Received: 4 May 2021 | Accepted: 13 October 2021

DOI: 10.1111/add.15730

RESEARCH REPORT

ADDICTION SSA

External validation of a machine learning classifier to identify unhealthy alcohol use in hospitalized patients

Yiqi Lin[1] | Brihat Sharma[2] | Hale M. Thompson[2] | Randy Boley[2] | Kathryn Perticone[3] | Neeraj Chhabra[4,5] | Majid Afshar[6] | Niranjan S. Karnik[1,2]

[1]Rush Medical College, Rush University, Chicago, IL, USA

[2]Department of Psychiatry and Behavioral Sciences, Rush Medical College, Rush University, Chicago, IL, USA

[3]Addiction Medicine, Cooper University Health Care, Camden, NJ, USA

[4]Department of Emergency Medicine, Rush Medical College, Rush University, Chicago, IL, USA

[5]Department of Emergency Medicine, John. H. Stroger, Jr. Hospital of Cook County, Chicago, IL, USA

[6]Division of Allergy, Pulmonary and Critical Care Medicine, Department of Medicine, School of Medicine and Public Health, University of Wisconsin, Madison, WI, USA

Correspondence
Niranjan S. Karnik, Department of Psychiatry and Behavioral Sciences, Rush University Medical Center, 1645 West Jackson Blvd., Suite 600, Chicago, IL, 60612, USA.
Email: niranjan_karnik@rush.edu

Funding information
Agency for Healthcare Research and Quality, Grant/Award Number: K12-HS026385; National Center for Advancing Translational Sciences, Grant/Award Numbers: KL2-TR002387, UL1-TR002389; National Institute on Alcohol Abuse and Alcoholism, Grant/Award Number: K23-AA024503; National Institute on Drug Abuse, Grant/Award Numbers: R01-DA041071, R01-DA051464, UG1-DA049467

Abstract

Background and Aims: Unhealthy alcohol use (UAU) is one of the leading causes of global morbidity. A machine learning approach to alcohol screening could accelerate best practices when integrated into electronic health record (EHR) systems. This study aimed to validate externally a natural language processing (NLP) classifier developed at an independent medical center.

Design: Retrospective cohort study.

Setting: The site for validation was a midwestern United States tertiary-care, urban medical center that has an inpatient structured universal screening model for unhealthy substance use and an active addiction consult service.

Participants/Cases: Unplanned admissions of adult patients between October 23, 2017 and December 31, 2019, with EHR documentation of manual alcohol screening were included in the cohort (n = 57 605).

Measurements: The Alcohol Use Disorders Identification Test (AUDIT) served as the reference standard. AUDIT scores ≥5 for females and ≥8 for males served as cases for UAU. To examine error in manual screening or under-reporting, a *post hoc* error analysis was conducted, reviewing discordance between the NLP classifier and AUDIT-derived reference. All clinical notes excluding the manual screening and AUDIT documentation from the EHR were included in the NLP analysis.

Findings: Using clinical notes from the first 24 hours of each encounter, the NLP classifier demonstrated an area under the receiver operating characteristic curve (AUCROC) and precision-recall area under the curve (PRAUC) of 0.91 (95% CI = 0.89–0.92) and 0.56 (95% CI = 0.53–0.60), respectively. At the optimal cut point of 0.5, sensitivity, specificity, positive predictive value (PPV), and negative predictive value (NPV) were 0.66 (95% CI = 0.62–0.69), 0.98 (95% CI = 0.98–0.98), 0.35 (95% CI = 0.33–0.38), and 1.0 (95% CI = 1.0–1.0), respectively.

Conclusions: External validation of a publicly available alcohol misuse classifier demonstrates adequate sensitivity and specificity for routine clinical use as an automated screening tool for identifying at-risk patients.

KEYWORDS
Addiction consultation service, data science, inpatient screening, machine learning, natural language processing, unhealthy alcohol use

INTRODUCTION

Globally, alcohol misuse is one of the most significant causes of morbidity and mortality with 3 million deaths attributed to alcohol each year and over 5% of the global disease burden attributed to this substance [1]. In the United States (US), an estimated 65.8 million people report binge drinking in the past month, and 24% of them also qualify as heavy drinkers with ≥ 5 days of binge drinking in a month [2]. Unhealthy alcohol use (UAU) contributes to ~95 000 deaths in the United States annually. Additionally, societal expenditure related to lost work productivity, healthcare, criminal justice, and motor vehicle crash amounted to $249 billion in 2010, of which 77% was attributable to binge drinking [3]. Recognizing that timely screening and appropriate intervention can improve UAU outcomes, regulatory entities such as the US Preventive Services Task Force (USPSTF) have recommended screening and brief counseling in primary care settings [4]. Given that the prevalence of UAU is even greater in the inpatient setting when compared to outpatient [5, 6], screening and intervention should be of high priority for hospital systems. However, implementation of screening protocols remains a challenge across various healthcare settings—common barriers include workload management, limited resources and support, non-standardized delivery [7], and patient under-reporting [8].

In lieu of manual collection methods for screening, a comparable source of information is the clinical narrative portion of the Electronic Health Record (EHR). Key information regarding alcohol use can be found in the social history and is routinely documented by providers [9]. An automated process that is capable of analyzing clinical notes from the EHR can help predict UAU across the patient cohort for an individual healthcare system, solving problems of inconsistent patient screening and under-reporting on self-report surveys. Natural language processing (NLP) with supervised machine learning is a promising tool that can be used to analyze unstructured data in the EHR. Through NLP methods, meaning can be generated from human created texts. Further application of machine learning algorithms can refine and improve predictions given preset parameters [9].

Although the use of NLP to extract clinical information has proven successful in various applications [10–12], use of NLP and supervised machine learning toward identification of UAU is relatively unexplored. Development of digital classifiers could produce several advantages in hospital-based screening programs. First, most hospitals lack sufficient staffing to screen patients using validated tools on a continuous 24-hour cycle. Second, a robust universal approach using digital technology has the potential to identify patients who might not report use on traditional surveys. Finally, in times of organization or clinical crisis (such as COVID-19), optional screenings are often suspended or dropped to save time and reduce staff burden.

We previously developed an NLP classifier for UAU that was trained against trauma patients who received the Alcohol Use Disorders Identification Test (AUDIT) [9]. The team subsequently internally validated the NLP classifier in a separate cohort of hospitalized patients at the same health system with a sensitivity and specificity above 90% [13]. Despite these promising findings there is a need to test the classifier in other contexts, to ensure that the classifier works on generalizable phenomena or experiences of alcohol use. Testing in a secondary health system is the first step toward a potential multi-site study with greater heterogeneity. We expect to find differences in the performance of the classifier based on the differing prevalence of UAU and demographics of the location populations.

The aim of our study was to examine the classifier's performance in a separate health system and provide external validation. The classifier was tested in an external validation cohort at Rush University Medical Center (RUMC), a tertiary-care academic health center that also services Chicago's West Side and the greater Chicago population. RUMC was deemed an ideal study site for external validation given its implementation in 2017 of standardized manual screening for UAU, including the 10-question AUDIT, which served as the reference standard [14] for this study. We hypothesized that the NLP classifier would achieve sensitivity and specificity greater than 80% in external validation at RUMC.

METHODS

Patient setting

In 2017, the Substance Use Intervention Team (SUIT) standardized a two-question alcohol and drug prescreen for all adult inpatients (age ≥ 18 years) across RUMC's 18 medical and surgical units. The Epic-based (Epic Systems Corporation) EHR workflow prompted nurses and social workers to ask eligible patients about their substance use in the past year, while automatically excluding those surveyed in the past 12 months [14, 15]. Any alcohol use on the prescreen triggered a follow-up evaluation with the AUDIT, a validated 10-item screening tool for alcohol misuse [16]. Patients are given progressively higher levels of care depending on the AUDIT zone [17] that they fall into and as reported elsewhere [14]. Adult inpatients of unplanned admissions between October 23, 2017 and December 31, 2019, with EHR documentation of the universal alcohol screen and AUDIT score if indicated, were included in the external validation cohort (Supporting information Figure S1).

Reference standard for UAU and *post hoc* error analysis

The AUDIT served as our reference standard for determining UAU. AUDIT scores ≥ 5 for females and ≥ 8 for males surpassed the lower-risk limit for any alcohol misuse and were labelled as cases for UAU [18]. To examine error in manual screening or patient under-reporting, *post hoc* error analysis was conducted to review discordance between the NLP classifier labels and the AUDIT-derived reference labels. The annotator (Y.L.) manually reviewed encounter-specific notes to identify possible reasons for discordance. The annotator was trained by a clinically certified nurse practitioner and a board certified psychiatrist, who are both specialists in addiction care and research (K.M.B. and N.S.K.). The inter-observer agreement reached Cohen's κ coefficient of 0.80, the minimum required with the trainer before independent review.

A Delphi process among content and research experts was applied to arrive at a Likert scale for determining likelihood of UAU during *post hoc* chart review. The following elements were examined in the EHR: (i) clinical notes that encompass the National Institute of Alcoholism and Alcohol Abuse (NIAAA) definition for drinking limits (e.g. "I drink a pint of whiskey a day" counted as 8.5 standard drinks per day and therefore, problematic behavior); (ii) laboratory values for blood alcohol content (BAC) at or above the legal limit of 80 mg/dL at earliest encounter screen; (iii) physical exam findings or nursing flowsheets on alcohol withdrawal symptoms per the Clinical Institute Withdrawal Assessment for Alcohol (CIWA) [19]; (iv) physician diagnoses and problem lists of alcohol-related injury and/or UAU; and (v) past history of unhealthy substance use and/or family history of UAU. Each patient was graded on the Likert scale for UAU as "definitely," "highly probable," "probable," or "definitely not" along with a summary of supporting evidence. Patients were categorized as "definitely" if they had at least one of the following: documented alcohol consumption quantity and frequency meeting the NIAAA misuse limits, BAC ≥ 80 mg/dL at admission, or current encounter diagnosis of UAU. Patients were designated as "highly probable" if they had past diagnosis of UAU and concurrent finding of one or more of the following elements: admission because of an alcohol-related injury, CIWA ≥ 8 at any point during the encounter, history of unhealthy substance use, or family history of UAU. Patients with documented past diagnosis of UAU and no other risk factors were graded as "probable." Those with documented alcohol consumption patterns that did not meet NIAAA limits were classified as "definitely not." Patients identified as "definitely" "highly probable," or "probable" were collectively categorized as cases of UAU. Conversely, patients labeled as "definitely not" were considered as non-cases without UAU.

Processing of clinical notes fed into NLP classifier

In the external validation cohort, all the clinical notes from the encounters of cases and non-cases were extracted and processed through the clinical Text Analysis and Knowledge Extraction System (cTAKES; http://ctakes.apache.org) NLP engine for analytics. In addition to cleaning and processing the notes, cTAKES was also applied to recognize words or phrases from the clinical notes as medical terms (named entity recognition) and map them to the Unified Medical Language System (UMLS) Metathesaurus [20] named entities (diseases, symptoms, anatomy, and procedures) as coded data. The coded data are concept unique identifiers (CUIs) and serve as inputs into machine learning models. Words or phrases constituting synonymous concepts were mapped to the same CUI. For example, "alcohol abuse," "ethanol abuse," and "problem drinking" were all grouped to the same CUI for alcohol abuse, C0085762. On the other hand, negated concepts (e.g. "no alcohol abuse") and meaningfully distinct concepts with overlapping vocabulary (e.g. "history of alcohol abuse") were separately mapped to different CUIs. CUIs may be directly or indirectly related to UAU. In some instances, the connections between the CUIs and its role to identify cases of UAU may be unclear because the vocabulary of concepts is over 54 000 unique CUIs. Before analysis, the CUIs were normalized through a term-frequency, inverse document-frequency transformation to account for commonly appearing terminology across all notes.

The NLP classifier previously developed by Afshar *et al.* (publicly accessible via: https://github.com/brihat9135/AlcoholNLP_Classifier) is a logistic regression classifier trained using CUIs to predict cases of UAU. The normalized CUIs were fed into an elastic net logistic regression classifier [13].

Analysis plan

Patient characteristics from the reference dataset of manually screened patients were conducted using the χ^2 test for proportions and Wilcoxon-Mann Whitney nonparametric tests for continuous variables between UAU and no unhealthy use groups. The International Classification of Disease (ICD) codes were grouped into disease categories based on the Elixhauser comorbidity disease classification [21]. Missing data analysis was performed to compare the manually screened hospitalizations to the hospitalized group as a whole. Data on the screened population as compared to the inpatient population is provided in Supporting information Table S1. Hypotheses were not pre-registered, and findings should be considered exploratory.

The primary outcome to examine the discrimination of the NLP classifier on the external validation cohort was measured by the area under the receiver operating characteristic curve (AUROC). The precision-recall area under the curve (PRAUC) was also generated to better examine performance in cohorts with a very low prevalence of cases [22]. A range of cut points from the AUROC curve were examined, including Youden indices, to identify the optimal cut point for screening [23]. Calibration was examined visually with calibration plots (Supporting information Figure S2). The screen characteristics were reported as sensitivity, specificity, negative predictive value (NPV), and positive predictive value (PPV) along with their 95% CI.

The median length of stay for hospitalized adult patients at RUMC between 2017 and 2019 was 3 days (interquartile range 2–6 days); therefore, we also ran the model on the first 24-hour of EHR notes, instead of the entire encounter, for a more pragmatic application that would allow screening to happen, but still provide ample time for screening, brief intervention, and referral to treatment (SBIRT). For external validation, the transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD) was followed [24, 25] (Supporting information Table S2). Analysis was completed using Python Version 3.6.5 (Python Software Foundation), RStudio Version 1.1.463 (RStudio Team), and R Version 4.0.3 (R Core Team). This study was approved by the Institutional Review Board as human subjects research and informed consent was waived for the validation of retrospective data and the manual chart review.

TABLE 1 Baseline patient characteristics and outcomes (n = 53 650)

Characteristics and outcomes	*Unhealthy alcohol use (n = 900)*	*No UAU (n = 52 750)*	*P value*
Age, median (IQR)	49 (39–59)	61 (45–71)	<0.001
Male sex, *n* (%)	624 (69.3)	21 329 (41.2)	<0.001
Race/ethnicity, *n* (%)			
Non-Hispanic White	376 (41.8)	22 415 (42.4)	<0.001
Non-Hispanic Black	302 (33.6)	17 545 (33.2)	
Hispanic White	56 (6.2)	2862 (5.4)	
Hispanic Black	1 (<1)	137 (<1)	
Other	165 (18.3)	9791 (18.5)	
AUDIT score (Mean score, IQR, *n* = 16 479)	20 (13–28)	1 (0–3)	<0.001
Lower n is because of the pre-screen negatives who do not get a full AUDIT			
Insurance, *n* (%)			
Medicare	126 (14.0)	20 082 (38.1)	
Medicaid	528 (58.6.2)	17 967 (34.1)	
Private	237 (26.3)	14 151 (26.8)	< 0.001
Other	9 (1)	550 (1)	
Elixhauser comorbidities, *n* (%)			
Hypertension, uncomplicated	356 (39.6)	17 180 (32.5)	<0.001
Renal failure	85 (9.4)	11 185 (21.2)	<0.001
Neurological disorders	193 (21.4)	8386 (15.8)	<0.001
Congestive heart failure	115 (12.8)	9750 (18.5)	<0.001
Diabetes mellitus, complicated	79 (8.7)	11 400 (21.6)	<0.001
Liver disease	306 (34.0)	3479 (6.6)	<0.001
Chronic lung disease	197 (21.9)	10 707 (20.3)	0.257
Diabetes mellitus, uncomplicated	66 (7.3)	3665 (6.9)	0.701
Psychoses	107 (11.8)	2139 (4.1)	<0.001
Depression	255 (28.3)	7920 (15.0)	<0.001
Hypertension, complicated	160 (17.7)	15 039 (28.5)	<0.001
Alcohol abuse	743 (82.5)	1073 (2.0)	<0.001
Drug abuse	210 (23.3)	1745 (3.3)	<0.001
AIDS/HIV	26 (2.9)	413 (<1)	<0.001
Discharge disposition, *n* (%)			
Home	572 (63.5)	30 600 (58.0)	<0.001
In-hospital death	9 (1.0)	581 (1.1)	
Long or shorter time care	126 (14.0)	7135 (13.5)	
Against medical advice	32 (3.5)	427 (<1)	
Other	161 (17.9)	14 007 (26.5)	

RESULTS

Patient and data characteristics

Within the study time frame, 82 881 unplanned adult inpatient admissions were identified and 69.5% (n = 57 605) contained EHR documentation of universal screen and/or AUDIT data, serving as the external validation cohort. Patient demographics in the external validation cohort were similar to that of the development cohort (which has been previously reported) [13], except for lower prevalence rate of UAU (Supporting information Table S3). In the validation cohort, the case rate of UAU was 1.68% (n = 900), as identified by qualifying AUDIT scores (Table 1). Compared to patients with no UAU (n = 52 750), the unhealthy use group was younger in median age and predominantly male (P < 0.001). Proportions for hypertension, fluid and electrolyte disorders, weight loss, liver disease, psychiatric conditions, depression, and coagulopathy were greater in the unhealthy use group (P < 0.001). The unhealthy use group also had a larger proportion with Medicaid insurance coverage (P < 0.001).

Alcohol misuse classifier performance

Encounter level data

The total number of clinical notes were 2 469 252 with 67 986 unique CUIs. Using clinical notes from the entire patient encounter, the NLP classifier demonstrated an AUCROC and PRAUC of 0.95 (95% CI = 0.94–0.96) and 0.68 (95% CI = 0.65–0.71), respectively. At the optimal cut point, the sensitivity, specificity, PPV, and NPV was 0.76 (95% CI = 0.73–0.78), 0.98 (95% CI = 0.98–0.98), 0.46 (95% CI = 0.44–0.49), and 0.99 (95% CI = 0.99–0.99), respectively (Table 2).

Using clinical notes from the first 24 hours of each encounter, the number of clinical notes were 644 710 and the number of unique CUIs were 54 799. The NLP classifier demonstrated an AUCROC and PRAUC of 0.91 (95% CI = 0.89–0.92) and 0.56 (95% CI = 0.53–0.60), respectively. At the optimal cut point of 0.5, sensitivity, specificity, PPV, and NPV was 0.66 (95% CI = 0.62–0.69), 0.98 (95% CI = 0.98–0.98), 0.35 (95% CI = 0.33–0.38), and 0.99 (95% CI = 0.99–0.99), respectively (Supporting information Table S4).

Error analysis of the NLP classifier (Table 3) identified misclassifications where the classifier's prediction contradicted the reference AUDIT labels in 1.88% (n = 1011) of total cases. False positives occurred in 1.48% (n = 792) of total cases, and false negatives occurred in 0.41% (n = 219). During *post hoc* chart review, 73.6% (n = 583) of false positives were deemed to be true positives, whereas 2.3% (n = 5) of false negatives were deemed to be true negatives ("definitely not"). In chart-reviewed cases designated as positive for UAU ("probable" or above), 67.3% (536/797) were male and 65.0% (518/797) had a history of tobacco use. Comparatively, those negative for UAU were 40.7% (87/214) male and 50.9% (109/214) had a history of tobacco use. When stratified by increasing levels of estimated misuse risk (from "definitely not" to "definitely"), there was a trending increase in the number of drinking activities per week (1.38 ± 1.44, 3.43 ± 2.75, 4.31 ± 2.74, 5.34 ± 2.28, P < 0.001); each level had significant missing data points, highest (87.9%) in the "definitely not" group and lowest (20.2%) in the "definitely" group. A similar trend was observed for the number of drinks in one sitting (2.03 ± 0.92, 5.31 ± 7.62, 4.38 ± 3.91, 6.63 ± 4.47, P < 0.001); again, many data points were missing at each level, highest (91.6%) in the "definitely not" group and lowest (24.3%) in the "definitely" group. Cases with evidence of withdrawal symptoms or admission because of alcohol-related injuries had higher levels of predicted probabilities for unhealthy use (Table 3).

DISCUSSION

Prevalence of UAU in our study population was 1.6%. This relatively low prevalence could be the result of our sample having an older average age. Because younger people tend to have the highest levels of UAU, this pattern would reduce our prevalence. Our study population had lower prevalence of alcohol misuse when compared to the national average of 6.3% with past month heavy alcohol use in the 18 or older population [26]. Much of the prevalence difference likely reflects the context of our manual screening program. The prescreen questions are asked at admission to the hospital. This is a point in time when patients are being admitted for a medical or surgical reason that generally does not link with UAU. In this situation, we would anticipate a high degree of under-reporting by patients. Calibration of our

TABLE 2 Test characteristics of unhealthy alcohol use classifier performance across a range of logistic regression cut points at the encounter level

Cut point	Sensitivity	Specificity	PPV	NPV
0.35	0.93 (0.91, 0.95)	0.76 (0.76, 0.76)	0.06 (0.06, 0.07)	1(1,1)
0.40	0.87 (0.85, 0.90)	0.92 (0.92, 0.92)	0.16 (0.15, 0.17)	1(1,1)
0.45	0.82 (0.79, 0.84)	0.97 (0.97, 0.97)	0.31 (0.29, 0.33)	1(1,1)
0.5	0.76 (0.73, 0.78)	0.98 (0.98, 0.98)	0.46 (0.44, 0.49)	1(1,1)
0.55	0.69 (0.65, 0.72)	0.99 (0.99, 1)	0.59 (0.55, 0.62)	1(1,1)

The cut point chosen for the classifier will be a trade-off between sensitivity and specificity as shown. In addition, PPV and NPV are shown for each cut point. PPV = positive predictive value; NPV = negative predictive value.

TABLE 3 Error analysis of classifications for unhealthy alcohol use (UAU)—patient characteristics and outcomes

	Likelihood of UAU per patient chart review (n = 1011)				
Characteristics and outcomes	***Definitely***	***Highly probable***	***Probable***	***Definitely not***	***P value***
n (%)	337 (33.3)	339 (33.5)	121 (12.0)	214 (21.2)	<0.001
Predictive probability (mean ± SD)	0.55 ± 0.18	0.56 ± 0.09	0.55 ± 0.11	0.54 ± 0.05	0.069
Age (mean ± SD)	51.06 ± 14.49	52.56 ± 12.39	50.43 ± 13.45	40.64 ± 15.66	<0.001
Male sex, *n* (%)	240 (71.2)	221 (65.2)	75 (62.5)	87 (40.7)	<0.001
Ethnicity, *n* (%)					
Non-Hispanic White	135 (40.1)	148 (43.7)	47 (38.8)	103 (48.1)	<0.001
Non-Hispanic Black	135 (40.1)	98 (28.9)	32 (26.4)	61 (28.5)	<0.001
Hispanic White	16 (4.7)	28 (8.3)	9 (7.4)	15 (7.0)	<0.001
Hispanic Black	1 (<1)	0 (<1)	0 (<1)	0 (<1)	<0.001
Other	0 (<1)	65 (19.2)	33 (27.3)	35 (16.4)	<0.001
AUDIT score (mean ± SD)	11.75 ± 7.75	7.03 ± 7.19	7.69 ± 6.72	2.02 ± 2.90	<0.001
AUDIT not measured, *n* (%)	84 (24.9)	202 (59.6)	71 (59.2)	148 (69.2)	<0.001
Blood alcohol content (BAC) (mean ± SD)	75.73 ± 110.83	1.00 ± 7.61	3.43 ± 2.75	1.38 ± 1.44	<0.001
BAC not measured, *n* (%)	206 (61.1)	259 (76.4)	97 (80.8)	188 (87.9)	<0.001
Frequency of drinking activities per week, (mean ± SD)	5.34 ± 2.28	4.31 ± 2.74	3.43 ± 2.75	1.38 ± 1.44	<0.001
Frequency of drinking activities missing, *n* (%)	68 (20.2)	279 (82.3)	97 (80.8)	188 (87.9)	
No. of drinks in one sitting (mean ± SD)	6.63 ± 4.47	4.38 ± 3.91	5.31 ± 7.62	2.03 ± 0.92	<0.001
No. of drinks missing, *n* (%)	82 (24.3)	293 (86.4)	104 (86.7)	196 (91.6)	
Binged within one month of encounter—yes, *n* (%)	278 (82.5)	5 (1.5)	2 (1.7)	0 (0.0)	<0.001
Binged within 1 month of encounter—unclear, *n* (%)	52 (15.4)	85 (26.0)	31 (34.2)	29 (13.6)	<0.001
Evidence of withdrawal symptoms, *n* (%)	129 (38.3)	30 (8.8)	7 (5.8)	2 (0.9)	<0.001
CIWA initiated for potential alcohol withdrawal, *n* (%)	175 (51.9)	53 (15.6)	15 (12.5)	16 (7.5)	<0.001
Max CIWA score documented when protocol initiated, *n* (%)	60 (34.3)	15 (28.3)	4 (26.7)	1 (6.3)	
Max CIWA score (mean ± SD)	9.38 ± 6.72	6.33 ± 6.09	1.75 ± 1.71	0.00 ± nan	NA
Previous history of alcohol misuse but not current, *n* (%)	0 (0.0)	315 (92.9)	100 (83.3)	0 (0.0)	<0.001
Previous history of alcohol misuse and current, *n* (%)	229 (68.0)	22 (6.5)	1 (0.8)	0 (0.0)	<0.001
No hx of alcohol misuse, *n* (%)	108 (32.0)	2 (0.6)	19 (15.8)	214 (100.0)	<0.001
Admission because of EtOH injuries, *n* (%)	41 (12.2)	8 (2.4)	2 (1.7)	1 (0.5)	<0.001
Family history of alcohol misuse, *n* (%)	38 (11.3)	65 (19.2)	4 (3.3)	43 (20.1)	<0.001
Family history not included in any notes, *n* (%)	60 (17.8)	40 (11.8)	19 (15.8)	27 (12.6)	<0.001
History of other substance use, *n* (%)					
Marijuana	100 (29.7)	82 (24.2)	14 (11.7)	76 (35.5)	<0.001
Opiates	39 (11.6)	50 (14.7)	4 (3.3)	46 (21.5)	<0.001
Cocaine	66 (19.6)	97 (28.6)	7 (5.8)	29 (13.6)	<0.001
Tobacco	222 (65.9)	277 (81.7)	19 (15.8)	109 (50.9)	<0.001
Other pain killer	0 (0.0)	0 (0.0)	0 (0.0)	0 (0.0)	NA
Amphetamines	7 (2.1)	6 (1.8)	1 (0.8)	6 (2.8)	0.643
Barbiturates	0 (0.0)	1 (0.3)	0 (0.0)	1 (0.5)	0.603
Benzodiazepines	5 (1.5)	14 (4.1)	1 (0.8)	11 (5.1)	0.026
Phencyclidine	3 (0.9)	4 (1.2)	0 (0.0)	7 (3.3)	0.046
No hx of other substance use	75 (22.3)	20 (5.9)	86 (71.7)	65 (30.4)	<0.001
Hx not available	3 (0.9)	0 (0.0)	5 (4.2)	0 (0.0)	<0.001

CIWA = Clinical Institute Withdrawal Assessment for Alcohol.

model was conducted considering the lower prevalence of alcohol misuse in the external validation cohort compared to the development cohort; however, the sensitivity and specificity did not improve.

Our NLP classifier for UAU demonstrated satisfactory discrimination in external validation, with sensitivity reaching 76% and specificity at 98% when analyzing encounter level EHR notes. When applied to the first 24 hours of clinical notes, sensitivity was moderately decreased to 66%, whereas specificity remained at 98%. In comparison to other well-validated screening methods for UAU in the hospital setting, the CAGE questionnaire demonstrated sensitivity and specificity of 77% and 99%, respectively, whereas the Michigan Alcoholism Screening Test (MAST) was 37% and 100%, respectively [27, 28]. Our classifier had similar performance to other self-report tools like the CAGE [27] and MAST [28] but application of the NLP classifier does not require additional hospital personnel work in screening and documenting results. Existing machine learning classifiers for UAU that also circumvent traditional screening methodology use specific biomarkers, personality traits, environmental influences, and/or self-documented reflections as predictors [29]. These elements are not routinely collected across all hospitalized patients and therefore, limit comprehensive screening.

The decrease in classifier sensitivity when applied to the first 24 hours of clinical notes was likely because of the reduction of data input from the clinical notes. In patients with shorter length of stays, the classifier could become an essential tool for identifying unhealthy use with increased sensitivity as data continues to be gathered before the care team can adequately screen. This is highly feasible as hospital census data from 2010 to 2015 indicate that the average length of stay was 6.1 days for 35 567 750 annual hospital admissions [30].

Post hoc chart review of the NLP classifier provided evidence that approximately three-quarters of false positives were re-evaluated as true positives. Concrete evidence in clinical notes regarding UAU and pertinent risk factors support classifier identification of UAU when AUDIT scores determined otherwise. Potential reasons for this discrepancy include patient under-reporting during AUDIT interviews and inconsistencies in administration of the AUDIT between rotating staff. Patient under-reporting is a common phenomenon in substance use screening, especially if patients lack readiness to change or when the interviewer is not an established member of the care team who is able to build rapport with the patient [31–33]. The *post hoc* chart review leads to the conclusion that the digital classifier may be performing better than the AUDIT screening during inpatient hospitalization. Confirmation of this theory would require a more rigorous study design with prospective assessment of a full cohort of patients using a gold standard assessment tool to index the AUDIT and digital classifier against.

There was another issue related to comorbidity that we noted. We found individuals with UAU were much younger with a median age of 49, whereas those without UAU in our sample tended to be older with a median age of 62. This difference may explain some of the variations in the Elixhauser comorbidities that were noted. For example, the younger UAU group tended to present with uncomplicated hypertension, whereas the no misuse group tended to have a greater proportion with complicated hypertension. In the case of hypertension, it is possible that the younger age individuals have not had time to progress to the more complicated pattern. Although these differences are certainly of interest and merit further research, the analyses required to understand these relationships reside beyond the scope of this paper. It is also possible that some of the false positives that we found in the *post hoc* analysis that the classifier identified include individuals with AUDIT scores <8. Examination on the effects of lower AUDIT threshold on the classifier performance would require re-training the model against the lower threshold and not relevant to the external validation work in this study.

Several limitations are present in the current study. The discrepancy in prevalence of UAU between our sample and the national average needs to be further understood in the context of the brief time frame of data that we used for assessment as well as the degree to which data reflect alcohol discussions done during hospitalization. Furthermore, the financial cost and equipment requirements for supporting an informatics team capable of processing clinical notes might be not feasible for some hospital systems. Implementation of the classifier at additional study sites can provide more conclusive data in terms of average costs required, effectiveness over current screening methods, as well as patient outcomes when the care team focuses intervention services on machine identified, at risk patients. Finally, it is possible that the universal screening program that was present at our hospital sensitized the staff to the potential presence of alcohol and substance use, and thereby encouraged them to have conversations with patients about these issues that then filter into the medical record and produce data that enables our classifier to work well.

For the clinician or specialist who might wish to consider a screening tool like this, several implications need to be considered. The digital classifier detected UAU from a general hospital inpatient cohort with a sensitivity of 0.76 and specificity of 0.98. This means that of 100 cases of UAU, the classifier will correctly ascertain 76 of those people. With regard to specificity, this means that for 100 cases without UAU the classifier will correctly exclude 98 individuals. This is done in a fully automated fashion using a combination of textual data that are charted within the first 24 hours of admission. Although the 0.75 sensitivity may be lower than desired, it is important to remember that these cases are being ascertained in a general hospital setting where patients are being admitted for primary reasons other than alcohol or substance use. Therefore, the opportunity here is to intervene with some patients that may not be aware of their UAU or may not yet have disclosed it. This classifier should not be used diagnostically and confirmation by more comprehensive clinical evaluation and/or use of a more accurate instrument is recommended.

In conclusion, the external validation of the alcohol machine learning classifier demonstrated adequate sensitivity, specificity, and may overcome limitations in manual screening fidelity. The trained classifier is publicly available, free for access, and may assist hospital systems in identifying at risk patients for targeted interventions in a timely manner and help improve patient outcomes.

DECLARATION OF INTERESTS

None.

ACKNOWLEDGEMENTS

This research was supported, in part, by the following: National Institute on Alcohol Abuse and Alcoholism (K23-AA024503 to M.A.); National Institute on Drug Abuse (R01-DA04171 to N.S.K., UG1-DA049467 to M.A. and N.S.K, and R01-DA051464 to M.A. and N.S.K.); National Center for Advancing Translational Sciences (UL1-TR002389 to M.A. and N.S.K. and KL2-TR002387 to N.S.K.); and Agency for Healthcare Research and Quality (K12-HS026385 to H.T.). The authors have no other conflicts of interest to report.

AUTHOR CONTRIBUTIONS

Yiqi Lin: Investigation; validation. **Brihat Sharma:** Data curation; formal analysis; investigation; methodology; software; validation; visualization. **Hale Thompson:** Formal analysis; investigation; project administration; supervision; validation. **Randy Boley:** Data curation; project administration; resources. **Kathryn Perticone:** Methodology; supervision. **Neeraj Chhabra:** Investigation; resources; validation. **Majid Afshar:** Conceptualization; data curation; formal analysis; funding acquisition; investigation; methodology; project administration; resources; software; supervision; validation. **Niranjan Karnik:** Conceptualization; formal analysis; funding acquisition; investigation; methodology; project administration; supervision; validation.

ORCID

Brihat Sharma https://orcid.org/0000-0003-0417-4553
Hale M. Thompson https://orcid.org/0000-0002-9704-934X
Randy Boley https://orcid.org/0000-0002-9365-8412
Neeraj Chhabra https://orcid.org/0000-0002-0945-1233
Majid Afshar https://orcid.org/0000-0002-6368-4652
Niranjan S. Karnik https://orcid.org/0000-0001-7650-3008

REFERENCES

1. Alcohol [Internet]. (2021) [cited 2021 Apr 20]. Available from: https://www.who.int/news-room/fact-sheets/detail/alcohol
2. Substance Abuse and Mental Health Services Administration. (2018) Key substance use and mental health indicators in the United States: Results from the 2018 National Survey on Drug Use and Health (HHS Publication No. PEP19–5068, NSDUH Series H-54) [Internet]. Rockville, MD:(2019) Center for Behavioral Health Statistics and Quality, Substance Abuse and Mental Health Services Administration; 2019. Available from: https://www.samhsa.gov/data/
3. Center for Disease Control and Prevention. (2019) Excessive Alcohol Use: A Drain on the American Economy [Internet]. 2019. Available from: https://www.cdc.gov/alcohol/onlinemedia/infographics/excessive-alcohol-economy.html
4. Bazzi A, Saitz R. Screening for unhealthy alcohol use. JAMA. 2018; 320(18):1869–71. https://doi.org/10.1001/jama.2018.16069
5. Rosón B, Monte R, Gamallo R, Puerta R, Zapatero A, Fernández-Solá J, et al. Prevalence and routine assessment of unhealthy alcohol use in hospitalized patients. Eur J Intern Med. 2010 Oct;21(5): 458–64.
6. Doering-Silveira J, Fidalgo TM, Nascimento CLES, Alves JB, Seito CL, Saita MC, et al. Assessing alcohol dependence in hospitalized patients. Int J Environ Res Public Health. 2014 Jun; 11(6):5783–91.
7. Johnson M, Jackson R, Guillaume L, Meier P, Goyder E. Barriers and facilitators to implementing screening and brief intervention for alcohol misuse: A systematic review of qualitative evidence. J Public Health. 2011 Sep 1;33(3):412–21.
8. Hoonpongsimanont W, Ghanem G, Chen Y, Sahota PK, Carroll C, Barrios C, et al. Underreporting of alcohol use in trauma patients: A retrospective analysis. Subst Abus. 2019 Oct;22:1–5. https://doi.org/10.1080/08897077.2019.1671936
9. Afshar M, Phillips A, Karnik N, Mueller J, To D, Gonzalez R, et al. Natural language processing and machine learning to identify alcohol misuse from the electronic health record in trauma patients: Development and internal validation. J am Med Inform Assoc JAMIA. 2019 Jan 2;26(3):254–61. https://doi.org/10.1093/jamia/ocy166
10. Van Vleck TT, Chan L, Coca SG, Craven CK, Do R, Ellis SB, et al. Augmented intelligence with natural language processing applied to electronic health Records for Identifying Patients with non-alcoholic fatty liver disease at risk for disease progression. Int J Med Inform. 2019 Sep;129:334–41. https://doi.org/10.1016/j.ijmedinf.2019.06.028
11. Gao H, Aiello Bowles EJ, Carrell D, Buist DSM. Using natural language processing to extract mammographic findings. J Biomed Inform. 2015 Apr;54:77–84.
12. Pons E, Braun LMM, Hunink MGM, Kors JA. Natural language processing in radiology: A systematic review. Radiology. 2016 Apr 18;279(2):329–43.
13. To D, Sharma B, Karnik N, Joyce C, Dligach D, Afshar M. Validation of an alcohol misuse classifier in hospitalized patients. Alcohol. 2020; 84:49–55. https://doi.org/10.1016/j.alcohol.2019.09.008
14. Thompson HM, Hill K, Jadhav R, Webb TA, Pollack M, Karnik N. The substance use intervention team: A preliminary analysis of a population-level strategy to address the opioid crisis at an academic health center. J Addict Med. 2019;13(6):460–63. https://doi.org/10.1097/ADM.0000000000000520
15. Thompson HM, Faig W, VanKim NA, Sharma B, Afshar M, Karnik NS. Differences in length of stay and discharge destination among patients with substance use disorders: The effect of substance use intervention team (SUIT) consultation service. PLoS ONE. 2020;15 (10):e0239761. https://doi.org/10.1371/journal.pone.0239761
16. 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. https://doi.org/10.1111/j.1530-0277.2002.tb02534.x
17. Johnson JA, Lee A, Vinson D, Seale JP. Use of AUDIT-based measures to identify unhealthy alcohol use and alcohol dependence in primary care: A validation study. Alcohol Clin Exp Res. 2013 Jan;37-(Suppl 1):E253–9. https://doi.org/10.1111/j.1530-0277.2012.01898.x
18. O'Connor EA, Perdue LA, Senger CA, Rushkin M, Patnode CD, Bean SI, et al. Screening and behavioral counseling interventions to reduce unhealthy alcohol use in adolescents and adults: Updated evidence report and systematic review for the US preventive services task force. JAMA. 2018 Nov 13;320(18):1910–28. https://doi.org/10.1001/jama.2018.12086
19. Sullivan JT, Sykora K, Schneiderman J, Naranjo CA, Sellers EM. Assessment of alcohol withdrawal: The revised clinical institute withdrawal assessment for alcohol scale (CIWA-Ar). Addiction. 1989 Nov;84(11):1353–7.
20. Information NC for B. Pike USNL of M 8600 R, MD B, Usa 20894. Metathesaurus [Internet]. UMLS® Reference Manual [Internet]. National Library of Medicine (US); 2021 2009 [cited 2021 Apr 20]. Available from: https://www.ncbi.nlm.nih.gov/books/NBK9684/
21. Elixhauser A, Steiner C, Harris DR, Coffey RM. Comorbidity measures for use with administrative data. Med Care. 1998 Jan;36(1): 8–27.

22. Saito T, Rehmsmeier M. The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. PLoS ONE. 2015 Mar 4;10(3):e0118432.
23. Ruopp MD, Perkins NJ, Whitcomb BW, Schisterman EF. Youden index and optimal cut-point estimated from observations affected by a lower limit of detection. Biom J. 2008 Jun;50(3):419–30.
24. Debray TPA, Vergouwe Y, Koffijberg H, Nieboer D, Steyerberg EW, Moons KGM. A new framework to enhance the interpretation of external validation studies of clinical prediction models. J Clin Epidemiol. 2015 Mar;68(3):279–89.
25. Collins GS, Reitsma JB, Altman DG, Moons KGM. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): The TRIPOD statement. Br J Surg. 2015 Feb; 102(3):148–58. https://doi.org/10.1002/bjs.9736
26. National Institute of Alcoholism and Alcohol Abuse. (2020) Alcohol Facts and Statistics [Internet]. 2020. Available from: https://www.niaaa.nih.gov/sites/default/files/publications/NIAAA_Alcohol_FactsandStats_102020_0.pdf
27. Ewing JA. Detecting alcoholism. The CAGE questionnaire. JAMA. 1984;252(14):1905–7. https://doi.org/10.1001/jama.252.14.1905
28. Selzer ML. Michigan alcoholism screening test (MAST): Preliminary report. Univ Mich Med Cent J. 1968 Jun;34(3):143–5.
29. Mak KK, Lee K, Park C. Applications of machine learning in addiction studies_ a systematic review. Psychiatry Res. 2019;275:53–60. https://doi.org/10.1016/j.psychres.2019.03.001
30. AHA. (2021) Annual Survey Database™|AHA Data [Internet]. [cited 2021 Apr 20]. Available from: https://www.ahadata.com/aha-annual-survey-database
31. 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. https://doi.org/10.1093/alcalc/agl022
32. 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. https://doi.org/10.1016/j.addbeh.2011.11.006
33. McNeely J, Kumar PC, Rieckmann T, Sedlander E, Farkas S, Chollak C, et al. Barriers and facilitators affecting the implementation of substance use screening in primary care clinics: A qualitative study of patients, providers, and staff. Addict Sci Clin Pract. 2018 Dec; 13(1):8. https://doi.org/10.1186/s13722-018-0110-8

SUPPORTING INFORMATION

Additional supporting information may be found in the online version of the article at the publisher's website.

How to cite this article: Lin Y, Sharma B, Thompson HM, Boley R, Perticone K, Chhabra N, et al. External validation of a machine learning classifier to identify unhealthy alcohol use in hospitalized patients. Addiction. 2022;117:925–33. https://doi.org/10.1111/add.15730