Original Paper

Natural Language Processing Supported by Annotation Process Artifacts to Detect Risky Alcohol Use in Electronic Health Records: Algorithm Development and Validation Study

Katherine Weber[1], PhD; Asher Strayhorn[1], BS; Xintong Ju[2], MS; Jessica Mellinger[3], MD, MSc; Gerald Scott Winder[4], MD, MSc; Anne C Fernandez[2*], PhD; VG Vinod Vydiswaran[1,5*], PhD

[1]Department of Learning Health Sciences, Medical School, University of Michigan, Ann Arbor, MI, United States

[2]Department of Psychiatry, Medical School, University of Michigan, Ann Arbor, MI, United States

[3]Division of Gastroenterology & Hepatology, Department of Internal Medicine, Henry Ford Hospital, Detroit, MI, United States

[4]Department of Psychiatry, Henry Ford Hospital, Detroit, MI, United States

[5]School of Information, University of Michigan, Ann Arbor, MI, United States

[*]these authors contributed equally

Corresponding Author:
VG Vinod Vydiswaran, PhD
Department of Learning Health Sciences
Medical School
University of Michigan
1111 E. Catherine Street
Room 209
Ann Arbor, MI, 48109-2054
United States
Phone: 1 734 647 1207
Fax: 1 734 647 3914
Email: vgvinodv@umich.edu

Abstract

Background: High-risk alcohol use and alcohol use disorder (AUD) are significant contributors to surgical risk and complications yet remain underidentified in surgical populations. While electronic health records (EHRs) contain relevant information about patients' alcohol use in clinical notes, manual chart review is time-intensive and impractical at scale. Natural language processing (NLP) offers a path to automated identification, but alcohol-related risk classes are highly imbalanced in clinical populations, making multiclass classification particularly challenging.

Objective: This study aims to develop an automated NLP pipeline for 4-class alcohol risk classification using clinical notes in a surgical population, addressing class imbalance by leveraging intermediate annotation evidence generated during the human labeling process.

Methods: We manually labeled 1000 surgical patients drawn from a large academic health system into ordinal alcohol risk classes (no-alcohol, low-risk, high-risk, and probable-dependence) and a fifth class indicating no mention of alcohol use. We used a structured annotation process with 6 trained annotators and expert-developed guidelines. We developed a multistage NLP pipeline operating on 3 years of clinical notes per patient. First, a rules-based preprocessing stage filtered notes to clinically relevant sections containing alcohol-related content. No-mention cases were excluded from further modeling. Second, a BioClinicalBERT-based multilabel classifier ("auto-annotator") was fine-tuned to reproduce the annotations created by human labelers at the section level, producing a structured 26-dimensional probability vector for each section. Finally, a bidirectional long short-term memory (LSTM) network classified patients into the 4 risk categories by processing each patient's ordered sequence of auto-annotation vectors, allowing the model to capture temporally accumulating evidence across clinical encounters.

Results: Human annotators achieved 93.7% first-pass agreement overall (Fleiss κ=0.8891), but agreement was as low as 67.5% for probable-dependence cases. The 4-class LSTM classifier achieved a weighted-average F_1-score of 0.86 and a macroaverage F_1-score of 0.77 on a held-out validation set of 195 patients. One-vs-rest area under the curve (AUC) scores were 0.97 for no-alcohol, 0.88 for low-risk, 0.90 for probable-dependence, and 0.71 for high-risk. Collapsed to a binary high/low risk

XSL•FO
RenderX

classification, the model achieved an accuracy of 0.92 and a macro-F_1-score of 0.80. Performance was comparable across biological sex without explicit sex-specific training.

Conclusions: We used a 2-stage NLP pipeline to identify 4 ordinal classes of alcohol use risk in surgical patients. Four-class alcohol-use classification presents a highly imbalanced problem that constrains the availability of high-importance but rare labels in the dataset. We addressed this data scarcity by using intermediate annotation evidence as well as final labels to enrich the feature space for the classification task. The pipeline was able to separate patient records into 4 strata of alcohol-use risk, with "high-risk" patients being most likely to be misclassified into an adjacent class.

(J Med Internet Res 2026;28:e91449) doi: 10.2196/91449

KEYWORDS

natural language processing; NLP; machine learning; electronic health record; EHR; alcohol-related disorder; neural network; computer; risk assessment; substance use disorder

Introduction

Background and Significance

Alcohol is the most widely used substance in the United States, and alcohol use disorder (AUD) is the most common type of substance use disorder in the nation [1]. Risky alcohol use (ie, consumption of alcohol above recommended daily, weekly, or per-occasion amounts) [2] is also common and increasing in some subpopulations [3,4]. Both risky alcohol use and AUD pose substantial surgical risk and affect a large proportion of surgical patients [5,6]. Risky alcohol use and AUD are linked to an increased risk of surgical complications, longer hospital stays, increased costs, and even death [6-15]. However, patients with these conditions remain alarmingly underidentified in surgical care [16-19].

Proper identification of surgical patients with risky alcohol use and AUD is needed to deliver appropriate surgical care, timely interventions, and inpatient alcohol withdrawal prophylaxis when necessary [20,21]. Implementation of interventions is hindered by suboptimal alcohol screening in real-world care. Past research indicates that electronic health records (EHRs) contain relevant alcohol information in clinical notes and other text that could yield insight into the presence of risky alcohol use and AUD even when formal diagnoses are lacking [22-24]. Access to this type of text information traditionally requires time-intensive manual chart review. Natural language processing (NLP) tools can make this information more accessible to clinicians and researchers earlier in the preoperative time course by using existing information that can supplement formal alcohol screening procedures. NLP is a core information extraction technique in health care because key information, such as clinical notes, social media health discussions, and clinical guidelines, is in the form of narrative text [25]. Because the definition of risky alcohol use extends beyond the specific criteria for AUD and because AUD is generally underdiagnosed, NLP approaches can examine the unstructured content of the clinical record for text-based indicators of alcohol-related issues [22,26].

Recent applications of NLP to substance use classification include a convolutional neural network (CNN) for screening for substance misuse (area under the receiver operating characteristic curve [AUROC] 0.97, 95% CI 0.96-0.98) [27] and a logistic regression classifier using concept unique identifiers (CUIs) generated by an NLP algorithm (AUROC=0.91, 95% CI 0.89-0.92) [23,28,29]. These classifiers produce a binary label for substance use disorder (or not), rather than a graded degree of risk. Another study found that only 44% of patients with mentions of AUD in clinical notes also had structured codes indicating these issues, highlighting the need for and potential contribution of text analysis [24].

There are a number of tools for identifying conditions related to substance use in the EHR, including individual, custom-developed algorithms [23] and processing using one or more publicly available NLP tools [30]. Recent efforts at identifying substance use include a regular-expression and rules-based approach that correctly identified 87% of human-labeled patients with risky alcohol use [22] and a combination of embeddings and a time-aware long short-term memory (TLSTM) network to detect evidence of alcohol and substance use disorders in clinical notes (AUROC=0.93) [31]. Some approaches use methods that extract computable entities from text, such as cTAKES [32], to produce features for machine learning classifiers [28]. Other NLP-driven approaches for locating evidence of substance use in notes include tools such as pretrained BERT (Bidirectional Encoder Representations from Transformers) models and universal sentence encoders [33,34] to infer context and meaning in free text. Several pretrained BERT models are tuned on health-related text corpora, including BioClinicalBERT [35], a fine-tuned XLM-RoBERTa approach [36], and transformer-based models [37]. To extract meaning from a large, unfiltered collection of notes and produce a graded risk label, we assembled a pipeline that combines several of these approaches.

Objective

The objective of the study is to develop an automated approach to classify alcohol risk using clinical text from EHRs and NLP. We address the challenge of locating clinically significant classes within a highly imbalanced dataset by leveraging both final labels and intermediate annotation information from the labeling process.

Methods

Risky Alcohol Use Class Definitions

Previous efforts to use NLP to identify risky alcohol use in clinical notes have returned a binary positive/negative

XSL•FO
RenderX

prediction. To provide health care professionals with a more nuanced view of a patient's alcohol risk levels, this study defines 4 ordinal risk categories (detailed annotation guidelines for risk classification criteria are available in Multimedia Appendix 1).

- No-alcohol: information in the clinical text explicitly indicates that the patient does not consume any alcohol.
- Low-risk: clinical text indicates that the patient consumes alcohol but does not exceed nationally recommended alcohol consumption limits (ie, no more than 4 drinks per day/occasion or 14 in a week for men, and no more than 3 drinks per day/occasion or 7 per week for women) [2,38], with no evidence of alcohol-related consequences or other criteria for AUD. This category also includes those who previously consumed alcohol or had AUD but have been sober (have not consumed alcohol) for more than 1 year.
- High-risk: the clinical text indicates that the patient exceeds nationally recommended alcohol consumption limits, has a diagnosis of mild or moderate AUD, and/or has evidence of 1 alcohol-related consequence, including specific events (eg, driving while intoxicated) or social, emotional, or family problems related to alcohol use.
- Probable-dependence: clinical text indicates that the patient is diagnosed with severe AUD, has specific alcohol-associated medical conditions, has attended treatment for alcohol problems, and/or has multiple alcohol-related consequences.

A fifth category, "no-mention," was used to label cases in which alcohol-related information is absent or scarce in the patient's available records. These cases would eventually be excluded from NLP-driven classification because of insufficient data.

Data Source

The study was conducted using data from a cohort of patients who underwent 1 or more surgical procedures and participated in the Michigan Genomic Initiative (MGI) [39], a large longitudinal cohort study conducted at a large tertiary academic health system in the Midwestern United States. The data extracted included inpatient and outpatient clinical notes, problem lists, visit summaries and letters, diagnostic codes, laboratory results, medications, and demographic information (including biological sex) for 3 years prior to the incident surgery for each patient.

The working cohort consisted of MGI patients (n=53,949) who met the inclusion criterion of having an index surgery date between May 29, 2012, and April 17, 2019. Patients with no structured data at all, without any clinical notes from encounters in the 3 years before surgery, or whose combined clinical notes contained fewer than 100 characters were excluded (n=320). From the remaining patients, 1000 patients were sampled for manual classification into the 4 risk classes or "no-mention." We aimed for a dataset that oversampled the rare cases, targeting 5% probable-dependence and 10% high-risk cases, and performed stratified sampling using keywords. The final dataset contains 4% and 8.8%, respectively. Table 1 describes the demographic characteristics of the patients in the dataset.

XSL•FO
RenderX

Table 1. Demographic characteristics of the manually labeled patient cohort used for training and evaluation of an NLPa classifier for risky alcohol use. Patients with elective surgeries between 2012 and 2019 were drawn from the Michigan Genomics Initiative.

Characteristic	Overall, n=1000	No-mention, n=24	No-alcohol, n=247	Low-risk, n=601	High-risk, n=88	Probable-dependence, n=40	Missing responses, n
Sex, n (%)							0
Female	483 (48.3)	12 (50)	147 (59.5)	280 (46.6)	31 (35.2)	13 (32.5)	
Male	517 (51.7)	12 (50)	100 (40.5)	321 (53.4)	57 (64.8)	27 (67.5)	
Race, n (%)							1
African American	57 (5.7)	—[b]	13 (5.3)	37 (6.2)	2 (2.3)	5 (12.5)	
American Indian or Alaska Native	8 (0.8)	1 (4.2)	2 (0.8)	4 (0.7)	—	1 (2.5)	
Asian	13 (1.3)	—	5 (2)	8 (1.3)	—	—	
White	893 (89.4)	20 (83.3)	223 (90.3)	534 (89)	83 (94.3)	33 (82.5)	
Native Hawaiian and other Pacific Islander	1 (0.1)	—	—	1 (0.2)	—	—	
Patient refused	1 (0.1)	—	—	1 (0.2)	—	—	
Unknown	10 (1)	1 (4.2)	2 (0.8)	7 (1.2)	—	—	
Other	16 (1.6)	2 (8.3)	2 (0.8)	8 (1.3)	3 (3.4)	1 (2.5)	
Ethnicity, n (%)							1
Hispanic or Latino	26 (2.6)	2 (8.3)	4 (1.6)	14 (2.3)	5 (5.7)	1 (2.5)	
Non-Hispanic or Latino	950 (95.1)	20 (83.3)	236 (95.5)	577 (96.2)	81 (92)	36 (90)	
Unknown	21 (2.1)	2 (8.3)	6 (2.4)	8 (1.3)	2 (2.3)	3 (7.5)	
Patient refused	2 (0.2)	—	1 (0.4)	1 (0.2)	—	—	
Marital status, n (%)							51
Married	567 (59.7)	16 (69.6)	128 (55.4)	354 (61.6)	47 (56)	22 (61.1)	
Unmarried	382 (40.3)	7 (30.4)	103 (44.6)	221 (38.4)	37 (44)	14 (38.9)	
Age (years), mean (SD)							0
At surgery	53.8 (16.0)	53.4 (20.2)	53.8 (17.2)	53.4 (15.5)	54.8 (15.8)	56.5 (14.9)	

[a]NLP: natural language processing.

[b]Not applicable.

Creating the Annotated Dataset

To facilitate manual classification of risky alcohol use in patients, a detailed annotation guideline document was created (Multimedia Appendix 1). The annotation guidelines included detailed definitions of the 4 classes and types of evidence that annotators should consider and annotate before assigning the risky alcohol use classification label. The document also included instructions on using the brat (brat rapid annotation tool) software tool [40] for annotation. The guideline document was developed and iteratively refined by 4 experts in addiction diagnosis and treatment, including 2 psychologists (ACF and FB) and 2 physicians (JLM and GSW). The experts based the annotation guidelines on published standards and research, including (1) national definitions for alcohol-related health risks at specific alcohol quantity and frequency cutoffs for men and women [41]; (2) AUD symptom type and timing as defined by the *DSM-5* (*Diagnostic and Statistical Manual of Mental Disorders* [Fifth Edition]) [38]; and (3) existing research specific to alcohol and surgical risk, which demonstrates increased surgical morbidity when active heavy drinking and AUD precede surgery, but reduced risk when patients abstain from alcohol use in the weeks and months leading up to surgery [10,12,42]. Twenty patients' complete notes were manually annotated by these experts following the annotation guidelines and were used to train annotators.

Annotation was conducted by 6 annotators, including students enrolled in advanced health-related degree programs. Two randomly selected annotators reviewed each chart in chronological order. They used the brat tool to mark segments of text relevant to alcohol risk classification by selecting 1 of 29 different annotation categories for each text segment (see Multimedia Appendix 2 for a list of annotation categories). Of those 29 categories, 3 were never used by annotators and were not carried forward into the analysis. Annotation categories included 2-part indicators expressing the kind of evidence (eg, a mention of the amount of alcohol consumed) and the degree of severity (eg, high-risk). After annotators read the entire chart, they recorded their final risky alcohol use label determination

XSL•FO
RenderX

(no-alcohol, low-risk, high-risk, or probable-dependence). If additional but identical alcohol use information was repeated in later clinical notes, annotators were not required to mark each repeated mention after the first.

When annotators disagreed on a label, they were informed that there was a disagreement, and they conducted a "second-pass" annotation independently (they remained blind to the other annotator's labels). If the disagreement persisted, the annotators met to resolve it through arbitration. During arbitration, the annotators met with the project manager, who moderated the discussion during which annotators provided one another with their rationales and examples that guided their labeling decisions. In cases where they still could not reach an agreement, a third annotator labeled the notes independently to break ties, and the principal investigators reviewed the final decision.

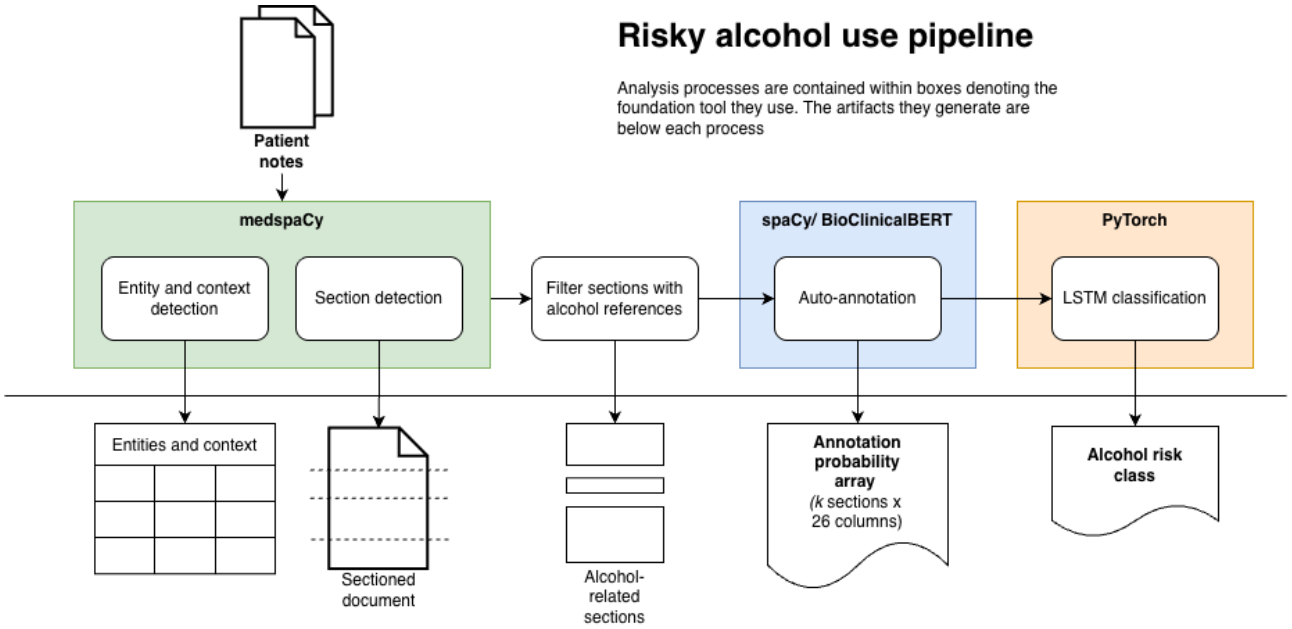
The final annotated dataset consisted of 1000 patient charts labeled into 1 of the risk classes, as well as all annotations placed on the notes. Eight hundred patients were sampled using stratified sampling to train the NLP pipeline models, and the remaining 200 patients were held out as the validation set. The 24 "no-mention" patients (19 in training and 5 in validation) were subsequently dropped from the pipeline and not used for training subsequent stages, leaving a training set of 781 patients and a validation set of 195 patients.

Model Design

Overview

To label alcohol use risk, we developed a multistage classification pipeline. The first stage, preprocessing and entity detection, standardizes the text input and splits the patient notes into relevant sections with mentions of alcohol use or alcohol-related problems. The second stage, auto-annotation, analyzes each section and returns a vector of probabilities indicating whether any of the 26 annotation labels apply to that section. The 26 possible labels (as well as the 3 that were never chosen by humans in the labeling process) are included in Multimedia Appendix 2. The final stage, classification, combines the array of auto-annotation vectors and emits a single class representing the level of alcohol use risk detected in the patient's notes. The overall classification workflow is summarized in Figure 1.

Figure 1. Schematic diagram of the multistage risky alcohol use classification workflow. LSTM: long short-term memory.



Preprocessing and Entity Detection

While patient notes are full of relevant health information, information about alcohol-related behavior and conditions is sparsely distributed throughout the patient notes. We narrowed the overall set of notes (n=64,703) by excluding all irrelevant patient instruction notes, retaining 45,550 notes about the 1000 patients in our dataset. The median length of these notes was 317 words. Next, as documents were imported for processing, we applied regular expressions to strip out templated text that instructed patients to avoid drinking alcohol because of a procedure or medication (eg, "heavy alcohol use may increase the risk of toxicity"). The Python script used to drive our rules is provided in Multimedia Appendix 3.

We then applied the medspaCy [43] toolkit to identify medical entities and context (relations between entities), applying the entity and context rules from a foundation of keywords and logic developed in Vydiswaran et al [22]. These rules included regular expressions, proximity calculations, and numerical analysis to identify text related to alcohol use and alcohol-related disorders. The context relations identified by medspaCy include indications of whether the alcohol-related concept ("entity") was in the past, related to family members, or negated.

Next, to filter out irrelevant sections from long documents, we broke the notes into sections (such as "social history" and "assessment and plan") using medspaCy section detection. This resulted in a total of 200,395 sections with a median length of 49 (IQR 15-137) words per section. Of all the sections, only 8417 contained at least 1 alcohol-related entity.

We experimented with a variety of ways to represent this information as features for classifiers, but the rare classes were

XSL•FO
RenderX

entirely ignored or overfitted. At best, macro-F_1-scores were 0.66. We developed the auto-annotation approach to provide our models with richer labeled data as an intermediate step.

Auto-Annotation

One of the main challenges of the 4-class classification task is that the classes are highly imbalanced, yet the most significant and valuable classes are those that are most infrequent. Models trained with only the final classification were prone to overfitting if we concentrated on the imbalanced classes. Instead of relying on a single label to train a risky alcohol use classifier based on a limited training set, we trained an "auto-annotator" algorithm to build a new feature set that reproduces the annotations created by human annotators as evidence to support a label. This helped address the limitations of a small training dataset, taking advantage of a much larger volume of annotations in the training data to identify key alcohol-risk features in an individual patient's record.

To train the auto-annotator, we built a multilabel classifier using the spaCy textcat_multilabel transformers pipeline. The model was trained on sections associated with training-set patients. A total of 6236 sections contained alcohol-related entities, and 128,886 sections had no annotations. We fine-tuned the BioClinicalBERT [35] transformer with a new multilabel classification head to return a probability array for each section indicating the probability that the section should trigger each of the 26 annotation classes. Because BERT models are limited to a maximum input sequence of 512 tokens [33], we truncated longer sections to 8000 characters and then to 512 tokens after tokenizing in the spaCy pipeline. Approximately 8% of sections were truncated before auto-annotation.

To validate the auto-annotator, we asked whether its output was functionally equivalent to human annotations as a feature set for downstream classification. We trained 4 scikit-learn classifiers (random forest, decision tree, support vector classifier [SVC], and k-nearest neighbors [KNN]) using auto-annotator output as features. For the best-performing classifier, SVC, we additionally trained the same model using human annotations as features. If the 2 feature sets produced comparable classification performance on the held-out validation set, we considered the auto-annotator a sufficient proxy for human annotation. SVC was the best-performing of the machine learning algorithms—the classifier returned a macro-F_1-score of 0.71 over human annotations and 0.69 over auto-annotations.

Patient Classification

Preliminary evaluation of these machine-learning models showed 2 key weaknesses, however. First, although precision was generally good overall (particularly for decision tree), recall for high-risk and probable-dependence cases was poor for the KNN, decision tree, and random forest classifiers (macro-F_1-scores of 0.51, 0.59, and 0.55, respectively). SVC achieved higher high-risk recall (0.61) than the selected long short-term memory (LSTM; 0.44) but lower overall performance (macro-F_1-scores 0.69 vs 0.77). Additionally, the machine-learning models effectively scored all patients as if they were assessed under the guidelines for male patients.

To generate a final alcohol risk classification for each patient, we turned to 3 neural network model architectures in the PyTorch [44] toolkit—a CNN, a multilayer Elman recurrent neural network (RNN), and a bidirectional LSTM model. When the model was given all sections from a patient's notes, we found that it triggered false positive labels based on sections that had nothing to do with alcohol consumption (anesthesia notes were especially prone to this). As a result, we only sent sections with alcohol-related entities to the classifier. Each model was fed the matrix of k × 26 output from the auto-annotator, where k denotes the number of alcohol-related sections in the patient's notes. Patients who had no alcohol-related sections were classified as "no-mention."

We experimented with a variety of optimizers, learning rates, and samplers, and chose the model that performed well on both the 4-class classification task and the more general 2-class classification into high-risk and low-risk categories. Training data were partitioned into 4 folds, and the model was trained with cross-validation over these folds. Hyperparameter optimization was performed with Weights & Biases [45]. The final model used a learning rate of 0.0094 (search was between 0.001 and 0.01), the Adam optimizer (Adam and stochastic gradient descent [SGD] were tested), and a hidden size of 123 (search between 96 and 172). We evaluated PyTorch's imbalanced dataset sampler as an alternative approach to class imbalance but found no meaningful improvement over the final model configuration.

We evaluated the final risky alcohol use classifier using both the weighted-average F_1-score and the macroaverage F_1-score. We also measured the 2-class performance on F_1-score.

To compare this model with binary classifiers for risky alcohol use, we also evaluated the model by rolling the risk classes up, with no-alcohol and low-risk as a single "low" risk class and high-risk and probable-dependence as a "high" risk class. We also split the validation set into male and female patients and evaluated them separately.

Ethical Considerations

This study has been reviewed and approved under human participants research guidelines by an institutional review board (study HUM00204651). The research was conducted under an informed consent waiver because it involves no more than minimal risk to the participants, the waiver will not adversely affect the rights or welfare of the participants, and the research could not be carried out without the waiver, as the starting patient dataset included data on thousands of patients. To protect patient privacy and confidentiality, synthetic identifiers ("file numbers") were generated for each participant. These file numbers were used as the participant identifier in all structured and unstructured data, and all other identifying structured columns were removed from the dataset. The cross-reference between file numbers and patient ID was kept in secure storage and segregated from storage used in the development and testing of models. All data storage and processing took place on a high-security research cluster designed to align with HIPAA (Health Insurance Portability and Accountability Act) data protection regulations. The text of clinical notes could not be

XSL•FO
RenderX

reliably deidentified; access to the notes was restricted to members of the research team with appropriate training for handling sensitive data. No compensation was offered to the participants.

Results

Key Results

To help us set appropriate expectations for model performance, we measured the frequency with which annotators were able to come to a consensus risk class during the labeling process.

Table 2 shows that human annotators were able to agree on the first pass 93.7% of the time, with Fleiss κ of 0.8891. After the second pass, the agreement rose to 97.9%, with Fleiss κ of 0.9630. While the overall agreement was substantial and "near perfect," especially after the second pass, human annotators struggled most with identifying high-risk and probable-dependent cases, with their initial agreement as low as 67.5% for the probable-dependent cases. Even after review, 10% of these patients' notes (as well as the notes of patients eventually labeled high-risk) were reviewed and arbitrated by a third rater.

Table 2. Interrater agreement between human annotator dyads for classification of risky alcohol use severity in clinical notes, stratified by class. Annotation was performed on 1000 patient records from the Michigan Genomics Initiative.

Label	First pass agreement, n (%)	Second pass agreement, n (%)	Required third-party review, n (%)
Probable-dependence (n=40)	27 (67.5)	36 (90)	4 (10)
High-risk (n=88)	68 (77.3)	79 (89.8)	9 (10.2)
Low-risk (n=601)	578 (96.2)	596 (99.2)	5 (0.8)
No-alcohol (n=247)	240 (97.2)	244 (98.8)	3 (1.2)
No-mention (n=24)	24 (100)	24 (100)	0 (0)
Total (n=1000)	937 (93.7)	979 (97.9)	21 (2.1)

The auto-annotator produced 8417 annotations for the 976 labeled patients. A total of 24 patients had no annotations because no alcohol-related note sections passed the filtering stage; they were labeled as "no-mention." We experimented with feature arrays with the original 26 classifications used by the annotators, as well as a condensed array of 5 features corresponding to the probability of the 4 severity suffix classes of the annotation categories (eg, "HighRisk" in "AlcoholAmount-HighRisk") and a fifth feature indicating whether the "Sober-LowRisk" annotation category was used. We found that the 26-class arrays provided more information to the final classifier. The final auto-annotator output is a 26-dimensional vector for each annotated section for every patient.

Table 3 contains the classification report for the best-performing LSTM classifier. The weighted-average F_1-score for the 4-class risky alcohol use classification model was 0.86, and the 4-class macroaverage F_1-score was 0.77. When we created a binary risk classifier (combining the bottom 2 and the top 2 risk categories), the accuracy of the 2-class model was 0.92.

XSL•FO
RenderX

Table 3. Classification performance (precision, recall, and F1-score) of the selected bidirectional LSTM[a] model for graded risky alcohol use detection from clinical notes, evaluated on a held-out validation set of 195 patients from the Michigan Genomics Initiative who underwent elective surgery, 2012-2019. Classes: no-alcohol, low-risk, high-risk, and probable-dependent.

Modeling approach	Precision	Recall	F_1-score	Support
Four-class				
Probable-dependence	0.83	0.62	0.71	8
High-risk	0.73	0.44	0.55	18
Low-risk	0.88	0.94	0.91	120
No-alcohol	0.90	0.90	0.90	49
Accuracy	—[b]	—	0.87	195
Macroaverage	0.83	0.73	0.77	195
Weighted average	0.87	0.87	0.86	195
Two-class				
High	0.82	0.54	0.65	26
Low	0.93	0.98	0.96	169
Accuracy	—	—	0.92	195
Macroaverage	0.88	0.76	0.80	195
Weighted average	0.92	0.92	0.92	195

[a]LSTM: long short-term memory.

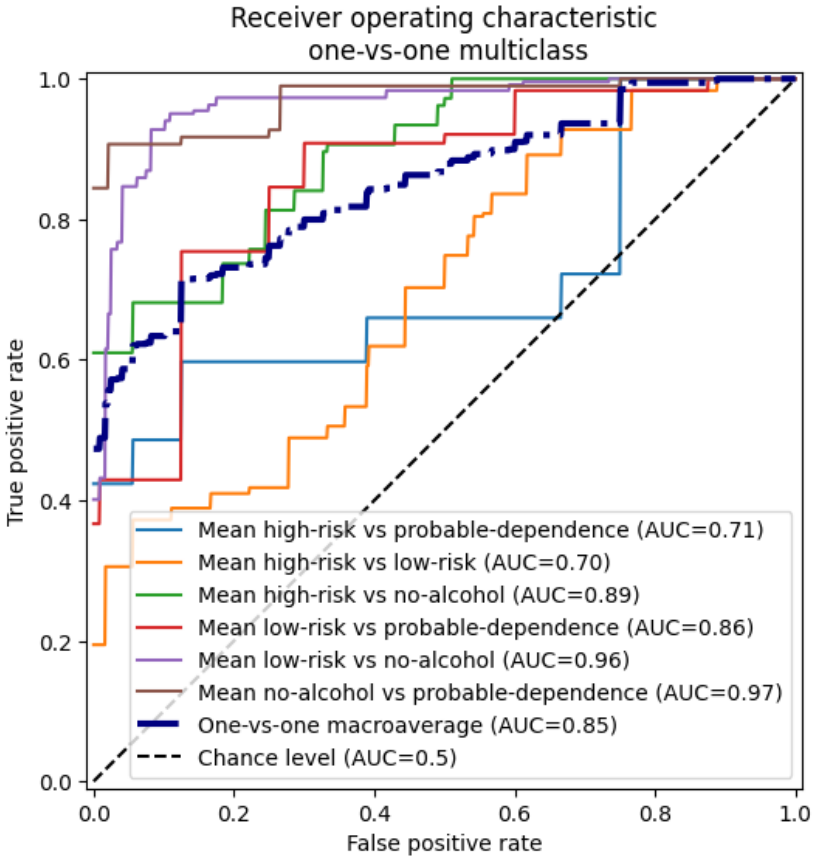
[b]Not applicable.

The bidirectional LSTM was the best-performing classifier among those we tested for patient classification. Complete classification reports and confusion matrices for all models reported in this study are provided in Multimedia Appendix 4.

Figures 2 and 3 present one-vs-one (OvO) and one-vs-rest (OvR) multiclass receiver operating characteristic (ROC) curves for the final LSTM classifier. The microaverage and macroaverage OvR area under the curve (AUC) scores were 0.94 and 0.87, respectively. Per-class OvR AUC scores were highest for no-alcohol (0.97) and probable-dependence (0.90), followed by low-risk (0.88) and high-risk (0.71). The OvO macro-average AUC was 0.85. Pairwise OvO analysis revealed that the lowest-performing class boundaries involved high-risk, both against low-risk (AUC=0.70) and against probable-dependence (AUC=0.71), while boundaries between other class pairs were substantially higher, ranging from 0.86 to 0.97. Due to the small number of validation cases in the minority classes, ROC curves for probable-dependence and high-risk should be interpreted with caution, as each step in the curve represents 1 or 2 patients.

XSL•FO
RenderX

Figure 2. One-vs-one receiver operating characteristic curves for the 4-class model imputing alcohol-use risk in a population of surgical patients drawn from the Michigan Genomics Initiative, 2012 to 2019. AUC: area under the curve.



XSL•FO
RenderX

Figure 3. One-vs-rest receiver operating characteristic (ROC) curves for the 4-class model imputing alcohol-use risk in a population of surgical patients drawn from the Michigan Genomics Initiative, 2012 to 2019. AUC: area under the curve.

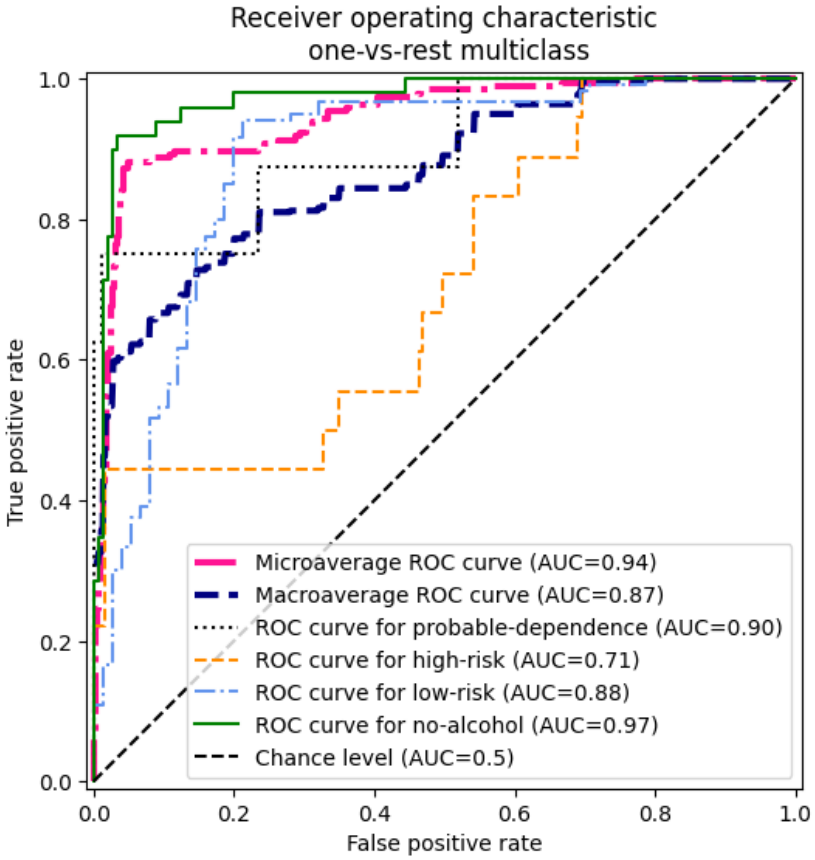
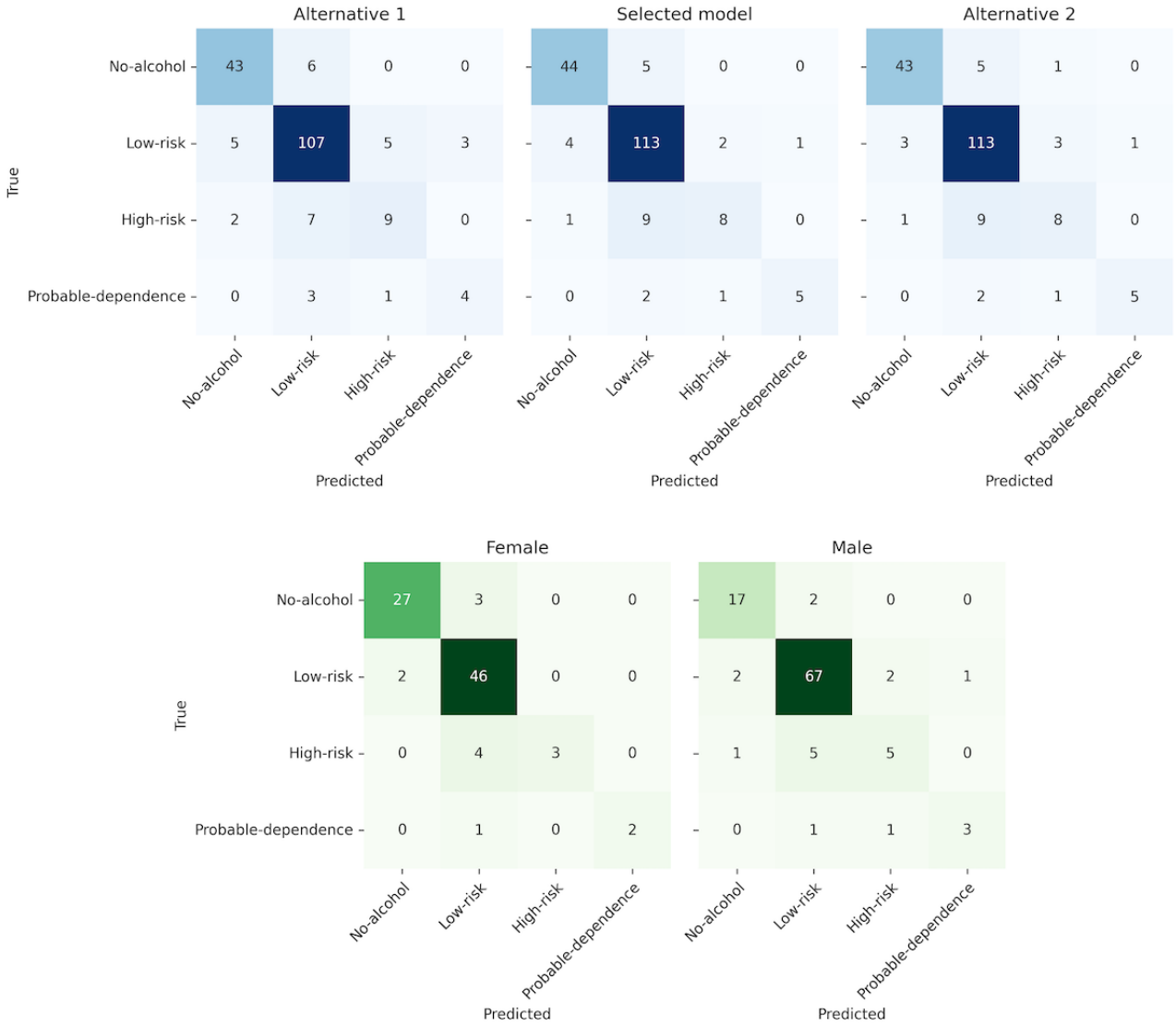


Figure 4 shows confusion matrices for the final selected model ("good-sweep-27"), as well as 2 other alternate candidate LSTM models. Although the final candidate models' performances were very similar (alternative 1 ["daily-sweep-51"]: macro-F_1-score 0.71; alternative 2 ["prime-sweep-56"]: macro-F_1-score 0.76), we selected our model based on the overall number of correct predictions as well as good off-diagonal performance for the high-risk class. Below these, we show the results for the selected model, split by the participant's biological sex. Macroaverage F_1-score is 0.80 for female patients and 0.74 for male patients.

XSL•FO
RenderX

Figure 4. Confusion matrices for 3 finalist models (alternative 1: “daily-sweep-51” F1-score 0.71; alternative 2: “prime-sweep-56” F1-score 0.76; selected model: “good-sweep-27” F1-score 0.77), as well as the final model split by sex over the validation set (n=195).



Alternative Model Architecture Comparisons

We also considered alternative approaches to building a patient classifier based on auto-annotator output. The CNN and RNN architectures did not approach the F_1-scores of the LSTM model on the validation set. The overall macroaverage F_1-scores of the CNN and RNN models were 0.55 and 0.63, respectively, in contrast to the LSTM’s score of 0.77. We believe the LSTM’s design, which assumes sequenced data, was able to better interpret evidence from auto-annotations as patient alcohol use changed over time.

Discussion

Principal Findings

Risky alcohol use and AUD are underdiagnosed in structured EHR data but are more frequently noted in clinical text. This important text information is not easily discovered by busy clinicians within a large body of clinical notes. However, this information is clinically important to patients’ surgical planning, recovery, health, and safety. Classifying risky alcohol use is an inherently imbalanced learning problem in which the classes of primary interest are a small minority of the cases available for training. It is difficult to amass enough training data to correctly infer its presence in notes. Our multistage pipeline uses the large volume of intermediate annotations produced by human labelers to enrich the dataset with better features that assist a final classifier.

Four-class labeling is complicated by sex-specific clinical rules and by the small number of examples of high-risk alcohol use within the dataset. When this model’s output is rolled up into a binary high/low risk classification, the pipeline identifies patients with risky alcohol use or AUD with 92% accuracy, high-risk sensitivity 0.54 (95% CI 0.35-0.71), and specificity 0.98 (95% CI 0.95-0.99; CIs calculated via the Wilson method [46,47]). Because the high-risk group comprises 13% of the validation set, accuracy is dominated by correct classification of low-risk patients. While the 4-class classification presented in this paper provides a more nuanced understanding of patients with AUD, the binary high/low risk classification is helpful when the classifier is intended to provide clinicians with a chart-based indicator for risky alcohol use prior to surgery.

XSL•FO
RenderX

Even with 1000 labeled patients, there are very few examples of high-risk and probable-dependent cases—even fewer when the dataset is split by patient sex. While developing this pipeline, we generated feature sets for notes that included a one-hot-encoded array of medspaCy-detected entities, risk scores calculated based on the entities detected, rulesets based on the detected entities, and auto-annotations generated from the severity suffix of annotations. None of these variations was able to support models that approached the performance of our final classifier. Instead of focusing on the final label, we opted to use more-plentiful human-generated evidence from the annotation process to build a richer set of features to aid classification.

In our model selection process, we opted to maximize the macro-average (vs the weighted average) so the performance on instances from infrequent classes was not lost among those from the more frequent classes. When models being evaluated together had similar macroaverage F_1-scores, we prioritized recall over precision in the low-count, high-risk classes, as recall has greater clinical significance. In the clinical context, a false positive would have less clinical risk than a false negative, given that a positive result is intended to lead to additional alcohol screening and a conversation between the clinician and their patient, but a false negative could miss a critical surgical risk factor [7,15].

We also split and evaluated the validation set by sex at birth for each model being considered, as we had noted in early trials that models tended to score all patients as if they were male. We had considered training separate models per sex but were concerned about losing even more signal when the rare classes were further subdivided. The final pipeline, a single auto-annotator and a single classification model, performs relatively well for both sexes without needing explicit sex-specific instruction. We believe the information is coming from the auto-annotator features, which we think may have identified a pattern linking pronouns, alcohol quantity references, and the annotation labels.

Comparison With Prior Work

Direct comparison between this pipeline and other approaches is difficult because our primary task was extracting 4, rather than 2 (positive and negative), classes of alcohol use risk and because of the diversity of reporting metrics. While we present results for a binary equivalent in Table 4 for the sake of comparison, our work primarily contributes to the field with its graded, 4-class output and our auto-annotation strategy that provides more information to a classifier in a limited, unbalanced dataset.

XSL•FO
RenderX

Table 4. Outcomes of NLP[a]-driven experiments that identify risky alcohol use or AUD[b] in clinical notes.

Study	Task/outcome	Dataset	Model type	Key reported metrics
This study	Four-class alcohol-use risk	1000 patients; 64,703 notes annotated; and held-out test, n=195 (26 positive)	BioClinicalBERT + LSTM[c] pipeline	Four-class, scored as binary: high-risk or probable-dependence vs the rest)—AUROC[d] 0.77 (95% CI 0.66-0.87), sensitivity 0.54, and specificity 0.98
Afshar et al [23]; trauma AUDIT[e] NLP	Binary alcohol misuse vs no misuse	1422 patients; 91,045 notes	ML[f] NLP classifier	AUROC=0.78 (95% CI 0.72-0.85), sensitivity 0.56, specificity 0.89, and rule-based keyword sensitivity of 0.18
To et al [29]; hospitalized validation	Alcohol misuse per NIAAA[g]	1000 hospitalized patients	Prior NLP classifier	AUROC=0.91 (95% CI 0.90-0.93), sensitivity 0.88, specificity 0.78, positive predictive value 0.85, and negative predictive value 0.82
Lin et al [28]; external validation unhealthy alcohol use	Unhealthy alcohol use (AUDIT)	57,605 encounters	Prior NLP classifier	AUROC=0.91, AUPRC[h] 0.56, sensitivity 0.66, specificity 0.98, positive predictive value 0.35, and negative predictive value 0.99 at a cut point of 0.5
Vydiswaran et al [22]; preoperation risky use, rule-based NLP	Risky alcohol use presurgery	53,629 notes; 100 manually labeled	Rule-based NLP	Sensitivity 0.87 vs *ICD*[i] 0.29 and specificity 0.84 vs *ICD* 0.90; NLP identified 3 times as many risky users as *ICD*
Mercurio et al [48]; ClinicalBERT AUD[j]	AUD from discharge summaries	MIMIC-III[k] (size not reported)	ClinicalBERT	Precision 0.90, recall 0.65, F_1-score 0.75, AUROC=0.97, AUPRC 0.86, approximately 4% overall underdocumentation, and up to 30% extra AUD vs *ICD* at high confidence
Afshar et al [27]; SMART-AI	Alcohol, opioid, and nonopioid misuse screening	Training, n=54,915; temporal validation, n=16,917; and external validation, n=1991	Multilabel CNN[l]	Temporal AUROC=0.97 (all misuse), external AUROC alcohol 0.88, opioid 0.94, and false-negative rates of 0.18-0.19 and false-positive rate of 0.03 across key racial groups

[a]NLP: natural language processing.

[b]AUD: alcohol use disorder.

[c]LSTM: long short-term memory.

[d]AUROC: area under the receiver operating characteristic curve.

[e]AUDIT: Alcohol Use Disorders Identification Test.

[f]ML: machine learning.

[g]NIAAA: National Institute on Alcohol Abuse and Alcoholism.

[h]AUPRC: area under the precision-recall curve.

[i]ICD: International Classification of Diseases.

[j]AUD: alcohol use disorder.

[k]MIMIC-III: Medical Information Mart for Intensive Care III.

[l]CNN: convolutional neural network.

Error Analysis

By design, the bidirectional LSTM's sequential approach mimics the way clinical evidence can grow and change over time. On deeper analysis of the errors made by the model, we noticed that for patients who had recently stopped drinking, their past behavior was often weighted more than a recent period of sobriety. This may be due to a very low number of examples of sobriety in the training data. Additional training and experimentation are needed to test whether biasing the model to favor recent information leads to better models overall.

This algorithm tends to downgrade high-risk patients to "low-risk," but this is offset by its much better performance in this aspect than other models we evaluated. The selected model misclassified 2 of the 8 "probable-dependence" validation cases as "low-risk." We chose to accept this performance trade-off because of the higher likelihood that a "probable-dependent" patient will have other indicators of AUD in the clinical record, such as diagnostic codes or alcohol-related comorbidities.

The issues in the high-risk class are also apparent in the ROC analysis: high-risk confusion occurs at both the high-low boundary as well as at the high-dependent boundary. We hypothesize that the difficulties between high- and low-risk arise in part from the difference in rules for male and female patients. Although this pipeline is about as performant for each sex, our decision to stay with a single model is probably responsible for some of these errors.

The model's difficulties with the more severe classes mirror the challenges experienced by the human annotators. Agreement between annotators was as low as 67.5% for the probable-dependent cases, and even after 1 round of conferring, 10% of the cases that were eventually labeled

XSL•FO
RenderX

"probable-dependence" had to be arbitrated by a third annotator. This underscores the challenge of the alcohol risk classification task using clinical text that can be unclear and ambiguous.

Limitations

The model has several limitations. First, it was trained exclusively on notes generated during clinical care in one health system. As local dialects, templates, documentation styles, and writing patterns change between institutions and communities, and with the introduction of writing assistants in documenting clinical encounters, the patterns distinguished by the auto-annotator over these notes may need to be adapted periodically and when applied in a different institutional setting. Some model elements cannot be released publicly due to privacy limitations.

Second, the 2-stage approach introduces the possibility of error accumulation; misclassification in the auto-annotation stage could propagate to the final classifier. Our functional equivalence validation partially addresses this concern, but a systematic error-case analysis was beyond the scope of this study and is identified as future work.

Third, the auto-annotator was trained without explicit class-balancing techniques such as weighted loss or undersampling of negative sections. Given the approximately 20:1 ratio of unannotated to annotated sections, this represents a limitation of the current implementation. However, the functional equivalence validation, which showed that downstream classification performance was comparable whether trained on auto-annotations or human annotations, provides evidence that the auto-annotator produced useful signal despite this imbalance.

Additionally, the pipeline assumes that initial alcohol-related entity detection had acceptable recall, as the ruleset used to detect alcohol entities was derived from an algorithm developed in an earlier experiment [22]. We note that a coding error (retained in Multimedia Appendix 3 for reference) prevented detection of the tokens "none," "abstained," "abstinent," "wks," and "quit" in a rule designed to capture a period of sobriety. Rules to capture a period of time since a last drink, or time since a person quit, were affected by failure to detect "wks" and "week."

Finally, the performance estimates for the probable-dependent class (n=8 in the validation set) are based on a very small sample and should be interpreted with caution; a change of 1 or 2 classifications meaningfully shifts per-class and macroaverage scores. The OvR AUC of 0.90 for this class may show that the pipeline can detect signals of alcohol dependency.

Future Directions

This model is currently being validated in a prospective study of presurgical patients who have provided detailed information about their alcohol intake and AUD symptoms. To determine the applicability of this model in clinical practice, we will also evaluate the use of the model as a predictor of adverse postsurgical outcomes across a large cohort of patients and their clinical notes.

Large language models have moved to the forefront of many NLP tasks. Our preliminary investigation found that the LLMs available had the following issues: foremost, at the time, no HIPAA-compliant LLM was offered at our institution. The costs of processing gigabytes of clinical notes are also a significant issue: the filtering part of this pipeline would need to stay in place. Finally, we did trials with deidentified data and found that even with careful prompting, the LLM tested was prone to applying creative interpretations of our classification rules. LLMs should be investigated more thoroughly for this task, however; a general-purpose model with strict instructions could be a better solution than a relatively brittle, dedicated pipeline.

Conclusion

In this paper, we present a 2-stage NLP pipeline that identifies ordinal classes of risky alcohol use in presurgical patients. To address the scarcity of labeled data for high-risk and high-importance classes, we included a BioClinicalBERT-based auto-annotator that mimics human annotators in identifying key evidence from clinical notes and outputs probabilities across 26 dimensions of alcohol use. These features are then fed into an LSTM model to determine the patient-level alcohol use risk in four categories—no-alcohol, low-risk, high-risk, and probable-dependence. The pipeline was able to separate patient records into 4 strata of alcohol-use risk, with "high-risk" patients being most likely to be misclassified into an adjacent class.

Data Availability

The datasets analyzed and models generated are not publicly available due to the presence of personally identifying information in the text that was processed but are available from the corresponding author on reasonable request.

Funding

Research reported in this publication was supported by the National Institute on Alcohol Abuse and Alcoholism of the National Institutes of Health (award number R33AA028315). The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

Generative AI Use

The authors declare that generative AI was not used in the preparation of this manuscript.

XSL•FO
RenderX

Authors' Contributions

VGVV and ACF are joint senior authors of the study and principal investigators of the grant that funded this research. VGVV and ACF designed the study protocol, developed annotation guidelines, and directed the annotation process. KW developed the analysis pipeline and deep learning models and analyzed the results. AS helped with data processing, developing preliminary models, and annotations. XJ helped with data analysis and discussion. JLM and GSW advised on annotation guidelines and validating the approaches for clinical significance. ACF led the overall project management and alcohol clinical expertise, and VGVV led the model development efforts. KW wrote the first draft of the manuscript, and all authors contributed to revising and editing the manuscript.

Conflicts of Interest

GSW consults for GlaxoSmithKline and Novo Nordisk. JM consults for GlaxoSmithKline, Novo Nordisk, and Boehringer Ingelheim. No other authors have any conflicts of interest to disclose at this time.

Multimedia Appendix 1

Annotation guidelines.
[DOCX File , 6019 KB-Multimedia Appendix 1]

Multimedia Appendix 2

Categories for annotating patient records containing evidence of alcohol use, also serving as category labels for the auto-annotator component of the pipeline.
[DOCX File , 16 KB-Multimedia Appendix 2]

Multimedia Appendix 3

Python script containing rules for identifying and preprocessing alcohol-related information in clinical notes.
[DOCX File , 21 KB-Multimedia Appendix 3]

Multimedia Appendix 4

Raw analysis output from the pipeline that applies rules, auto-annotation, and a long short-term memory (LSTM) classifier to patient notes to detect risky alcohol use.
[DOCX File , 22 KB-Multimedia Appendix 4]

References

1. Key Substance Use and Mental Health Indicators in the United States: Results From the 2016 National Survey on Drug Use and Health. Substance Abuse and Mental Health Services Administration (SAMHSA). URL: https://www.samhsa.gov/data/report/key-substance-use-and-mental-health-indicators-united-states-results-2016-national-survey [accessed 2023-03-13]
2. Drinking Levels Defined. National Institute on Alcohol Abuse and Alcoholism. URL: https://www.niaaa.nih.gov/alcohol-health/overview-alcohol-consumption/moderate-binge-drinking [accessed 2024-02-21]
3. Grant BF. Prevalence and correlates of alcohol use and DSM-IV alcohol dependence in the United States: results of the National Longitudinal Alcohol Epidemiologic Survey. J Stud Alcohol. 1997;58(5):464-473. [doi: 10.15288/jsa.1997.58.464] [Medline: 9273910]
4. Lechner WV, Sidhu NK, Jin JT, Kittaneh AA, Laurene KR, Kenne DR. Increases in risky drinking during the COVID-19 pandemic assessed via longitudinal cohort design: associations with racial tensions, financial distress, psychological distress and virus-related fears. Alcohol Alcohol. 2021;56(6):702-707. [FREE Full text] [doi: 10.1093/alcalc/agab019] [Medline: 33765131]
5. Fernandez AC, Waljee JF, Gunaseelan V, Brummett CM, Englesbe MJ, Bicket MC. Prevalence of unhealthy substance use and associated characteristics among patients presenting for surgery. Ann Surg. 2023;278(4):e740-e744. [doi: 10.1097/SLA.0000000000005767] [Medline: 36538617]
6. Harris AHS, Frey MS, Debenedetti A, Bradley K. Alcohol Misuse Prevalence and Associations with Post-Operative Complications in US Surgical Patients: A Review. Open J Surg. 2008;2:50-58. [FREE Full text] [doi: 10.2174/1874300500802010050]
7. Eliasen M, Grønkjær M, Skov-Ettrup LS, Mikkelsen SS, Becker U, Tolstrup JS, et al. Preoperative alcohol consumption and postoperative complications: a systematic review and meta-analysis. Ann Surg. 2013;258(6):930-942. [doi: 10.1097/SLA.0b013e3182988d59] [Medline: 23732268]
8. Harris AHS, Reeder R, Ellerbe L, Bradley KA, Rubinsky AD, Giori NJ. Preoperative alcohol screening scores: association with complications in men undergoing total joint arthroplasty. J Bone Joint Surg Am. 2011;93(4):321-327. [doi: 10.2106/JBJS.I.01560] [Medline: 21325583]

XSL•FO
RenderX

9. Bradley KA, Rubinsky AD, Sun H, Bryson CL, Bishop MJ, Blough DK, et al. Alcohol screening and risk of postoperative complications in male VA patients undergoing major non-cardiac surgery. J Gen Intern Med. 2011;26(2):162-169. [FREE Full text] [doi: 10.1007/s11606-010-1475-x] [Medline: 20878363]
10. Rubinsky AD, Bishop MJ, Maynard C, Henderson WG, Hawn MT, Harris AH, et al. Postoperative risks associated with alcohol screening depend on documented drinking at the time of surgery. Drug Alcohol Depend. 2013;132(3):521-527. [doi: 10.1016/j.drugalcdep.2013.03.022] [Medline: 23683792]
11. Tønnesen H, Petersen K, Højgaard L, Stokholm KH, Nielsen HJ, Knigge U, et al. Postoperative morbidity among symptom-free alcohol misusers. Lancet. 1992;340(8815):334-337. [doi: 10.1016/0140-6736(92)91405-w] [Medline: 1353805]
12. Oppedal K, Møller AM, Pedersen B, Tønnesen H. Preoperative alcohol cessation prior to elective surgery. Cochrane Database Syst Rev. 2012;(7):CD008343. [doi: 10.1002/14651858.CD008343.pub2] [Medline: 22786514]
13. Lowery EM, Yong M, Cohen A, Joyce C, Kovacs EJ. Recent alcohol use prolongs hospital length of stay following lung transplant. Clin Transplant. 2018;32(6):e13250. [FREE Full text] [doi: 10.1111/ctr.13250] [Medline: 29620796]
14. Spies CD, Nordmann A, Brummer G, Marks C, Conrad C, Berger G, et al. Intensive care unit stay is prolonged in chronic alcoholic men following tumor resection of the upper digestive tract. Acta Anaesthesiol Scand. 1996;40(6):649-656. [doi: 10.1111/j.1399-6576.1996.tb04505.x] [Medline: 8836256]
15. Fernandez AC, Bohnert KM, Bicket MC, Weng W, Singh K, Englesbe M. Adverse surgical outcomes linked to co-occurring smoking and risky alcohol use among general surgery patients. Ann Surg. 2023;278(2):201-207. [doi: 10.1097/SLA.0000000000005735] [Medline: 36268706]
16. Shourie S, Conigrave KM, Proude EM, Ward JE, Wutzke SE, Haber PS. Pre-operative screening for excessive alcohol consumption among patients scheduled for elective surgery. Drug Alcohol Rev. 2007;26(2):119-125. [doi: 10.1080/09595230601146595] [Medline: 17364846]
17. Fernandez A. Alcohol-Related Clinical Practices in Surgical Healthcare: Qualitative Findings From Patients and Healthcare Providers. Stanford University. May 06, 2019. URL: https://med.stanford.edu/content/dam/sm/s-spire/documents/WIP/WIP_5.6.19_Alcohol-Use-and-Surgical-Health-Optimization_Fernandez.pdf [accessed 2026-10-06]
18. Cooley S, Bicket MC, Mohammed H, Lai Y, Evilsizer S, Brummett CM, et al. Worth the risk? Standardized screening to identify substance use among patients before surgery. Ann Surg. 2025;282(6):991-997. [doi: 10.1097/SLA.0000000000006358] [Medline: 38814073]
19. Kip MJ, Neumann T, Jugel C, Kleinwaechter R, Weiss-Gerlach E, Guill MM, et al. New strategies to detect alcohol use disorders in the preoperative assessment clinic of a German university hospital. Anesthesiology. 2008;109(2):171-179. [doi: 10.1097/ALN.0b013e31817f5be3] [Medline: 18648225]
20. Chiang PP. Perioperative management of the alcohol-dependent patient. Am Fam Physician. 1995;52(8):2267-2273. [Medline: 7484720]
21. Tønnesen H, Møller AM. The smoking and drinking patient. Ugeskr Laeger. 2006;168(49):4293-4296. [Medline: 17164055]
22. Vydiswaran V, Strayhorn A, Weber K, Stevens H, Mellinger J, Winder G, et al. Automated-detection of risky alcohol use prior to surgery using natural language processing. Alcohol Clin Exp Res (Hoboken). Jan 2024;48(1):153-163. [FREE Full text] [doi: 10.1111/acer.15222] [Medline: 38189663]
23. 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. 2019;26(3):254-261. [FREE Full text] [doi: 10.1093/jamia/ocy166] [Medline: 30602031]
24. Chen E, Garcia-Webb M. An analysis of free-text alcohol use documentation in the electronic health record: early findings and implications. Appl Clin Inform. 2014;5(2):402-415. [FREE Full text] [doi: 10.4338/ACI-2013-12-RA-0101] [Medline: 25024757]
25. Hao T, Huang Z, Liang L, Weng H, Tang B. Health natural language processing: methodology development and applications. JMIR Med Inform. 2021;9(10):e23898. [FREE Full text] [doi: 10.2196/23898] [Medline: 34673533]
26. Liao KP, Cai T, Savova GK, Murphy SN, Karlson EW, Ananthakrishnan AN, et al. Development of phenotype algorithms using electronic medical records and incorporating natural language processing. BMJ. 2015;350:h1885. [FREE Full text] [doi: 10.1136/bmj.h1885] [Medline: 25911572]
27. Afshar M, Sharma B, Dligach D, Oguss M, Brown R, Chhabra N, 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(6):e426-e435. [FREE Full text] [doi: 10.1016/S2589-7500(22)00041-3] [Medline: 35623797]
28. 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(4):925-933. [FREE Full text] [doi: 10.1111/add.15730] [Medline: 34729829]
29. 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. [FREE Full text] [doi: 10.1016/j.alcohol.2019.09.008] [Medline: 31574300]

XSL•FO
RenderX

30. Vydiswaran VGV, Strayhorn A, Zhao X, Robinson P, Agarwal M, Bagazinski E, et al. Hybrid bag of approaches to characterize selection criteria for cohort identification. J Am Med Inform Assoc. 2019;26(11):1172-1180. [FREE Full text] [doi: 10.1093/jamia/ocz079] [Medline: 31197354]
31. Miranda O, Fan P, Qi X, Wang H, Brannock MD, Kosten TR, et al. DeepBiomarker2: prediction of alcohol and substance use disorder risk in post-traumatic stress disorder patients using electronic medical records and multiple social determinants of health. J Pers Med. 2024;14(1):94. [FREE Full text] [doi: 10.3390/jpm14010094] [Medline: 38248795]
32. Savova GK, Masanz JJ, Ogren PV, Zheng J, Sohn S, Kipper-Schuler KC, et al. Mayo clinical text analysis and knowledge extraction system (cTAKES): architecture, component evaluation and applications. J Am Med Inform Assoc. 2010;17(5):507-513. [FREE Full text] [doi: 10.1136/jamia.2009.001560] [Medline: 20819853]
33. Devlin J, Chang M, Lee K, Toutanova K. BERT: pre-training of deep bidirectional transformers for language understanding. ArXiv. Preprint posted online on October 11, 2018. [doi: 10.48550/arXiv.1810.04805]
34. Cer D, Yang Y, Kong S, Hua N, Limtiaco N, Constant N, et al. Universal Sentence Encoder. ArXiv. Preprint posted online on March 29, 2018. [FREE Full text] [doi: 10.48550/arXiv.1803.11175]
35. Alsentzer E, Murphy J, Boag W, Weng W, Jindi D, Naumann T, et al. Publicly available clinical BERT embeddings. 2019. Presented at: Proceedings of the 2nd Clinical Natural Language Processing Workshop; June 7, 2019; Minneapolis, Minnesota. [doi: 10.18653/v1/w19-1909]
36. Kim HK, Park Y, Park Y, Choi E, Kim S, You H, et al. Identifying alcohol-related information from unstructured bilingual clinical notes with multilingual transformers. IEEE Access. 2023;11:16066-16075. [doi: 10.1109/access.2023.3245523]
37. Khandve SI, Wagh VK, Wani AD, Joshi IM, Joshi RB. Hierarchical neural network approaches for long document classification. USA.; 2022. Presented at: Proceedings of the 2022 14th International Conference on Machine Learning and Computing (ICMLC '22); February 18-21, 2022:115-119; Guangzhou China. [doi: 10.1145/3529836.3529935]
38. American Psychiatric Association. Diagnostic and Statistical Manual of Mental Disorders. 5th Edition. Arlington, Virginia. American Psychiatric Publishing; 2013.
39. Zawistowski M, Fritsche LG, Pandit A, Vanderwerff B, Patil S, Schmidt EM, et al. The Michigan Genomics Initiative: a biobank linking genotypes and electronic clinical records in Michigan Medicine patients. Cell Genom. 2023;3(2):100257. [FREE Full text] [doi: 10.1016/j.xgen.2023.100257] [Medline: 36819667]
40. Stenetorp P, Pyysalo S, Topi? G, Ohta T, Ananiadou S, Tsujii J. brat: a web-based tool for NLP-assisted text annotation. 2012. Presented at: Proceedings of the Demonstrations at the 13th Conference of the European Chapter of the Association for Computational Linguistics; April 23-27, 2012:102-107; Avignon, France.
41. Understanding alcohol drinking patterns. National Institute on Alcohol Abuse and Alcoholism. 2026. URL: https://www.niaaa.nih.gov/alcohols-effects-health/alcohol-drinking-patterns [accessed 2026-05-26]
42. Rajapakse N, Rubenzahl E, Saripella A, Lee JW, Mansouri S, Fan S, et al. Preoperative alcohol use disorders and adverse outcomes in surgical patients: a systematic review and meta-analysis. J Clin Anesth. 2026;111:112194. [FREE Full text] [doi: 10.1016/j.jclinane.2026.112194] [Medline: 41922128]
43. Eyre H, Chapman AB, Peterson KS, Shi J, Alba PR, Jones MM, et al. Launching into clinical space with medspaCy: a new clinical text processing toolkit in Python. AMIA Annu Symp Proc. 2021;2021:438-447. [FREE Full text] [Medline: 35308962]
44. Paszke A, Gross S, Massa F, Lerer A, Bradbury J, Chanan G, et al. Pytorch: an imperative style, high-performance deep learning library. 2019. Presented at: Advances in Neural Information Processing Systems 32 (NeurIPS 2019); December 8-14, 2019; Vancouver, British Columbia. URL: https://proceedings.neurips.cc/paper_files/paper/2019/hash/bdbca288fee7f92f2bfa9f7012727740-Abstract.html
45. Biewald L. Experiment tracking with Weights and Biases. Weights & Biases. 2020. URL: https://wandb.com [accessed 2026-10-06]
46. Wilson EB. Probable inference, the law of succession, and statistical inference. Journal of the American Statistical Association. 1927;22(158):209-212. [doi: 10.1080/01621459.1927.10502953]
47. Newcombe RG. Two-sided confidence intervals for the single proportion: comparison of seven methods. Stat Med. 1998;17(8):857-872. [doi: 10.1002/(sici)1097-0258(19980430)17:8<857::aid-sim777>3.0.co;2-e] [Medline: 9595616]
48. Mercurio L, Garcia A, Ruest S, Duffy SJ, Eickhoff C. One third of alcohol use disorder diagnoses are missed by ICD coding. Subst Use Addctn J. 2025;46(2):328-336. [doi: 10.1177/29767342241288112] [Medline: 39511779]

Abbreviations

AUC: area under the curve
AUD: alcohol use disorder
AUROC: area under the receiver operating characteristic curve
BERT: Bidirectional Encoder Representations from Transformers
CNN: convolutional neural network
CUI: concept unique identifier
DSM-5: Diagnostic and Statistical Manual of Mental Disorders (Fifth Edition)

XSL•FO
RenderX

EHR: electronic health record
HIPAA: Health Insurance Portability and Accountability Act
KNN: k-nearest neighbors
LSTM: long short-term memory
MGI: Michigan Genomic Initiative
NLP: natural language processing
OvO: one-vs-one
OvR: one-vs-rest
RNN: recurrent neural network
ROC: receiver operating characteristic
SGD: stochastic gradient descent
SVC: support vector classifier
TLSTM: time-aware long short-term memory

Edited by A Coristine; submitted 14.Jan.2026; peer-reviewed by V Jawa, M Elbattah, S Zeng; comments to author 13.Apr.2026; revised version received 23.Sep.2026; accepted 24.Sep.2026; published 20.Nov.2026

Please cite as:
Weber K, Strayhorn A, Ju X, Mellinger J, Winder GS, Fernandez AC, Vydiswaran VGV
Natural Language Processing Supported by Annotation Process Artifacts to Detect Risky Alcohol Use in Electronic Health Records: Algorithm Development and Validation Study
J Med Internet Res 2026;28:e91449
URL: https://www.jmir.org/2026/1/e91449
doi: 10.2196/91449
PMID:

©Katherine Weber, Asher Strayhorn, Xintong Ju, Jessica Mellinger, Gerald Scott Winder, Anne C Fernandez, VG Vinod Vydiswaran. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 20.Nov.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX