Using Clinical Notes and Natural Language Processing for Automated HIV Risk Assessment

Daniel J. Feller, BA, Jason Zucker, MD,† Michael T. Yin, MD,† Peter Gordon, MD,† and Noémie Elhadad, PhD**

Objective: Universal HIV screening programs are costly, labor intensive, and often fail to identify high-risk individuals. Automated risk assessment methods that leverage longitudinal electronic health records (EHRs) could catalyze targeted screening programs. Although social and behavioral determinants of health are typically captured in narrative documentation, previous analyses have considered only structured EHR fields. We examined whether natural language processing (NLP) would improve predictive models of HIV diagnosis.

Methods: One hundred eighty-one HIV+ individuals received care at New York Presbyterian Hospital before a confirmatory HIV diagnosis and 543 HIV negative controls were selected using propensity score matching and included in the study cohort. EHR data including demographics, laboratory tests, diagnosis codes, and unstructured notes before HIV diagnosis were extracted for modeling. Three predictive algorithms were developed using machine-learning algorithms: (1) a baseline model with only structured EHR data, (2) baseline plus NLP topics, and (3) baseline plus NLP clinical keywords.

Results: Predictive models demonstrated a range of performance with F measures of 0.59 for the baseline model, 0.63 for the baseline + NLP topic model, and 0.74 for the baseline + NLP keyword model. The baseline + NLP keyword model yielded the highest precision by including keywords including "msm," "unprotected," "hiv," and "methamphetamine," and structured EHR data indicative of additional HIV risk factors.

Conclusions: NLP improved the predictive performance of automated HIV risk assessment by extracting terms in clinical text indicative of high-risk behavior. Future studies should explore more advanced techniques for extracting social and behavioral determinants from clinical text.

Key Words: predictive analytics, social determinants of health, HIV, natural language processing, prevention

(*J Acquir Immune Defic Syndr* 2018;77:160–166)

Received for publication July 14, 2017; accepted October 13, 2017.

From the *Department of Biomedical Informatics, Columbia University, New York, NY; and †Division of Infectious Diseases, Department of Medicine, Columbia University, New York, NY.

Supported by National Library of Medicine—T15 LM007079: "Training in Biomedical Informatics at Columbia University," National Institutes of Health—T32 AI007531: "Training in Pediatric Infectious Diseases at Columbia University."

The authors have no conflicts of interest to disclose.

Correspondence to: Daniel J. Feller, BA, Department of Biomedical Informatics, Columbia University, 622 West 168th Street, New York, NY 10032 (e-mail: djf2150@cumc.columbia.edu).

BACKGROUND

Public health programs aimed at preventing the spread of HIV and other sexually transmitted infections (STIs) rely on the identification of individuals at elevated risk of infection and HIV+ persons unaware of their diagnoses.[1] HIV screening guidelines including those promulgated by the CDC and USPSTF recommend both universal screening and targeted screening of high-risk individuals as part of a comprehensive approach to infection identification.[2,3] Current guidelines suggest that all patients receive at least 1 lifetime HIV test, whereas those who exhibit HIV risk factors and should be screened for HIV annually. Universal HIV screening programs are, however, unlikely by themselves to fully identify all individuals at risk of active HIV infection, as they may miss incident cases when screening intervals are extended or missed altogether.[4–10] Universal HIV screening programs are also costly and labor intensive, 2 attributes that make such initiatives difficult to effectively implement in resource-constrained settings.[11–15] Targeted HIV screening is recognized as an essential component of comprehensive population HIV screening programming given its cost-effectiveness and potential ability to identify incident HIV infections. Targeted screening programs rely on the ability to elucidate social and behavioral risk factors associated with an increased risk of the acquisition of HIV. Although clinical prediction rules exist, these are resource intensive to manually administer and are seldom used.[16–19] Internet-based risk assessments exist as an alternative to clinical surveys but rarely generate significant participation.[20,21] There exists an unmet need for innovative approaches to HIV testing that are cost-effective.

Although electronic health records (EHRs) have the potential to improve the effectiveness and efficiency of preventive care by automating risk assessment, limited research has explore its use in the context of assessing risk of HIV and other sexually transmitted diseases.[16,17] At present, the most sophisticated method for identifying candidates for HIV screening executes simple logic on previous STI testing history and select diagnosis codes.[18] Krakower et al[19] recently proposed the use of machine learning to identify individuals at high risk of HIV using routinely collected clinical data. Their preliminary model relied on structured EHR fields including diagnosis codes and laboratory tests that capture a limited amount of information regarding HIV risk factors. Explicit details of social and behavioral determinants such as sexual orientation, and

sexual activity are typically collected in a narrative or seminarrative format within the social history section of clinical notes.[20,21] Natural language processing (NLP), a subdiscipline of computer science concerned with the use of computers to extract meaning from human language, is a well-established means of extracting information from clinical notes.[22] This suggests an opportunity to leverage NLP to extract critical HIV risk information from unstructured EHR data. Several NLP techniques have successfully leveraged clinical documentation in tasks such as the prediction of hospital readmission or chronic kidney disease progression.[22–25]

Our present study was motivated by the need to understand whether information found in clinical records can identify individuals at elevated risk of HIV infection. We compare the prognostic ability of several machine-learning approaches that feature NLP for HIV risk assessment. We trained predictive models using structured EHR data and content extracted from clinical notes including automatically identified clinical keywords predictive of future HIV infection and automatically identified topics inferred by a latent Dirichlet allocation (LDA), a probabilistic graphical model that discovers linguistic themes in clinical documentation.

METHODS AND MATERIALS

Data Set

We queried the Clinical Data Warehouse (CDW) at New York Presbyterian Hospital—Columbia University Medical Center, a large academic medical center in New York City, which has collected clinical data from about 5 million patients since 1995. The medical center includes 2 hospitals and a collection of outpatient clinics in the New York metropolitan area. Because the CDW began integrating data from the hospital's EHR system and multiple ancillary systems in mid 2006, we limited our data extraction from January 1, 2007, to December 31, 2015.

Five hundred seventy-seven persons were diagnosed with HIV at the medical center during this timeframe. Two hundred seventy-eight of the aforementioned group had evidence of health care encounters before a confirmatory HIV diagnosis at CUMC. Ninety-seven patients from this group were excluded from the HIV+ cohort because their initial visits to CUMC came within 7 days of a concurrent HIV/AIDS diagnosis.

Propensity score matching was used to identify a sample of 543 HIV-uninfected individuals with similar health care utilization patterns compared with the sample of 181 HIV-infected individuals, resulting in a 3:1 ratio of cases to controls. Our matching algorithm considered the number of emergency department (ED), inpatient, ambulatory surgery, and outpatient visits as independent variables and used these data to assess the similarity of HIV-infected and HIV-uninfected controls. The nonparametric "nearest-neighbor" algorithm included in the MatchIt package in R version 3.0.1 (The R Foundation for Statistical Computing, Vienna, Austria) was used for propensity score matching.

EHR Processing

We extracted all clinical data generated more than 5 days before confirmatory HIV diagnosis for individuals in the HIV-infected sample and all longitudinal data for uninfected controls. All encounters 5 days and fewer before a confirmatory HIV diagnosis were excluded to ensure that no information regarding positive HIV testing and diagnosis was made informative to models. An overview of the extraction of data from EHRs is outlined in Figure 1. Because the machine-learning approach automatically selected variables predictive of HIV, there was no cost to including a large number of preliminary variables.

Demographics

We obtained age, sex, race/ethnicity, marital status, and insurance payer from patient records and analyzed these data retrospective to the patient's most recent health care encounter.

Visit History

Each patient's longitudinal visit history at New York Presbyterian was queried from the CDW. We considered the count of each patient's outpatient, inpatient, ambulatory surgery, and ED visits as independent variables.

Diagnoses

We extracted International Classification of Diseases—Clinical Modification, ninth edition (*ICD-9*-CM) diagnoses codes from the CDW. *ICD-9*-CM codes were truncated to a whole number without rounding, as has been done in previous studies.[24] Each *ICD* variable represented the number of times each truncated code was documented in the patient's longitudinal record.

Laboratory Tests

Physicians with domain expertise identified several laboratory tests potentially suggestive of high-risk sexual activity including screening for HIV, serum hepatitis B surface antigen or antibody, hepatitis C antibody and syphilis by RPR/FTA positivity, and gonorrhea/chlamydia by nucleic acid amplification at any site. It should be noted that we considered the evidence of testing (not specific results) of each STI laboratory test as a discrete binary variable (eg, gonorrhea/chlamydia test performed). Knowledge about a test order has been found an informative variable in several predictive tasks.[26,27]

Clinical Notes

The notes included in this study were a collection of select physician, nursing, and social work note types within New York Presbyterian's EHR that were identified as potentially capturing indicators of HIV risk. Note types included in the study were admission notes, discharge summaries, outpatient primary care notes, and inpatient progress notes. We excluded many other inpatient notes and hand-offs (aka "sign-outs"). We preprocessed clinical documents to remove punctuation, numbers, English stop words, and section headers.

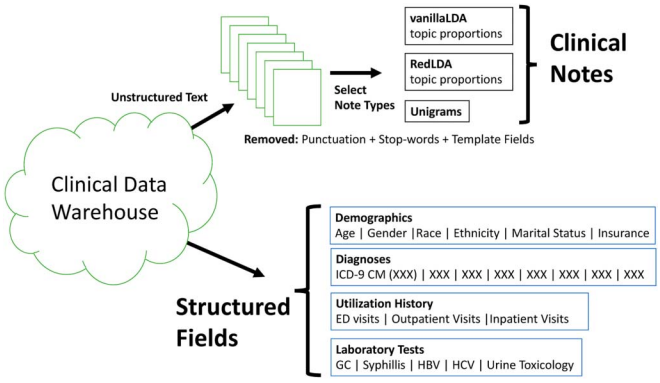

FIGURE 1. Overview of EHR feature engineering process.

Natural Language Processing

We compared 2 NLP methods to extract variables from clinical notes that both empirically identify words and learn themes that are predictive of future HIV diagnosis.

Automated Keyword Identification

Individual words contained in clinical notes may be predictive of future HIV infection. We identified words with possible high information value by representing each word according to their respective term frequency-inverse document frequency weight. This is essential to the analysis of clinical documentation, as many words occur in a large proportion of clinical notes such as "allergies" or "symptoms".[24,28] The vocabulary of our corpus consisted of approximately 100,000 unique words and we used univariate χ^2 tests to identify a smaller subset of words indicative of high-risk behavior. We calculated Pearson $\chi 2$ test statistic for each word and selected 300 with highest measure of association. From these, 37 clinical keywords related to established HIV risk factors were manually selected for inclusion in the predictive model.

Automated Topic Modeling

Topic modeling is a frequently used text mining method that can discover abstract "topics" within a collection of documents. A topic consists of a collection of words that frequently occur together. Topic modeling has proven to be useful technique for analyzing clinical notes and has demonstrated utility for predictive modeling.[23,29–32] We trained topic models using LDA, a robust and established method for unsupervised topic modeling.[33–35] LDA takes as input a corpus of notes and learns K clusters, where each cluster is represented as a distribution of words in the corpus. Given a new document, LDA can infer the presence or the absence of the previously learned K topics. Accordingly, each document processed using topic modeling is represented as a distribution over the K learned topics. Domain experts characterized a model with 250 topics as producing the most coherent topics. The genism library in Python version 2.7 was used to fit LDA models.

Variable Selection and Statistical Modeling

Because the inclusion of a large number of uninformative variables can limit the accuracy of machine-learning algorithms, we identified a subset of the most useful variables using mutual information criteria.[36] Mutual information quantifies the dependence of 2 random variables and can account for both linear and nonlinear associations. Preprocessing yielded 1,583 variables from structured EHR fields including demographics, diagnoses, and laboratory tests, 35 unigram variables, and 250 topics generated by LDA. For each model independently, we selected 150 variables by

TABLE 1. Cohort Identified Using Propensity Score Matching

	HIV+ (N = 181)	HIV− (N = 543)	Cohort (N = 724)	χ^2†, *P*
Age				<0.005
0–20	4 (2.2%)	91 (16.7%)	95 (13.1%)	
20–30	67 (37.0%)	166 (30.5%)	233 (32.2%)	
30–40	47 (26.0%)	149 (27.4%)	196 (27.1%)	
40–50	39 (21.5%)	107 (19.7%)	146 (20.2%)	
50+	8 (4.4%)	30 (5.5%)	38 (5.2%)	
Sex				<0.005
Male	124 (68.5%)	184 (33.9%)	308 (42.5%)	
Female	41 (22.6%)	359 (66.1%)	400 (55.2%)	
Ethnicity				0.88
Hispanic	55 (30.4%)	168 (30.9%)	223 (30.8%)	
Non-Hispanic	126 (69.6%)	375 (69.0%)	501 (69.2%)	
Utilization history*				
# ED visits	2	1	1	
# Inpatient visits	0	1	1	
# Outpatient visits	2	4	4	

*Median.
†Chi-square statistic.

TABLE 2. Performance Using 30 Bootstrap Iterations of 10-fold Cross-Validation

	Precision (%)	Recall (%)	F1 (%)
Baseline	66.3 (64.2–68.1)	53.7 (50.7–56.6)	59.2 (56.6–61.8)
Unigrams	79.1 (77.1–80.2)	68.6 (67.2–69.9)	73.3 (72.1–74.5)
LDA (K = 250)	66.9 (62.6–80.7)	62.1 (59.2–64.9)	64.3 (61.6–67.1)

identifying the variables with the highest mutual information computed against instance labels (HIV+/HIV−).

Subsequent to variable selection, we used random forest classifiers to generate models because they are easy to tune, robust to overfitting, and provide a measure of variable importance and therefore enable interpretation.[37,38] Given the relatively small sample size, we used cross-validation to generate a robust estimate of the predictive models. We evaluated each model using precision, recall, and the F measure. In the context of HIV acquisition, precision (or positive predictive value) quantifies the proportion of individuals predicted to be HIV infected (by the algorithm) that were truly HIV infected. By contrast, recall (or sensitivity) quantifies the proportion of HIV-infected individuals among all predicted to be HIV infected (by the algorithm). The F measure is the weighted average of precision and recall and thus provides a high-level overview of model performance. We used a precision-recall plot to compare the 3 models, which is preferable to a receiver operating characteristic curve when evaluating classifiers trained on imbalanced data.[39] The algorithm was trained using the scikit-learn library in Python.

For the sake of interpretation, we also ranked the variables in each model according to their importance, as determined by mean decrease in node purity (measured using the Gini index) averaged across all trees.[40] We used sci-kit learn in Python version 2.7 to develop and evaluate all models.

RESULTS

The analysis included 181 HIV-infected individuals who received health services before a confirmatory HIV Western blot test and 543 matched uninfected controls (Table 1). All individuals in the cohort received care at Columbia University Medical Center between 2007 and 2015. Individuals in the uninfected sample were significantly older than those in the HIV-infected sample ($P < 0.05$). HIV-infected individuals had an average record length of 410 days (±629) and an average of 28.9 unique encounters, whereas uninfected individuals had an average record length of 982 days (±1,007) and on average 37.8 unique encounters. HIV-infected individuals were associated with 4,808 notes (average 26.5 notes per MRN) and uninfected individuals were associated with 27,147 notes (average. 49.9 notes per MRN).

Clinical keywords identified using NLP and listed in Table 3 reflect established risk factors for HIV diagnosis including sexual orientation ("homosexual" and "msm"), high-risk sexual activities ("anal" and "unprotected"), and previous testing and/or diagnosis of STIs ("sti," "hiv," "testing," "tested," "cervical," and "meningitis"). In addition, the model also included a number of terms related to drug use (amphetamine, cocaine, and meth), housing instability ("homeless"), and psychological comorbidities ("psychiatrist," "psychology," and "psychiatrist"). Structured EHR data including demographics (sex), diagnoses (*ICD*: 079, 79, 296), and health care utilization history (# past visits) also had high variable importance.

We trained predictive models using a 25% prevalence sampling and model performance is displayed in Table 2. We observed F measures of 0.59 for the baseline model, 0.69 for the baseline plus NLP topic modeling, and 0.74 for the baseline plus NLP clinical keyword model. The baseline + NLP clinical keyword model displayed the highest precision (0.81) and recall (0.67), whereas the baseline model exhibited the lowest precision and recall (0.68 and 0.53, respectively). A plot of precision and recall across predicted probabilities from the Random Forest algorithms is presented in Figure 2. Each model was trained using 150 variables selected using mutual

TABLE 3. Strongest Positive and Negative Predictors Within Predictive Models

Baseline Model	Unigram Model*	LDA Model
Ethnicity—Hispanic (0.41)	"hiv" (0.27)	Urine toxicology test (0.41)
# Previous outpatient visits (0.097)	# Past visits (0.085)	*ICD* 304—Drug dependence (0.077)
ICD 304—Drug dependence (0.036)	*ICD* 079—Viral and chlamydial infection (0.032)	*ICD* 424—Other diseases endocardium (0.026)
ICD 305—Nondependent drug use (0.036)	Sex = Female (0.025)	Topic 136 (0.025)†
# Previous ED visits (0.034)	Sex = Male (0.024)	# Previous outpatient visits (0.020)
ICD 292—Drug-induced mental disorders (0.029)	*ICD* 79—Fracture and dislocation (0.018)	Topic 28 (0.017)
# Previous inpatient visits (0.028)	"msm" (0.017)	*ICD* 305—Nondependent drug use (0.01)
ICD 217—Benign neoplasm (0.025)	"cervical" (0.015)	Topic 182 (0.009)
Urine toxicology test (0.024)	"lymph" (0.015)	*ICD* 440—Atherosclerosis (0.008)
ICD 719—Other joint disorders (0.016)	*ICD* 296—Episodic mood disorder (0.013)	Topic 94 (0.007)†
ICD 249—Diabetes mellitus (0.011)	"unprotected" (0.013)	Age (0.006)
F1 = 59.2%	**F1 = 73.3%**	**F1 = 64.3%**

*Additional clinical keywords: amphetamine, anal, cocaine, condom, crack, crisis, enlarged, hepatitis, homeless, homosexual, ivd, lymphadenopathy, male, man, men, meningitis, mens, meth, neurosyphillis, psychiatrist, seronegative, sex, sexual, sti, strep, tb, tested, testing, transgender, viral, psychology, and psychiatrist.
†Negative predictor.

information criteria. The 11 strongest negative and positive predictors for each model (variables with the highest mean decrease in node impurity) are detailed in Table 3, and we list 4 topics with high variable importance in Table 4. However, it is important to note that the Random Forest models used all 150 variables to discriminate low- and high-risk individuals. Diagnosis codes and demographics were chosen as predictors in each model, although the baseline + NLP clinical keyword model relied most heavily on information extracted from clinical notes.

DISCUSSION

Our findings suggest that clinical notes exist as a valuable source of information on HIV risk factors including drug use and high-risk sexual activity. Extracting variables using NLP improved the performance of predictive models for HIV risk compared with a model using only structured EHR data. The clinical keyword model achieved the highest performance by identifying terms in clinical notes indicative of high-risk behavior. Established HIV risk factors including sexual orientation, high-risk sexuality activity, and history of STIs were included in the keyword model. Social determinants of health are increasingly recognized as predictors of HIV infection and were also included in the model through the inclusion of terms related to drug use, housing instability, and psychological comorbidities.[41,42] However, structured EHR data also had high variable importance in the predictive models and therefore unstructured clinical text and structured EHR data exist as complementary sources of information for automated HIV risk assessment.

An NLP technique called topic modeling was used to discover themes in clinical text capable of broadly distinguishing low- and high-risk individuals. Two topics with high variable importance were negatively correlated with future HIV diagnosis (topic 94; older women and topic 136; cardiovascular disease) and likely represented clinical notes associated with older, acutely ill individuals who lacked established HIV risk factors. Clinical experts (J.Z., M.T.Y., and P.G.) failed to identify a strong interpretation of topics 136 and 28 (pneumonia and intensive care unit care) which may reflect the fact that the baseline + NLP topic model was optimized for prognostic ability rather than interpretation. In addition, previous studies have found topic modeling to provide varying levels of clinical interpretability when evaluated by physicians.[43,44] The inferior interpretability and prognostic ability of the baseline + NLP topic model compared with the baseline + NLP keyword model suggests that future studies should focus on extracting individual HIV risk factors from clinical text rather linguistic topics.[45]

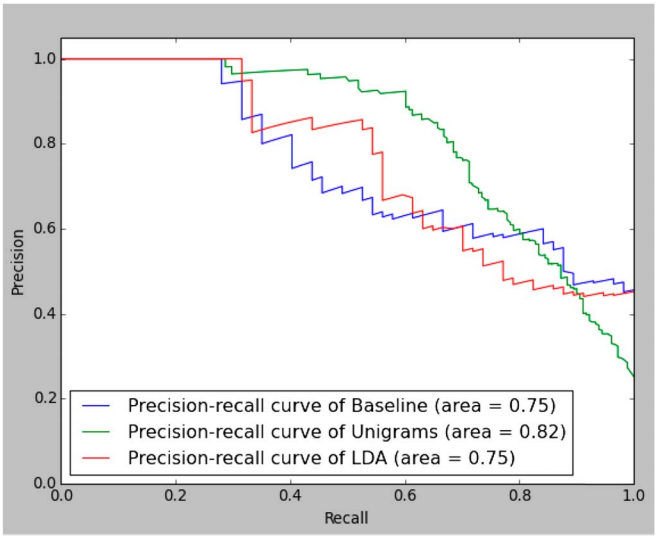

FIGURE 2. Precision and recall for the 3 modeling approaches (area = area under the curve).

Figure 2 demonstrates the precision–recall curve of the 3 models. Precision in this context is the percentage of cases identified by a predictive model as being high risk that are actually at risk of imminent HIV diagnosis. Recall is the percentage of individuals actually at risk of imminent HIV diagnosis that are identified as such by the predictive model. Balancing precision and recall is critical when considering how predictive analytics will be used in clinical practice or public health initiatives. For example, a universal HIV screening program in an ED would likely benefit from a model with relatively high recall given the low cost of screening and an emphasis on reducing missed opportunities for HIV diagnosis. By contrast, a predictive model with high precision would be desired for interventions that feature direct outreach to high-risk individuals by electronic means such as an alert through a personal health record or text messaging. Concerns about harming the patient–provider relationship and retaining patient engagement require targeted messaging that is precise and relevant to each patient.[46,47] Other potential scenarios for targeted HIV prevention exist and require a careful balancing of whether to prioritize precision or recall.

This study has several limitations that should be considered. First, preliminary analyses found that K = 250 produced the most coherent topics as characterized by a domain expert. However, we identified 250 and 300 topics as displaying the highest relative F1 score, and thus, performance may not improve with an empirical method.[35] Second, unigram model did not account for lexical variants by using lemmatization or a biomedical lexicon similar to the commonly used Specialist Lexicon within the publicly available Unified Medical Language System.[48] Third, because notes often contain negative findings such as "the patient denies use of illicit drugs and alcohol," model performance may have been improved by performing word-sense disambiguation or considering negation.[25] Fourth, it is important to note that a considerable amount of information was lost because of the use of templated notes at Columbia University Medical Center. We encountered more than a 100 different note templates enumerated with various HTML strings such as "Current Drug Use." Our findings suggest that EHR vendors and health care providers may benefit from unstructured notes to support text mining initiatives. Fifth, we did not externally validate our HIV prediction model within another institution's EHR. The current lack of interoperability among EHR

TABLE 4. Topics With Strong Positive and Negative Predictive Values

Topic #136* (Cardiovascular)	Topic #28 (IVDU)	Topic #182* (Intensive Care Unit)	Topic #94* (Older Women)
Valve	Nebs	Intubated	Uterus
Mitral	Augmentin	Sedated	Adnexal
Aortic	Baclofen	Grasps	Receive
Regurgitation	Albuterol	Bacteremic	Spirometry
Ventricular	Methadone	Alert	Infrarenal
Stenosis	Vodka	Admitted	Functional capacity

*Negative predictor.

platforms challenges the portability of predictive models and additional studies are needed to determine whether the model is generalizable. Last, population health analyses are challenging to perform in the New York City metropolitan area due to the fact that patients often receive care from multiple health care providers who do not exchange clinical data. Future analyses would likely benefit from the use of data from public Health Information Exchanges similar to those in New York City.[49]

CONCLUSIONS

Ending the HIV epidemic will require identifying high-risk individuals for HIV testing and referral to comprehensive prevention services. Universal screening methods hold enormous value for identifying individuals infected with HIV but may become less cost-effective, as the prevalence of undiagnosed HIV infection decreases. In contrast to contemporary targeted HIV screening programs that are outside the scope of many medical practices, predictive analytics could automate HIV risk assessment and be integrated into EHRs through the use of alerts and reminders. We hypothesize that predictive analytics represent a novel approach to HIV prevention and may reduce missed opportunities for screening among individuals that exhibit HIV risk factors. Future studies should explore whether comprehensive data from Health Information Exchanges and more advanced NLP techniques can improve model performance and drive innovative HIV prevention interventions.

REFERENCES

1. Vassall A, Pickles M, Chandrashekar S, et al. Cost-effectiveness of HIV prevention for high-risk groups at scale: an economic evaluation of the Avahan programme in south India. *Lancet Glob Health*. 2014;2:e531–e540.
2. Chou R, Selph S, Dana T, et al. *Screening for HIV: Systematic Review to Update the U.S. Preventive Services Task Force Recommendation*. Rockville, MD: Agency for Healthcare Research and Quality (US); 2012.
3. HIV screening and testing | guidelines and recommendations|HIV/AIDS|CDC. Available at: https://www.cdc.gov/hiv/guidelines/testing.html. Accessed May 16, 2017.
4. Hsieh YH, Kelen GD, Beck KJ, et al. Evaluation of hidden HIV infections in an urban ED with a rapid HIV screening program. *Am J Emerg Med.* 2016;34:180–184.
5. Zucker J, Cennimo D, Sugalski G, et al. Identifying areas for improvement in the HIV screening process of a high-prevalence emergency department. *AIDS Patient Care STDs*. 2016;30:247–253.
6. Liao C, Golden WC, Anderson JR, et al. Missed opportunities for repeat HIV testing in pregnancy: implications for elimination of mother-to-child transmission in the United States. *AIDS Patient Care STDs*. 2017;31:20–26.
7. Myers JJ, Modica C, Dufour MSK, et al. Routine rapid HIV screening in six community health centers serving populations at risk. *J Gen Intern Med.* 2009;24:1269–1274.
8. Cunningham CO, Doran B, DeLuca J, et al. Routine opt-out HIV testing in an urban community health center. *AIDS Patient Care STDs*. 2009;23:619–623.
9. Weis KE, Liese AD, Hussey J, et al. A routine HIV screening program in a South Carolina community health center in an area of low HIV prevalence. *AIDS Patient Care STDs*. 2009;23:251–258.
10. Haukoos JS, Hopkins E, Conroy AA, et al. Routine opt-out rapid HIV screening and detection of HIV infection in emergency department patients. *JAMA*. 2010;304:284–292.
11. Sanders GD, Anaya HD, Asch S, et al. Cost-effectiveness of strategies to improve HIV testing and receipt of results: economic analysis of a randomized controlled trial. *J Gen Intern Med.* 2010;25:556–563.
12. Walensky RP, Weinstein MC, Smith HE, et al. Optimal allocation of testing dollars: the example of HIV counseling, testing, and referral. *Med Decis Mak Int J Soc Med Decis Mak.* 2005;25:321–329.
13. Holtgrave DR. Costs and consequences of the US Centers for Disease Control and Prevention's recommendations for opt-out HIV testing. *PLoS Med.* 2007;4:e194.
14. Sanders GD, Bayoumi AM, Sundaram V, et al. Cost-effectiveness of screening for HIV in the era of highly active antiretroviral therapy. *N Engl J Med.* 2005;352:570–585.
15. Paltiel AD, Weinstein MC, Kimmel AD, et al. Expanded screening for HIV in the United States–an analysis of cost-effectiveness. *N Engl J Med.* 2005;352:586–595.
16. Amarasingham R, Patzer RE, Huesch M, et al. Implementing electronic health care predictive analytics: considerations and challenges. *Health Aff Proj Hope.* 2014;33:1148–1154.
17. Bates DW, Saria S, Ohno-Machado L, et al. Big data in health care: using analytics to identify and manage high-risk and high-cost patients. *Health Aff Proj Hope*. 2014;33:1123–1131.
18. Felsen UR, Cunningham CO, Hey M, et al. Expanded HIV testing strategy leveraging the electronic medical record uncovers undiagnosed infection among hospitalized patients. *J Acquir Immune Defic Syndr.* 2017;1999:27–34.
19. Krakower D, Gruber S, Menchaca JT, et al. Automated identification of potential candidates for human immunodeficiency virus pre-exposure prophylaxis using electronic health record data. *Open Forum Infect Dis.* 2016;3.
20. Adler NE, Stead WW. Patients in context—EHR capture of social and behavioral determinants of health. *N Engl J Med.* 2015;372:698–701.
21. Chen ES, Manaktala S, Sarkar IN, et al. Multi-site content analysis of social history information in clinical notes. *AMIA Annu Symp Proc.* 2011;2011:227–236.
22. Demner-Fushman D, Elhadad N. Aspiring to unintended consequences of natural language processing: a review of recent developments in clinical and consumer-generated text processing. *IMIA Yearb.* 2016;10:224–233.
23. Perotte A, Ranganath R, Hirsch JS, et al. Risk prediction for chronic kidney disease progression using heterogeneous electronic health record data and time series analysis. *J Am Med Inform Assoc JAMIA*. 2015;22:872–880.
24. Walsh C, Hripcsak G. The effects of data sources, cohort selection, and outcome definition on a predictive model of risk of thirty-day hospital readmissions. *J Biomed Inform.* 2014;52:418–426.
25. Topic models for mortality modeling in intensive care units—GhassemiNaumannICML2012.pdf. Available at: http://www.cs.uml.edu/~arum/publications/GhassemiNaumannICML2012.pdf. Accessed April 2, 2017.
26. Exploiting missing clinical data in Bayesian network modeling for predicting medical problems. *J Biomed Inform.* 2014;41:1–14.
27. Pivovarov R, Albers DJ, Sepulveda JL, et al. Identifying and mitigating biases in EHR laboratory tests. *J Biomed Inform.* 2014;51:24–34.
28. Feature selection: finding distinctive words—text analysis with topic models for the humanities and Social Sciences. Available at: https://de.dariah.eu/tatom/feature_selection.html. Accessed April 2, 2017.

29. Pivovarov R, Perotte AJ, Grave E, et al. Learning probabilistic phenotypes from heterogeneous EHR data. *J Biomed Inform.* 2015;58: 156–165.
30. Cohen R, Aviram I, Elhadad M, et al. Redundancy-aware topic modeling for patient record notes. *PLoS One.* 2014;9:e87555.
31. Arnold C, Speier W. A topic model of clinical reports. Proceedings of the 35th International ACM SIGIR Conference on Research and Development in Information Retrieval 1031–1032 (ACM, 2012). doi: 10.1145/2348283.2348454.
32. Ghassemi M, Naumann T, Doshi-Velez F, et al. Unfolding physiological state: mortality modelling in intensive care units. Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining 75–84 (ACM, 2014). doi: 10.1145/2623330.2623742.
33. Blei DM, Ng AY, Jordan MI. Latent Dirichlet allocation. *J Mach Learn Res.* 2003;3:993–1022.
34. Blei DM. Probabilistic topic models. *Commun ACM.* 2012;55:77–84.
35. Chang J, Gerrish S, Wang C, et al. Reading tea leaves: how humans interpret topic models. 2009:288–296.
36. Hira ZM, Gillies DF. A review of feature selection and feature extraction methods applied on microarray data. *Adv Bioinforma.* 2015;2015: 198363.
37. Breiman L. Random forests. *Mach Learn.* 2001;45:5–32.
38. Do we need hundreds of classifiers to solve real world classification problems—Google Search. Available at: https://www.google.com/search?q=Do+we+need+hundreds+of+classifiers+to+solve+real+world+classification+problems&ie=utf-8&oe=utf-8. Accessed April 25, 2017.
39. Saito T, Rehmsmeier M. The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. *PLoS One* 2015;10.:e0118432
40. 3.2.4.3.1. sklearn.ensemble. RandomForestClassifier—scikit-learn 0.18.1 documentation. Available at: http://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html. Accessed June 3, 2017.
41. Facente SN, Pitcher CD, Hartogensis WE, et al. Performance of risk-based criteria for targeting acute HIV screening in San Francisco. *PLoS One.* 2011;6:e21813.
42. Haukoos JS, Lyons MS, Lindsell CJ, et al. Derivation and validation of the Denver Human Immunodeficiency Virus (HIV) risk score for targeted HIV screening. *Am J Epidemiol.* 2012;175:838–846.
43. Arnold CW, Oh A, Chen S, et al. Evaluating topic model interpretability from a primary care physician perspective. *Comput Methods Programs Biomed.* 2016;124:67–75.
44. Resnik P, Armstrong W, Claudino L, et al. Beyond LDA: exploring supervised topic modeling for depression-related language in twitter. 2015;99–107. CLPsych@ HLT-NAACL
45. Stubbs A, Kotfila C, Xu H, et al. Identifying risk factors for heart disease over time: overview of 2014 i2b2/UTHealth shared task Track 2. *J Biomed Inform.* 2015;58:S67–S77.
46. Wells S, Rozenblum R, Park A, et al. Personal health records for patients with chronic disease. *Appl Clin Inform.* 2014;5:416–429.
47. Saparova D. Motivating, influencing, and persuading patients through personal health records: a scoping review. *Perspect Health Inf Manag AHIMA Am Health Inf Manag Assoc.* 2012;9:1.
48. The Specialist lexicon. Available at: https://lsg3.nlm.nih.gov/LexSysGroup/Projects/lexicon/current/web/index.html. Accessed May 11, 2017.
49. Healthix|Public Health Information Exchange (HIE). *Healthix*. Available at: http://healthix.org/. Accessed August 2, 2016.