Journal of the American Medical Informatics Association, 26(3), 2019, 254–261
doi: 10.1093/jamia/ocy166
Advance Access Publication Date: 2 January 2019
Research and Applications

Research and Applications

Natural language processing and machine learning to identify alcohol misuse from the electronic health record in trauma patients: development and internal validation

Majid Afshar,[1,2,3] Andrew Phillips,[4] Niranjan Karnik,[5] Jeanne Mueller,[6] Daniel To,[1] Richard Gonzalez,[6] Ron Price,[2] Richard Cooper,[4] Cara Joyce,[2,4] and Dmitriy Dligach[2,3]

[1]Health Sciences Division, Burn and Shock Trauma Research Institute, Stritch School of Medicine, Loyola University, Maywood, Illinois, USA, [2]Health Sciences Division, Center for Health Outcomes and Informatics Research, Loyola University, Maywood, Illinois, USA, [3]Department of Public Health Sciences, Stritch School of Medicine, Loyola University, Maywood, Illinois, USA, [4]Department of Computer Science, Loyola University, Chicago, Illinois, USA, [5]Department of Psychiatry, Rush University Medical Center, Chicago, Illinois, USA, and [6]Department of Surgery, Loyola University Medical Center, Maywood, Illinois, USA

Corresponding Author: Majid Afshar, MD, MSCR, Center for Translational Research and Education, 2160 S First Avenue, Building 115, Room 447, Maywood, IL 60153, USA (majid.afshar@lumc.edu).

Received 2 August 2018; Revised 4 November 2018; Editorial Decision 15 November 2018; Accepted 17 November 2018

ABSTRACT

Objective: Alcohol misuse is present in over a quarter of trauma patients. Information in the clinical notes of the electronic health record of trauma patients may be used for phenotyping tasks with natural language processing (NLP) and supervised machine learning. The objective of this study is to train and validate an NLP classifier for identifying patients with alcohol misuse.

Materials and Methods: An observational cohort of 1422 adult patients admitted to a trauma center between April 2013 and November 2016. Linguistic processing of clinical notes was performed using the clinical Text Analysis and Knowledge Extraction System. The primary analysis was the binary classification of alcohol misuse. The Alcohol Use Disorders Identification Test served as the reference standard.

Results: The data corpus comprised 91 045 electronic health record notes and 16 091 features. In the final machine learning classifier, 16 features were selected from the first 24 hours of notes for identifying alcohol misuse. The classifier's performance in the validation cohort had an area under the receiver-operating characteristic curve of 0.78 (95% confidence interval [CI], 0.72 to 0.85). Sensitivity and specificity were at 56.0% (95% CI, 44.1% to 68.0%) and 88.9% (95% CI, 84.4% to 92.8%). The Hosmer-Lemeshow goodness-of-fit test demonstrates the classifier fits the data well ($P = .17$). A simpler rule-based keyword approach had a decrease in sensitivity when compared with the NLP classifier from 56.0% to 18.2%.

Conclusions: The NLP classifier has adequate predictive validity for identifying alcohol misuse in trauma centers. External validation is needed before its application to augment screening.

Key words: natural language processing, machine learning, artificial intelligence, phenotyping, alcohol misuse, trauma, cTAKES

INTRODUCTION

Alcohol misuse is an attributable cause for 1 in 10 deaths in the United States, and prevalence rates of misuse rose 9% between 2002 and 2012.[1,2] As many as 33% of patients with trauma encounters have alcohol misuse.[3] Screening, brief intervention, and referral to treatment (SBIRT) programs at trauma centers have been shown to reduce alcohol consumption and decrease injury recurrence by nearly 50%.[4–6] However, significant barriers exist to implementation of current screening methods.[7] Collection of data with self-report questionnaires requires building new forms and procedures into electronic health record (EHR) systems and hiring staff to implement and administer the tools. Despite advances in health technology, screening remains a resource-intensive process that imposes significant costs on a health system.[8]

Information in the clinical narrative collected on admission is a potentially rich source of data. Documentation of a social history including substance use is part of training and routine care by providers in clinical settings. Natural language processing (NLP) is a set of computational or rule-based methods for deriving meaning from human-generated texts. Machine learning algorithms can use the derived features from NLP to learn and predict,[9] and it has been successfully used in clinical practice and research.[10–12] In particular, the most powerful NLP methods rely on supervised learning, a type of machine learning that takes advantage of current reference standards to make predictions about unseen cases.[13] The role of NLP for case identification of alcohol misuse is in its infancy; to our knowledge, no peer-reviewed publications have yet examined NLP for this purpose.

An NLP classifier could potentially provide an automated and comprehensive approach for identification of patients with alcohol misuse and improve implementation fidelity in SBIRT programs. The goal of this phase of our research is to develop a tool leveraging NLP that may be used by SBIRT programs to identify patients with alcohol misuse. Using EHR data, we hypothesize that a NLP classifier using notes available in the first 24 hours of presentation to the emergency department (ED) will have adequate discrimination with an area under the receiver-operating characteristic (AUC ROC) curve above 0.70 for alcohol misuse and outperform a rule-based keyword approach.

MATERIALS AND METHODS

Patient selection and environment

We performed an observational cohort study of 1422 consecutive patients that were screened for alcohol misuse and at least 18 years of age who were admitted to a Level I Trauma Center between April 2013 and November 2016. Patients with a primary admission for trauma were evaluated, and patients admitted for nontrauma injuries were excluded. All admissions, injury characteristics, and dates of injury were verified by dedicated trauma registrar coders. As part of the American College of Surgeons Certification, an SBIRT program was in place since 2013 and the Alcohol Use Disorders Identification Test (AUDIT) was used to screen for alcohol misuse. Screening results and reasons for screen fails were maintained in a separate database by 2 full-time prevention nurses dedicated to the task. Additional clinical variables were extracted from the EHR by linkage of the trauma registry and AUDIT registry to our institute's clinical research database. Linkage could not be performed in 6.0% ($n = 163$) of patients, and they were removed from analysis (Figure 1).

Reference standard for alcohol misuse

The 10-item AUDIT is the screening questionnaire developed by the World Health Organization to identify alcohol consumption above the lower risk limits[14] and is currently 1 of the recommended screening tools for EDs and trauma centers.[15,16] AUDIT scores range between 0 and 40 and have been validated for[14,17] sex-specific cutpoints for alcohol misuse.[18] An AUDIT score above cutpoints of ≥ 5 and ≥ 8 for women and men, respectively, represent the lower risk limit for any alcohol misuse and cutpoints of ≥ 13 and ≥ 16 for severe misuse.[19]

Given the known phenomenon of underreporting on alcohol questionnaire data,[20,21] post hoc error analysis was performed in the validation cohort to examine possible misclassifications by the AUDIT questionnaire. In this process, chart review was performed to identify reasons for discordance of the NLP classifier with the AUDIT. An SBIRT-certified annotator (D.T.) supervised by a critical care physician with expertise in alcohol misuse (M.A.) performed chart reviews after a kappa score of >0.75 was achieved. The following criteria were used as an operational definition for defining alcohol misuse during the chart review process: (1) National Institute of Alcoholism and Alcohol Abuse definition for drinking limits captured in the notes by mention of frequency or quantity of alcoholic beverages[22]; (2) alcohol-related trauma injuries defined as arriving with levels of blood alcohol above the legal limit within 6 hours of trauma occurrence[23]; (3) symptoms of alcohol withdrawal; and (4) physician diagnosis for alcohol misuse.

Processing of clinical text and feature extraction

Linguistic processing of clinical notes was performed using the clinical Text Analysis and Knowledge Extraction System (cTAKES) (http://ctakes.apache.org).[24] The spans of Unified Medical Language System (UMLS)–named entity mentions (diseases, symptoms, anatomy, procedures) were identified. Each named entity mention was mapped to a UMLS concept unique identifier (CUI). For instance, the named entity mention for "alcohol abuse" is assigned C0085762 as its CUI. The named entity mention of "alcohol abuse" in the text of the note is mapped to a separate CUI than "history of alcohol abuse," which is C0221628. Each named entity mention is subsequently analyzed to determine its negation status (eg "no alcohol abuse"). This method of data processing mitigates lexical variations between providers. Additional UMLS semantic types were included to accommodate items relevant to the task at hand, such as food (T168) (eg, wine, beer, whiskey) utilizing the latest dictionary lookup module from Apache cTAKES. The full list of UMSL semantic types are shown in Supplementary Material S1, and the source code is available in Apache cTAKES SVN repository with associated documentation. A term-frequency, inverse document-frequency (tf-idf) transformation was used to weigh the CUIs into normalized values for machine learning classifiers. The TfidfVectorizer class from scikit-learn was used to compute the weights. https://scikit-learn.org/stable/modules/generated/sklearn.feature_extraction.text.TfidfVectorizer.html.

Baseline model using rule-based keyword approach

A handcrafted, rule-based algorithm with keywords was developed a priori based on content expertise (M.A., N.K.) for comparison to the NLP classifier. The algorithm used the following keywords to identify alcohol misuse: *alcohol dependence*, *alcohol abuse*, *alcohol withdrawal*, and *alcoholic*.

Analysis with supervised machine learning

Descriptive statistics for those with and without alcohol misuse were calculated. Differences in continuous variables were assessed for statistical significance using Wilcoxon rank sum tests or Kruskal-Wallis test, and the categorical variables were analyzed using chi-square tests. The primary analysis was the binary classification of alcohol misuse using clinical notes within 24 hours of admission to the ED. The sample was divided into 80% ($n = 1137$)

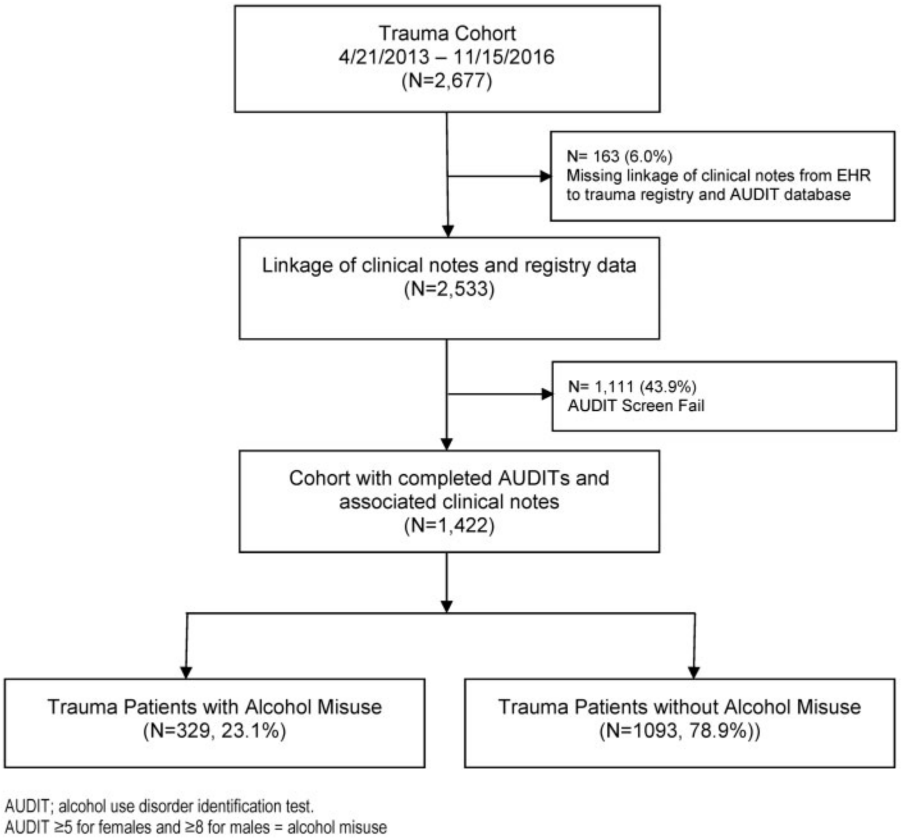

Trauma Cohort
4/21/2013 – 11/15/2016
(N=2,677)
N= 163 (6.0%)
Missing linkage of clinical notes from EHR to trauma registry and AUDIT database
Linkage of clinical notes and registry data
(N=2,533)
N= 1,111 (43.9%)
AUDIT Screen Fail
Cohort with completed AUDITs and associated clinical notes
(N=1,422)
Trauma Patients with Alcohol Misuse
(N=329, 23.1%)
Trauma Patients without Alcohol Misuse
(N=1093, 78.9%))
AUDIT; alcohol use disorder identification test.
AUDIT ≥5 for females and ≥8 for males = alcohol misuse

Figure 1. CONSORT diagram. Alcohol misuse was rated as Alcohol Use Disorders Identification Test (AUDIT) score ≥5 for women and ≥8 for men. EHR: electronic health record.

for training and 20% (n = 285) for internal validation for all machine learning classifiers.

CUIs were inputs to machine learning classifiers, and classifier hyperparameters were tuned to the highest AUC ROC curve using 10-fold cross-validation. The primary framework used was scikit-learn,[25] which provides a number of classifiers and other functionality to facilitate model creation and optimization. The primary classifiers from scikit-learn were LogisticRegression, PassiveAggressiveClassifier, support vector machine, and SGDClassifier. A grid search with 10-fold cross-validation was performed on the training dataset with examination of several classifiers and tuning performed within promising hyperparameter ranges. Then the AUC ROC curve scores were compared with find the best classifiers, and the hyperparameters further tuned. A tournament-style process was used to reduce the number of classifiers and hyperparameters through several iterations, until only a single best classifier and set of hyperparameters remained. This was used to create a final model for testing. During this process, word n-grams (sequence of adjacent words of length n) were also evaluated and noted to perform no better than CUIs so they were abandoned in favor of simpler and potentially more robust CUI-based model. Furthermore, examination of CUIs from all hospital notes versus the first 24-hours and the addition of expression-based algorithms for blood alcohol concentration were examined as well.

Discrimination of the model was evaluated with the AUC ROC curve and 95% confidence interval (CI). Model calibration was measured visually with calibration plots and formally tested with the Hosmer-Lemeshow goodness-of-fit test. Test characteristics including total accuracy, sensitivity, specificity, negative predictive value (NPV), and positive predictive value (PPV) were examined to compare between NLP classifiers and the rule-based keyword approach. Adding all available notes during hospitalization to the classifier, and adding expression-based algorithms to target blood alcohol concentration (BAC) values to the classifier were also examined for improved performance using the net reclassification improvement (NRI) and integrated discrimination improvement (IDI) measures.[25]

A learning curve was generated to investigate the effect of sample size on classifier performance as an approach to assess adequacy of statistical power. We demonstrated a peak effect on AUC ROC curve in a sample size approaching the 1200 patients used for training (Figure 2). Analysis was performed using Python version 3.6.5 and SAS version 9.4; https://www.python.org (SAS Institute, Cary, NC). The Institutional Review Board of Loyola University Chicago approved this study.

RESULTS

Patient and data characteristics

The data corpus comprised 91 045 EHR notes and 16 091 CUI features (including negation) from 1422 patients. The count decreased to 22 642 EHR notes and 11 813 CUI features using notes available in first 24 hours from patients presenting to the ED. In the cohort of patients who completed the AUDIT, 22.9% (n = 329) reported any level of misuse, and severe misuse was present in 28.0% (n = 92) of those with any level of alcohol misuse. Only 17.4% (n = 16) of the patients with severe alcohol misuse had a discharge diagnosis for alcohol dependence or abuse. Baseline characteristics and outcomes between alcohol misuse and nonmisuse patients are detailed in Table 1.

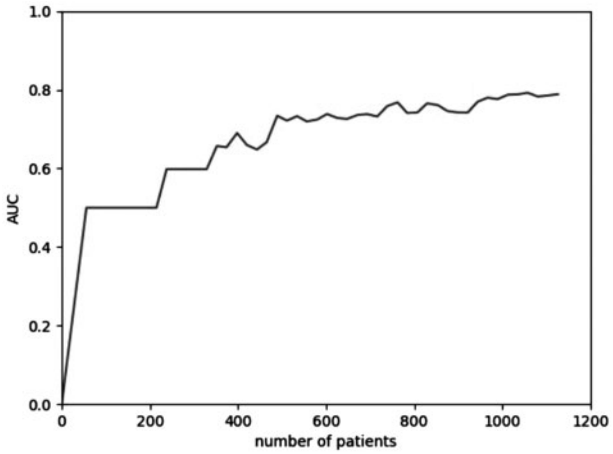


Figure 2. Learning curve demonstrating peak effect on area under the curve (AUC) in sample size up to 1137 patients used for the development cohort.

PATIENT CHARACTERISTICS IN THE COHORT WITHOUT SCREENING

Approximately 43.9% (n = 1111) of the 2533 trauma patients with clinical notes in our EHR did not receive alcohol screening with the AUDIT (Figure 1). The most common reasons reported by screeners for not performing the AUDIT were unavailability of staff (41.1%, n = 457) and patient's inability to communicate (19.6%, n = 217) (Table 2). BAC levels in those without an AUDIT score were higher than in those with an AUDIT score (197 mg/dL vs 157 mg/dL; $P < .01$). Patient characteristics and outcomes between those with and without AUDIT scores are described in Supplementary Material S2.

Development of NLP classifier

Using notes available in first 24 hours from patients presenting to the ED, the NLP classifier that produced the highest AUC ROC curve in the development cohort was a logistic regression model with least absolute errors loss function and regularization. From the list of 11 814 CUI features, the classifier retained 16 CUI features. The following were the top positive CUIs in logistic regression: thiamine, intoxication, neglect, drinking problems, drinking, liver imaging, sexually active, marijuana, and alcohol or drug abuse. The complete list of CUIs with their logistic regression coefficients are

Table 1. Demographics, alcohol information, and outcomes

Characteristic	Total (N = 1422)	No misuse (n = 1093)	Any misuse (n = 329)	*P* value
Age, y	44 (27–61)	46 (28–64)	38 (26–53)	<.001
Male	1007 (70.8)	748 (68.4)	259 (78.7)	<.001
White race	764 (53.7)	597 (54.6)	167 (50.8)	.22
Hispanic ethnicity	253 (17.8)	170 (15.6)	83 (25.3)	<.001
Tobacco use[a]	266 (18.7)	176 (16.1)	90 (27.4)	<.001
Alcohol dependence	23 (1.6)	7 (0.6)	16 (4.9)	<.001
Diabetes	94 (6.6)	79 (7.2)	15 (4.6)	.09
Hypertension	259 (18.2)	211 (19.3)	48 (14.6)	.05
Coronary heart disease	32 (2.3)	29 (2.5)	5 (1.5)	.31
Admission SBP <90 mm Hg	44 (3.1)	34 (3.1)	10 (3.0)	.95
Admission BAC (n = 343)	157 (83–229)	105 (60–176)	193 (135–252)	<.001
Urine toxicology positive (yes)	950 (66.8)	699 (64.0)	251 (76.3)	<.001
ISS	9 (5–14)	9 (5–14)	9 (5–14)	.57
Mechanism of Injury				
MVC/MCC	509 (35.8)	401 (36.7)	108 (32.8)	
Fall	385 (27.1)	303 (27.7)	82 (24.9)	
Assault	69 (4.9)	45 (4.1)	24 (7.3)	.07
GSW/stabbing	219 (15.4)	168 (15.4)	51 (15.5)	
Other[b]	240 (16.8)	176 (16.1)	64 (19.5)	
Mechanical ventilation	192 (13.5)	134 (12.3)	58 (17.6)	.01
ICU stay, d	2 (0–4)	2 (0–4)	2 (0–4)	.27
LOS, d	4.9 (2.3–9.4)	4.9 (2.4–8.9)	4.5 (2.3–10.3)	.88
Disposition[c]				
Home	938 (66.0)	708 (64.8)	230 (69.9)	
Acute care	218 (15.3)	169 (15.5)	49 (149)	.005
Chronic care	207 (14.6)	177 (16.2)	30 (9.1)	
Other	46 (3.2)	32 (2.9)	14 (4.3)	
In-hospital death	13 (0.9)	7 (0.6)	6 (1.8)	

Note: Values are presented as median (interquartile range) or n (%).

BAC: blood alcohol concentration; GSW: gunshot wound; ICU: intensive care unit; ISS: injury severity score; LOS: length of stay; MCC: motorcycle; MVC: motor vehicle collision; SBP: systolic blood pressure.

[a] Alcohol dependence, tobacco use, diabetes, hypertension, and coronary heart disease based off of International Classification of Diseases–Ninth/Tenth Revision codes.

[b] Other = AMA, Jail/Prison, Other; Positive Urine Toxicology = amphetamines, barbiturates, benzo, cannabis, cocaine, opiates, phencyclidine.

[c] Acute Care = Inpatient rehab, inpatient psych, short-term hospital; Chronic Care = SNF: Skilled Nursing Facility; LTAC: Long Term Acute Care.

Table 2. Reasons AUDIT not done (n = 1111)

Language/deaf/jaw wired shut/trach	153 (13.7)
ICU/vent	32 (2.9)
Pain/sleeping/OOR	32 (2.9)
Police hold	6 (0.5)
TBI/not oriented/agitated/psych	217 (19.6)
Patient declined AUDIT	56 (5.0)
Patient died	138 (12.4)
Staff unavailable (weekend admission, vacation, patient discharged before screen)	457 (41.1)
Other/AMA	20 (1.8)

Note: Values are presented as n (%).

AMA: against medical advice; AUDIT: Alcohol Use Disorders Identification Test; ICU: intensive care unit; OOR: out of room; TBI: traumatic brain injury.

Table 3. Features (n = 16) from natural language processing classifier using notes available in first 24 hours from patient presenting to the emergency department

Positive CUI features (β coefficients from logistic regression classifier)	c0039840 (THIAMINE), 9.99
	c0085762 (drinking problems), 6.17
	c0241028 (SEXUALLY ACTIVE), 1.89
	c0521874 (neglecting), 6.38
	c0684271 (drinking), 5.78
	c0728899 (INTOXICATION), 6.59
	c0740858 (alcohol or drug abuse), 1.10
	c0024808 (marijuana), 1.49
	c0034131 (Purified Protein Derivative of Tuberculin), 0.44
	c0034606 (isotope studies), 2.61
	c0012383 (2; 3 Dithiopropan 1 o1), 0.26
Negative CUI features (β coefficients from logistic regression classifier)	c0035345 (retired), −4.77
	c0015663 (fasted state), −1.15
	c0018681 (Cephalodynia), −1.98
	negated c0042963 (hanyas), −0.92
	negated c0234425 (Level of consciousness), −1.40

CUI: concept unique identifier.

listed in Table 3. The NLP classifier produced an average AUC ROC of 0.78 (95% CI, 0.68 to 0.89) across 10-fold cross-validation.

Discrimination and calibration in validation cohort

In the validation cohort of 285 patients, the NLP classifier had an AUC ROC curve of 0.78 (95% CI, 0.72 to 0.85). Discrimination of the NLP model is shown with the AUC ROC curve in Figure 3a, and Figure 3b is the corresponding calibration plot across 5 strata of predicted probabilities for alcohol misuse. The Hosmer-Lemeshow goodness-of-fit test demonstrated that the NLP classifier fit the data well across the 5 strata (*P* =.17). There is a linear trend for increasing levels of AUDIT score with increasing stratum for predicted probabilities (*P* <.001) (Figure 4).

Comparison with keyword approach and other NLP classifiers

Additional feature engineering was performed to extract BAC values embedded in the notes and added into the NLP classifier, but this more complex model neither increased the AUC ROC curve nor improved reclassification, as measured by a NRI of 0.13 (95% CI, −0.05 to 0.31; *P* =.15) and an IDI of 0.02 (95% CI, −0.02 to 0.05; *P* =.33). Furthermore, with a median length of stay of 4.9 days (interquartile range, 2.3–9.4 days), expanding the data corpus to include all notes during hospitalization also neither increased the AUC ROC curve nor improved the reclassification of the classifier, with a NRI of −0.12 (95% CI, −0.32 to 0.07; *P* =.20) and an IDI of 0.01 (95% CI, −0.03 to 0.05; *P* =.62). Test characteristics between the NLP classifier using the first 24 hours of notes and the more complex models with BAC data and additional notes are shown in Table 4. The simpler rule-based keyword approach had a decrease in sensitivity when compared with the NLP classifier from 56.0% to 18.2%.

Number needed to evaluate and error analysis

The workup to detection ratio of the NLP classifier (also known as number needed to evaluate) is 1.67. In other words, about 2 patients reaching alert threshold would need to be evaluated to detect 1 case of alcohol misuse. In the cohort of patients that missed screening with an AUDIT (n = 1111), the NLP classifier would have detected another 257 patients with alcohol misuse or 23.1% of the cohort.

In error analysis, chart review was performed in the validation cohort to identify reasons for discordance of the NLP classifier with the AUDIT. Forty-four (15.4%) of the cases and noncases for alcohol misuse were relabeled from the original AUDIT determination, with 55% (n = 24) of the relabeled cases due to underreporting on the AUDIT. In this scenario, the NLP classifier had minimal improvement in test characteristics with a sensitivity of 57.0% (95% CI, 46.4% to 69.2%), specificity of 89.4% (95% CI, 85.2% to 93.5%), PPV of 62.9% (95% CI, 50.9% to 74.9%), and NPV of 86.6% (95% CI, 82.1% to 91.0%).

DISCUSSION

In nearly 4 years of alcohol screening at a Level I trauma center, full-time dedicated screeners administered the AUDIT to about half of the patients. Failure to screen was attributable mainly to staffing or communication barriers. Using clinical documentation within the first 24 hours of the encounter, our NLP classifier demonstrated adequate discrimination and calibration as a tool to identify patients with alcohol misuse. Discrimination did not improve with a larger quantity of notes or inclusion of blood alcohol data. At a workup to detection ratio of approximately 2, the NLP classifier provides an automated approach to potentially overcome staffing and patient barriers for SBIRT programs at trauma centers.

The role of NLP in identifying alcohol misuse is in its infancy; herein, we show adequate discrimination for classifying cases and noncases. The NLP classifier performed better than a rule-based keyword approach that was wholly unsuitable for this purpose with a sensitivity below 20%. Physician diagnoses using claims data is not time sensitive and had similarly poor sensitivity, consistent with prior evidence.[26] Over a quarter of the patients have risk levels on the AUDIT that are indicative of an alcohol use disorder. Few false positives occurred from the NLP classifier, likely because many of the cases of alcohol misuse are severe and contain considerable information about unhealthy alcohol use in the notes.[27] Our results support these findings by showing better calibration in the group with higher predicted probabilities for alcohol misuse. Adding notes beyond the first 24 hours or adding blood alcohol data to the model proved unnecessary, with no improvement in the NRI or IDI. This suggests that additional provider notes add noise to the classifier

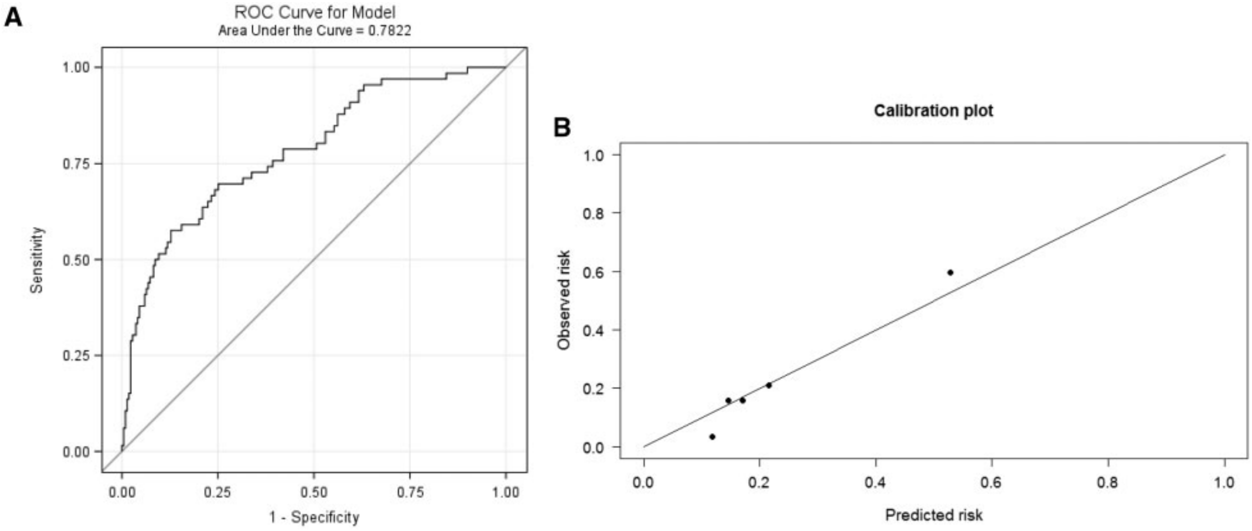
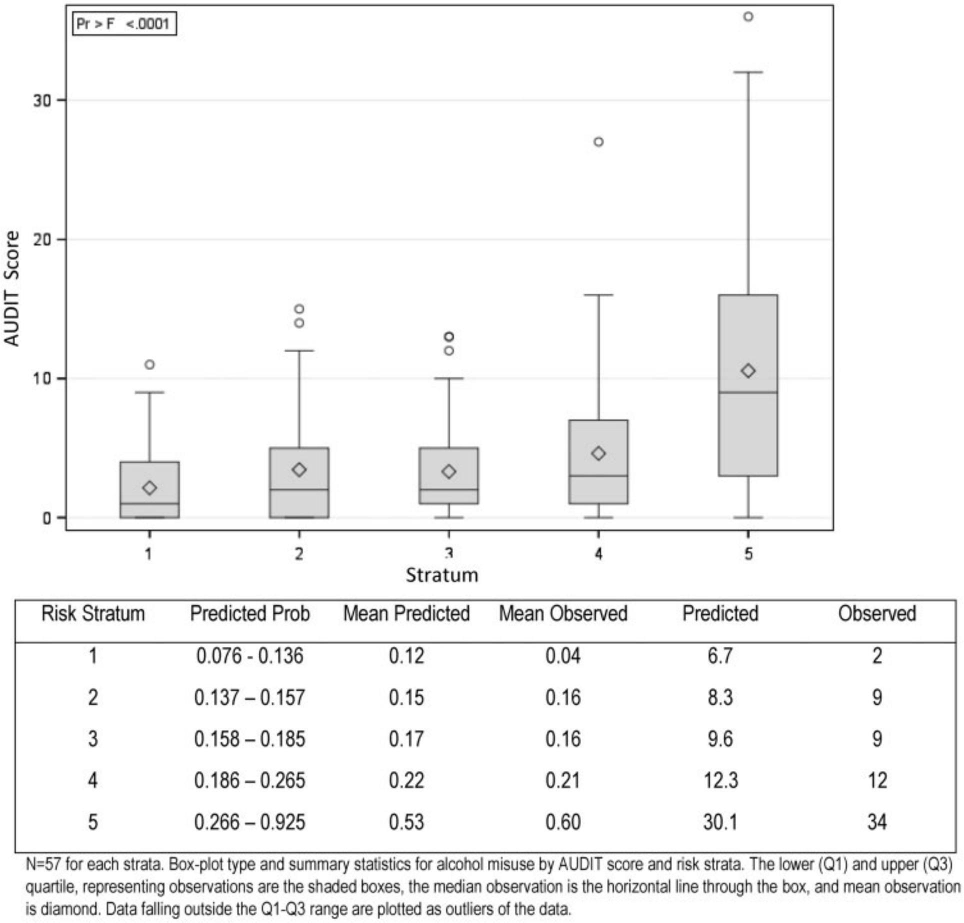


Figure 3. (A) Discrimination for alcohol misuse with receiver-operating characteristic (ROC) area under the curve. (B) Calibration plot across 5 strata of predicted probabilities with n = 57 in each strata.



Risk Stratum	Predicted Prob	Mean Predicted	Mean Observed	Predicted	Observed
1	0.076 - 0.136	0.12	0.04	6.7	2
2	0.137 – 0.157	0.15	0.16	8.3	9
3	0.158 – 0.185	0.17	0.16	9.6	9
4	0.186 – 0.265	0.22	0.21	12.3	12
5	0.266 – 0.925	0.53	0.60	30.1	34

N=57 for each strata. Box-plot type and summary statistics for alcohol misuse by AUDIT score and risk strata. The lower (Q1) and upper (Q3) quartile, representing observations are the shaded boxes, the median observation is the horizontal line through the box, and mean observation is diamond. Data falling outside the Q1-Q3 range are plotted as outliers of the data.

Figure 4. Comparison between Alcohol Use Disorders Identification Test (AUDIT) score across 5 strata of predicted probabilities. n = 57 for each stratum. Boxplot type and summary statistics for alcohol misuse by AUDIT score and risk strata. The lower and upper quartiles, representing observations are the shaded boxes, the median observation is the horizontal line through the box, and mean observation is diamond. Data falling outside the lower to upper quartile range are plotted as outliers of the data.

Table 4. Test characteristics of selected algorithms in the validation cohort

Model	AUC ROC	Sensitivity (%)	Specificity (%)	PPV (%)	NPV (%)
Rule-based keyword approach	NA	18.2 (8.9–27.5)	94.9 (91.2–97.4)	52.2 (31.8–72.6)	79.4 (74.5–84.3)
NLP classifier using notes in first 24 h of encounter	**0.78 (0.72–0.85)**	**56.0 (44.1–68.0)**	**88.9 (84.4–92.8)**	**60.0 (47.5–71.9)**	**87.0 (82.6–91.4)**
NLP classifier with notes and blood alcohol levels in first 24 h	0.76 (0.69–0.83)	54.6 (42.5–66.6)	85.8 (81.2–90.5)	53.7 (41.8–65.7)	86.2 (81.7–90.8)
NLP classifier using notes from entire hospital encounter	0.78 (0.72–0.85)	34.9 (23.3–46.3)	95.4 (92.7–98.2)	69.7 (54.0–85.4)	82.3 (78.3–87.6)

Note: The 95% confidence interval data are presented in parentheses. Keyword approach used the following terms to identify alcohol misuse: *alcohol dependence*, *alcohol abuse*, *alcohol withdrawal*, *alcoholic*. AUC: area under the curve; NA: not applicable; NLP: natural language processing; NPV: negative predictive value; PPV: positive predictive value; ROC: receiver-operating characteristic. Bold = most parsimonious model.

and the greatest utility derives from the admission notes that routinely contain a social history information.

Underreporting alcohol consumption is a known limitation of self-report tools like the AUDIT, and rates of upwards of 30% have been reported.[28,29] In a survey on patient attitudes, over 90% of respondents state they would give an honest answer about their alcohol misuse to their provider,[30] presumably motivated by the desire to insure that they receive proper care. During error analysis, the NLP classifier captured a handful of false negative cases, but the classifier also did not capture a handful of true positive cases leading to little change in sensitivity. Nevertheless, the high specificity of the classifier highlights an opportunity for data collected in clinical notes and processed with NLP to better target patients for interventions to reduce alcohol consumption.

The classifier features not only demonstrate adequate predictive validity but also represent a structured and interoperable approach that may be used by other centers. In addition, the classifier features support good face validity by representing concepts that are associated with alcohol use such as vitamin deficiencies, co-substance use, and neglect.[31–34] The final 16 CUIs selected by machine learning are from a list of over 10 000 CUIs and a data corpus of nearly 100 000 notes. Processing large amounts of data illustrates the strengths of NLP and machine learning to streamline a phenotyping task from unstructured data.

The open-access availability, scalability, and portability of modern clinical NLP engines such as cTAKES allows for an efficient and reproducible pipeline to convert complex and dense clinical free text from the EHR into a "bag-of-words" representation. Furthermore, mapping to structured ontologies from the National Library of Medicine provides a standardized approach to machine learning. This study highlights methods in NLP that have previously been shown to be effective,[10–12] but our application in the context of alcohol use for identifying patients in a hospitalized setting has not been previously described. In this regard, many trauma patients have not previously received care at the trauma center they arrive at, so fewer notes are available than for patients with established care and prior encounters. Herein, we show benefit in NLP and machine learning in the first 24 hours of notes from a single encounter to identify patients with alcohol misuse. Accordingly, our NLP classifier may augment human screeners and improve our center's current SBIRT program.

During routine care for trauma, the clinician at the bedside does not typically prioritize alcohol misuse into the treatment plan. The American College of Surgeons (ACS) addresses this gap in treatment with a recommendation to provide SBIRT.[35] However, despite implementation of an SBIRT program at our center, fidelity remains an issue with nearly half of the trauma encounters missing a screen. Nonevening or weekend staffing hours and patient's inability to communicate are common barriers for screening. The identification of an additional 250 patients by the NLP classifier from the cohort without AUDIT screens highlights its potential impact. Our NLP classifier leverages not only the provider's documentation but also proxy reports, embedded medication and laboratory data, and additional notes from other medical staff that are captured in the first 24 hours. Many of the unhealthy behaviors documented in the EHR may largely go unnoticed without the aid of additional clinical support tools such as the NLP classifier described in this study.

Several limitations are present in this study. First, this is a single-center study and the NLP classifier will need external validation and possibly require revisions at other institutions before application. Although we used concepts to account for lexical variation between providers, there may be concepts that are unique to our center. Second, patients may be discharged before processing of the first 24 hour of notes for the NLP classifier. These patients may require SBIRT postdischarge if their length of stay was under 24 hours. Processing of the notes and machine learning algorithms requires expertise that many centers may not currently have. Additional customizations or inquiries may be required to account for concepts that are not seemingly relevant. For instance, the CUI for "2; 3 Dithiopropan" was 1 of the features in the NLP classifier, but it represents an anti-gas warfare agent. This appears illogical but when examined more closely we noted that 1 of the qualifying synonyms for the CUI is BAL (British Anti-Lewsite), which is also an acronym for blood alcohol level and a frequent term in the notes. Last, we did not perform any misspelling normalization to account for typing errors that may have occurred.

CONCLUSION

The NLP classifier has adequate predictive validity for identifying alcohol misuse in the trauma setting. Trauma staffing models could be modified by using the NLP classifier as a predictive enrichment strategy so that targeted groups of patients receive SBIRT by fewer screeners across more shifts. Application of our classifier outside our health system at other trauma centers in the United States may provide external validation for a standardized and comprehensive approach to augment screening for alcohol misuse.

FUNDING

This research was supported in part by the National Institute of Alcoholism and Alcohol Abuse grant number K23AA024503 (M.A.), National Institute on Drug Abuse grant numbers R01DA04171 and UL1TR002398, and an intramural award from Loyola's Center for Health Outcomes and Informatics Research.

CONTRIBUTORS

Drs. Afshar and Dligach are the guarantors of the manuscript. Concept and design: M. Afshar, D. Dligach, C. Joyce, A. Phillips, R. Gonzalez, N. Karnik. Acquisition, analysis, or interpretation of data: M. Afshar, A. Phillips, C. Joyce, Cooper, D. Dligach, D. To. Final approval of the article: M. Afshar, A. Phillips, R. Cooper, C. Joyce, D. Dligach, R. Gonzalez, J. Mueller, N. Karnik, R. Price, D. To. Administrative, technical, or logistic support: M. Afshar, A. Phillips, R. Price, D. Dligach, C. Joyce.

SUPPLEMENTARY MATERIAL

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

Conflict of interest statement. None declared.

REFERENCES

1. Stahre M, Roeber J, Kanny D, Brewer RD, Zhang X. Contribution of excessive alcohol consumption to deaths and years of potential life lost in the United States. *Prev Chronic Dis* 2014; 11: E109.
2. Dawson DA, Goldstein RB, Saha TD, Grant BF. Changes in alcohol consumption: United States, 2001–2002 to 2012–2013. *Drug Alcohol Depend* 2015; 148: 56–61.
3. Afshar M, Netzer G, Murthi S, Smith GS. Alcohol exposure, injury, and death in trauma patients. *J Trauma Acute Care Surg* 2015; 79 (4): 643–8.
4. Field C, Walters S, Marti CN, Jun J, Foreman M, Brown C. A multisite randomized controlled trial of brief intervention to reduce drinking in the trauma care setting: how brief is brief? *Ann Surg* 2014; 259 (5): 873–80.
5. Zatzick D, Donovan DM, Jurkovich G, *et al*. Disseminating alcohol screening and brief intervention at trauma centers: a policy-relevant cluster randomized effectiveness trial. *Addiction* 2014; 109 (5): 754–65.
6. Gentilello LM, Rivara FP, Donovan DM, *et al*. Alcohol interventions in a trauma center as a means of reducing the risk of injury recurrence. *Ann Surg* 1999; 230 (4): 473–80. Discussion 480–473.
7. Marjoua Y, Bozic KJ. Brief history of quality movement in US healthcare. *Curr Rev Musculoskelet Med* 2012; 5 (4): 265–73.
8. Barbosa C, Cowell AJ, Landwehr J, Dowd W, Bray JW. Cost of screening, brief intervention, and referral to treatment in health care settings. *J Subst Abuse Treat* 2016; 60: 54–61.
9. Demner-Fushman D, Chapman WW, McDonald CJ. What can natural language processing do for clinical decision support? *J Biomed Inform* 2009; 42 (5): 760–72.
10. Jones BE, South BR, Shao Y, *et al*. Development and validation of a natural language processing tool to identify patients treated for pneumonia across VA emergency departments. *Appl Clin Inform* 2018; 9 (1): 122–8.
11. Castro VM, Dligach D, Finan S, *et al*. Large-scale identification of patients with cerebral aneurysms using natural language processing. *Neurology* 2017; 88 (2): 164–8.
12. 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 (12): 1057–64.
13. Gonzalez-Hernandez G, Sarker A, O'Connor K, Savova G. Capturing the patient's perspective: a review of advances in natural language processing of health-related text. *Yearb Med Inform* 2017; 26 (1): 214–27.
14. Saunders JB, Aasland OG, Babor TF, de la Fuente JR, Grant M. Development of the alcohol use disorders identification test (AUDIT): WHO collaborative project on early detection of persons with harmful alcohol consumption–II. *Addiction* 1993; 88 (6): 791–804.
15. MacKenzie D, Langa A, Brown TM. Identifying hazardous or harmful alcohol use in medical admissions: a comparison of audit, cage and brief mast. *Alcohol Alcohol* 1996; 31 (6): 591–9.
16. Donovan DM, Dunn CW, Rivara FP, Jurkovich GJ, Ries RR, Gentilello LM. Comparison of trauma center patient self-reports and proxy reports on the Alcohol Use Identification Test (AUDIT). *J Trauma* 2004; 56 (4): 873–82.
17. Reinert DF, Allen JP. The alcohol use disorders identification test: an update of research findings. *Alcohol Clin Exp Res* 2007; 31 (2): 185–99.
18. Neumann T, Neuner B, Gentilello LM, *et al*. Gender differences in the performance of a computerized version of the alcohol use disorders identification test in subcritically injured patients who are admitted to the emergency department. *Alcohol Clin Exp Res* 2004; 28 (11): 1693–701.
19. Reinert DF, Allen JP. The Alcohol Use Disorders Identification Test (AUDIT): a review of recent research. *Alcohol Clin Exp Res* 2002; 26 (2): 272–9.
20. Bajunirwe F, Haberer JE, Boum Y 2nd, *et al*. Comparison of self-reported alcohol consumption to phosphatidylethanol measurement among HIV-infected patients initiating antiretroviral treatment in southwestern Uganda. *PLoS One* 2014; 9 (12): e113152.
21. Boniface S, Kneale J, Shelton N. Drinking pattern is more strongly associated with under-reporting of alcohol consumption than socio-demographic factors: evidence from a mixed-methods study. *BMC Public Health* 2014; 14: 1297.
22. National Institute on Alcohol Abuse and Alcoholism. *NIAAA Council Approves Definition of Binge Drinking*. NIAAA Newsletter, No. 3, Winter 2004. Available at: . Accessed July 19, 2018.
23. Cherpitel CJ, Ye Y, Bond J, Borges G, Monteiro M. Relative risk of injury from acute alcohol consumption: modeling the dose-response relationship in emergency department data from 18 countries. *Addiction* 2015; 110 (2): 279–88.
24. 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.
25. Pedregosa F, Varoquanox G, Gramfort A, *et al*. Scikit-learn: marchine learning in Python. *J Mach Learn Res* 2011; 12: 2825–30.
26. Steyerberg EW, Vickers AJ, Cook NR, *et al*. Assessing the performance of prediction models: a framework for traditional and novel measures. *Epidemiology* 2010; 21 (1): 128–38.
27. Boscarino JA, Moorman AC, Rupp LB, *et al*. Comparison of ICD-9 codes for depression and alcohol misuse to survey instruments suggests these codes should be used with caution. *Dig Dis Sci* 2017; 62 (10): 2704–12.
28. Korcha RA, Cherpitel CJ, Moskalewicz J, Swiatkiewicz G, Bond J, Ye Y. Readiness to change, drinking, and negative consequences among Polish SBIRT patients. *Addict Behav* 2012; 37 (3): 287–92.
29. Miller PM, Thomas SE, Mallin R. Patient attitudes towards self-report and biomarker alcohol screening by primary care physicians. *Alcohol Alcohol* 2006; 41 (3): 306–10.
30. Day E, Bentham P, Callaghan R, Kuruvilla T, George S. Thiamine for Wernicke-Korsakoff Syndrome in people at risk from alcohol abuse. *Cochrane Database Syst Rev* 2004; (1): CD004033.
31. Smothers BA, Yahr HT. Alcohol use disorder and illicit drug use in admissions to general hospitals in the United States. *Am J Addict* 2005; 14 (3): 256–67.
32. Grant BF, Saha TD, Ruan WJ, *et al*. Epidemiology of DSM-5 drug use disorder: results from the national epidemiologic survey on alcohol and related conditions-III. *JAMA Psychiatry* 2016; 73 (1): 39–47.
33. Cheng T, Johnston C, Kerr T, Nguyen P, Wood E, DeBeck K. Substance use patterns and unprotected sex among street-involved youth in a Canadian setting: a prospective cohort study. *BMC Public Health* 2016; 16 (4): 1–7.
34. Doran KM, Rahai N, McCormack RP, *et al*. Substance use and homelessness among emergency department patients. *Drug Alcohol Depend* 2018; 188: 328–33.
35. Bonevski B, Regan T, Paul C, Baker AL, Bisquera A. Associations between alcohol, smoking, socioeconomic status and comorbidities: evidence from the 45 and up Study. *Drug Alcohol Rev* 2014; 33 (2): 169–76.
36. Treatment Improvement Protocol (TIP) Series 16. *Alcohol and other drug screening of hospitalized trauma patients*. DHHS Publication No. (SMA) 95–3041. Rockville, MD: U.S. Department of Health and Human Services, Substance Abuse and Mental Health Services Administration, 1995.