Title

Development and Validation of an Algorithm to Identify Patients Newly Diagnosed with HIV Infection from Electronic Health Records

Journal
AIDS Research and Human Retroviruses, 30(7)

ISSN
0889-2229

Authors
Goetz, Matthew Bidwell
Hoang, Tuyen
Kan, Virginia L
et al.

Publication Date
2014-07-01

DOI
10.1089/aid.2013.0287

Peer reviewed

Development and Validation of an Algorithm to Identify Patients Newly Diagnosed with HIV Infection from Electronic Health Records

Matthew Bidwell Goetz, MD[1]; Tuyen Hoang, PhD[1]; Virginia L. Kan, MD[2]; David Rimland, MD[3]; Maria Rodriguez-Barradas, MD[4]

1. VA Greater Los Angeles Healthcare System and David Geffen School of Medicine, University of California, Los Angeles, CA
2. Washington DC VA Medical Center and George Washington University School of Medicine, Washington, DC
3. Atlanta VA Medical Center and Emory University School of Medicine, Atlanta, GA
4. Michael E. DeBakey VA Medical Center and Baylor Unviersity School of Medicine, Houston, TX

Source of Funding and Conflicts of Interest: Veterans Health Administration, Health Services Research & Development Service (RRP 12-220). Noneof the authors have any conflicts of interest.

Running Title: An Algorithm to Identify New HIV Diagnoses

Key words: Diagnosis, HIV infections, Continuum of care, Electronic health records

Word count 3039

Corresponding author and requests for reprints: Dr. Goetz

Matthew Bidwell Goetz, M.D.
Chief, Infectious Diseases Section (111-F), VA Greater Los Angeles Healthcare System
11301 Wilshire Blvd.
Los Angeles, CA 90073
Tel: 310-268-3015
Fax: 310-268-4928
Email: matthew.goetz@va.gov or mgoetz@ucla.edu

ABSTRACT

Objective: To develop an algorithm that identifies patients with new diagnoses of HIV infection by use of electronic health records.

Methods: An algorithm was developed based on the sequence of HIV diagnostic tests, entry of ICD-9-CM diagnostic codes and measurement of HIV-1 plasma RNA levels in persons undergoing HIV testing from 2006 – 2012 at four large urban Veterans Health Administration (VHA) facilities. Source data was obtained from the VHA National Corporate Data Warehouse. Chart review was done by a single trained abstractor to validate site-level data regarding new diagnoses.

Results: 1,153 patients were identified as having a positive HIV diagnostic test within the VHA. Of these, 57% were determined to have prior knowledge of their HIV status from testing at non-VHA facilities. An algorithm based on the sequence and results of available laboratory tests and ICD-9-CM entries identified new HIV diagnoses with a sensitivity of 83%, specificity of 86%, positive predictive value of 85% and negative predictive value of 90%. There were no meaningful demographic or clinical differences between newly diagnosed patients who were correctly or incorrectly classified by the algorithm.

Conclusions: We have validated a method to identify cases of new diagnosis of HIV infection in large administrative datasets. This method, which has a sensitivity of 83%, specificity of 86%, positive predictive value of 85% and negative predictive value of 90% can be used in analyses of the epidemiology of newly diagnosed HIV infection.

INTRODUCTION

Approximately 18% of the 1.1 million HIV-infected persons in the United States (US) do not know their status and therefore cannot benefit from life-saving and restoring treatment. [1] This gap has led the Veterans Health Administration (VHA), Centers for Disease Control and Prevention (CDC), American College of Physicians (ACP) and U.S. Preventive Services Task Force to recommend that routine, voluntary HIV testing be offered to adults.[2-5]

The VHA has committed substantial resources to promote HIV testing [6-8]. These efforts have borne fruit.[6;9;10] However, it is not known whether the ultimate goal of expanded HIV testing has been achieved, namely, to what degree does expanded HIV testing identify patients with previously unknown HIV infection and, even more crucially, whether newly diagnosed patients are being promptly linked to appropriate medical care. Timely care linkage is associated with meaningful improvements in clinical outcomes[11-14] but does not occur in approximately 25% of newly diagnosed HIV-Infected patients in the US.[1;15;16] Variations in linkage help to explain gender and racial/ethnic disparities in HIV treatment outcomes.[13;14;17] The importance of linkage has prompted the National HIV/AIDS Strategy (NHAS) to call for action to ensure that 85% of newly diagnosed patients are linked to care within 3 months of diagnosis.[18]

Both assessment of linkage to care for newly diagnosed patients and evaluation of the efficiency of routine HIV testing require the ability to discriminate between newly diagnosed patients and patients with previously known HIV infection who undergo repeated testing upon transfer of care from one healthcare system to another or for other reasons. This distinction is critical as previously diagnosed patients are often actively seeking care and thus may be more motivated to engage in care than are newly diagnosed patients.

The ability to identify newly diagnosed HIV-infected patients in large cohorts would enable analyses of variances in diagnostic rates of previously undetected HIV infection, as well as permit evaluation of subsequent progression through the continuum of care. However, there is no validated procedure for using data from medical records to identify newly diagnosed HIV-infected patients within large observational cohorts. To address this deficit, we determined the ability of using the sequential timing of laboratory results and diagnostic codes to identify newly diagnosed cases as recorded in electronic health records (EHR) of HIV-infection among persons receiving care within the VHA.

METHODS

The cohort of this study included all patients seen at four large urban Veterans Health Administration (VHA) facilities in Los Angeles, Houston, Washington DC, and Atlanta who had positive HIV antibody tests from August 2006 to August 2012. This time period was selected to obtain a representative sample of patients undergoing HIV testing before and after implementation of routine HIV testing in VHA in August 2009; previous to that time VHA policy recommended risk-based HIV testing.

We obtained data spanning the period from 2000 to 2012 from the VHA National Corporate Data Warehouse that included patient visits, demographics, laboratory test results, International Classification of Diseases, Ninth Revision and Clinical Modification (ICD-9-CM) diagnostic codes and factors used by a risk-based prompt for HIV testing used by some VHA facilities [6]. In addition, we obtained direct access to facility-level EHR of the cohort of patients with confirmed positive HIV tests from the four study sites.

We distinguished newly diagnosed patients from patients who were previously diagnosed with HIV infection and who underwent repeat or confirmatory HIV testing by assessing the temporal relationship of the ordering of HIV antibody tests and plasma HIV-1 RNA measurement, and the recording of ICD-9-CM codes for HIV infection. We reasoned that persons for whom there was no prior evidence of HIV infection (i.e. newly diagnosed patients) measurement of HIV-1 plasma RNA levels and recording of HIV-specific ICD-9-CM diagnoses would occur after the performance of HIV antibody testing, whereas for persons who were already known to be HIV infected, plasma HIV-1 RNA testing and ICD-9-CM code entry would be recorded either before or at the same time as HIV antibody testing was performed. Finally, we reasoned that persons with undetectable or low levels of plasma HIV-1 RNA were likely to be receiving antiretroviral therapy and thus previously diagnosed.

Based on the aforementioned principles, the following sequential procedures were used to classify patients based on HIV-1 antibody tests, Western Blot tests, HIV-1 plasma RNA levels, and ICD-9-CM codes contained within EHR data available in the VA Corporate Data Warehouse. Further details are provided in the Figure and Appendix.

Stage 1: Identify patients with a non-negative HIV antibody test from August 2006 to August 2012. Non-negative HIV antibody tests included instances where result was recorded as “positive”, “reactive”, “comment” or similar language that did not identify the test as being negative.

Stage 2: Exclude from all instances where a prior HIV antibody test was positive.

Stage 3: Identify patients with confirmed HIV infection. Confirmation of HIV infection required a positive antibody test plus either a diagnostic HIV Western Blot or quantifiable HIV-1 RNA in plasma. Plasma HIV-1 RNA values were sought only in persons in whom Western Blot test results were unavailable, or the result was recorded as “indeterminate” or “comment” or similar language was used. For such persons, if the HIV-1 plasma RNA level was recorded as being below the lower limit of quantification or resuls were not available, the patient was categorized as having “unconfirmed HIV status”. Persons with negative Western Blots were considered to be uninfected and were excluded from further analyses.

Stage 4: Determine whether patients with confirmed HIV infection were newly diagnosed or had previously been identified as being HIV-infected. Patients who underwent standard blood-based HIV-1 testing were considered as having been previously known to be HIV infected if the date of the first HIV-1 plasma RNA determination or ICD-9-CM record of HIV diagnosis was on or before the date the HIV antibody test was ordered, or if the first HIV-1 plasma RNA level was below the lower limit of quantification. Patients in whom there was neither a HIV-1 plasma RNA determination nor an entry of a HIV related ICD-9-CM code were considered to be un-classifiable.

We made an exception for persons undergoing HIV Point of Care (i.e. Rapid) testing, wherein clinicians had rapid access to preliminarily positive HIV test results and therefore often ordered HIV-1 plasma RNA tests and recorded the appropriate ICD-9-CM code on the same date as the Point of Care test was ordered. Consequently, patients who underwent HIV-1 Point of Care testing were considered to have been previously known to be HIV infected only if the date of the first HIV-1 plasma RNA determination or ICD-9-CM record of HIV diagnosis was before the date the HIV antibody test was ordered.

Site investigators reviewed the medical records and other locally available information for each patient at their facility with a positive HIV antibody test to determine whether each such positive test represented a new diagnosis of HIV infection or confirmation of a previously established infection.   To confirm that these classifications were correct, a single trained abstractor used the VA Compensation and Pension Record Interchange (CAPRI) system to review the EHR of all patients identified as having newly diagnosed HIV infection. In addition, records of patients identified as having previously diagnosed HIV infection at each site were re-reviewed. A prior diagnosis of HIV infection was confirmed in 287 of 289 patients at the first two sites evaluated. Based on this experience a limited, random sampling of 30 patients determined to have been previously diagnosed by each of the other two local site investigators were re-reviewed; all 60 of these patients were confirmed as having had previous diagnoses of HIV infection.   A senior investigator re-reviewed all charts where there was disagreement between the assignment made by the local investigator and the chart abstractor and made a final determination as to whether the HIV test result represented a newly made diagnosis. This final adjudicated chart review served as the reference standard in assessments of the validity of the algorithm.

The statistical methods include frequencies to summarize the distributions of demographic and clinical characteristics of the patients across four study sites (Table 1). Measures to validate the algorithm include sensitivity which is the proportion of patients classified as new diagnoses by the algorithm among the new diagnoses based on chart review, specificity which is the proportion of patients classified as non-VHA diagnoses by the algorithm among the non-VHA diagnoses based on chart review, positive predictive value which is the proportion of new diagnoses based on chart review among the new diagnoses classified by the algorithm, and negative predictive value which is the proportion of non-VHA diagnoses based on chart review among the non-VHA diagnoses classified by the algorithm (Tables 2 and 4). Finally, frequencies and chi-square tests were used to compare demographics, clinical characteristics, and care utilization between correctly-classified and misclassified diagnoses (Table 3).The study protocol was approved by the VA Central Institutional Review Board and by the Research and Development Committees at each of the participating facilities.

RESULTS

Each site contributed 239 to 386 unique patients with positive HIV antibody tests to the study cohort. Five patients from Site 2 and one patient from Site 4 were missing from the VA National Corporate Data Warehouse; therefore there were a total of 1,147 patients comprised the final analytical cohort available for algorithm validation analysis (Table 1). As per the adjudicated chart review, the first positive HIV antibody test represented a new diagnosis 43% of the time; this varied from 30% to 64% across the sites. While gender and age distributions were similar across the sites (>95% male, mean age=52 years), race and ethnicity varied with African-Americans constituting from 52% to 82% of new diagnoses. The correlation between local site designation and central chart review of patients as having new HIV diagnoses ranged from 87% (Site 3) to 100% (Site 1); data not shown.

Table 2 provides the sensitivity, specificity, positive predictive value (PPV) and negative predictive value (NPV) of the algorithm. Overall the algorithm achieved a sensitivity of 83%, specificity of 86%, positive predictive value of 85% and negative predictive value of 90%. The most frequent contributor to decreased sensitivity of the algorithm were instances wherein patients underwent standard blood-based HIV antibody testing and had a measurement of HIV-1 plasma RNA levels done the same day. Revision of the algorithm to classify such patients as being newly diagnosed increased the sensitivity of the algorithm to 89% but reduced the specificity to 58%. Chart review indicated that the most common circumstance leading to categorization of a previously diagnosed patient as being newly diagnosed was due to the provider not being aware that the patient had a previous diagnostic test at a non-VHA facility and therefore not recording a diagnostic code for HIV infection or ordering HIV-1 RNA levels when antibody testing was done. Evidence concerning these prior HIV diagnoses was found in chart notes that were often written well after the date of the HIV testing. Thirty-seven other patients could not be classified by the algorithm; these included cases with a positive HIV antibody test and Western Blot but no plasma HIV-1 RNA test or ICD-9-CM code (unclassifiable, n=20), cases with positive HIV antibody tests but no identifiable confirmatory Western Blot or plasma HIV-1 RNA (unconfirmed status, n=12), and cases without records of positive antibody tests in the VA Corporate Data Warehouse (n=5). We found that these misclassified cases were more likely than classificable cases to not have any visits to the VA in the year after HIV diagnostic testing was performed, and had fewer documented HIV-associated risk factors (49% versus 59-65%) or mental health conditions (43% versus 53-65%; Table 3).

To ascertain whether the algorithm mischaracterized patients at random or whether particular groups of patients were liable to be misclassified, we compared the patient-level demographics, clinical characteristics, and frequency of care utilization between the 409 correctly-classified and 60 misclassified new diagnoses, and also the 566 correctly-classified and 75 misclassified non-VA diagnoses. Pair-wise comparison showed no meaningful differences between the 409 persons correctly-classified as being newly diagnosed with HIV infection and the 60 newly diagnosed patients who were misclassified as having a previously known HIV diagnoses. However due to the disproportionate use of HIV rapid testing by one site (Site 3), we found that the 75 previously diagnosed persons who were misclassified as being new diagnoses were more likely to be African-American than were the 566 persons correctly classified as having been previously diagnosed with HIV infection (72% vs. 58%).

We conducted several sensitivity analyses. First, we examined how plasma HIV-1 RNA thresholds or elimination of virological criteria affected the performance of the algorithm (Table 4). We found that the predictive power of the algorithm was optimal when the virological threshold was set as less than the lowest quantifiable plasma HIV-1 RNA value as determined by the locally used laboratory assay or when virological criteria were not used to discriminate between new and previously diagnosed patient with sensitivity and positive predictive values of 83% and 86%, and 84% and 85%, respectively. On the first test that was done, 5.1% of the 491 news diagnoses by chart review had unquantifiable plasma HIV-1

RNA (i.e., had a value below the limits of detection of the assay that was used). We next evaluated whether the initial CD4+ lymphocyte count discriminated between new and previously diagnosed patients. For newly diagnosed patients the median initial CD4+ lymphocyte count was 73 cells/□L (interquartile range 20-382) whereas for previously diagnosed patients the median was 326 (IQR 90-565). As shown in table 4, addition of CD4+ cell criteria to the base case virological criteria substantially reduced the sensitivity of the algorithm. Finally as CD4 cell counts were generally performed at the same times as viral load measurements these values provided no independent information.

DISCUSSION

We developed an algorithm using electronic health records to differentiate between persons with newly diagnosed HIV infection and those with previously established infection who underwent repeated testing. This algorithm correctly identified newly diagnosed patients with 83% sensitivity and an 85% positive predictive value. In contrast, only 43% (491) of the 1153 positive HIV antibody tests in these four VHA facilities represented a new diagnosis of HIV infection; the remaining positive tests were performed in persons who were already aware of their HIV status.

The algorithm was based on the sequential timing of HIV antibody tests, measurements of HIV-1 plasma RNA levels, and recording of diagnostic codes. In clinical practice HIV-1 plasma RNA levels are rarely performed for persons whose HIV status is not known; exceptions are generally limited to persons with acute HIV seroconversion syndromes or to persons in clinical circumstances in which HIV infection is highly likely (e.g., acute opportunistic infection). However, in sensitivity analyses we found that use of HIV-1 plasma RNA values did not improve the ability of the algorithm to distinguish between prior and new HIV diagnoses. In addition, we found that the value of timing or value of of CD4+ lymphocyte counts did not improve the accuracy of the algorithm.

We identified two issues that were the main contributors to decreased accuracy of the algorithm. The first issue involved HIV Point of Care testing. Because results of Point of Care tests could be read within 20 minutes after performance, providers often ordered measurements of plasma HIV-1 RNA on the same date as a positive Point of Care test was reported. As information in the laboratory results database did not consistently specify that HIV-1 Point of Care testing had been performed, patients undergoing HIV rapid testing were often misidentified as having undergone standard blood-based antibody testing. Consequently, when HIV-1 plasma RNA levels were determined on the same day, the algorithm incorrectly classified them as having previously established diagnoses of HIV infection. The second issue involved patients who were diagnosed prior to entering VHA care and who did not disclose their HIV status until after they underwent repeat HIV testing. For these patients, VHA providers ordered plasma HIV-1 RNA measurements after their positive antibody tests were confirmed; consequently, the algorithm incorrectly classified them as new diagnoses. This misclassification reduced the specificity of the algorithm from 97% to 86%.

The strengths of the algorithm validation are that it relied upon a rigorous adjudication process to identify newly diagnosed patients and that the validation was done using patients from four geographically diverse sites with differing demographics, clinical characteristics, HIV testing methodologies, recording of laboratory information and frequency of new diagnoses. The consistency of performance of the algorithm in the face of such diversity strengthens confidence in the use of this tool to identify new diagnoses in other settings. In addition, no differences were found in the characteristics of patients who were properly or improperly classified as new HIV diagnoses.

Limitations of our work include that our results may not be generalizable outside of a VHA setting, particularly as very few women were included in our cohort, and will not be feasible for cohorts that do not record ICD-9-CM codes or laboratory results as directly analyzable fieldsin EHR. Alternative methods to distinguish between persons with positive HIV antibody tests that represent a newly diagnosed infection versus confirmation of a previously established diagnosis will need to be developed for these cohorts. In addition, we were hampered by an inability to consistently distinguish between standard blood based HIV testing and Point of Care testing. This misclassification, which had a major affect on assay performance as 81% of patients diagnosed on the basis of a positive point of care test had a plasma HIV-1 RNA assay on the same day as the point of care test, could have been resolved if key words such as ‘rapid’ or “point of care” were to be explicitly indicated in the laboratory test names. Alternatively, accurate and reliable time stamps that included the time of day as well as the date of test performance or natural language processing of text notes could be used to determine whether orders for virologic tests were concurrent with or followed point of care diagnostic tests. The other major source of misclassification, was due to the reliance solely on administrative and laboratory data and the exclusion of information contained innarrative notes. While natural language processing methods could be used to extract narrative data and overcome this limitation, there are many challenges due to the unstructured nature of how and when information regarding the timing of HIV infection is recorded. Finally, the algorithm does not identify newly diagnosed persons who present with acute HIV infection and lack detectable antibody or who are elite controllers with low or undetectable HIV-1 plasma RNA levels.

This work has clinical implications for future epidemiologic research regarding the frequency and epidemiology of new diagnoses of HIV infection and for evaluation in the patterns and outcomes of care in such individuals. A validated method to identify cases of new diagnosis of HIV infection will enable a more thorough delineation of the frequency of newly diagnosed HIV infection in differing communities and demographic settings and thus inform the development of programs to increase the efficiency and utility of efforts to promote HIV testing among the most vulnerable patients. In this regard the the CDC has recommended that 80% of expanded HIV testing efforts should be focused on sites with a ≥ 2% rate of positive HIV tests.[19] Furthermore, identification of the newly diagnosed patients will allow for detailed analyses of the effectiveness of programs to ensure that these patients are linked to care in a timely manner. At present, timely care linkage occurs in less than 65% of newly diagnosed HIV-Infected patients in the US.[15] Variations in linkage to care are likely to explain much of the gender and racial/ethnic disparities in HIV treatment outcomes[13;14;17] and underscore the importance of the NHAS goal to ensure that 85% of newly diagnosed patients are linked to care within 3 months of diagnosis.[18]

As recommended by the CDC,[19] additional efforts should be taken to assure that the highest risk patients are being tested for HIV infection. Both the need to develop such programs and the design of such an intervention requires knowledge of the distribution of newly diagnosed HIV-Infected patients and thus a systematic approach to detect such patients. The algorithm we have developed provides means to acquire this information. Such a tool is also needed to assess geographical-, institutional- and patient-level variations in linkage to medical care of newly diagnosed HIV-Infected patients; such data will inform the development of interventions to reduce such variations, improve overall linkage and assist the medical care system in meeting NHAS goals.[18;20;21]

Acknowledgements

Author participation: All authors gave final approval for this manuscript. Matthew B. Goetz contributed to the conception of the study, interpretation of the data and drafting the article. Tuyen Hoang contributed to the conception and design of the study, analysis and interpretation of data, and drafting the article. Virgnia Kan, David Rimland and Maria Rodriguez-Barradas contributed to the conception and design of the study and revising the article for critical intellectual content.

We also acknowledge the invaluable assistance and contributions of Michael Fletcher and Nui Brown.

Figure Legend

Characterization and distribution of patients with non-negative HIV antibody tests. Non-negative HIV antibody tests included instances where result was recorded as “positive”, “reactive”, “comment” or similar text. Of the 672 Western Blot results that were neither negative nor positive, the Western Blot result was reported as “comment” in 606 cases and as “I” in 8 cases; in 58 cases no Western Blot result was found. HIV-1 RNA was detected in 3 of 5 tests performed when the Western Blot was reported as “comment”, in 1 of the 7 tests performed when the Western Blot was reported as “I” and in 36 of the 47 tests performed when no Western Blot result was found.*LLQ = lower limit of quantification

** If HIV rapid testing was done, patients were classified as having a prior non-VA diagnosis if the first viral load or ICD-9-CM entry is before the date that rapid test was ordered

Appendix

To accurately distinguish newly diagnosed patients from non-VA diagnosed patients, we paid careful attention to data cleaning of lab tests and ICD-9-CM codes. Unlike ICD-9-CM codes which are standardized data, laboratory tests are free-text data entered by local lab staff, leading to increased data entry errors such as misspelling, obscure abbreviations, and obscure test results. In addition, some ordered tests were cancelled and some test results were not reported or reported in delayed time. To resolve these problems, we cleaned the data as follows:

- ICD-9-CM codes: we used the following ICD-9-CM codes: AIDS (042), asymptomatic HIV (V08), and HIV-related codes (042.0-.2, 042.9, 043.0-.3, 043.9, 044.9, 079.53).
- Lab names: We searched for key words 'HIV' and 'immunodeficient' to search for HIV tests in general. Then we printed out the test names, had an HIV specialist review these test names and classify them into Antibody tests, Western Blot confirmatory tests, and Plasma HIV-1 RNA tests. Then we wrote programming codes to classify these tests based on their names. For example, Antibody test names contained key words such as 'antibody,' 'AB,' 'Rapid HIV,' 'point of contact,' and 'needle stick'; Western blot test names contained key words such as 'western blot,' 'WB,' 'reflex,' and 'confirm' while Plasma HIV-1 RNA test names contained key words such as 'Plasma HIV-1 RNA,' 'PCR,' 'RNA,' and 'quant.' Of note, we found Western Blot test names without key words 'HIV.' For example, one facility named its Western Blot test as 'WB int' without key word 'HIV.'
- Lab results: for qualitative tests, we classified the tests into three groups by their results: positive tests were indicated by key words 'positive,' 'present,' 'reactive,' and 'detected'; negative tests were indicated by key words 'negative' and 'non-reactive' while indeterminate tests were indicated by key words 'indeterminate,' 'comment,' and 'pending.' For quantitative tests, we kept tests with valid

numeric results. We deleted tests of which results contained key words ‘cancel,’ ‘not performed,’ ‘not required,’ and ‘N/A.’