Journal of the American Medical Informatics Association, 28(1), 2021, 104–112
doi: 10.1093/jamia/ocaa220
Advance Access Publication Date: 5 November 2020
Research and Applications

Research and Applications

Development of a predictive model for retention in HIV care using natural language processing of clinical notes

Tomasz Oliwa,[1] Brian Furner,[1] Jessica Schmitt,[2,3] John Schneider,[2,3] and Jessica P. Ridgway[2,3]

[1]Center for Research Informatics, University of Chicago, Chicago, Illinois, USA, [2]Department of Medicine, University of Chicago, Chicago, Illinois, USA, and [3]Chicago Center for HIV Elimination, University of Chicago, Chicago, Illinois, USA

Corresponding Author: Tomasz Oliwa, PhD, Center for Research Informatics, University of Chicago, The Shoreland, 5454 S. Shore Dr., Suite 1D, Chicago, IL 60615, USA (toliwa@bsd.uchicago.edu)

Received 15 May 2020; Revised 10 July 2020; Editorial Decision 17 August 2020; Accepted 24 August 2020

ABSTRACT

Objective: Adherence to a treatment plan from HIV-positive patients is necessary to decrease their mortality and improve their quality of life, however some patients display poor appointment adherence and become lost to follow-up (LTFU). We applied natural language processing (NLP) to analyze indications towards or against LTFU in HIV-positive patients' notes.

Materials and Methods: Unstructured lemmatized notes were labeled with an LTFU or Retained status using a 183-day threshold. An NLP and supervised machine learning system with a linear model and elastic net regularization was trained to predict this status. Prevalence of characteristics domains in the learned model weights were evaluated.

Results: We analyzed 838 LTFU vs 2964 Retained notes and obtained a weighted F1 mean of 0.912 via nested cross-validation; another experiment with notes from the same patients in both classes showed substantially lower metrics. "Comorbidities" were associated with LTFU through, for instance, "HCV" (hepatitis C virus) and likewise "Good adherence" with Retained, represented with "Well on ART" (antiretroviral therapy).

Discussion: Mentions of mental health disorders and substance use were associated with disparate retention outcomes, however history vs active use was not investigated. There remains further need to model transitions between LTFU and being retained in care over time.

Conclusion: We provided an important step for the future development of a model that could eventually help to identify patients who are at risk for falling out of care and to analyze which characteristics could be factors for this. Further research is needed to enhance this method with structured electronic medical record fields.

Key words: HIV, natural language processing, retention in care, machine learning, lost to follow-up

INTRODUCTION

Background and significance

Retention in care is essential for the health of people living with HIV (PLWH) and for public health. PLWH who are retained in care are more likely to receive antiretroviral therapy and experience improved health outcomes compared to PLWH not retained in care.[1,2] Moreover, PLWH retained in care and virally suppressed have effectively no risk of transmitting HIV to others.[3] Despite the clear health benefits of retention in care, less than 50% of PLWH in the United States are retained in care.[4] Improving retention in care is a key component of public health initiatives to eliminate HIV transmission in the United States.[5]

Public health agencies and researchers have recently shown interest in utilizing electronic medical record (EMR) data to improve retention

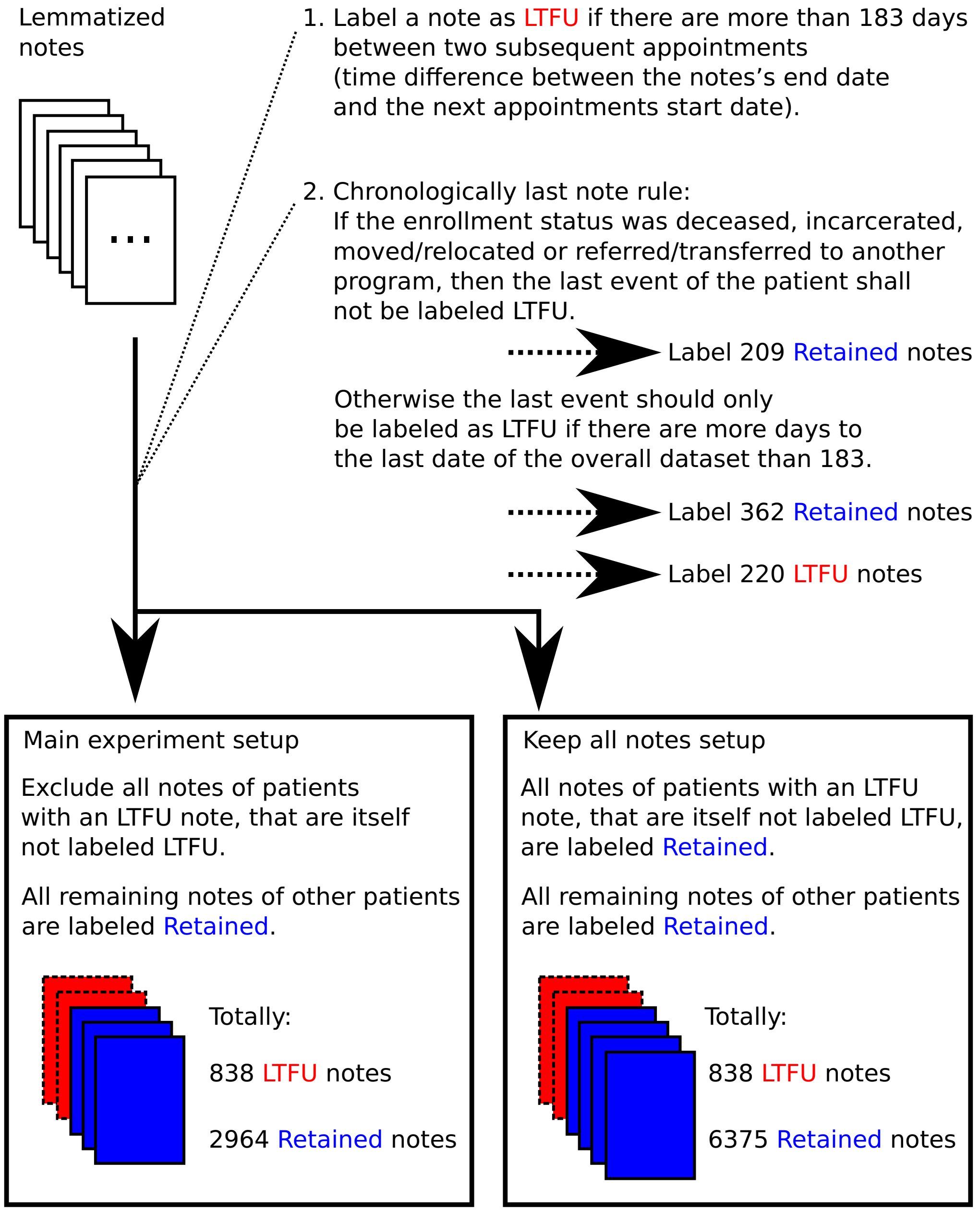
Lemmatized notes
...
1. Label a note as LTFU if there are more than 183 days between two subsequent appointments (time difference between the notes's end date and the next appointments start date).
2. Chronologically last note rule:
If the enrollment status was deceased, incarcerated, moved/relocated or referred/transferred to another program, then the last event of the patient shall not be labeled LTFU.
Label 209 Retained notes
Otherwise the last event should only be labeled as LTFU if there are more days to the last date of the overall dataset than 183.
Label 362 Retained notes
Label 220 LTFU notes
Main experiment setup
Exclude all notes of patients with an LTFU note, that are itself not labeled LTFU.
All remaining notes of other patients are labeled Retained.
Totally:
838 LTFU notes
2964 Retained notes
Keep all notes setup
All notes of patients with an LTFU note, that are itself not labeled LTFU, are labeled Retained.
All remaining notes of other patients are labeled Retained.
Totally:
838 LTFU notes
6375 Retained notes

Figure 1. Overview of the automated labeling rules to label a note as LTFU or Retained.

in care and other HIV care continuum outcomes.[6–8] Electronic medical records have rapidly expanded in the last decade, with over 95% of US hospitals and 86% of US ambulatory clinics utilizing EMRs.[9] EMRs contain vast amounts of data that can be analyzed to understand and predict HIV outcomes. By characterizing the future risk of an adverse outcome such as disengagement from care, HIV care continuum prediction models provide the opportunity to intervene beforehand to prevent the adverse outcome from occurring.

To date, retention in care prediction models have utilized data documented in structured fields of the EMR, such as "past medical history," diagnoses, and laboratory values. While these structured fields may capture some characteristics or risk factors for disengagement from care, such as race, age, low CD4 count, and drug use, they do not necessarily capture social and behavioral determinants of health (SBDH), such as life stressors, psychosocial and behavioral factors, and structural barriers that may contribute to an individual's disengagement from care. Natural language processing (NLP) of unstructured fields (ie, free text within provider and support staff notes) can detect early warning signs of potential disengagement from care, including descriptions of housing instability, limited social support, physician mistrust, stigma, and structural barriers. Thus, NLP of EMR-based clinical notes has the potential to enhance predictive models of retention in care among PLWH.

OBJECTIVE

NLP of EMR-based clinical notes has been used to predict various health outcomes, including mortality among patients in the intensive care unit,[10] suicide attempts among patients with psychiatric illness,[11] and readmission after hospitalization.[12,13] Feller et al[14] used

Table 1. Labeling counts and patient demographics

Description	LTFU	Retained[a]
Number of patients	470	321
Number of notes	838	2964
Age at first appointment mean	42.95	44.16
Age at first appointment standard deviation	13.1	14.2
Gender male	59.79 %	68.97 %
Race/Ethnicity		
Black	76.60%	83.49%
White	17.02%	8.72%
Latino	3.19%	3.74%
Other/unknown	3.19%	4.05%
Insurance		
Private	47.02%	38.00%
Medicaid	34.26%	39.88%
Medicare	17.45%	20.87%
Self-pay/None	1.28%	1.25%
Time from first to last visit in years mean	2.52	1.61
Time from first to last visit in years standard deviation	1.88	1.8
Time from first to last visit in years max	5.62	5.40
Time from first to last visit in years min[b]	0	0

Abbreviation: LTFU: lost to follow-up.

[a] Two patients from the set of notes classified as Retained did not have a DOB and gender in our dataset.

[b] 0 reflects cases of patients with a single visit.

NLP of clinical notes to predict risk for HIV acquisition among HIV-negative clients. However, no prior studies to date have utilized NLP of EMR data to predict retention in care or other HIV care continuum outcomes and analyzed the most predictive textual features. Accordingly, the purpose of this study was to utilize supervised machine learning and NLP of EMR-based clinical notes to predict retention in care among people living with HIV.

MATERIALS AND METHODS

Study sample

PLWH 18 years of age and older who attended at least 1 medical appointment at the University of Chicago adult HIV care clinic between July 1, 2010 and October 31, 2016 were included in the study. The University of Chicago adult HIV care clinic is located on the south side of Chicago, a major US HIV epicenter. For eligible patients who saw an HIV provider, we extracted the text of all clinical notes related to an outpatient HIV care medical appointment during the study period from the University of Chicago Clinical Research Data Warehouse. We included notes written by physicians, advanced practice providers, nurses, social workers, and case managers. This study was approved by the University of Chicago Biological Sciences Division Institutional Review Board, and a waiver of consent was granted.

Preprocessing

The raw notes were stored as plain text files. To increase generalizability and provide dimensionality reduction by excluding words and phrases unrelated to this research, we applied PhysioNet deid,[15,16] MIST,[17] and Stanford NER[18] and removed the protected health information from the notes that these software tools found. We applied further common NLP dimensionality reduction[10,19] as follows: All notes were tokenized and lemmatized with spacy (https://spacy.io). Through lemmatization, inflected forms of a word were reduced to the same canonical form. Numeric tokens were replaced by the token "number," all tokens were lowercased, and tokens without any alphanumeric content were removed.

Notes setup

The initial dataset contained 8970 notes, the majority of which were progress notes (8225), with the remainder being addendum and ancillary notes. Notes of the same patient, that had the same start and end date and the same billing number, were concatenated into single notes, which led to a dataset of 7213 notes, each representing an encounter/event in time among 791 unique patients.

Outcome/labeling

Retention in care was defined as having no more than 6 months (183 days) between sequential attended appointments. This measure of retention in care, also referred to as a "6-month gap" is associated with clinical outcomes including viral suppression.[20] Patients with more than 183 days between sequential appointments were considered "not retained in care" at that time. Patients were followed for a minimum of 6 months prior to determining if they were lost to follow-up (LTFU) or retained in care for that period of time. Figure 1 shows how we algorithmically labeled each note for the main experiment. A "keep all notes" setup is also defined to determine the classification impact of excluding notes. Our clinic staff includes a robust social work team who reach out to patients who miss appointments or may be lost to follow-up and confirm with patients if they are in care elsewhere or have moved. With our approach, we also aim to address the reported need[21] for methods to account for gaps in treatment and uninterrupted therapy among patients that are categorized as LTFU.

Table 1 provides the labeling counts and patient demographics for the main experiment. As can be seen in the notes row, only a subset of the 7213 notes resulted in a LTFU or "Retained" label, and there was a class imbalance in favor of "Retained," as most remaining notes were categorized as retained in care. In the keep all notes setup, this class imbalance was even larger.

Classification

With scikit-learn,[22] we created a supervised machine learning pipeline for this binary classification task. N-grams features up to size 3 were extracted from the preprocessed notes, and the commonly used[10,19] term frequency-inverse document frequency (TF-IDF) weighting with L2 row normalization was applied on these high-dimensional bag-of-words vectors. With TF-IDF, every n-gram's count is weighted by a heuristic that takes into account its term frequency and its inverse document frequency. As the classifier to predict the class label LTFU or Retained for a note, we followed and chose[10] a linear model with stochastic gradient descent and elastic net regularization (see Supplementary Materials for details).

We performed a hyperparameter optimization with 10-fold cross-validation. With this, the pipeline was tested on the classification task with a varying regularization parameter *alpha* to find the best one, and the 10-fold cross-validation aimed to reduce sample bias.

Analysis

To address class imbalance, the metric of the hyperparameter optimization evaluating the 10-fold cross-validation was scikit-learn's weighted F1 metric.[23] F1 is a commonly used metric in machine

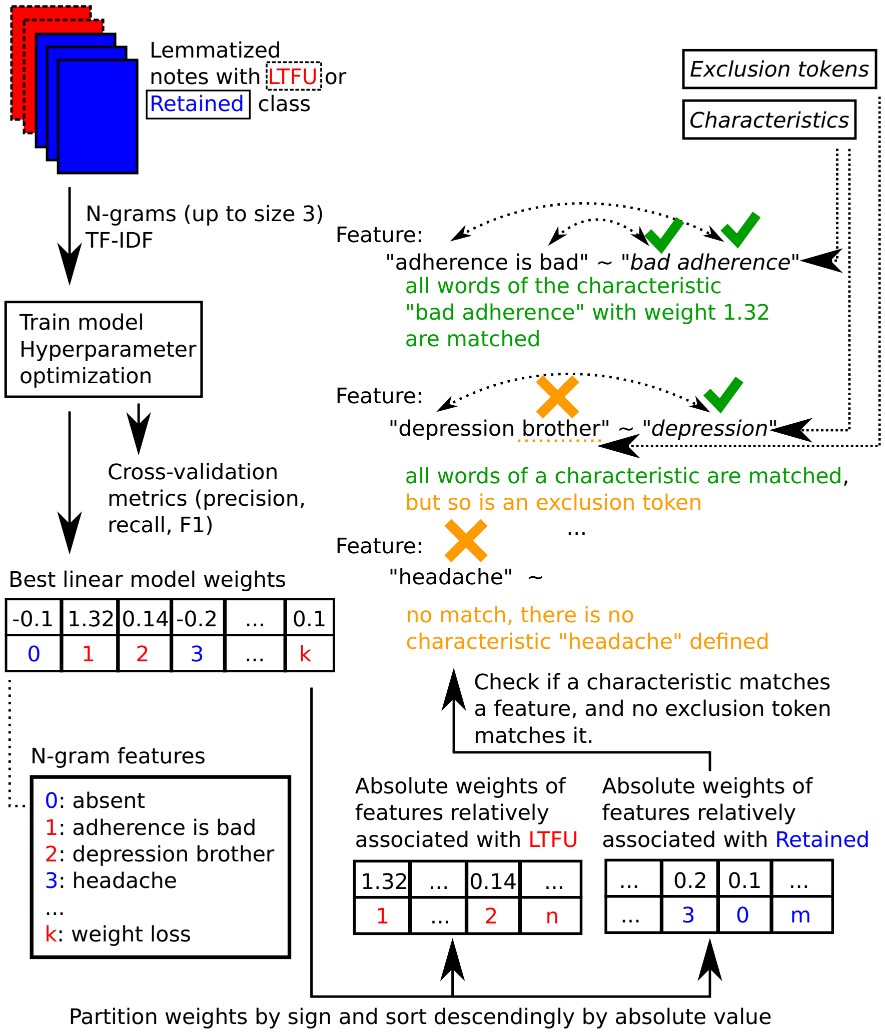

Figure 2. An overview of the natural language processing (NLP) and machine learning approach. As an example with synthetic data, "adherence is bad" is shown to be matched correctly with the characteristic "bad adherence."

learning and is obtained by the harmonic mean of precision and recall.[23] Similar to Feller et al,[14] we are using these performance metrics in the presence of class imbalance. We calculated a classification report, based on the predicted label for each note when it was in the cross-validation test set, and provide it together with the best *alpha*.

As recommended by Müller & Guido,[23] we also ran a nested cross-validation (with both inner and outer folds set to 10) to estimate the model's generalizability and present the mean and standard deviation of the weighted F1 metric.

Figure 2illustrates our methods and depicts our model's weights application as relative feature importances that we analyzed for the main experiment. This approach has been similarly published before for text classification: In Mladenić et al,[19] a feature selection for linear classifiers using the model's weights is described, where each document is also set up as a TF-IDF bag-of-words representation normalized to unit length, and features with smaller weights close to zero are identified to have less influence on the prediction than features with larger absolute value weights. For an elastic net text classifier in Marafino et al,[10] the relative feature influence has been interpreted "as proportional to the absolute values of the parameter estimates." Similarly, we extracted the weights (coefficients) of the linear classifiers for each corresponding n-gram feature, separated them into numerical positive and numerical negative weights, sorted each separated list descendingly according to their absolute value, and interpreted these as relative feature importances and ranks.

For this work, an infectious disease physician defined a set of characteristics, categorized by domains that were subsequently lemmatized, as shown in Table 2. Figure 2 gives examples of how a specific characteristic word/phrase is considered to match an n-gram feature.

If the words from a characteristic were all contained in the n-gram feature (eg, the characteristic is "bad adherence," the feature is "adherence is bad"), and the feature did not contain an exclusion token, (eg, the characteristic is "depression," the feature is "depression brother"), it was defined as a match. Exceptions to this were characteristics that themselves contained exclusion tokens (eg, "no miss dose"), in these cases, the specific exclusion token ("no") was not a reason to dismiss a match. The exclusion tokens are also given in Table 2.

For the main experiment setup, we obtained a results table, similar to Marafino et al,[10] with selected high-ranking features that matched characteristics. In the Supplementary Materials, we provide summary approaches and give their technical descriptions.

Table 2. The lemmatized characteristics by domain and the exclusion words

Domain	Characteristics words / phrases
HIV genotype mutation	"k103," "k103n," "m184v," "resistant hiv," "resistant virus"
Congenital HIV	"congenital hiv," "congenital infection," "perinatal," "since birth," "vertical transmission"
IV drug use	"heroin," "idu," "intravenous drug use," "ivdu"
Opportunistic infection	"burkitt," "candidiasi," "candidiasis," "cmv," "cmv retinitis," "cryptococcal," "cryptococcus," "cryptosporidium," "cytomegalovirus," "jc virus," "kaposi," "ks," "lymphoma," "mac," "mai," "mycobacterium avium," "pcp pneumonia," "pjp pneumonia," "pml," "pneumocystis," "thrush," "toxo," "toxoplasmosis"
Comorbidities	"hbv," "hcv," "hepatitis b," "hepatitis c," "tb," "tuberculosis"
Poor adherence	"bad adherence," "difficulty with adherence," "do not fill," "frequent miss dose," "medication access," "miss appointment," "miss appt," "no show," "non adherence," "non adherent," "non compliant," "nonadherence," "nonadherent," "noncompliant," "not adherent," "off art," "off haart," "out of med," "poor adherence," "poor compliance," "run out," "sometimes forget," "unable to fill," "without med"
Condomless sex	"condom use no," "no condom," "not use condom," "unprotected," "unprotected sex," "without condom"
Good adherence	"adherent," "compliant," "do well on," "excellent adherence," "good adherence," "never miss," "no miss dose," "well on art"
Sex with condoms	"condom use yes," "use condom"
Sexual and gender minorities	"lgbt," "lgbtq," "man who have," "msm," "transwoman"
Heterosexual	"hetero," "heterosexual"
Life stressors and markers of socioeconomic status	"adap," "afc," "case management," "case manager," "court date," "death in family," "disability," "disclosure," "ed visit," "emergency room," "financial support," "fmla," "homeless," "homophobia," "house arrest," "incarcerate," "incarceration," "insurance issue," "insurance lapse," "jail," "kick out," "lawyer," "life alone," "live alone," "medicaid lapse," "new phone number," "no family," "not disclose," "not discuss status," "partner unaware," "phone break," "prison," "recent death," "recent ed visit," "redetermination," "rent assistance," "stigma," "stress," "stressor," "sw call pt," "transportation," "unemployed," "uninsured," "unstable housing," "ventra," "vital bridge," "wic," "without insurance coverage"
Mental illness	"anxiety," "anxious," "behavioral health," "bh referral," "bipolar," "c2p," "care prevent," "care2prevent," "cry," "deny si," "depress," "depressed," "depression," "emotional," "grief," "insomnia," "panic," "passive si," "psychiatrist," "psychiatry," "psychosis," "psychotic," "sad," "schizophrenia," "sleepless," "suicidal," "suicide," "tearful," "therapist"
Pregnancy	"c section," "cesarean section," "pregnancy," "pregnant," "vaginal delivery"
Preventive health services	"administer pneumococcal," "colonoscopy," "flu shot," "human papillomavirus vaccine," "influenza vaccine," "mammogram," "menactra," "pap," "pap smear," "pcv13," "pneumovax," "prevnar," "psv23," "quadrivalent," "tdap," "trivalent," "vaccine"
Married/partnered	"husband," "married," "marry," "monogamous," "wife"
Social support	"church," "prayer," "support group"
STI	"benzathine," "bicillin," "chancroid," "chlamydia," "chlamydia trachomatis," "crab," "crab louse," "gc," "genital herpe," "gonorrhea," "granuloma inguinale," "haemophilus ducreyi," "herpes," "herpes simplex," "hpv," "hsv," "hsv1," "hsv2," "human papillomavirus," "klebsiella granulomatis," "lgv," "lymphogranuloma venereum," "neisseria gonorrhoeae," "pediculosis pubis," "pelvic inflammatory disease," "penicillin g," "pid," "positive rpr," "pthirus pubis," "pubic lice," "std," "sti," "syphilis," "treponema pallidum," "trich," "trichomona vaginalis," "trichomoniasis," "valacyclovir," "valtrex"
Substance use disorder	"aa meeting," "alcohol abuse," "alcoholic," "amphetamine abuse," "beer," "cocaine," "crack," "crystal meth," "drug treatment program," "drunk," "haymarket," "heroin," "intravenous drug user," "ivdu," "ivu," "marijuana," "meth," "methadone," "methamphetamine abuse," "na meeting," "narcotic," "sober," "substance abuse," "substance use," "take drug"
Exclusion	"no," "not," "none," "neg," "never," "negative," "non," "deny," "denies," "father," "dad," "mother," "mom," "brother," "brothers," "sister," "sisters," "sibling," "siblings," "cousin," "cousins," "aunt," "aunts," "uncle," "uncles," "grandmother," "grandparent," "grandparents," "grandfather," "grandchild," "grandchildren," "grandson," "grandsons," "granddaughter," "granddaughters," "wife," "spouse," "husband," "child," "children," "offspring," "progeny," "son," "sons," "daughter," "daughters," "nephew," "nephews," "niece," "nieces," "kin"

Note: In "Life stressors and markers of socioeconomic status," ADAP could be a marker of lower income, and case manager a sign of possibly difficult life situations. In this table, commas have been placed before the closing quotation marks for stylistic reasons. In the code, the token matching is performed without these commas.

RESULTS

In the main experiment, the elastic net reduced the initial bag-of-words feature size from 290 309 to 7177 non-zero weight features, therefore only 2.47% of the features were found useful by the model.

The best hyperparameter *alpha* was 5e-05. The results from the 10-fold cross-validation with this *alpha* are displayed in Table 3. Overall, a strong performance can be observed, with micro and weighted F1 above 0.9.

The nested cross-validation weighted F1 mean was 0.912 with a standard deviation of 0.017.

Adding the additional Retained notes resulted in a major decrease of performance in the 10-fold cross-validation for the LTFU class with a F1 of 0.273 (from 0.801). The F1 of the Retained class

Table 3. 10-fold cross-validation metrics results of the classification task, based on the predicted label for each note when it was in the cross-validation test set

Experiment	Metric	Precision	Recall	F1-score	Support
Main	Retained	0.931	0.966	0.948	2964
	LTFU	0.861	0.748	0.801	838
	micro average	0.918	0.918	0.918	3802
	macro average	0.896	0.857	0.875	3802
	weighted average	0.916	0.918	0.916	3802
Keep all notes	Retained	0.903	0.925	0.914	6375
	LTFU	0.302	0.248	0.273	838
	micro average	0.846	0.846	0.846	7213
	macro average	0.603	0.586	0.593	7213
	weighted average	0.834	0.846	0.839	7213

Note: LTFU = lost to follow-up; tp = The number of true positives; fp = The number of false positives; fn = The number of false negatives; precision = tp/(tp + fp); recall = tp/(tp + fn); F1-score = 2 * (precision * recall)/(precision + recall); support = The number of occurrences in a class; micro = Globally calculate metrics by using the total true positives, false negatives, and false positives; macro = Get the unweighted average of the metrics for each label; weighted = Calculate metrics for each label and get their average weighted by support. These metrics definitions were taken from scikit-learn, the result table was obtained through scikit-learn's classification report method. The linear model with stochastic gradient descent and elastic net regularization was evaluated with the best found hyperparameter alpha of 5e-05 for the main experiment, and 1e-05 for the "keep all notes" setup.

dropped slightly to 0.914 (from 0.948). The nested cross-validation weighted F1 mean was 0.839 with a standard deviation of 0.008.

Selected matched characteristics with NLP phrases relatively associated with LTFU and Retention are shown in Table 4. Notable phrases associated with LTFU included "MSM," "HCV," "substance abuse," and "unemployed." The phrase "well on ART" was a highly ranked feature in the "Retained" class with a rank of 3. Other phrases associated with retention in care included "pregnancy," "cesarean section," "excellent adherence," and "congenital HIV." Of note, certain domains contained phrases that were associated with disparate outcomes. For example, in the mental illness domain, the phrases "tearful" and "bipolar" were associated with LTFU, but "depression" and "schizophrenia" were associated with retention. In the opportunistic infection domain, "KS," "Burkitt," and "CMV retinitis" were associated with LTFU, but "toxo" and "cryptococcal" were associated with retained status.

DISCUSSION

The classification results showed a strong predictive quality in the main experiment. Our relatively low LTFU recall can be partly explained by the comparatively smaller number of LTFU notes: the model had only 838 LTFU notes to learn and generalize from vs 2964 Retained notes.

Keeping all notes substantially lowered the performance on the LTFU class. One factor likely contributing to this was that the majority of notes from patients with some LTFU notes were now in the opposite Retained class, and it seems probable that notes of the same patient would contain many shared clinical concepts such as comorbidities and other words and phrases, which collectively negatively affected the classifier metrics. An additional factor we think influenced the result was the general increase of the label imbalance to 838 vs 6375 notes.

Pence et al[6] used structured EMR fields to predict one measure of retention in care, missed visits, among PLWH. Their model had an area under the curve of 0.7 and included variables related to prior missed visits, demographics, clinical, and psychosocial characteristics. While our results are not directly comparable, given the different performance measures and experiment setups, we demonstrated a solid predictive quality with purely textual features, and see promising opportunities for hybrid models combining NLP and structured EMR fields. By using NLP on clinical notes, we were able to identify phrases representing characteristics for LTFU that are not captured in structured EMR fields or nuances in the provider–patient relationship and the patient's presentation at the time of the visit (eg, "tearful," "unemployed").

Comparable to our results, the dimensionality reduction of n-gram features through the elastic net model was also observed in Marafino et al.[10] Many characteristics previously shown to be associated with retention in care corresponded with phrases in our NLP model that were positively associated with retention. For example, the phrase "Well on ART," indicating the patient was doing well on antiretroviral therapy, was strongly associated with retention in care. Prior studies have shown that ART adherence is strongly linked to retention in care.[20] In addition, we found that phrases related to pregnancy were associated with retention in care. Because pregnant patients often have more frequent medical visits and receive more resources, patients experiencing pregnancy may be more likely to be retained in care than those who are not.[24,25] The phrase "congenital HIV" was also associated with retention in care. While others have found that patients with perinatally acquired HIV have had poor retention in care, these studies were predominantly performed among adolescents and often around the transition from pediatric to adult HIV care.[26–29] Our study was performed only in an adult HIV clinic, so the group of patients with perinatally acquired HIV in our clinic have already successfully transitioned from pediatric HIV care to adult HIV care. Therefore, our clinic population with perinatally acquired HIV may be more inclined to retention in care than the general population of patients with perinatally acquired HIV.

Characteristics for LTFU previously described in the literature also corresponded with phrases in our model that were associated with LTFU. The phrase "HCV" was a highly ranked feature in the LTFU class, likely reflecting that patients with comorbid HIV and Hepatitis C are less likely to be retained in care. Giordano et al[30] also previously described HCV as a risk factor for lack of retention in care.

Several feature domain categories included phrases that were associated with disparate retention outcomes. It is important to note that higher prevalence of a characteristic in the Retained class does not necessarily mean that LTFU patients are comparatively less affected by it. Rather, this particular characteristic may be less frequently mentioned in notes of LTFU patients compared to notes of retained patients. For example, among phrases associated with mental health disorders, the phrases "tearful," and "bipolar" were associated with LTFU, but "depression" and "schizophrenia" were associated with retention. Our model does not detect whether a patient suffers from mental health disorder, but only whether a phrase associated with mental health was mentioned in the clinical note. This could be counterintuitive from a clinician's perspective. A possible intuition for this is that while mentions have been put into a note for a specific reason, and our approach also tries to capture sur-

Table 4. Selection of matched characteristics with natural language processing (NLP) phrases associated with LTFU and Retention (comprehensive table in the Supplementary Material)

Matched characteristics relatively associated with LTFU	Likely expanded phrase	Domain	Rank	Weight
KS	Kaposi's Sarcoma	Opportunistic infection	23	3.655
syphilis		STI	25	3.613
Burkitt	Burkitt's lymphoma	Opportunistic infection	41	3.348
HCV	Hepatitis C virus	Comorbidities	61	3.116
MSM	Man who has sex with men	Sexual and gender minorities	111	2.705
CMV retinitis		Opportunistic infection	155	2.439
K103N		HIV genotype mutation	177	2.319
substance abuse		Substance use disorder	188	2.282
chlamydia		STI	223	2.182
tearful		Mental illness	232	2.149
Human papillomavirus		STI	259	2.053
HBV	Hepatitis B virus	Comorbidities	271	2.039
cocaine		Substance use disorder	294	1.967
influenza vaccine		Preventive health services	298	1.963
m184v		HIV genotype mutation	401	1.74
Bipolar		Mental illness	407	1.717
heroin		Substance use disorder	432	1.673
Not use condom		Condomless sex	552	1.484
unemployed		Life stressors and markers of socioeconomic status	747	1.273
Matched characteristics relatively associated with Retention	Likely expanded phrase	Domain	Rank	Weight
well on art		Good adherence	3	–4.698
Depression		Mental illness	13	–3.845
Marijuana		Substance use disorder	71	–2.562
HSV	Herpes simplex virus	STI	91	–2.329
toxo	toxoplasmosis	Opportunistic infection	109	–2.229
Alcohol abuse		Substance use disorder	117	–2.191
cryptococcal		Opportunistic infection	131	–2.091
congenital HIV		Congenital HIV	146	–2.043
Pap	Papaliconaou smear	Preventive health services	158	–2.008
Cesarean section		Pregnancy	321	–1.529
schizophrenia		Mental illness	680	–1.037
Excellent adherence		Good adherence	948	–0.778
Pregnancy		Pregnancy	1011	–0.736

Note: LTFU = lost to follow-up; Weights and ranks show the relative importance of a feature that matched the characteristic, sorted as illustrated in Figure 2. Positive weights are relatively related to LTFU, negative weights are relatively related to Retention. Weight ranks start at 0 (zero). The results are from the classifier trained on the main experiment setup. The model intercept was: -1.155.

rounding negations and family mentions in the n-grams, a wider range of these and other contexts not addressed by our method is possible. Prior literature suggests that mental health disorders, particularly untreated mental health disorders, are often associated with poor retention in care.[31] Clinic notes with "depression" mentioned may indicate that the provider discussed depression screening and referred the patient for depression treatment, whereas clinic notes without depression mentioned may include instances in which patients actually had depression but providers did not identify depression or address depression treatment.

Prior studies have found that substance use is associated with lack of retention.[32,33] We similarly found that the phrases "substance abuse," "cocaine," and "heroin" were associated with LTFU. However, other substance-associated phrases including "marijuana," "alcohol abuse," and "meth" were positively associated with retention in care. This finding was surprising given that others have found that methamphetamine and alcohol misuse are associated with reduced engagement in care among PLWH.[34,35] It may be that marijuana, alcohol, and methamphetamine use are less likely to be associated with LTFU than cocaine and heroin. Another possible explanation for this finding is that documentation of marijuana, alcohol, and methamphetamine use within the text of the note indicates a careful and thorough provider who took the time to screen and document prior and current substance use, and the positive physician–provider relationship is associated with improved retention. Finally, our approach did not model a difference in "history of" vs active use.

Furthermore, in our main experiment setup, there were no common patients between the LTFU and Retained labeled notes. In reality, patients often transition between being LTFU and retained in care over time.[36] With this in mind, it seems that the classifier could better learn to distinguish between patients and their features to optimize the classification metrics, for which the characteristics play a role but together with many other HIV unrelated or tangentially related words and concepts. While we strove for generalizability by applying both a deidentification and lemmatization approach, ideally a larger corpus of notes would be desirable. Because some patients with well-controlled HIV may only be scheduled for appointments once per year, we also investigated a LTFU threshold of 365 days. However, we observed a reduction in model performance, possibly due to an even stronger class imbalance. Moreover, within our clinic, the vast majority of patients are routinely scheduled for HIV care every 3–4 months, not once yearly, and the labeling approach would have to be readjusted in clinics in which visits are annual by design. Furthermore, despite nested cross-validation, our high-performance results could partly be influenced by overfitting, and it would be desirable to test the models on an external infectious disease patient notes corpus from another institution.

Limitations to this approach also include that the characteristics here are word lists, and synonyms have to be explicitly encoded. An alternative would be to apply ontology matching to extract concept identifiers from notes, such as capturing Unified Medical Language System (UMLS http://www.nlm.nih.gov/research/umls/) concepts via approaches like cTAKES.[37] In our case, cTAKES did not identify a substantial number of our defined characteristics, and most likely an augmentation of the ontology would have to be performed to alleviate this. Another limitation is that, despite steps taken to achieve greater generalization, we may have not captured negations or concepts relating to family history (ie, a note mentions that a patient's son has depression, not the patient themselves) outside of our defined n-gram range. We also did not aim to detect the difference between a history of substance use vs active use or to encode the viral load,[20] which could lead to beneficial features for retention modeling and a more precise measuring of characteristics. Furthermore, as also noted in Marafino et al,[10] the elastic net classifier does not explicitly account for correlations between features, and, beyond the common TF-IDF weighting with L2 row normalization, our approach does not perform further feature standardizations. Also, some providers possibly used their own individually customized templates; while we tried to mitigate the effect of template n-grams on classification by a data-driven stop word removal (see Supplementary Material), such influence on the results cannot be ruled out.

Our approach could be extended and generalized for other clinical tasks in a straightforward way by including different characteristics domains with word/phrases lists. For example, in Carson et al,[11] NLP is used to identify suicidal behavior from electronic health records. Protective and risk factor word lists from that study could be integrated into our method. In future research using unsupervised learning, NLP methods such as topic modeling[14] could also bring additional insights into retention in care for PLWH.

CONCLUSION

Using NLP of clinical notes from the EMR for PLWH, we developed a predictive model to differentiate patients' notes that were chronologically last before an LTFU event vs notes of patients that kept their Retained status and showed a strong performance based on (nested) cross-validation. To our knowledge, this is the first use of NLP of clinical notes to predict retention in HIV care and provide an interpretation of n-gram/features into characteristics relatively associated with an LTFU or Retained status.

In another experiment, we additionally included the notes of patients with an LTFU event that were themselves not labeled LTFU to the Retained notes set, and the classifier performance was substantially lower. It seems probable that notes of the same patient would contain many shared clinical concepts, such as comorbidities, which likely negatively affected the classifier metrics.

In the future, we think an NLP-based model of retention would gain from augmentation with structured EMR data, including viral loads, the modeling of contexts related to history vs current characteristics, and validation through another notes set. Then, it could help to identify PLWH who are at risk for falling out of care and to determine which characteristics are more prevalent in their clinical notes. We provided an important step for such a model that could be used to aid in directing resources of retention interventions to prevent at risk patients from becoming LTFU.

FUNDING

This work was supported by NIH 1K23MH121190-01 and the NIH-funded Third Coast Center for AIDS Research (CFAR) (P30 AI117943). The Center for Research Informatics is funded by the Biological Sciences Division, the Institute for Translational Medicine/CTSA (NIH UL1 TR000430) at the University of Chicago.

AUTHOR CONTRIBUTIONS

JR, JAS, and TO conceived the study. Data collection was performed by JR and JS. TO performed the research implementation, including NLP machine learning. BF gave overall approach guidance and feedback. JR obtained funding and oversaw the work. All authors contributed to interpretation of the results and the writing of the manuscript.

SUPPLEMENTARY MATERIAL

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

ACKNOWLEDGMENTS

We thank Dr Julie A. Johnson from the Center for Research Informatics (CRI) at the University of Chicago. Computational infrastructure was provided by the Applications Development team maintained by the CRI. Data from this study were provided by the Clinical Research Data Warehouse (CRDW) maintained by the CRI.

CONFLICT OF INTEREST STATEMENT

JR has received grant funding from Gilead Sciences for research unrelated to the current manuscript. Others: none declared.

REFERENCES

1. Ulett KB, Willig JH, Lin HY, *et al.* The therapeutic implications of timely linkage and early retention in HIV care. *AIDS Patient Care STDS* 2009; 23 (1): 41–9.
2. Skarbinski J, Rosenberg E, Paz-Bailey G, *et al.* Human immunodeficiency virus transmission at each step of the care continuum in the United States. *JAMA Intern Med* 2015; 175 (4): 588–96.
3. The Lancet HIV. *U=U taking off in 2017. Lancet HIV* 2017; 4 (11): e475.
4. Understanding the HIV Care Continuum. 2019. https://www.cdc.gov/hiv/pdf/library/factsheets/cdc-hiv-care-continuum.pdf Accessed September 18, 2020.
5. Status of HIV in America. 2019. https://www.hiv.gov/federal-response/ending-the-hiv-epidemic/key-strategies Accessed September 18, 2020.

6. Pence BW, Bengtson AM, Boswell S, *et al.* Who will show? Predicting missed visits among patients in routine HIV primary care in the United States. *AIDS Behav* 2019; 23 (2): 418–26.
7. Olatosi B, Zhang J, Weissman S, *et al.* Using big data analytics to improve HIV medical care utilisation in South Carolina: a study protocol. *BMJ Open* 2019; 9 (7): e027688.
8. Ridgway JP, Schmitt J, Almirol E, *et al.* Electronic data sharing between public health department and clinical providers improves accuracy of HIV retention data. *Open Forum Infect Dis* 2017; 4 (Suppl 1): S421–22.
9. Myrick KL, Ogburn DF, Ward BW. Percentage of office-based physicians using any electronic health record (EHR)/electronic medical record (EMR) system and physicians that have a certified EHR/EMR system, by U.S. *state: National Electronic Health Records Survey*, 2017. National Center for Health Statistics. January 2019. https://www.cdc.gov/nchs/data/nehrs/2017_NEHRS_Web_Table_EHR_State.pdf Accessed September 18, 2020.
10. Marafino BJ, Boscardin WJ, Dudley RA. Efficient and sparse feature selection for biomedical text classification via the elastic net: Application to ICU risk stratification from nursing notes. *J Biomed Inform* 2015; 54: 114–20.
11. Carson NJ, Mullin B, Sanchez MJ, *et al.* Identification of suicidal behavior among psychiatrically hospitalized adolescents using natural language processing and machine learning of electronic health records. *PLoS One* 2019; 14 (2): e0211116.
12. Rumshisky A, Ghassemi M, Naumann T, *et al.* Predicting early psychiatric readmission with natural language processing of narrative discharge summaries. *Transl Psychiatry* 2016; 6 (10): e921.
13. Greenwald JL, Cronin PR, Carballo V, *et al.* A novel model for predicting rehospitalization risk incorporating physical function, cognitive status, and psychosocial support using natural language processing. *Med Care* 2017; 55 (3): 261–6.
14. Feller DJ, Zucker J, Yin MT, *et al.* Using clinical notes and natural language processing for automated HIV risk assessment. *J Acquir Immune Defic Syndr* 2018; 77 (2): 160–6.
15. Neamatullah I, Douglass MM, Lehman LW, *et al.* Automated de-identification of free-text medical records. *BMC Med Inform Decis Mak* 2008; 8 (1): 32.
16. Goldberger AL, Amaral LA, Glass L, *et al.* PhysioBank, PhysioToolkit, and PhysioNet: components of a new research resource for complex physiologic signals. *Circulation* 2000; 101 (23): E215–20.
17. Aberdeen J, Bayer S, Yeniterzi R, *et al.* The MITRE Identification Scrubber Toolkit: design, training, and assessment. *Int J Med Inform* 2010; 79 (12): 849–59.
18. Finkel JR, Grenager T, Manning C. Incorporating non-local information into information extraction systems by Gibbs sampling. In: proceedings of the 43rd Annual Meeting on Association for Computational Linguistics; June 25–30, 2005; Ann Arbor, MI.
19. Mladenić D, Brank J, Grobelnik M, *et al.* Feature selection using linear classifier weights: interaction with classification models. In: proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval; July 25–29, 2004; Sheffield.
20. Mugavero MJ, Westfall AO, Zinski A, *et al.* Measuring retention in HIV care: the elusive gold standard. *J Acquir Immune Defic Syndr* 2012; 61 (5): 574–80.
21. Tweya H, Feldacker C, Estill J, *et al.* Are they really lost? “true” status and reasons for treatment discontinuation among HIV infected patients on antiretroviral therapy considered lost to follow up in Urban Malawi. *PLoS One* 2013; 8 (9): e75761.
22. Pedregosa F, Varoquaux G, Gramfort A, *et al.* Scikit-learn: machine learning in python. *J Mach Learn Res* 2011; 12: 2825–30.
23. Müller AC, Guido S. *Introduction to Machine Learning with Python: A Guide for Data Scientists*. Sebastopol, CA: O’Reilly Media, Inc.; 2016.
24. Meade CM, Badell M, Hackett S, *et al.* HIV care continuum among postpartum women living with HIV in Atlanta. *Infect Dis Obstet Gynecol* 2019; 2019: 8161495.
25. Recommendations for the Use of Antiretroviral Drugs in Pregnant Women with HIV Infection and Interventions to Reduce Perinatal HIV Transmission in the United States https://aidsinfo.nih.gov/contentfiles/lvguidelines/PerinatalGL.pdf Accessed July 2020
26. Judd A, Sohn AH, Collins IJ. Interventions to improve treatment, retention and survival outcomes for adolescents with perinatal HIV-1 transitioning to adult care: moving on up. *Curr Opin HIV AIDS* 2016; 11 (5): 477–86.
27. Kim SH, Gerver SM, Fidler S, Ward H. Adherence to antiretroviral therapy in adolescents living with HIV: systematic review and meta-analysis. *AIDS* 2014; 28 (13): 1945–56.
28. Agwu AL, Fairlie L. Antiretroviral treatment, management challenges and outcomes in perinatally HIV-infected adolescents. *J Int AIDS Soc* 2013; 16 (1): 18579.
29. Idele P, Gillespie A, Porth T, *et al.* Epidemiology of HIV and AIDS among adolescents: current status, inequities, and data gaps. *J Acquir Immune Defic Syndr* 2014; 66 (Suppl 2): S144–53.
30. Giordano TP, Hartman C, Gifford AL, *et al.* Predictors of retention in HIV care among a national cohort of US veterans. *HIV Clin Trials* 2009; 10 (5): 299–305.
31. Rooks-Peck CR, Adegbite AH, Wichser ME, *et al.* Mental health and retention in HIV care: a systematic review and meta-analysis. *Health Psychol* 2018; 37 (6): 574–85.
32. Bulsara SM, Wainberg ML, Newton-John TRO. Predictors of adult retention in HIV care: a systematic review. *AIDS Behav* 2018; 22 (3): 752–64.
33. Hartzler B, Dombrowski JC, Williams JR, *et al.* Influence of substance use disorders on 2-year HIV care retention in the United States. *AIDS Behav* 2018; 22 (3): 742–51.
34. Cohen JK, Santos GM, Moss NJ, *et al.* Regular clinic attendance in two large San Francisco HIV primary care settings. *AIDS Care* 2016; 28 (5): 579–84.
35. Amirkhanian YA, Kelly JA, DiFranceisco WJ, *et al.* Predictors of HIV care engagement, antiretroviral medication adherence, and viral suppression among people living with HIV infection in St. Petersburg, Russia. *AIDS Behav* 2018; 22 (3): 791–9.
36. Lee H, Wu XK, Genberg BL, *et al.* Beyond binary retention in HIV care: predictors of the dynamic processes of patient engagement, disengagement, and re-entry into care in a US clinical cohort. *AIDS* 2018; 32 (15): 2217–25.
37. Savova GK, Masanz JJ, Ogren PV, *et al.* Mayo clinical Text Analysis and Knowledge Extraction System (cTAKES): architecture, component evaluation and applications. *J Am Med Inform Assoc* 2010; 17 (5): 507–13.