Development and validation of a computer-based algorithm to identify foreign-born patients with HIV infection from the electronic medical record

J. Levison[1,2,3,4,5]; V. Triant[1,2,3,5] ; E. Losina[2,3,5,6,7]; K. Keefe[2,3,5]; K. Freedberg[1,2,3,5,6,7]; S. Regan[2,3,5]

[1]Massachusetts General Hospital, Division of Infectious Diseases, Boston, Massachusetts, United States;
[2]Massachusetts General Hospital, Division of General Internal Medicine, Boston, Massachusetts, United States;
[3]Massachusetts General Hospital, Medical Practice Evaluation Center, Boston, Massachusetts, United States;
[4]Brigham and Women's Hospital, Division of Infectious Diseases, Boston, Massachusetts, United States;
[5]Harvard Medical School, Boston, Massachusetts, United States;
[6]Boston University School of Public Health, Departments of Biostatistics and Epidemiology, Boston, Massachusetts, United States;
[7]Harvard University Center for AIDS Research, Harvard University, Boston, Massachusetts, Unites States

Keywords
Foreign-born, immigrant health, HIV, classification, electronic medical record

Summary

Objective: To develop and validate an efficient and accurate method to identify foreign-born patients from a large patient data registry in order to facilitate population-based health outcomes research.

Methods: We developed a three-stage algorithm for classifying foreign-born status in HIV-infected patients receiving care in a large US healthcare system (January 1, 2001-March 31, 2012) (n = 9,114). In stage 1, we classified those coded as non-English language speaking as foreign-born. In stage 2, we searched free text electronic medical record (EMR) notes of remaining patients for keywords associated with place of birth and language spoken. Patients without keywords were classified as US-born. In stage 3, we retrieved and reviewed a 50-character text window around the keyword (i.e. token) for the remaining patients. To validate the algorithm, we performed a chart review and asked all HIV physicians (n = 37) to classify their patients (n = 957).We calculated algorithm sensitivity and specificity.

Results: We excluded 160/957 because physicians indicated the patient was not HIV-infected (n = 54), "not my patient" (n = 103), or had unknown place of birth (n = 3), leaving 797 for analysis. In stage 1, providers agreed that 71/95 foreign language speakers were foreign-born. Most disagreements (23/24) involved patients born in Puerto Rico. In stage 2, 49/50 patients without keywords were classified as US-born by chart review. In stage 3, token review correctly classified 55/60 patients (92%), with 93% (CI: 84.4, 100%) sensitivity and 90% (CI: 74.3, 100%) specificity compared with full chart review. After application of the three-stage algorithm, 2,102/9,114 (23%) patients were classified as foreign-born. When compared against physician response, estimated sensitivity of the algorithm was 94% (CI: 90.9, 97.2%) and specificity 92% (CI: 89.7, 94.1%), with 92% correctly classified.

Conclusion: A computer-based algorithm classified foreign-born status in a large HIV-infected cohort efficiently and accurately. This approach can be used to improve EMR-based outcomes research.

Correspondence to:
Julie Levison, MD, MPhil, MPH
Program in HIV Epidemiology and Outcomes Research
Medical Practice Evaluation Center
Massachusetts General Hospital
50 Staniford St, 9th Floor
Boston, Massachusetts 02114
Phone: 617–724–4698
Fax: 617–726–2691
Email: jlevison@partners.org

Appl Clin Inform 2014; 5: 557–570
http://dx.doi.org/10.4338/ACI-2014-02-RA-0013
received: February 16, 2014
accepted: May 5, 2014
published: June 18, 2014
Citation: Levison J, Triant V, Losina E, Keefe K, Freedberg K, Regan S. Development and validation of a computer-based algorithm to identify foreign-born patients with HIV infection from the electronic medical record. Appl Clin Inf 2014; 5: 557–570
http://dx.doi.org/10.4338/ACI-02-RA-0013

1. Introduction

At 40.2 million individuals, foreign-born persons represent a substantial and rising proportion of the US population [1]. Foreign-born status can confer a number of vulnerabilities to poorer health outcomes including limited language proficiency, low socioeconomic position, and limited access to health insurance [2, 3].

The challenges these factors present may be heightened for those infected with HIV. Isolation from the community due to minority status or socioeconomic position, and from social support due to sexual identity and HIV status, can influence utilization of HIV services [4-6]. For example, lower rates of HIV testing and delayed entry into HIV care have been observed among foreign-born compared to US-born persons [7, 8]. Important goals of HIV care, such as receipt of antiretroviral treatment and virologic suppression, may therefore be more difficult to achieve in foreign-born patients. Given the relevance of foreign-born status to achieving clinical and public health outcomes, it is important to integrate patient place of birth as well as primary language with appointment history and clinical outcome data.

Patient data registries derived from electronic medical records (EMRs) can serve as vital data sources on clinical outcomes and patient characteristics. However, much of the valuable information held in EMRs is embedded in free text notes rather than explicitly coded fields. Extracting information about place of birth from an EMR by reading free text notes is labor-intensive and not feasible for population-based research using registries comprising thousands of patients. Methods, including natural language processing (NLP), to extract clinical information from free text notes exist for some conditions (e.g. asthma, smoking status, hepatocellular carcinoma) [9-13]. While foreign-born represent a rising portion of the US population for whom this status can have a major impact on health outcomes, no method exists to capture foreign-born status from the EMR. Our objective was to develop an efficient method to identify foreign-born patients in a large patient registry that would be primarily computer-based and would minimize the human effort required to review notes. We then validated the algorithm using physician knowledge as the external standard.

2. Methods

2.1. Data source

The Partners' Research Patient Data Registry (RPDR) is a clinical research database that comprises 4.5 million patients who receive care in the largest healthcare system in Massachusetts and encompasses over 1 billion clinical data points derived from hospital EMRs, administrative databases, and billing data. The RPDR's data elements include demographics (e.g. self-identified race/ethnicity and address), clinical encounters (inpatient and outpatient), diagnoses (International Classification of Diseases, Ninth revision, Clinical Modification (ICD-9-CM) codes and locally-specified codes), medications, procedures, laboratory results, imaging, progress notes, and discharge summaries. We used the RPDR's online query tool to establish our study cohort. We identified 11,113 patients with at least 1 encounter associated with an ICD-9 CM billing code for HIV infection (V08 or 042) from either Brigham and Women's Hospital or Massachusetts General Hospital, both members of the Partners HealthCare System. Because this cohort was intended for use in a study of predictors of loss to follow-up from outpatient care, we excluded patients who did not have at least 1 outpatient encounter between January 1, 2001 and March 31, 2012 (n = 1,999), leaving a final cohort of 9,114 patients. The Partners Human Research Committee approved all aspects of analyses.

2.2. Development and application of computer-based algorithm

Since there is no coded field for country of origin in the RPDR, we developed a three-stage computer-based algorithm to apply to the EMR to identify foreign-born individuals (▶Figure 1). In stage 1, we classified those clearly coded as non-English language speaking, based on EMR-coded demographics, as foreign-born (n = 971). In stage 2, we conducted searches using structured query language (SQL) of the free text EMR notes of the remaining 8,143 patients for specific keywords po-

tentially associated with political context (e.g. "asylum"), language (e.g. "interpreter"), geography (e.g. country of origin) or migration (e.g. "emigrated from") (▶Table 1). There were a total of 94 keywords or phrases that we applied to the free text notes of the EMR. The keyword search included all countries in Africa, North, South, and Central America (including the Caribbean); China; and Portugal due to the predominance of these countries of origin among HIV-infected foreign-born persons in Massachusetts [14, 15]. In order to maximize the yield of identifying HIV-infected foreign-born persons, the keyword strategy also included keywords related to political status (e.g. "green card"), language ("Spanish"), geography (e.g. "born in"), and migration (e.g. "emigrate"). For keywords whose stem was found in more than one word (e.g. Mexico and Mexican, immigrant and immigrate), we used the actual word stem (e.g. "Mexic" and "immigra") to efficiently capture more than one word in a keyword search. We also accounted for potential spelling errors in documentation (e.g. "interpretor" and "interpreter").

We searched all free text notes and hospital discharge summaries for the presence of any keyword. For each instance of a keyword, we extracted the patient identifier, date of the note, and a 50-character window of text around the keyword (termed a "token"). Those with no occurrence of keywords were classified as US-born (n = 4,167) (▶Figure 1).

Patients with at least one instance of a keyword (n = 3,976) proceeded to stage 3, in which we reviewed the tokens to determine place of birth and language spoken. For each individual patient, we reviewed all available tokens and categorized the place of birth as US, foreign, or unknown, and categorized primary language spoken as English, other language, or unknown. We based our definitions of foreign- and US-born on US Census definitions [16]. We considered a patient US-born if place of birth was documented as US mainland, Puerto Rico, or other US territory. In stage 3, patients were classified as foreign-born if a token contained documentation that the patient was foreign-born or if no token indicated where the patient was born but at least one token documented that the patient's primary language was not English. All others were classified as US-born. In addition, we specifically recorded whether patients were born in Puerto Rico. These patients are US-born but share certain characteristics with, and therefore may be confused with, foreign-born patients.

2.3. Validation

We assessed the algorithm by validating all three algorithm steps: the identification of coded foreign language speakers, the search for keywords, and the review of tokens. We then validated the final classification based on the application of the full algorithm.

2.3.1 Coded language validation: Stage 1.

In stage 1 of the algorithm, patients who were coded as foreign language speakers were classified as foreign-born and not considered further. To estimate the adequacy of this definition, we performed chart review and physician survey; the latter is described below under "Full algorithm validation."

2.3.2 Keyword search validation: Stage 2

In stage 2 of the algorithm, patients whose free text notes did not contain any keywords (i.e. keyword-negative) were classified as US-born and not considered further. To estimate how many foreign-born patients we missed using this assumption, we conducted a full chart review of 50 keyword-negative patients, among which we randomly interspersed 14 patients who were keyword-positive and classified as foreign-born on token review. The chart reviewer was blind to the patients' keyword status.

2.3.3. Token review validation: Stage 3

In stage 3 of the algorithm, a human reviewer read all available tokens for a patient and classified the patient. We assessed the performance of this limited token review compared to a full chart review. Sixty patients were randomly selected from among those keyword-positive, stratified by their status-based token review: 20 foreign-born, 20 US-born, and 20 unknown place of birth. Their full charts were reviewed by a reviewer who did not participate in the token review and was blind to the patients' token review status. In the chart review, patients whose place of birth were not recorded in the

EMR but were documented as having a primary language other than English were considered to be foreign-born.

2.3.4. Full algorithm validation

In the final validation, we compared the algorithm results to the judgment of the patients' primary physician. We linked patients to physicians using the RPDR, which provides a ranking of up to three physicians for each patient based on the frequency of encounters. We identified a subset of the 9,114 patients who were linked to at least one of the 37 HIV physicians from the Brigham and Women's Hospital and Massachusetts General Hospital outpatient HIV practices. If a patient saw more than one HIV physician, we assigned the provider with the highest rank. We were able to link 2,813 of 9,114 patients to the 37 physicians. Each provider received a list of their linked patients to review. Providers with <50 linked patients received their complete lists; the rest (n = 10) received a list of 49 patients randomly selected from among those linked to them. The lists contained a total of 957 patients. Providers were asked to classify each patient's status as foreign-born, US-born, or Puerto Rican. In addition, providers were asked to rate their confidence in their response (confident vs. not confident) and indicate if a patient was not theirs or not HIV-infected. No limitations were placed on the sources of information that could be used.

2.4. Statistical Analysis

The proportion of US and foreign-born HIV-infected individuals was calculated based on the proportion of the classified population divided by sample size of the eligible cohort (n = 9,114). In the primary analysis, we calculated sensitivity, specificity, negative and positive predictive values with 95% confidence intervals (CI); we used physician designation as the external standard.

We performed three secondary analyses on the calculation of sensitivity and specificity. First, we excluded patients when physicians indicated they were "not confident" of their classifications. Second, because individuals born in Puerto Rico are technically US-born but may identify as non-English speaking, we explored a possible improvement in the algorithm by evaluating patients, who in stage 1, were coded as primary Spanish-language speakers. We then reviewed those tokens whose free text included mention of Puerto Rico and classified these patients' place of birth as US-born or foreign-born based on the information in the token. Third, we excluded patients for whom the physician judgment was incorrect when compared with repeat token review by study staff. This process involved two reviewers (JHL and SR), blinded to the results of the original token review, who independently re-reviewed the tokens for patients for whom the algorithm classification differed from physician judgment. We considered the physician judgment incorrect if all three token reviews (the original assessment and the two repeat reviews) were in agreement and place of birth was conclusively documented (i.e. as foreign- or US-born and not "unknown"). Analyses were repeated excluding these provider "errors". All statistical analyses were performed using Stata statistical software (StataCorp, 2008. Stata Statistical Software: Release 10. College Station, TX: Stata Corporation.).

3. Results

We identified 9,114 individuals with at least 1 HIV diagnosis code and at least 1 HIV primary care visit between January 1, 2001 and December 31, 2012, and the algorithm was applied to this group (▶Figure 1). In stage 1 of the algorithm, we found 971 patients (11%) coded as foreign language speakers and classified them as foreign-born. We applied the keyword search to the remaining 8,143 patients, of whom 4,167 (46%) were keyword-negative and classified as US-born in stage 2. At least 1 instance of a keyword was found for 3,976 patients (44%), who progressed to stage 3. These patients were classified after review of their tokens. Token review took on average one minute per patient. After full application of the algorithm, 2,102/9,114 patients, or 23%, were classified as foreign-born.

3.1. Coded language validation: Stage 1

We examined the provider classifications of those classified as foreign-born in stage 1 of the algorithm because they were coded as non-English speakers in the EMR. Of the 797 patients classified by their providers, 95 were classified as foreign-born by the algorithm in stage 1. The provider agreed that 71/95 (75%) of these patients were foreign-born. Almost all of the disagreements (23/24, 96%) occurred among patients classified by providers as born in Puerto Rico (and therefore US-born).

3.2. Keyword search validation: Stage 2

The algorithm classifies patients with no keywords as US-born. On full chart review, 1/50 (2%, CI: 0, 6.0%) patients with no keywords was found to be foreign-born. All 14 patients with keywords and classified as foreign-born by token review were confirmed as foreign-born on full chart review. For 163 of the patients reviewed by the providers, no keyword was found at stage 2 of the algorithm and they were therefore classified as US-born. As a replication of the keyword search validation above, we examined the provider assessment of these patients. Only 1/163 (1%) of the keyword-negative patients was judged to be foreign-born on physician review.

3.3. Token review validation: Stage 3

Using full chart review as the standard, token review correctly classified 55/60 patients (92%), with 93% (CI: 84.4, 100%) sensitivity and 90% (CI: 74.3, 100%) specificity.

3.4. Full algorithm validation

Responses were received from all 37 outpatient HIV providers in the infectious disease clinics of the two major academic hospitals in the health system (▶Figure 2). We excluded 160 patients because the provider indicated the patient was "not HIV-infected" (n = 54) or "not my patient" (n = 103), or did not know the patient's place of birth (n = 3). The provider classified foreign-born status in the remaining 797 patients, of whom 579 were US-born (73%) and 218 were foreign-born (27%). The algorithm correctly classified 92% using physician categorization as the external standard. The algorithm's sensitivity was 94% (CI: 90.9, 97.2%), specificity 92% (CI: 89.7, 94.1%), positive predictive value of 81% (CI: 76.5, 86.2%) and negative predictive value of 98% (CI: 96.3, 98.9%).

3.5. Secondary analyses

Providers were confident in 717/797 (90%) classifications. In analyses limited to patients for whom providers were confident about their response, the algorithm performed slightly better, with a sensitivity of 97% (CI: 94.5, 99.4%), up from 94.0%, and a specificity of 93% (CI: 90.2, 94.8%), up from 92% (▶Table 2).

We then re-examined a possible improvement on the classification of foreign-born status for coded foreign language speakers (stage 1) who had mention of Puerto Rico by keyword search of the EMR. Of the 971 coded foreign language speakers, 191 patients were coded Spanish-language speakers and had mention of Puerto Rico by keyword search of their EMR. We examined tokens for all 191 patients and classified each patient as either foreign-born or US-born. Based on this adjustment the number of patients classified as foreign-born decreased from 2,102 (23%), in the primary analysis, to 1,911 (21%), in this secondary analysis. Of the 191 patients, 27 were included in the provider review. When compared against the physician classification, the accuracy of the alternative algorithm improved marginally to 93% from 92% in the primary analysis. The estimated sensitivity decreased from 94% to 90% (CI: 85.1, 93.3), and specificity increased from 92% to 95% (CI:92.7, 96.4).

Providers disagreed with the algorithm classifications for 60/797 (8%) patients. After repeat token review by two reviewers, we considered 24/60 patients to have been misclassified by the pro-

vider. Excluding these 24 patients resulted in slightly improved sensitivity (97%, CI: 94.3, 99.1%) and specificity (95%, CI: 93.0, 96.7%).

4. Discussion

We assessed a hybrid, three-stage computer-based algorithm using coded field data, keyword search, and human review to determine foreign-born status for a large cohort of HIV patients drawn from an EMR-based patient registry. Using this algorithm we efficiently assigned foreign-born status to over 9,000 patients, limiting human review to very small portions of the medical record for fewer than half of them. The algorithm embodies three assumptions that we specifically tested and that were supported: 1) patients coded as non-English speakers are foreign-born, with the exception of those born in Puerto Rico, 2) the vast majority of patients whose EMR notes do not include certain terms related to place of birth are US-born, and 3) reading only a small, 50-character portion of the notes surrounding those terms produces similar results to reading the entire medical record. We further demonstrated that the algorithm performs well when measured against the standard of physician classification, even though physicians are free to bring sources of information beyond the EMR to bear in their assessment.

The computer algorithm identified the proportion of HIV-infected foreign-born persons accessing outpatient HIV care between 2001 and 2012 within a large Massachusetts-based healthcare system as 23%. The Massachusetts Department of Public Health reported that 29% of HIV-infected persons in the state were foreign-born [17]. This proportion is based on 2008 seroprevalence data for HIV diagnoses and does not reflect those patients that have linked to HIV care, which is likely lower.

We found some evidence that physician knowledge of the patient's place of birth was incomplete even though they were free to use the EMR if they chose, and the number of patients they were asked to review was small enough to make this feasible. They were not confident in 10% of their responses and when those responses were excluded, the apparent performance of the algorithm improved slightly. We found that for half of the disagreements between the algorithm and physician review, the medical record contained documentation supporting the algorithm and not the physician. Physicians may not obtain or may forget a patient's place of birth, making physician review something short of a gold standard.

Patients born in Puerto Rico were commonly misclassified by the algorithm. Nearly half of patients (23/48) identified as of Puerto Rican birth by their provider were coded foreign language speakers and classified as foreign-born by the algorithm in stage 1 and not considered further. In a secondary analysis, we assessed a possible refinement of the algorithm to address the challenge in classifying patients from Puerto Rico. We reviewed tokens for coded foreign language speakers who had the keyword "Puerto Ric" appearing in their medical record and classified foreign-born status. This method only modestly improved the overall accuracy of the algorithm since the number of patients was small (n = 191) in proportion to the total cohort size (n = 9,114). In populations where the proportion of primary Spanish-speaking Puerto Ricans is larger, this adjusted classification may improve the overall accuracy of the algorithm.

In the United States, HIV epidemiology of foreign-born persons has primarily come from seroprevalence data [18-20]. However, outcomes research examining the relationship between foreign-born status and HIV clinical outcomes have been limited by infrequent reporting of foreign-born status, and studies that have relied on the medical chart for confirmation of foreign-born status have been based on small cohort studies identified from HIV clinics [8, 19, 21, 22]. A recent HIV outcomes study used the absence of a valid social security number as a surrogate for undocumented immigrant status, another indirect measurement in light of no coded variables for foreign-born identity [23]. A Veterans Administration-based study similarly intended to improve detection of homelessness from the EMR rather than rely solely on administrative coding for diagnostic capture [12]. The investigators used human review to create a learning set to train an NLP-based tool. Application of the model took three months and precision was 70% based on the error analysis by human review of the medical documentation. Rather, we retain the component of human review, particularly crucial for conditions where documentation is non-standard and diagnostic status implied in the free

text notes of the EMR, and apply a three-stage methodology to achieve a highly efficient and accurate mechanism to extract clinical information from the EMR. To our knowledge, our report is the first to examine the application of an algorithm to a large EMR-based patient registry for the classification of patients' foreign-born status to facilitate HIV outcomes research.

We developed the algorithm with the goal that it would be applicable to other EMR-based clinical registries given that it requires minimal computer code. We also anticipate that the algorithm could be adapted to identify other variables of interest that are under-coded or not explicitly coded, such as underreported diseases or conditions, medication side effects, and social history items (e.g. domestic violence). The study cohort was sizable and drawn from the largest healthcare provider in Massachusetts. We employed ICD-9 diagnosis codes to identify HIV-infected patients and relied on clinical notes and hospital discharge summaries as the source of keyword search. The countries included in the keyword search reflect the HIV-infected foreign-born population in the Boston metropolitan area, but could easily be adjusted for other registry locations based on location and specific purpose.

This study has several limitations. Since information on foreign-born status is often expressed in the EMR in a non-standard fashion and requires human review to limit misclassification, we did not compare the performance of this method to alternative machine-based strategies. Due to feasibility concerns we did not validate the token review (stage 3) against full chart review for each patient; instead we selected a random sample of the cohort for validation. The cohort was restricted to patients who are HIV-infected and live in a Northeastern state, two characteristics that may limit the generalizability, including predictive values, in our results. These factors also contribute to a relatively high proportion of foreign-born patients in the cohort, which may affect the performance of the algorithm in other settings, but is modifiable based on HIV prevalence and foreign-born demographics of the study setting.

Conclusions and Clinical Relevance

Large clinical data registries are underutilized sources to characterize patterns of HIV care utilization by foreign-born persons and risk factors for poor clinical outcomes. Techniques to mobilize the EMR for efficient and accurate identification of patients, such as foreign-born status and other conditions, will enhance the value of EMRs for comparative effectiveness research.

Conflicts of Interest

All authors have declared that no competing interests exist.

Human Subjects Protections

Study procedures were approved by Partners HealthCare Human Research Committee (Boston, Massachusetts, USA).

Acknowledgments

The authors appreciate the efforts of Dr. Shawn Murphy and the Partners HealthCare Research Patient Data Registry group in facilitating use of the database.

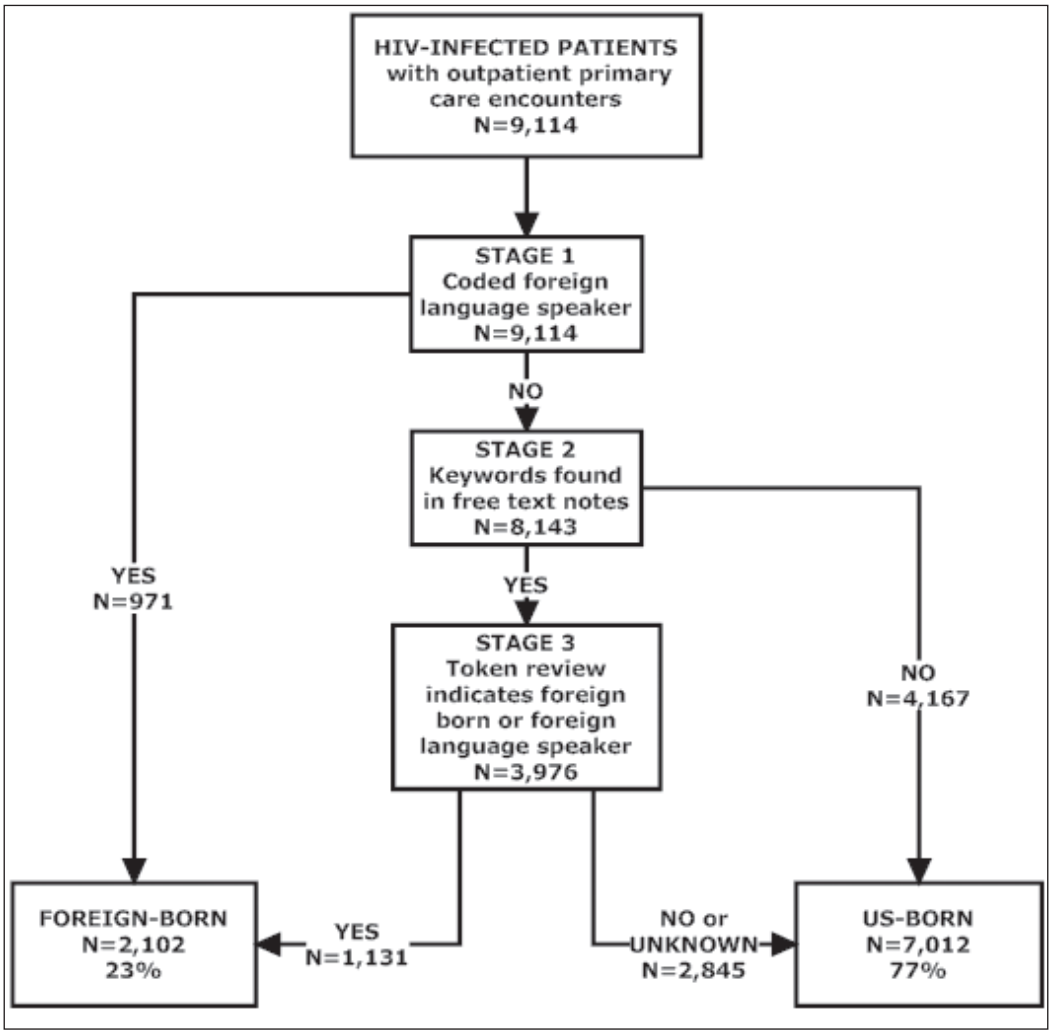

Fig. 1 Flow chart for identification of foreign-born individuals from a large electronic patient data registry.

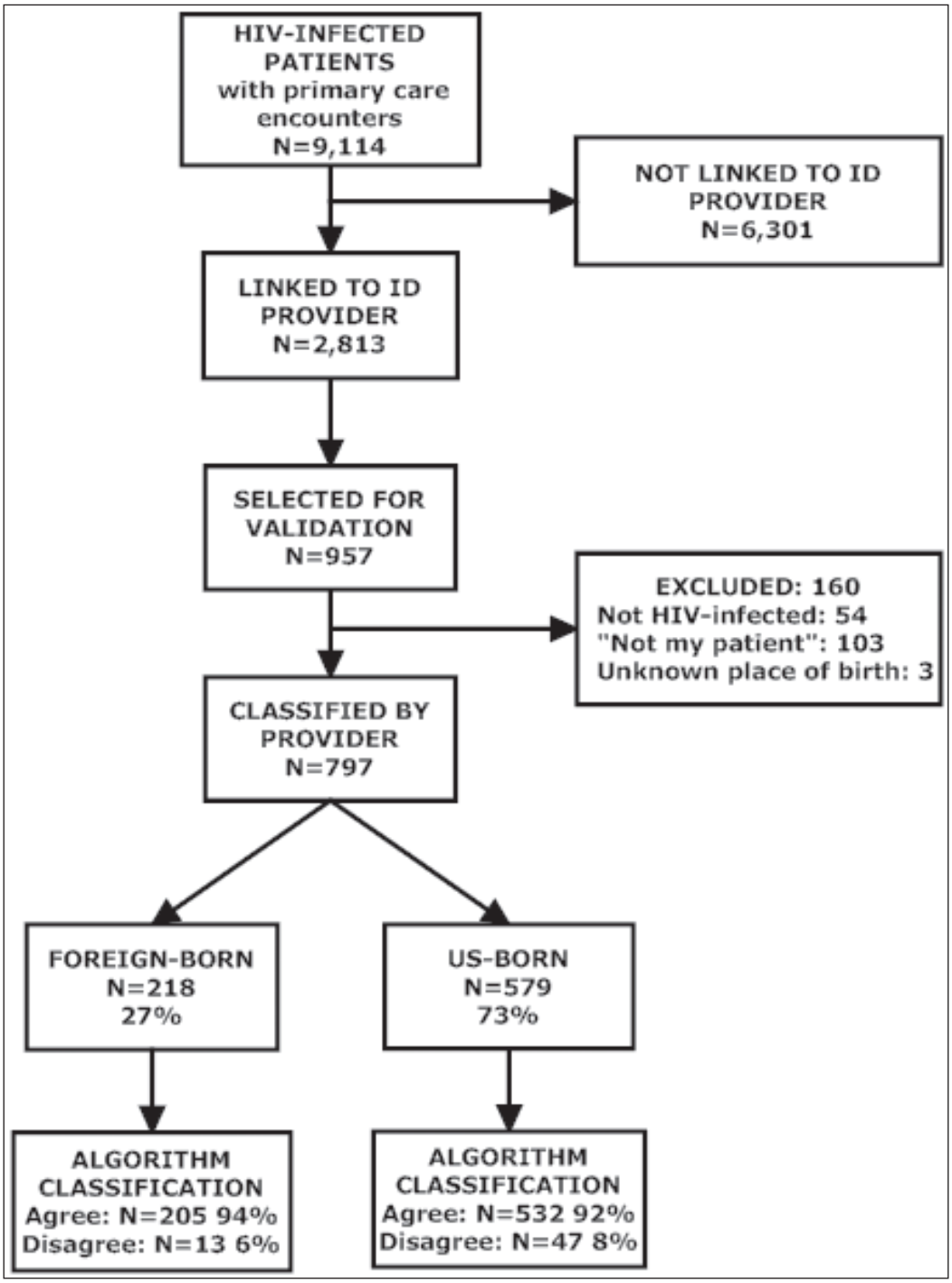

Fig. 2 Flow chart for validation of algorithm against HIV provider classification. Abbreviation: ID, infectious disease.

Algeria	green card	Sao Tome
Angola	Guatemal	Senegal
Argentin	Guiana	Sierra Leone
Asylum	Guinea	Somali (included with Mali)
Belize	Guyana	South Africa
Benin	Haiti	Spanish
Bolivia	Hondura	Sudan
born and raised	immigra	Surinam
born in	interpreter	Swaziland
Botswana	interpretor	Tanzania
Brazil	Ivoire	Togo
Burkina Faso	Kenya	Torture
Burundi	Lesotho	Translate
Cameroon	Liberia	Tunisia
Camp	Libya	Uganda
Canad	Madagascar	Urugua
Cape verde	Malawi	Venezuela
Chad	Mali	visa
Chile	Maurit	Zambia
Chinese	Mexic	Zimbab
Colombia	Morocc	
Comoros	moved to the US	
Congo	moved to US	
Costa Ric	Mozambiqu	
creole	Namibia	
Djibouti	Nicaragua	
Dominican	Niger (includes Nigeria)	
Ecuador	originally from	
Egypt	Panama	
emigrat	Paragua	
Eritrea	Peru	
Ethiopia	Portug	
foreign	Puerto Ric	
Gabon	refugee	
Gambia	Rwanda	
Ghana	Salvador	

Table 1 Comprehensive List of Keywords Included in Computerized Search of the Electronic Medical Record for HIV-infected Foreign-born Patients.

Table 2 Test Characteristics for a Computer-based Algorithm to Identify Foreign-born Patients in the Electronic Medical Record.

	Sensitivity (95% CI)	Specificity (95% CI)
Primary analysis	94% (90.9, 97.2%)	92% (89.7, 94.1%)
Secondary analyses		
Only confident physician responses	97% (94.5, 99.4%)	93% (90.2, 94.8%)
Adjustment for foreign language speakers with key-words positive for Puerto Rico	90% (85.1, 93.3%)	95% (92.7, 96.4%)
Exclusion of provider-misclassified patients	97% (94.3, 99.1%)	95% (93.0, 96.7%)

Abbreviation: CI, confidence interva

References

1. Bureau of the Census, US Department of Commerce. State and County QuickFacts. Washington, DC: Bureau of the Census; 2013. Available at http://quickfacts.census.gov/qfd/states/00000.html. Accessed on February 16, 2014.
2. Derose KP, Escarce JJ, Lurie N. Immigrants and health care: sources of vulnerability. Health Aff (Millwood) 2007 Sep-Oct;26(5):1258-68.
3. Martinez O, Wu E, Sandfort T, Dodge B, Carballo-Dieguez A, Pinto R, Rhodes S, Moya E, Chavez-Baray S. Evaluating the Impact of Immigration Policies on Health Status Among Undocumented Immigrants: A Systematic Review. J Immigr Minor Health Dec 28.
4. Wohl AR, Galvan FH, Myers HF, Garland W, George S, Witt M, Cadden J, Operskalski E, Jordan W, Carpio F. Social support, stress and social network characteristics among HIV-positive Latino and African American women and men who have sex with men. AIDS Behav 2010 Oct;14(5):1149-58.
5. Wohl AR, Galvan FH, Myers HF, Garland W, George S, Witt M, Cadden J, Operskalski E, Jordan W, Carpio F, Lee ML. Do social support, stress, disclosure and stigma influence retention in HIV care for Latino and African American men who have sex with men and women? AIDS Behav 2011 Aug;15(6):1098-110.
6. Keesee MS, Natale AP, Curiel HF. HIV positive Hispanic/Latinos who delay HIV care: analysis of multilevel care engagement barriers. Soc Work Health Care 2012;51(5):457-78.
7. Gilbert PA, Rhodes SD. HIV testing among immigrant sexual and gender minority Latinos in a US region with little historical Latino presence. AIDS Patient Care STDS 2013 Nov;27(11):628-36.
8. Akinsete OO, Sides T, Hirigoyen D, Cartwright C, Boraas C, Davey C, Pessoa-Brandao L, McLaughlin M, Kane E, Hall J, Henry K. Demographic, clinical, and virologic characteristics of African-born persons with HIV/AIDS in a Minnesota hospital. AIDS Patient Care STDS 2007 May;21(5):356-65.
9. Uzuner O, Goldstein I, Luo Y, Kohane I. Identifying patient smoking status from medical discharge records. J Am Med Inform Assoc 2008 Jan-Feb;15(1):14-24.
10. Zeng QT, Goryachev S, Weiss S, Sordo M, Murphy SN, Lazarus R. Extracting principal diagnosis, co-morbidity and smoking status for asthma research: evaluation of a natural language processing system. BMC Med Inform Decis Mak 2006;6:30.
11. Sada Y, Hou J, Richardson P, El-Serag H, Davila J. Validation of case finding agorithms for hepatocellular cancer from administrative data and electronic health records using natural language processing. Med Care Aug 6.
12. Gundlapalli AV, Carter ME, Palmer M, Ginter T, Redd A, Pickard S, Shen S, South B, Divita G, Duvall S, Nguyen TM, D'Avolio LW, Samore M. Using natural language processing on the free text of clinical documents to screen for evidence of homelessness among US veterans. AMIA Annual Symposium proceedings / AMIA Symposium. AMIA Symposium 2013;2013:537-46.
13. Kavuluru R, Hands I, Durbin EB, Witt L. Automatic extraction of ICD-O-3 primary xites from cancer pathology reports. AMIA Summits on Translational Science proceedings AMIA Summit on Translational Science 2013;2013:112-6.
14. City of Boston, Mayor's Office of New Bostonians. New Bostonians Demographic Report. Boston: Mayor's Office of New Bostonians; 2004. Available at http://www.cityofboston.gov/newbostonians/pdfs/dem_report.pdf. Accessed February 16, 2014.
15. Massachusetts Department of Health and Human Services. Refugee Arrivals to Massachusetts by Country of Origin. Boston: Massachusetts Department of Health and Human Services; 2013. Available at http://www.mass.gov/eohhs/gov/departments/dph/programs/id/public-health-cdc-refugee-arrivals.html. Accessed February 16, 2014.
16. Bureau of the Census, US Department of Commerce. American Community Survey, 5-Year Estimates. Updated every year. Washington, DC: Bureau of the Census; 2013. Available at http://quickfacts.census.gov/qfd/meta/long_POP645210.htm. Accessed February 16, 2014.
17. Massachusetts Department of Public Health Office of HIV/AIDS. Massachusetts HIV/AIDS Data Fact Sheet. People Born Outside the United States. Available at http://www.mass.gov/eohhs/docs/aids/2010-profiles/born-outside-us.pdf. Accessed on February 16, 2014.
18. Kent JB. Impact of foreign-born persons on HIV diagnosis rates among Blacks in King County, Washington. AIDS Educ Prev 2005 Dec;17(6 Suppl B):60-7.
19. Harawa NT, Bingham TA, Cochran SD, Greenland S, Cunningham WE. HIV prevalence among foreign- and US-born clients of public STD clinics. Am J Public Health 2002 Dec;92(12):1958-63.
20. Marc LG, Patel-Larson A, Hall HI, Hughes D, Alegria M, Jeanty G, Eveillard YS, Jean-Louis E. HIV among Haitian-born persons in the United States, 1985-2007. AIDS Aug 24;24(13):2089-97.

21. Beckwith CG, DeLong AK, Desjardins SF, Gillani F, Bazerman L, Mitty JA, Ross H, Cu-Uvin S. HIV infection in refugees: a case-control analysis of refugees in Rhode Island. Int J Infect Dis 2009 Mar;13(2):186-92.
22. Antiretroviral Therapy Cohort Collaboration (ART-CC). Influence of geographical origin and ethnicity on mortality in patients on antiretroviral therapy in Canada, Europe, and the United States. Clin Infect Dis 2013 Jun;56(12):1800-9.
23. Poon KK, Dang BN, Davila JA, Hartman C, Giordano TP. Treatment outcomes in undocumented Hispanic immigrants with HIV infection. PLoS One;8(3):e60022.