to the data repository, manual chart review of the clinical documentation was performed.

Following the initial review of laboratory names, the list of pertinent permutations required a second manual review by study personnel in order to codify the results. For HIV antibody testing, only "positive" was considered to be positive, whereas "negative" and "indeterminate" were considered to be negative and did not contribute to a diagnosis. In rare instances, this approach does have the potential to introduce false negatives if the patient has acute HIV infection, as a positive HIV PCR may not have been performed if a negative or indeterminate HIV antibody test was recorded. For HIV PCR testing, we included both quantitative measures of HIV viral load and results from HIV genotyping and resistance pattern

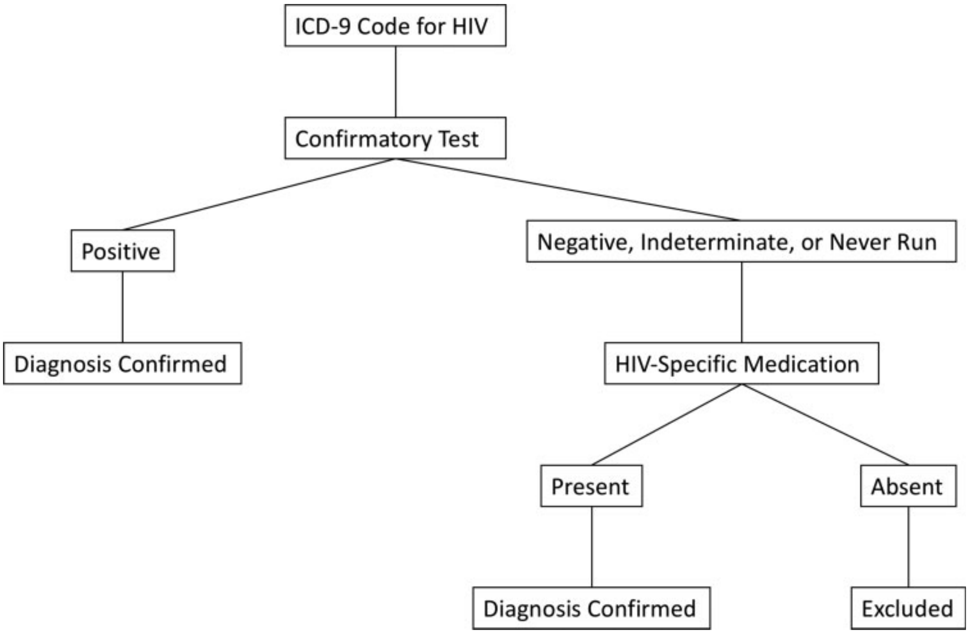

Figure 2. Computable phenotype Algorithm 2. This algorithm is designed to capture individuals lacking components of Algorithm 1 required for lab-based diagnosis. Should a confirmatory test be required, the algorithm returns to this common confirmation pathway outlined in Algorithm 1.

Table 2. ICD-9 Codes for HIV-related illnesses

ICD-9 Code	Description
042	Human immunodeficiency virus (HIV) disease
042.0	HIV and specific infection
042.1	HIV causing other infection
042.2	HIV with neoplasm
042.9	Unspecified acquired immunodeficiency syndrome (AIDS)
043	HIV causing condition necrotizing enterocolitis (NEC)
043.0	HIV lymphadenopathy
043.1	HIV causing central nervous system (CNS) disease
043.2	HIV causing other disorders involving the immune mechanism
043.3	HIV causing disease NEC
043.9	AIDS-related complex not otherwise specified (NOS)
044	Other HIV infection
044.0	HIV with acute infection
044.9	HIV infection NOS
079.53	HIV, type 2
795.78	Positive serologic findings; HIV
V08	Asymptomatic HIV infection status
795.71	Nonspecific serologic evidence of HIV

Comprehensive list of HIV-specific ICD-9 codes and their descriptions used by Algorithm 2 for case detection.

analyses as equivalent tests for case detection purposes. For viral load, a detectable level was considered "positive" regardless of level, and an undetectable level was considered "negative." This definition also has the potential to introduce false negatives, as an undetectable viral load can be seen in patients on effective antiretroviral therapy or in "elite controllers," HIV-infected persons with undetectable viremia off antiretrovirals. For genotyping, a reported result was considered to be "positive," as this test can only be performed when there is detectable HIV present.

Once all components of the 2 algorithms were codified, we applied the algorithms to the data for our study population. Data collected outside of the 5-year enrollment period for the cohort were included in the analysis for those subjects enrolled in the cohort.

Validation

Following implementation of our computed algorithms, a validation step was performed to determine test characteristics, including sensitivity and specificity. Additional information, including potential strengths and weaknesses of each algorithm, were also identified. The gold standard used for comparison in the validation step was physician manual chart review. Because the prevalence of HIV infection in the general population is low, simple random sampling from the entire population would require a prohibitively large sample size. In order to decrease the required sample size for a precision of 0.05% around our estimates, we stratified our population. We believe that the algorithms will perform differently across strata, and this performance variability drives the sampling scheme. Patients were stratified by the number of algorithms positive (both, one, or none), and those with no algorithm positive were further stratified by meeting any algorithm components as well as high vs low risk for HIV. High-risk individuals were defined as those with an ICD-9 code for hepatitis C, any sexually transmitted infection, or tuberculosis (see Supplementary Table A1). We then made assumptions for the true positive fraction and false positive fraction for each stratum. Our true positive fraction estimates were 0.975 for both algorithms positive and 1 algorithm positive strata, and 0.000 for single component positive, high-risk, and no component positive strata. Our false positive fraction estimates were 0.000 for both algorithms positive, 0.950 for 1 algorithm positive, 0.010 for single component positive and high risk, and 0.00000001 for no component positive strata. This estimate for 1 algorithm positive is high, signifying a strong belief in the independence of the algorithms to correctly identify HIV-infected individuals. Using these estimates, a stratified random sample size of 171 charts was selected to provide an overall point estimate for prevalence of HIV infection of ±0.05%. Our sampling

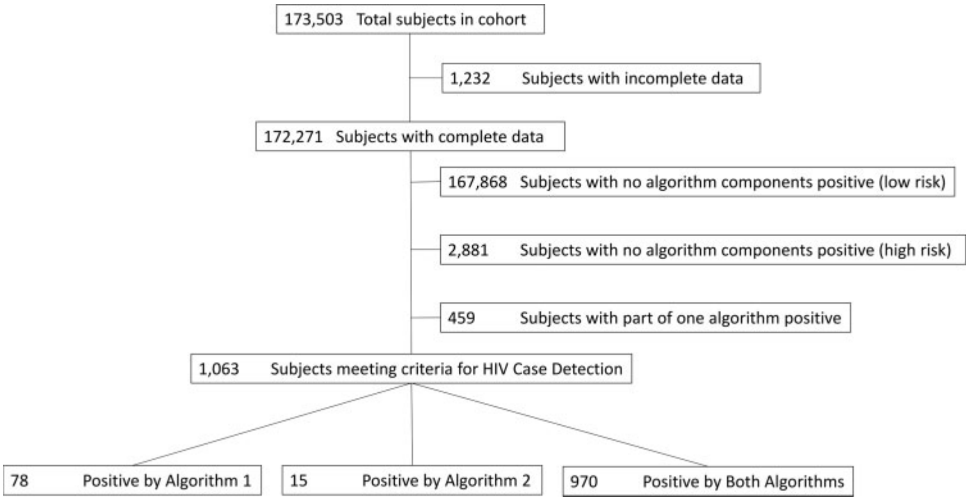
173,503 Total subjects in cohort
1,232 Subjects with incomplete data
172,271 Subjects with complete data
167,868 Subjects with no algorithm components positive (low risk)
2,881 Subjects with no algorithm components positive (high risk)
459 Subjects with part of one algorithm positive
1,063 Subjects meeting criteria for HIV Case Detection
78 Positive by Algorithm 1
15 Positive by Algorithm 2
970 Positive by Both Algorithms

Figure 3. Flow diagram of case detection through the study. Of the 172 271 subjects with complete data, 1063 met criteria by 1 or both algorithms for HIV case detection. The contribution of each algorithm to overall HIV case detection is outlined.

approach, designed to optimize estimation of sensitivity and specificity, naturally induces verification bias,[18] which we accounted for when estimating the operating characteristics of each phenotype by using methods from Begg and Greenes.[19] By reweighing our estimates at the stratum level to correspond to the weights derived in our sampling scheme, we corrected for the selection bias that we introduced. Simulation studies have been performed on these approximations for sample sizes of 100, 500, and 1000, with the coverage probabilities being close to the nominal level, which we expect should cover a sample size such as ours within this range.[19]

Seven physician study team members participated in the review. Reviewers were given full access to the clinical EMRs of all patients, which included access to results posted prior to EMR implementation, as well as data linked to other health care systems via Epic's Care Everywhere function and to records scanned in from outside facilities that may not have been accessible by the algorithms. Charts were coded as definitely HIV-infected (positive HIV antibody or PCR), possibly HIV-infected (HIV on problem list plus HIV medication or 2+ infectious disease clinic notes with HIV as a diagnosis), or HIV-negative (not meeting the criteria for definite or possible). These criteria were approved by our physician panel in order to maximize specificity and minimize false positives. Each chart was reviewed separately by 2 investigators blinded to both the results of the algorithms and the decision of the other reviewer. All cases of disagreement were adjudicated by a third reviewer blinded in similar fashion to algorithm results and individual reviewer responses but unblinded as to status as a third reviewer for adjudication. Final HIV status was determined by the third reviewer in cases of disagreement. For the purpose of statistical analysis, a binary construct of "HIV-positive" and "HIV-negative" was created, with definitely and possibly HIV-infected results being combined into the "HIV-positive" category. The sensitivity and specificity of each algorithm was calculated with the chart review being treated as the gold standard.

RESULTS

Algorithms

We identified 172 271 individual patients with complete data, comprising our study population (Figure 3). An additional 1232 individuals had insufficient data to be included in the pool of participants. After applying the finalized algorithms to our data extract, we identified 1063 patients who met the criteria for the HIV-infected case definition. Seventy-eight were positive by Algorithm 1 alone, 15 were identified using Algorithm 2 alone, and 970 met the criteria for HIV infection by both algorithms. An additional 459 patients, as outlined in Figure 4, had at least some positive component of the HIV algorithms but did not meet the full definition for being HIV-infected by either algorithm. Performance of each component of the 2 algorithms is outlined in Figure 4, and details of the 459 patients with only a component positive are described in Table 3. When utilized to detect cases of HIV infection in our cohort of subjects in DUHS, we found an estimated HIV prevalence of 0.6%.

Validation process

Comparing the results of our algorithms to the gold standard of clinician manual chart review, we identified a sensitivity and specificity of 78% (95% confidence interval [CI], 71.7-83.6%) and 99% (95% CI, 99.9-100%), respectively, for Algorithm 1 and 77% (95% CI, 70.7-82.9%) and 100% (95% CI, 100-100%) for Algorithm 2. We subsequently reviewed all cases that were misclassified by our algorithms. In total, there were 12 false negatives and 6 false positives (Table 4). Those individuals misclassified as HIV-negative by both algorithms (false negatives) were missed due the critical diagnostic components of their records not being accessible to our algorithms, due to data being stored either in scanned form from an outside facility or in an outside EMR identified manually through Epic's Care Everywhere feature. Other misclassified results were due to incomplete metadata in our algorithms, with some individual HIV PCR results missed by our algorithm. Those individuals incorrectly classified as HIV-infected by the algorithms (false positives) were identified as such based on misclassification of medications prescribed for hepatitis B or PEP/PrEP as HIV treatment medications (Table 5).

DISCUSSION

The widespread adoption of EMRs for clinical documentation in the United States has led to unprecedented opportunities for large-scale epidemiologic research. Traditional retrospective studies that previously took months or years to complete can now be performed in a fraction of the time. EMR-based computed phenotypes allow for rapid identification of patients, and have the added benefit of being adaptable to different health care systems, allowing for the development of multicenter cohorts of patients for research, clinical care, and public health initiatives. Computed phenotypes for disease case detection have been previously published for diabetes, cardiovascular disease, and asthma, with robust results.[2–4] In order to optimally utilize the benefits of these new systems for HIV research, we have developed novel computable phenotype algorithms to detect HIV-infected patients within a single institution. Our combined algorithms leverage the strengths of current guideline-based clinical diagnostic strategies while improving upon traditional chart review methodologies for case-finding by applying a multistep diagnostic approach. Implementing our algorithms to health system data identified a local estimated prevalence of HIV of 0.6%. This is consistent with the official estimate of HIV prevalence in Durham County, North Carolina, which in 2011 was reported to be 0.5%.[20,21]

Our novel algorithmic approach to the identification of HIV infection in an EMR system contributes to the current literature in several ways. Our objective of developing algorithms that can be applied to an entire nongovernmental health care system to create a virtual cohort of all individuals living with HIV infection is an approach with unique implications for research and clinical care. Similarly, our methodology of using a 2-step confirmation process instead of relying on administrative data alone improved the performance of our algorithms, as seen in Table 4. Finally, combining a current guideline-based clinical approach with a non–laboratory-based 2-step confirmation pathway is unique to the published literature in the area and allowed for detection of cases in our cohort that would otherwise have been missed, representing an important population of individuals with fragmented health care access patterns.

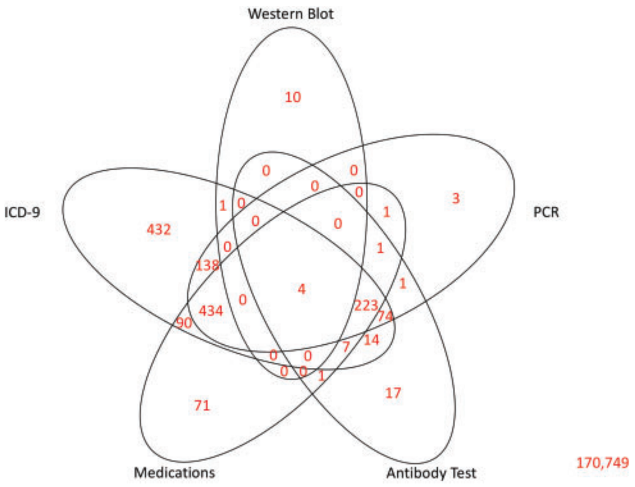

Figure 4. Venn diagram of the contribution of each component of the 2 final algorithms to the case detection of HIV-infected patients in the complete cohort. Areas of overlap highlight complementary data components, where nonoverlapping regions emphasize areas for further refinement. The total number of subjects with no components positive in the dataset was 170 749 individuals.

When we compare the results of our current study to previous work by Goetz et al., our sensitivity for HIV case detection is lower. This was an intentional tradeoff for much higher specificity, which was the goal for our algorithms in the development phase.[8] This may also reflect differences in the objectives of each study. The Goetz study aimed to identify new diagnoses of HIV infection in a Veterans Health Administration cohort, whereas the goal of this study was to identify all patients with HIV infection in a nongovernment setting.[8]

Applying our algorithms sequentially, we identified several issues that contribute to the misclassification of HIV infection status. The most common circumstance leading to mischaracterization of a patient as HIV-negative was incomplete metadata, where some medications and lab results were present in the chart but not identified by our algorithms. A secondary source of false negative results was identified if the diagnosis was obtained outside of our institution and listed solely in the narrative notes. This data could only be identified by manual extraction of data from the records. We also found that 4.2% of patients (6 out of 143) were misclassified as HIV-infected when they were actually HIV-negative. This was mainly secondary to antiretroviral medications used to treat hepatitis B or PEP, falsely categorizing patients as HIV-infected (Table 5).

Our overall approach and final algorithms have several strengths. First, both an international guideline-based diagnostic approach and a more traditional ICD-9 code and billing information component were incorporated into the algorithm. Doing so allowed us to identify subjects that would have been missed by either algorithm alone. Second, the individual data components of our algorithm should be readily accessible for use in other health systems that have access to and experience using clinical data repositories by adapting individual algorithm components to local systems. The individual component domains of our algorithms (medications, ICD-9 codes, laboratory results) have been used by others in collaborative research, such as the Phenotype KnowledgeBase (PheKB) project from eMERGE creating EMR phenotypes for a variety of diseases.[22] It should be possible to take the individual components of our algorithm and build a PheKB HIV infection phenotype, which could then be implemented at partnering PheKB sites. Translation of our algorithms into a common data standard, either Observational Health Data Sciences and Informatics or an equivalent, may be necessary to permit portability between institutions. Finally, our study was performed in a large academic institution with experience in collecting data pre- and post-implementation of the Epic EMR. This setting allows for greater generalizability to other nongovernment health care systems that may use a similar medical record system.

There are limitations to our current approach. The initial review of the metadata failed to identify several laboratory results and medi-

Table 3. Details of single component positive patients

Component Positive	Number of Patients	Explanation
ICD-9 code	432	ICD-9 code for HIV present, but no confirmatory test or medications identified in metadata
HIV antibody	17	Antibody result identified and positive, but no confirmatory component identified in metadata
Western blot	10	Western blot result identified and positive, but no confirmatory component identified in metadata

Results of the 459 patients who were identified to have a positive component of the algorithms but did not meet the full criteria for a positive algorithm.

Table 4. Validation results

Source	Strata	N (172 271)	N (171)	Results			
				D+		D−	
				R+	R−	R+	R−
Algorithm 1	Both algorithms positive	970	12	12	0	0	0
	One algorithm positive	93	10	3	1[a]	6[a]	0
	Component only positive	459	21	0	12[a]	0	9
	High-risk individuals	2881	119	0	0	0	119
	No components positive	167 868	9	0	0	0	9
Algorithm 2	Both algorithms positive	970	12	12	0	0	0
	One algorithm positive	93	10	1	3[a]	0	6
	Component only positive	459	21	0	12[a]	0	9
	High-risk individuals	2881	119	0	0	0	119
	No components positive	167 868	9	0	0	0	9

Results of the validation step. Manual review of 171 charts (*n*) selected at random from each of 5 strata was performed, with the results listed according to the algorithm used. Results are listed according to algorithm result R+ or R− as well as gold standard result D+ or D−.

[a] False negative and false positive results.

Table 5. False positive and false negative results

Result	Algorithm	*N*	Explanation
False positive	Algorithm 1	6	2 results were unexplainable
	Algorithm 2	0	2 subjects on PEP following a needle-stick injury
			1 subject on therapy for hepatitis B
			1 subject with inaccurate clinical documentation of AZT/3TC (Combivir) prescription
False negative	Algorithm 1	13	10 subjects misclassified due to incomplete metadata
	Algorithm 2	15	Remaining subjects misclassified due to results only available in outside records or documentation

Explanation of false positive (6) and false negative (16) results from validation step, by algorithm.

cations, yielding an inaccurate output that underestimated the prevalence of disease. A subsequent manual review of the data and software development was required to identify all pertinent lab and medication results missed. Moreover, there was difficulty interpreting the metadata output of laboratory tests due to confusion related to reference lab results, where multiple lines of redundant text were uploaded into the system and had to be manually reviewed and refined prior to being interpreted with our algorithm. Both of these issues may be unique to our local EMR implementation, and generalizability to other health care systems should be investigated. A further limitation involves our validation step results. Because of the small number of false negatives identified in our manual chart review, there is the potential to overestimate sensitivity with small changes in the number of false negatives. A final limitation is the changes in clinical practice since our cohort closed for enrollment in 2011. Since that time, billing has transitioned from ICD-9 to ICD-10, and new clinical diagnostic algorithms for HIV testing have been published.[5] Both of these changes should be accounted for as these algorithms are put into practice.

Efforts are now being made to merge complementary diagnostic algorithms for other chronic diseases with our novel algorithms for HIV to develop a multitiered case finding system. By developing a repository of EMR phenotypes for different diagnoses with their associated test characteristics, we will be able to rapidly identify large cohorts of subjects with specific diagnoses accurately. Given the proliferation of EMRs across the United States, we anticipate that the process we have outlined can be readily adapted to other sites as well by using our completed algorithms as macro programs for statistical analysis software packages.

CONCLUSIONS

In this study, we have developed and validated novel HIV case-detection algorithms with good sensitivity and specificity. These novel algorithms allow for rapid case detection of HIV-infected individuals in large datasets using methodology that can be extrapolated to other health care systems, allowing for collaborative, cross-institution cohort creation. While this represents a significant step forward, future studies should focus on applying these algorithms to multiple datasets and determining their test characteristics in different settings.

ACKNOWLEDGMENTS

The authors would like to thank Allie Frear, Riti Shah, and Amber Knight for their assistance with this study.

FUNDING

This work was supported by National Institutes of Health grant no. NIH R56 HL126537 01. This publication resulted (in part) from research supported by the Duke University Center for AIDS Research, an NIH-funded program (5P30 AI064518).

COMPETING INTERESTS

The authors have no competing interests to declare.

CONTRIBUTORS

JB, MK, DWP, MM, IR, DMM, and LGQ significantly contributed to the conception and design of the study. NBN, MP, and JL acquired and analyzed the data. DWP, MC, IR, MM, DMM, MAH, and LGQ performed the clinical chart reviews. All authors were involved in data interpretation. DWP, MP, MM, and LGQ drafted the manuscript, and all authors revised it critically for important intellectual content. MP, NBN, and LGQ had access to the data in the study and take responsibility for data integrity and accuracy. All authors gave final approval of this version to be submitted.

SUPPLEMENTARY MATERIAL

Supplementary material is available at *Journal of the American Medical Informatics Association* online.

REFERENCES

1. Allison JJ, Wall TC, Spettell CM, *et al*. The art and science of chart review. *Jt Comm J Qual Patient Saf*. 2000;26:115–36.
2. Afzal Z, Engelkes M, Verhamme KMC, *et al*. Automatic generation of case-detection algorithms to identify children with asthma from large electronic health record databases: Automated case-detection algorithms. *Pharmacoepidemiol Drug Saf*. 2013;22:826–33.
3. Liao KP, Ananthakrishnan AN, Kumar V, *et al*. Methods to develop an electronic medical record phenotype algorithm to compare the risk of coronary artery disease across 3 chronic disease cohorts. *PloS One*. 2015;10:e0136651.
4. Richesson RL, Rusincovitch SA, Wixted D, *et al*. A comparison of phenotype definitions for diabetes mellitus. *J Am Med Inform Assoc*. 2013;20:e319–26.
5. US Centers for Disease Control and Prevention, Bernard MB, Association of Public Health Laboratories, et al. *Laboratory Testing for the Diagnosis of HIV Infection: Updated Recommendations*. Centers for Disease Control and Prevention; 2014. https://stacks.cdc.gov/view/cdc/23447. Accessed July 12, 2016.
6. Thornton C, Fasciano N, Turner B, *et al*. Methods for identifying AIDS cases in Medicare and Medicaid claims data. *Health Care Financ Admin*. 1997.
7. Fasciano NJ, Cherlow AL, Turner BJ, *et al*. *Profile of Medicare beneficiaries with AIDS: application of an AIDS case-finding algorithm*. Health Care Finance Administration; 1998. https://archive.org/details/methodsforidenti00thor. Accessed July 12, 2016.
8. Goetz MB, Hoang T, Kan VL, *et al*. Development and validation of an algorithm to identify patients newly diagnosed with HIV infection from electronic health records. *AIDS Res Hum Retroviruses*. 2014;30:626–33.
9. 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;4:1015–24.
10. McGinnis KA, Fine MJ, Sharma RK, *et al*. Understanding racial disparities in HIV using data from the veterans aging cohort 3-site study and VA administrative data. *Am J Public Health*. 2003;93:1728–33.
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:S25–30.
12. Peabody JW, Luck J, Jain S, *et al*. Assessing the accuracy of administrative data in health information systems. *Med Care*. 2004;42:1066–72.
13. Levison J, Triant V, Losina E, *et al*. 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;5:557–70.
14. Felsen UR, Bellin EY, Cunningham CO, *et al*. Development of an electronic medical record–based algorithm to identify patients with unknown HIV status. *AIDS Care*. 2014;26:1318–25.
15. Spratt SE, Batch BC, Davis LP, *et al*. Methods and initial findings from the Durham Diabetes Coalition: Integrating geospatial health technology and community interventions to reduce death and disability. *J Clin Transl Endocrinol*. 2015;2:26–36.
16. Horvath MM, Winfield S, Evans S, *et al*. The DEDUCE Guided Query tool: providing simplified access to clinical data for research and quality improvement. *J Biomed Inform*. 2011;44:266–76.
17. CDC. Interpretation and use of the western blot assay for serodiagnosis of human immunodeficiency virus type 1 infections. *MMWR Morb Mortal Wkly Rep*. 1989;38:1–7.
18. Cronin AM, Vickers AJ. Statistical methods to correct for verification bias in diagnostic studies are inadequate when there are few false negatives: a simulation study. *BMC Med Res Methodol*. 2008;8:75–83.
19. Begg CB, Greenes RA. Assessment of diagnostic tests when disease verification is subject to selection bias. *Biometrics*. 1983;39:207–15.
20. *North Carolina 2011 HIV/STD Surveillance Report*. Raleigh, NC, Communicable Disease Branch, NC Division of Public Health, NC Department of Health and Human Services; 2012. http://epi.publichealth.nc.gov/cd/stds/figures/std11rpt.pdf.
21. National Center for Health Statistics. *Vintage 2015 postcensal estimates of the resident population of the United States (April 1, 2010, July 1, 2010–July 1, 2015), by year, county, single-year of age (0,1,2,.., 85 years and over), bridged race, Hispanic origin, and sex*. Prepared under a collaborative arrangement with the US Census Bureau; 2015. https://www.cdc.gov/nchs/nvss/bridged_race/data_documentation.htm#vintage2015. Accessed July 27, 2016.
22. Kirby JC, Speltz P, Rasmussen LV, *et al*. PheKB: a catalog and workflow for creating electronic phenotype algorithms for transportability. *J Am Med Inform Assoc*. 2016;23:1046–52.