Title

Natural Language Processing and Machine Learning to Identify People Who Inject Drugs in Electronic Health Records

Permalink

https://escholarship.org/uc/item/2kg2g8bz

Journal

Open Forum Infectious Diseases, 9(9)

ISSN

2328-8957

Authors

Goodman-Meza, David

Tang, Amber

Aryanfar, Babak

et al.

Publication Date

2022-09-02

DOI

10.1093/ofid/ofac471

Peer reviewed

Open Forum Infectious Diseases
MAJOR ARTICLE

Natural Language Processing and Machine Learning to Identify People Who Inject Drugs in Electronic Health Records

David Goodman-Meza,[1,2] Amber Tang,[3] Babak Aryanfar,[2] Sergio Vazquez,[4] Adam J. Gordon,[5,6] Michihiko Goto,[7,8] Matthew Bidwell Goetz,[2,3] Steven Shoptaw,[9] and Alex A. T. Bui[10]

[1]Division of Infectious Diseases, David Geffen School of Medicine, University of California, Los Angeles, Los Angeles, California, USA, [2]Veterans Affairs Greater Los Angeles Healthcare System, Los Angeles, California, USA, [3]Department of Internal Medicine, David Geffen School of Medicine, University of California, Los Angeles, Los Angeles, California, USA, [4]Undergraduate Studies, Dartmouth College, Hanover, New Hampshire, USA, [5]Informatics, Decision-Enhancement, and Analytic Sciences Center, Veterans Affairs Salt Lake City Health Care System, Salt Lake City, Utah, USA, [6]Division of Epidemiology, Department of Internal Medicine, University of Utah School of Medicine, Salt Lake City, Utah, USA, [7]Department of Internal Medicine, University of Iowa, Iowa City, Iowa, USA, [8]Center for Access and Delivery Research and Evaluation, Iowa City Veterans Affairs Medical Center, Iowa City, Iowa, USA, [9]Department of Family Medicine, David Geffen School of Medicine, University of California, Los Angeles, Los Angeles, California, USA, and [10]Medical and Imaging Informatics Group, Department of Radiological Sciences, University of California, Los Angeles, Los Angeles, California, USA

Background. Improving the identification of people who inject drugs (PWID) in electronic medical records can improve clinical decision making, risk assessment and mitigation, and health service research. Identification of PWID currently consists of heterogeneous, nonspecific *International Classification of Diseases* (*ICD*) codes as proxies. Natural language processing (NLP) and machine learning (ML) methods may have better diagnostic metrics than nonspecific *ICD* codes for identifying PWID.

Methods. We manually reviewed 1000 records of patients diagnosed with *Staphylococcus aureus* bacteremia admitted to Veterans Health Administration hospitals from 2003 through 2014. The manual review was the reference standard. We developed and trained NLP/ML algorithms with and without regular expression filters for negation (NegEx) and compared these with 11 proxy combinations of *ICD* codes to identify PWID. Data were split 70% for training and 30% for testing. We calculated diagnostic metrics and estimated 95% confidence intervals (CIs) by bootstrapping the hold-out test set. Best models were determined by best F-score, a summary of sensitivity and positive predictive value.

Results. Random forest with and without NegEx were the best-performing NLP/ML algorithms in the training set. Random forest with NegEx outperformed all *ICD*-based algorithms. F-score for the best NLP/ML algorithm was 0.905 (95% CI, .786–.967) and 0.592 (95% CI, .550–.632) for the best *ICD*-based algorithm. The NLP/ML algorithm had a sensitivity of 92.6% and specificity of 95.4%.

Conclusions. NLP/ML outperformed *ICD*-based coding algorithms at identifying PWID in electronic health records. NLP/ML models should be considered in identifying cohorts of PWID to improve clinical decision making, health services research, and administrative surveillance.

Keywords. EHR; machine learning; NLP; PWID.

Injection drug use (IDU) and its complications continue to rise in the United States (US). The estimated number of people who inject drugs (PWID) nearly quintupled from 2011 to 2018 [1] and there was an 8-fold increase in PWID-related overdoses from 2000 to 2018 [2]. With the rise of IDU, epidemiologic studies reported an increase in bacterial infections related to IDU (eg, skin and soft tissue infections, endocarditis) [3–5] and viral pathogens (human immunodeficiency virus [HIV] and hepatitis C virus [HCV]) [6, 7]. Accurate identification of PWID is important clinically for prognostic reasons, risk assessment, and risk mitigation, as well as to provide correct estimates in health services research, and for administrative purposes like resource allocation. However, current methods for identification of PWID in electronic health record (EHR)–based studies are inaccurate [8–13].

An inherent problem with estimates of PWID are the use of *International Classification of Diseases* (*ICD*) codes to identify cases. In clinical practice, *ICD* codes are used for the purpose of billing clinical encounters and are often specified by healthcare providers or extracted by trained human coders who review clinical notes [14]. An important caveat regarding PWID is that there is no *ICD* code for IDU [8, 10]. Both in clinical and epidemiologic studies, authors have used different and heterogeneous combinations of nonspecific codes as proxies to identify PWID [3, 15, 16]. Yet, recent studies have

Received 04 August 2022; editorial decision 06 September 2022; accepted 08 September 2022; published online 12 September 2022

Correspondence: David Goodman-Meza, MD, MAS, David Geffen School of Medicine at UCLA, 10833 Le Conte Ave, CHS 52-215, Los Angeles, CA, 90095-1688 (dgoodman@mednet.ucla.edu)

Open Forum Infectious Diseases®
Published by Oxford University Press on behalf of Infectious Diseases Society of America 2022.

https://doi.org/10.1093/ofid/ofac471

demonstrated the inaccuracies of *ICD*-based approaches [8–12]. Marks et al showed that in 229 cases of endocarditis, an *ICD*-based approach had a sensitivity of 65% compared to manual review [9]. When analyzing the effect of medications for opioid use disorder (OUD) on patient-directed discharges and mortality, results were nonsignificant when using the *ICD*-based approach, compared to protective when using only the manually reviewed cases. McGrew et al showed a sensitivity of 70.3%–81.9% and specificity of 97.8%–98.9% in identifying IDU in 321 cases with endocarditis using an *ICD*-based approach compared to manual review [11]. In a systematic review, authors identified 3 other manuscripts that used *ICD* codes to identify PWID. Their summary finding was that there is wide heterogeneity in the combinations of codes used to identify PWID [10].

Natural language processing (NLP) and machine learning are potential options to identify cohorts from EHR data [17]. NLP is a field of computer science that concerns the processing, understanding, and generation of spoken or written language. There are many tasks that NLP deals with, largely around the identification of concepts and relationships; but for this article we deal with the methods of automated data extraction to convert unstructured free text to structured, encoded data by applying probabilistic or rule-based algorithms to combinations of terms [18]. Previously, NLP has been used in many research tasks to identify problems related to opioid use and its related harms in diverse types of EHRs. Algorithms have been developed and evaluated to identify "problem opioid use" [19], opioid misuse [20, 21], opioid-related aberrant behaviors [22], OUD [23, 24], and to characterize opioid use [25]. NLP has also been used to identify overdoses in both hospital records and coroner's reports [26–30]. To date, there has been no published attempt at developing an algorithm to identify PWID in EHRs.

Given the continued rise of IDU-related complications in parallel to the ongoing drug use epidemic in the US [1, 31], ongoing epidemiologic and clinical research are needed to both describe the problem and evaluate solutions to improve outcomes in care. An important initial step in these two types of research is having an accurately identified cohort. Based on the recent literature, *ICD* codes are insufficient for that purpose. Our objective was to develop and evaluate an NLP/machine learning algorithm to identify PWID in the EHR and compare these algorithms to *ICD* codes. We hypothesized that we could use methods from NLP and machine learning to identify a PWID cohort more accurately from the unstructured clinical notes contained in EHR compared to an *ICD* code–based strategy.

METHODS

Data

This was a retrospective EHR study. The study was approved by the University of California, Los Angeles Institutional Review Board (IRB), the Veterans Health Administration (VHA) Greater Los Angeles Healthcare System Research and Development Committee, the University of Iowa IRB, and the Iowa City VHA Research and Development Committee. Informed consent was waived for this study. We used a previously assembled cohort developed by Goto et al [32]. This dataset contained a cohort of 36 868 cases of patients diagnosed with *Staphylococcus aureus* bacteremia at 124 VHA hospitals between 1 January 2003 and 31 December 2014. This dataset covered the 48 continental states, District of Columbia, and Puerto Rico (Alaska and Hawaii are not included as the VHA did not have acute inpatient units in these states). This dataset was selected due to its extensive characterization, wide geographic distribution, large numbers, and high probability of including PWID. This study is reported following the Transparent Reporting of a Multivariable Prediction Model for Individual Prognosis or Diagnosis (TRIPOD) reporting guideline [33].

We recreated the cohort and extracted data for the cases from the VHA Corporate Data Warehouse within the VHA's Veterans Informatics and Computing Infrastructure (VINCI) framework. We extracted admission notes for each case during their hospitalization for *S aureus* bacteremia. Additionally, we extracted demographic variables (age, sex, race, ethnicity, hospital), laboratory features related to urine toxicology, and *ICD, Ninth Revision* (*ICD*-9) codes. We used keyword matching within the notes and positive urine drug screens (opiates, cocaine, methamphetamines) to subset the cases with the highest risk for being PWID. We randomly selected 1000 individual cases and split the data 70%/30% between a training and testing dataset (held-out). Final evaluation comparing *ICD* code algorithms and NLP algorithms were done on the testing dataset. Figure 1 depicts the overall study design.

Reference Standard

Our reference standard was manual case review. We annotated all text data from the admission notes in VINCI ChartReview [34]. B. A. was trained by D. G.-M. in annotating text based on a predefined annotator guide (Supplementary Table 1). B. A. annotated all charts after he achieved an interannotator agreement (κ) with D. G.-M. >0.80 on a subset of 100 cases. Once B. A. completed all annotations, D. G.-M. verified the accuracy of all classification of notes pertaining to PWID.

ICD Code Algorithms

We extracted *ICD* code–based algorithms to identify PWID from Ball et al [8]. These algorithms used different combinations of codes to denote opioid use dependence/abuse, hepatitis C, HIV, other substance use disorders, substance use, and homelessness. These codes were published using *ICD, Tenth Revision* (*ICD-10*) codes. During the study period of our EHR data (2003–2014), the VHA used *ICD-9* codes (the

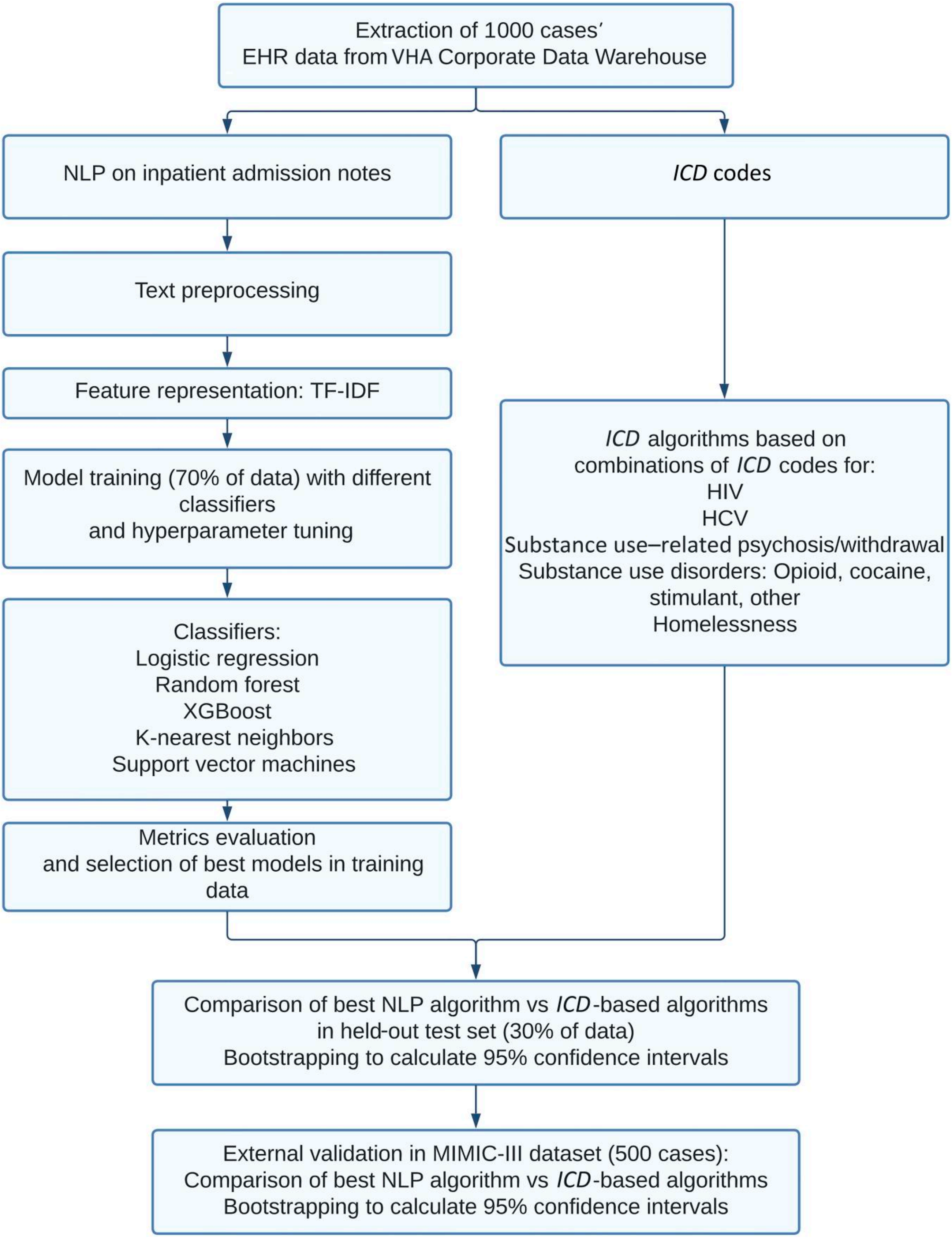
Extraction of 1000 cases'
EHR data from VHA Corporate Data Warehouse
NLP on inpatient admission notes
ICD codes
Text preprocessing
Feature representation: TF-IDF
Model training (70% of data) with different classifiers and hyperparameter tuning
Classifiers:
Logistic regression
Random forest
XGBoost
K-nearest neighbors
Support vector machines
Metrics evaluation and selection of best models in training data
ICD algorithms based on combinations of *ICD* codes for:
HIV
HCV
Substance use–related psychosis/withdrawal
Substance use disorders: Opioid, cocaine, stimulant, other
Homelessness
Comparison of best NLP algorithm vs *ICD*-based algorithms in held-out test set (30% of data)
Bootstrapping to calculate 95% confidence intervals
External validation in MIMIC-III dataset (500 cases):
Comparison of best NLP algorithm vs *ICD*-based algorithms
Bootstrapping to calculate 95% confidence intervals

Figure 1. Study design. Abbreviations: EHR, electronic health record; HCV, hepatitis C virus; HIV, human immunodeficiency virus; ICD, *International Classification of Diseases*; NLP, natural language processing; TF-IDF, term frequency–inverse document frequency; VHA, Veterans Health Administration.

Table 1. Descriptive Characteristics of Sampled Cases for Annotation Stratified by Inclusion in the Training and Testing Datasets

Characteristic	Overall (N = 1000)	Train (n = 700)	Test (n = 300)	*P* Value
Age, y, median (IQR)	56 (52–61)	57 (52–62)	56 (50–61)	.028
Sex				.075
Male	971 (97)	685 (98)	286 (96)	
Female	29 (2.9)	16 (2.3)	13 (4.3)	
Race				.14
White	536 (54)	374 (53)	162 (54)	
Black/African American	386 (39)	277 (40)	109 (36)	
American Indian/Alaska Native	16 (1.6)	7 (1.0)	9 (3.0)	
NHPI	5 (0.5)	4 (0.6)	1 (0.3)	
Asian	1 (0.1)	0 (0)	1 (0.3)	
Unknown	56 (5.6)	39 (5.6)	17 (5.7)	
Ethnicity				.2
Not Latino	888 (89)	615 (88)	273 (91)	
Latino	64 (6.4)	47 (6.7)	17 (5.7)	
Unknown	48 (4.8)	39 (5.6)	9 (3.0)	
District				.4
Pacific	258 (26)	173 (25)	85 (28)	
North Atlantic	238 (24)	170 (24)	68 (23)	
Southeast	212 (21)	154 (22)	58 (19)	
Continental	173 (17)	126 (18)	47 (16)	
Midwest	119 (12)	78 (11)	41 (14)	
Substances reported[a]				
Cocaine or crack	294 (29)	200 (29)	94 (31)	.4
Heroin	129 (13)	86 (12)	43 (14)	.4
Methamphetamine/amphetamine	44 (4.4)	27 (3.9)	17 (5.7)	.2
Prescription opioid	78 (7.8)	55 (7.8)	23 (7.7)	>.9
Cannabis/marijuana	84 (8.4)	60 (8.6)	24 (8.0)	.8
Benzodiazepine	7 (0.7)	5 (0.7)	2 (0.7)	>.9
Comorbidities[b]				
HIV	449 (45)	309 (44)	140 (47)	.4
Hepatitis C virus	554 (55)	390 (56)	164 (55)	.8
Substance use disorder	680 (68)	475 (68)	205 (69)	.8
Substance-related psychosis/withdrawal	254 (25)	177 (25)	77 (26)	.9
Homelessness	283 (28)	211 (30)	72 (24)	.053
Injection drug use status[a]				.9
Non-PWID	716 (72)	501 (71)	215 (72)	
PWID	284 (28)	200 (29)	84 (28)	

Data are presented as No. (%) unless otherwise indicated.

Abbreviations: HIV, human immunodeficiency virus; IQR, interquartile range; NHPI, Native Hawaiian and other Pacific Islander; PWID, person who injects drugs.

[a] Identified by manual review.

[b] Identified by *International Classification of Diseases* codes.

VHA transitioned to *ICD-10* on 1 October 2015) [35]. As the dataset that we used in our analyses predated *ICD-10*, we translated the different algorithms' *ICD-10* codes to *ICD-9* codes using general equivalence mapping. Supplementary Tables 2 and 3 show codes, mappings, and combination *ICD* code algorithms. We used all *ICD* codes for each case in the EHR for analysis.

NLP Algorithms

We split the annotated dataset into training (70%) and testing (30%) datasets. A general NLP pipeline consists of preprocessing data, converting text into numerical features, modeling, and evaluating results. First, during data preprocessing for all notes, we converted all case to lower case; removed numbers, dates, names, and general English stop words; and filtered out words with <10 mentions in the text. Additionally in the preprocessing stage, we compared between the use of NegEx [36], a rule-based algorithm to remove negated terms within clinical text; the use of regular general expressions (RegEx) to abbreviate certain terms related to IDU (eg, "IV drug abuse" to "IVDA"); and the use of neither of these approaches. We also compared different combinations of N-grams. N-grams are continuous sequences of words. We compared the use of N-grams of length 1 (unigrams), 2 (bigrams), and 3 (trigrams). For example, in the sentence "The patient denied drug use," individually each word is a unigram (eg, "the," "patient," "denied"), combinations of 2 words are bigrams (eg, "the patient," "patient denied," "denied drug"), and combinations of 3 words are trigrams (eg, "the patient denied," "patient denied drug," "denied drug use").

Second, we converted textual features to numeric vectors using term frequency–inverse document frequency (TF-IDF). TF-IDF is a frequency-based numeric representation of each word that provides a higher weight to rarer words that are likely to provide more meaning within a text. TF-IDF is calculated as the product of the TF (number of times a word appears in each observation) and IDF. The IDF is calculated as the log of the number of documents divided by the number of documents that contain the word or words in question.

Third, we trained and evaluated several machine learning classifiers on their predictive ability to identify PWID based on the annotator's classification (see Reference Standard above). Machine learning classifiers and different combination of their hyperparameters were trained on 70% of the annotated data (training dataset) using 10-fold cross-validation. Machine learning classifiers included logistic regression, K-nearest neighbor, support vector machines, random forest, and XGBoost. Hyperparameters were tuned based on an automated grid search. We selected the best classifier based on the mean F-score of the 10-fold cross-validation. We used the mean F-score instead of accuracy or area under the receiver operating characteristic (ROC) curve due to the imbalance in the outcome, and the F-score not being influenced by the outcome's prevalence.

Final Evaluation

We compared metrics for the *ICD*-based approaches and the NLP-based approaches to evaluate the best-performing strategy. Using only the held-out test set (30% of the annotated dataset), we calculated diagnostic metrics to include accuracy, F-score, sensitivity (recall), specificity, positive predictive value (precision), negative predictive value, and the area under the ROC curve. We calculated 95% confidence intervals (CIs) by bootstrapping resampling. We bootstrapped the testing set with replacement 1000 times and calculated diagnostic metrics for each resample. We report the 2.5th and 97.5th percentile as the lower and upper end of the CI, respectively, and the 50th percentile as the mean. We performed a manual error analysis of the false-positive and false-negative predictions from the best-performing NLP models. All statistical analyses were performed in R 4.1.2 software.

Validation Analyses

We performed an external validation analysis. Here, we compared the best-performing NLP models from the previous analyses to the *ICD* code algorithms in the MIMIC-III dataset [37]. This dataset contains de-identified data from admissions to the intensive care unit at Beth Israel Deaconess Medical Center in Boston, Massachusetts, between 2001 and 2012. De-identified discharge summaries were available, as well as *ICD*-9 codes. Similar to our original analysis, cases were selected based on a high likelihood of pertaining to PWID based on keyword matching. Then, we annotated text documents to identify PWID and non-PWID as the gold standard. A. T. and S. V. classified documents after attaining an interannotator agreement >0.80 on a subset of 50 cases. All notes pertaining to PWID were then verified by D. G.-M. *ICD* algorithms tested were the same as the algorithms described above, and we used the best-performing NLP model as a comparison.

RESULTS

Text contained a median of 821 words (interquartile range [IQR], 479–11 856 words). The median number of characters per text was 5271 (IQR, 2956–7632 characters). There were no significant differences in the number of words or characters between notes describing PWID and non-PWID.

Table 1 shows descriptive statistics for the annotated cases. Median age was 56 years (IQR, 52–61 years), 97% were male, 54% White, 39% Black, and 6% Latino. Cocaine (29%) was the most mentioned substance in the notes, followed by heroin (13%), prescription opioids (8%), and methamphetamine (4%). Based on *ICD* codes, HIV was present in 45%, HCV in 55%, a substance use disorder in 68%, and homelessness in 28%. Notes were classified as pertaining to a PWID in 28% and non-PWID in 72%.

Table 2 shows F-scores for the NLP algorithms in 10-fold cross-validation of the training dataset. Tree-based models

Table 2. Mean F-Scores Based on 10-Fold Cross-Validation for Machine Learning Models in Testing Dataset (n = 700)

Model	N-Grams	No NegEx F-Score	NegEx F-Score
Random forest	Unigram	0.829	0.908
Random forest	Bigram	**0.871**	**0.913**
Random forest	Trigram	0.866	**0.913**
XGBoost	Unigram	0.807	0.901
XGBoost	Bigram	0.843	0.909
XGBoost	Trigram	0.839	0.907
SVM	Unigram	0.527	0.568
SVM	Bigram	0.485	0.576
SVM	Trigram	0.494	0.551
Logistic regression	Unigram	0.371	0.366
Logistic regression	Bigram	0.332	0.346
Logistic regression	Trigram	0.362	0.319
KNN	Unigram	0.195	0.460
KNN	Bigram	0.223	0.421
KNN	Trigram	0.151	0.319

Values in bold indicate the best-performing models based on F-scores.

Abbreviations: KNN, K-nearest neighbor; SVM, support vector machine.

Table 3. Comparison of *International Classification of Diseases*–Based Algorithms and Best-Performing Natural Language Processing/Machine Learning Model in Held-Out Test Set (n = 300)

Algorithm	Accuracy	AUROC	F-Score	κ	Negative Predictive Value	Positive Predictive Value/ Precision	Sensitivity/Recall	Specificity
ICD algorithms								
Algorithm 6	0.565 (.515–.615)	0.675 (.629–.716)	0.55 (.513–.585)	0.247 (.181–.313)	0.939 (.887–.98)	0.391 (.361–.422)	0.931 (.872–.977)	0.418 (.358–.479)
Algorithm 7	0.528 (.478–.575)	0.648 (.605–.688)	0.53 (.495–.562)	0.204 (.142–.265)	0.929 (.868–.977)	0.37 (.343–.398)	0.93 (.872–.977)	0.367 (.302–.433)
Algorithm 8	0.562 (.508–.615)	0.593 (.533–.651)	0.464 (.402–.522)	0.147 (.053–.241)	0.796 (.742–.845)	0.357 (.312–.403)	0.665 (.558–.756)	0.52 (.456–.581)
Algorithm 9	0.504 (.458–.548)	0.646 (.612–.68)	0.53 (.505–.556)	0.194 (.145–.246)	0.971 (.931–1)	0.364 (.343–.386)	0.977 (.942–1)	0.316 (.251–.377)
Algorithm 10	0.637 (.588–.684)	0.722 (.677–.765)	0.592 (.55–.632)	0.334 (.259–.404)	0.943 (.898–.982)	0.437 (.399–.475)	0.92 (.86–.977)	0.524 (.46–.586)
Algorithm 11	0.537 (.485–.585)	0.651 (.605–.693)	0.531 (.495–.566)	0.211 (.144–.274)	0.921 (.861–.968)	0.374 (.345–.403)	0.918 (.849–.965)	0.385 (.321–.451)
Algorithm 12	0.488 (.445–.532)	0.635 (.6–.669)	0.522 (.498–.547)	0.177 (.128–.227)	0.969 (.925–1)	0.356 (.335–.378)	0.977 (.942–1)	0.293 (.228–.353)
Algorithm 13	0.548 (.502–.595)	0.663 (.62–.703)	0.541 (.505–.574)	0.228 (.163–.29)	0.935 (.882–.978)	0.381 (.354–.411)	0.931 (.872–.977)	0.394 (.335–.456)
Algorithm 14	0.501 (.455–.545)	0.644 (.608–.677)	0.528 (.503–.553)	0.191 (.139–.242)	0.971 (.93–1)	0.362 (.34–.385)	0.977 (.942–1)	0.311 (.247–.372)
Algorithm 15	0.485 (.442–.528)	0.632 (.598–.666)	0.52 (.496–.545)	0.173 (.124–.223)	0.969 (.924–1)	0.354 (.333–.376)	0.977 (.942–1)	0.288 (.228–.349)
Algorithm 16	0.461 (.419–.505)	0.616 (.583–.649)	0.509 (.486–.533)	0.149 (.105–.198)	0.965 (.915–1)	0.344 (.326–.365)	0.977 (.942–1)	0.255 (.195–.312)
NLP algorithms								
Random forest—No NegEx	0.881 (0.826–0.93)	0.933 (0.877–0.977)	0.766 (0.612–0.863)	0.688 (0.512–0.815)	0.889 (0.819–0.957)	0.866 (0.727–1)	0.697 (0.484–0.889)	0.954 (0.89–1)
Random forest—NegEx	0.945 (.890–.981)	0.959 (.919–.991)	0.905 (.786–.967)	0.867 (.716–.954)	0.970 (.884–1.00)	0.890 (.810–.969)	0.926 (.697–1.00)	0.954 (.914–.988)

Numbers in parentheses denote 95% confidence intervals based on bootstrapping.

Abbreviations: AUROC, area under the receiver operating characteristic curve; *ICD, International Classification of Diseases*; NLP, natural language processing.

(random forest and XGBoost) were the best-performing models. Support vector machines, logistic regression, and k-nearest neighbors performed poorly. Models that incorporated bigrams were better than their unigram or trigram counterparts. Adding NegEx improved F-scores for all models, except logistic regression. The best-performing model was random forest with bigrams and NegEx (mean F-score 0.913, mean sensitivity 0.945, mean specificity 0.950). Without NegEx, random forest with bigrams had the best performance (mean F-score 0.871, mean sensitivity 0.945, mean specificity 0.910). Full diagnostic metrics can be seen in Supplementary Tables 4 and 5.

Table 3 shows performance metrics for the best NLP/machine learning algorithms compared to the *ICD*-based algorithms in the testing dataset. The NLP/machine learning algorithms outperformed all *ICD*-based algorithms based on F-scores. *ICD*-based algorithms had a high sensitivity (range of means, 0.665–0.977) but a low specificity (range of means, 0.255–0.524). The NLP/machine learning model that incorporated NegEx outperformed its counterpart that did not incorporate NegEx. The former had both excellent sensitivity (mean, 0.926 [95% CI, .697–1.00]) and specificity (mean, 0.954 [95% CI, 0.914–0.988]), whereas the latter had a poor sensitivity (mean, 0.697 [95% CI, .484–.889]) but excellent specificity (mean, 0.954 [95% CI, .89–1.00]). ROC and precision-recall (PR) curves can be seen in Supplementary Figures 1 and 2.

In our manual error analysis, we reviewed 13 cases where the NLP algorithm had false predictions. The NLP model predicted 3 cases as false negatives, and 10 cases as false positives. We evaluated age and race/ethnicity as potential sources of errors and did not find any specific pattern in the cases with errors. Reasons for false negatives included missing phrases such as "injecting oxycontin," "past IVD+," and "patient admits to IV drug abuse." Reasons for the false positive mostly included the inability for the NegEx algorithm to remove a negated mention of IDU. For example, keywords denoting IDU were not removed in phrases such as "claims he has not used any illicit drugs [cocaine/heroin/amphetamine]" and "denies any other drug use to include IV drug use."

Validation Analysis

In the external validation analysis, the NLP (ie, random forest with NegEx) model outperformed all *ICD*-based algorithms in the MIMIC-III dataset. The F-score for the NLP model was 0.929 (95% CI, .901–.954) compared to 0.775 (95% CI, .740–.809) for the best *ICD*-based algorithm (ie, algorithm 10). Of note, *ICD* algorithms performed better in the MIMIC-III dataset than in the VHA dataset. Descriptive statistics for the MIMIC-III sample can be seen in Supplementary Table 6 and full diagnostic metrics can be seen in Supplementary Table 7; ROC and PR curves can be seen in Supplementary Figure 3.

DISCUSSION

We showed that NLP/machine learning outperformed *ICD*-based algorithms, scoring a mean of ≥0.9 in every technical performance metric. *ICD*-based algorithms had a high sensitivity but a very poor specificity, leading to an overestimation of cases as PWID. Additionally, we externally validated the best-performing algorithm, again finding excellent diagnostic metrics for classification of cases as PWID or non-PWID.

We also showed that machine learning models alone were not sufficient in attaining excellent performance. Adding a rule-based algorithm for negation (NegEx) and regular expressions improved model performance significantly—specifically the sensitivity of the NLP/machine learning models. This is likely due to models keying on the appearance of singular words like IVDU [intravenous drug use], IVDA, or heroin. For example, if a common phrase like "denied smoking, alcohol use or IVDA" or "denied any recent heroin use" appeared, the non-NegEx model would mark this as pertaining to a PWID. It is possible that the use of longer N-grams could capture these phrases as significant for the negative class, but this would also increase the computational complexity significantly. Adding a simple rule-based strategy increased the models' performance dramatically.

Based on our present work, and that of others, combinations of *ICD* codes to identify PWID are nonspecific and inaccurate. Use of these codes in research risks identifying a high number of false-positive cases. In epidemiologic studies, this may inflate the number of cases pertaining to PWID, overestimating the burden of the behavior. In clinical effectiveness studies, inclusion of false positives may bias toward the null when examining the effects of important interventions such as medications for OUD. This was the case in the study published by Marks et al [13], where use of *ICD* codes alone led to no difference in negative outcomes based on medications for OUD status. However, when manually reviewing cases and analyzing a smaller and more accurate cohort, they showed that medication for OUD status was protective of all-cause mortality.

NLP/machine learning algorithms could be embedded within clinical contexts to improve patient care. For example, real-time screening of daily admission notes could trigger alerts to addiction medicine teams to intervene and improve the quality of care by offering medication for substance use disorders and other services. If not in real time, screening of notes could trigger alerts to backend teams for case review and posterior intervention, like the VHA's Project STORM (Stratification Tool for Opioid Risk Management) [38]. Nevertheless, future applications of NLP algorithms to trigger clinical care need to be rigorously evaluated to determine if they improve patient outcomes before widespread implementation.

NLP/machine learning should be considered for future research or for administrative purposes related to PWID. For example, NLP/machine learning could be used to evaluate trends of complications in PWID (eg, endocarditis or other deep-seeded infections) or in comparative effectiveness studies evaluating different interventions (eg, medications for OUD or other harm reduction services). However, certain pros and cons would need to be considered. Use of NLP/machine learning would require access to patients' notes and expertise in open-source software such as R or Python. Manual review is still the most accurate but requires a large amount of time to review the EHR compared to our NLP/machine learning algorithm. *ICD*-based identification of PWID may be the most accessible as *ICD* codes are available in most EHRs and many administrative databases used for epidemiologic research. Many administrative databases do not have access to patients' notes, and *ICD* codes are the only viable option. Yet, *ICD*-based identification has the drawback of inaccuracies that we demonstrated here. A solution could be a specific *ICD* code for PWID. Until then, the inaccuracies are likely to persist.

This study has several limitations. First, correct identification of PWID by an NLP algorithm is first based on clinicians documenting the behavior in their clinical notes. However, many issues may arise that include poor documentation from providers or failure to disclose the behavior from patients out of fear of stigma [39]. Second, the dataset used is from 2003 to 2014. Since then, the IDU epidemic has shifted away from prescription opioids and heroin to synthetic opioids like fentanyl. Based on our experience reviewing contemporary charts in the hospital, terms such as "IVDU" and "IVDA" are still present when describing fentanyl use via injection. However, this must be tested empirically, and continuous model development and evaluation are necessary to ensure accuracy of these models as substance use patterns change. Third, the study lacks generalizability to other settings and populations. The data were entirely from the VHA, all cases had *S aureus* bacteremia, and the sample was almost entirely male, reflective of the US Veteran population. To these points, the algorithm was validated in MIMIC-III—a cohort of cases admitted to critical care units at a large tertiary care hospital inclusive of both men and women, as well as cases other than *S aureus* bacteremia. Fourth, we used the NegEx algorithm to filter out negated terms. Newer negation algorithms that are either lexical or syntax based, such as ConTextNLP [40] and DEEPEN [41], or machine learning based [42] are available and may improve negation and provide other contextual findings. However, none of these were implemented in R at the time of our algorithm development. Additionally, other more advanced NLP techniques are also available that include the use of medical ontologies (eg, Unified Medical Language System) to normalize text, word embeddings, and transfer learning from medical-specific models, or deep learning.

CONCLUSIONS

We showed that NLP/machine learning–based algorithms outperformed *ICD* code–based algorithms at identifying PWID in a large national dataset from the VHA and externally validated this performance in an external dataset. We add to the body of literature that shows that *ICD* codes are inaccurate and insufficient for the purpose of identifying PWID [8–12]. The continued use of *ICD* codes for the purpose of identifying PWID may lead to inaccurate estimates with a major concern being overestimation of the burden of PWID. Further and continued research is necessary to validate NLP/machine learning models in the landscape of changing substance use patterns and incorporate NLP/machine learning in clinical decision support and future research or surveillance projects. Other research directions using NLP in substance use research include developing named entity recognition models to identify specific substances used or other ancillary information such as chronicity or use of treatments such as methadone or buprenorphine.

Supplementary Data

Supplementary materials are available at *Open Forum Infectious Diseases* online. Consisting of data provided by the authors to benefit the reader, the posted materials are not copyedited and are the sole responsibility of the authors, so questions or comments should be addressed to the corresponding author.

Notes

Acknowledgments. We appreciate the support of Olga Patterson and Lacey Lewis for their assistance with ChartReview; Quan Douc and Anita Yuan for their support in extracting admission notes from the Corporate Data Warehouse; and Simon Hans for his implementation of NegEx in R.

Data availability. Due to the inability to protect participant confidentiality within text notes used to develop and validate the models, source data are not made publicly available. All code and trained models are available at https://github.com/davigood1/PWID_NLP.

Disclaimer. The views expressed in this article are those of the authors and do not necessarily reflect the position or policy of the Department of Veterans Affairs, the United States government, the National Institutes of Health, or the authors' affiliated universities. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Financial support. This material is the result of work supported with resources and the use of facilities at the VHA Greater Los Angeles Healthcare System. D. G.-M. was supported by the US National Institute on Drug Abuse (grant number K08DA048163). S. S. was supported by the Center for HIV Identification, Prevention, and Treatment Services (grant number P30MH58107), the UCLA Center for AIDS Research (grant number 5P30AI028697), and the UCLA Clinical Translational Science Institute (grant number UL1TR001881).

Potential conflicts of interest. All authors: No reported conflicts.

All authors have submitted the ICMJE Form for Disclosure of Potential Conflicts of Interest. Conflicts that the editors consider relevant to the content of the manuscript have been disclosed.

References

1. Bradley H, Hall E, Asher A, et al. Estimated number of people who inject drugs in the United States [manuscript published online ahead of print 6 July 2022]. Clin Infect Dis **2022**. doi:10.1093/cid/ciac543
2. Hall EW, Rosenberg ES, Jones CM, Asher A, Valverde E, Bradley H. Estimated number of injection-involved drug overdose deaths, United States, 2000–2018. Drug Alcohol Depend **2022**; 234:109428.
3. Wurcel AG, Anderson JE, Chui KK, et al. Increasing infectious endocarditis admissions among young people who inject drugs. Open Forum Infect Dis **2016**; 3: ofw157.
4. Sredl M, Fleischauer AT, Moore Z, Rosen DL, Schranz AJ. Not just endocarditis: hospitalizations for selected invasive infections among persons with opioid and

stimulant use diagnoses—North Carolina, 2010–2018. J Infect Dis **2020**; 222(Suppl 5):S458–464.

5. See I, Gokhale RH, Geller A, et al. National public health burden estimates of endocarditis and skin and soft-tissue infections related to injection drug use: a review. J Infect Dis **2020**; 222(Suppl 5):S429–36.
6. Powell D, Alpert A, Pacula RL. A transitioning epidemic: how the opioid crisis is driving the rise in hepatitis C. Health Aff (Millwood) **2019**; 38:287–94.
7. Strathdee SA, Kuo I, El-Bassel N, Hodder S, Smith LR, Springer SA. Preventing HIV outbreaks among people who inject drugs in the United States: plus ça change, plus ça même chose. AIDS **2020**; 34:1997–2005.
8. Ball LJ, Sherazi A, Laczko D, et al. Validation of an algorithm to identify infective endocarditis in people who inject drugs. Med Care **2018**; 56:e70–5.
9. Marks LR, Nolan NS, Jiang L, Muthulingam D, Liang SY, Durkin MJ. Use of *ICD*-10 codes for identification of injection drug use–associated infective endocarditis is nonspecific and obscures critical findings on impact of medications for opioid use disorder. Open Forum Infect Dis **2020**; 7:ofaa414.
10. McGrew K, Homco J, Garwe T, et al. Validity of *International Classification of Diseases* codes in identifying illicit drug use target conditions using medical record data as a reference standard: a systematic review. Drug Alcohol Depend **2020**; 208:107825.
11. McGrew KM, Carabin H, Garwe T, et al. Validity of *ICD*-based algorithms to estimate the prevalence of injection drug use among infective endocarditis hospitalizations in the absence of a reference standard. Drug Alcohol Depend **2020**; 209: 107906.
12. Barnes E, Peacock J, Bachmann L. *International Classification of Diseases* (*ICD*) codes fail to accurately identify injection drug use associated endocarditis cases. J Addict Med **2022**; 16:27–32.
13. Marks LR, Nolan NS, Jiang L, Muthulingam D, Liang SY, Durkin MJ. Use of *ICD*-10 codes for identification of injection drug use-associated infective endocarditis is nonspecific and obscures critical findings on impact of medications for opioid use disorder. Open Forum Infect Dis **2020**; 7:ofaa414.
14. Rios A, Kavuluru R. Supervised extraction of diagnosis codes from EMRs: role of feature selection, data selection, and probabilistic thresholding. IEEE Int Conf Healthc Inform **2013**; 2013:66–73.
15. Cooper HLF, Brady JE, Ciccarone D, Tempalski B, Gostnell K, Friedman SR. Nationwide increase in the number of hospitalizations for illicit injection drug use-related infective endocarditis. Clin Infect Dis **2007**; 45:1200–3.
16. Hartman L, Barnes E, Bachmann L, Schafer K, Lovato J, Files DC. Opiate injection-associated infective endocarditis in the southeastern United States. Am J Med Sci **2016**; 352:603–8.
17. Hayes CJ, Cucciare MA, Martin BC, et al. Using data science to improve outcomes for persons with opioid use disorder. Subst Abus **2022**; 43:956–63.
18. Nadkarni PM, Ohno-Machado L, Chapman WW. Natural language processing: an introduction. J Am Med Inform Assoc **2011**; 18:544–51.
19. Carrell DS, Cronkite D, Palmer RE, et al. Using natural language processing to identify problem usage of prescription opioids. Int J Med Inform **2015**; 84: 1057–64.
20. Afshar M, Sharma B, Bhalla S, et al. External validation of an opioid misuse machine learning classifier in hospitalized adult patients. Addict Sci Clin Pract **2021**; 16:19.
21. Afshar M, Sharma B, Dligach D, et al. Development and multimodal validation of a substance misuse algorithm for referral to treatment using artificial intelligence (SMART-AI): a retrospective deep learning study. Lancet Digit Health **2022**; 4: e426–35.
22. Lingeman JM, Wang P, Becker W, Yu H. Detecting opioid-related aberrant behavior using natural language processing. AMIA Annu Symp Proc **2017**; 2017: 1179–85.
23. Blackley SV, MacPhaul E, Martin B, Song W, Suzuki J, Zhou L. Using natural language processing and machine learning to identify hospitalized patients with opioid use disorder. AMIA Annu Symp Proc **2020**; 2020:233–42.
24. Zhu VJ, Lenert LA, Barth KS, et al. Automatically identifying opioid use disorder in non-cancer patients on chronic opioid therapy. Health Informatics J **2022**; 28: 14604582221107808.
25. Poulsen MN, Freda PJ, Troiani V, Davoudi A, Mowery DL. Classifying characteristics of opioid use disorder from hospital discharge summaries using natural language processing. Front Public Health **2022**; 10:850619.
26. Ward PJ, Rock PJ, Slavova S, Young AM, Bunn TL, Kavuluru R. Enhancing timeliness of drug overdose mortality surveillance: a machine learning approach. PLoS One **2019**; 14:e0223318.
27. Badger J, LaRose E, Mayer J, Bashiri F, Page D, Peissig P. Machine learning for phenotyping opioid overdose events. J Biomed Inform **2019**; 94:103185.
28. Hazlehurst B, Green CA, Perrin NA, et al. Using natural language processing of clinical text to enhance identification of opioid-related overdoses in electronic health records data. Pharmacoepidemiol Drug Saf **2019**; 28:1143–51.
29. Harris DR, Eisinger C, Wang Y, Delcher C. Challenges and barriers in applying natural language processing to medical examiner notes from fatal opioid poisoning cases. Proc IEEE Int Conf Big Data **2020**; 2020:3727–36.
30. Goodman-Meza D, Shover CL, Medina JA, Tang AB, Shoptaw S, Bui AAT. Development and validation of machine models using natural language processing to classify substances involved in overdose deaths. JAMA Netw Open **2022**; 5: e2225593.
31. Ciccarone D. The rise of illicit fentanyls, stimulants and the fourth wave of the opioid overdose crisis. Curr Opin Psychiatry **2021**; 34:344–50.
32. Goto M, Schweizer ML, Vaughan-Sarrazin MS, et al. Association of evidence-based care processes with mortality in *Staphylococcus aureus* bacteremia at Veterans Health Administration hospitals, 2003-2014. JAMA Intern Med **2017**; 177:1489–97.
33. Collins GS, Reitsma JB, Altman DG, Moons KG. Transparent Reporting of a Multivariable Prediction Model for Individual Prognosis or Diagnosis (TRIPOD): the TRIPOD statement. Ann Intern Med **2015**; 162:55–63.
34. Department of Veterans Affairs. ChartReview. **2016**. https://github.com/department-of-veterans-affairs/ChartReview. Accessed February 2, 2022.
35. Weems S, Heller P, Fenton SH. Results from the Veterans Health Administration *ICD*-10-CM/PCS coding pilot study. Perspect Health Inf Manag **2015**; 12:1b.
36. Chapman WW, Bridewell W, Hanbury P, Cooper GF, Buchanan BG. A simple algorithm for identifying negated findings and diseases in discharge summaries. J Biomed Inform **2001**; 34:301–10.
37. Johnson AEW, Pollard TJ, Shen L, et al. MIMIC-III, a freely accessible critical care database. Sci Data **2016**; 3:160035.
38. Strombotne KL, Legler A, Minegishi T, et al. Effect of a predictive analytics-targeted program in patients on opioids: a stepped-wedge cluster randomized controlled trial [manuscript published online ahead of print 2 May 2022]. J Gen Intern Med **2022**. doi:10.1007/s11606-022-07617-y
39. Biancarelli DL, Biello KB, Childs E, et al. Strategies used by people who inject drugs to avoid stigma in healthcare settings. Drug Alcohol Depend **2019**; 198: 80–6.
40. Harkema H, Dowling JN, Thornblade T, Chapman WW. Context: an algorithm for determining negation, experiencer, and temporal status from clinical reports. J Biomed Inform **2009**; 42:839–51.
41. Mehrabi S, Krishnan A, Sohn S, et al. DEEPEN: a negation detection system for clinical text incorporating dependency relation into NegEx. J Biomed Inform **2015**; 54:213–9.
42. Wu S, Miller T, Masanz J, et al. Negation's not solved: generalizability versus optimizability in clinical natural language processing. PLoS One **2014**; 9:e112774.