# Using natural language processing to identify opioid use disorder in electronic health record data

Jade Singleton [a,b,*], Chengxi Li [c], Peter D. Akpunonu [d], Erin L. Abner [a], Anna M. Kucharska-Newton [a,e]

[a] *Department of Epidemiology, College of Public Health, University of Kentucky, Lexington, KY 40536, United States*
[b] *University of Kentucky Healthcare IT Department, Business Intelligence, Lexington, KY 40517, United States*
[c] *Department of Computer Science, College of Engineering, University of Kentucky, Lexington, KY 40526, United States*
[d] *Emergency Medicine & Medical Toxicology, University of Kentucky Hospital, Lexington, KY 40536, United States*
[e] *Department of Epidemiology, The Gillings School of Global Public Health, University of North Carolina at Chapel Hill, Chapel Hill, NC 27514, United States*

A R T I C L E I N F O

*Keywords:*
Opioid use disorder
Natural language processing
Electronic healthcare records

A B S T R A C T

*Background:* As opioid prescriptions have risen, there has also been an increase in opioid use disorder (OUD) and its adverse outcomes. Accurate and complete epidemiologic surveillance of OUD, to inform prevention strategies, presents challenges. The objective of this study was to ascertain prevalence of OUD using two methods to identify OUD in electronic health records (EHR): applying natural language processing (NLP) for text mining of unstructured clinical notes and using ICD-10-CM diagnostic codes.
*Methods:* Data were drawn from EHR records for hospital and emergency department patient visits to a large regional academic medical center from 2017 to 2019. International Classification of Disease, 10th Edition, Clinic Modification (ICD-10-CM) discharge codes were extracted for each visit. To develop the rule-based NLP algorithm, a stepwise process was used. First, a small sample of visits from 2017 was used to develop initial dictionaries. Next, EHR corresponding to 30,124 visits from 2018 were used to develop and evaluate the rule-based algorithm. A random sample of the results were manually reviewed to identify and address shortcomings in the algorithm, and to estimate sensitivity and specificity of the two methods of ascertainment. Last, the final algorithm was then applied to 29,212 visits from 2019 to estimate OUD prevalence.
*Results:* While there was substantial overlap in the identified records (n = 1,381 [59.2 %]), overall n = 2,332 unique visits were identified. Of the total unique visits, 430 (18.4 %) were identified only by ICD-10-CM codes, and 521 (22.3 %) were identified only by NLP. The prevalence of visits with evidence of an OUD diagnosis in this sample, ascertained using only ICD-10-CM codes, was 1,811/29,212 (6.1 %). Including the additional 521 visits identified only by NLP, the estimated prevalence of OUD is 2,332/29,212 (7.9 %), an increase of 29.5 % compared to the use of ICD-10-CM codes alone. The estimated sensitivity and specificity of the NLP-based OUD classification were 81.8 % and 97.5 %, respectively, relative to gold-standard manual review by an expert addiction medicine physician.
*Conclusion:* NLP-based algorithms can automate data extraction and identify evidence of opioid use disorder from unstructured electronic healthcare records. The most complete ascertainment of OUD in EHR was combined NLP with ICD-10-CM codes. NLP should be considered for epidemiological studies involving EHR data.

## 1. Introduction

### 1.1. Background

Electronic health records (EHR) are a rich source of data that can be leveraged to inform strategies for measuring and addressing the ongoing opioid crisis in the United States [23,26]. Accurate and timely identification of patients with opioid use disorder (OUD) is an important step in any such effort. International Classification of Diseases (ICD) codes are commonly used for this purpose due to their widespread use in medical record coding and their accessibility to researchers [2,11]. The limitations of ICD codes, including low sensitivity and specificity for many

* Corresponding author.
*E-mail address:* jade.singleton@seattlechildrens.org (J. Singleton).

https://doi.org/10.1016/j.ijmedinf.2022.104963
Received 27 June 2022; Received in revised form 25 October 2022; Accepted 31 October 2022
Available online 10 December 2022

conditions, have been well-documented [9,14,19].

Several authors have investigated the strengths and limitations of ICD codes for identifying opioid use disorder (OUD) in EHRs and related data systems. Ranapurwala et al conducted an expert review of 166 patient records for 2014 to 2017, from EHRs in four large healthcare systems in the southern United States. They reported sensitivity of 59 % and specificity of 93 % for ICD code-based identification of OUD based on two clinic experts manual review of 166 cases [20]. Palumbo et al analyzed ICD codes for OUD for a sample of 16,253 patients enrolled in a medication monitoring program (MMP) within a single health system from 2001 through 2017 [16]. They also conducted a manual review of medical records to identify DSM-5 criteria for OUD in a randomly selected subset of 200 patients. They reported that 2 % of the 16,253 patients had an ICD code for OUD in their MMP record, whereas the medical record review detected evidence of moderate-to-severe OUD in 73 % of the 200 reviewed patients [16]. Chartash et al developed and tested a computable phenotype aimed at identifying patients with OUD for inclusion in pragmatic clinical trials [4]. The phenotype was based on ICD codes plus terms identified from the chief complaint. For the purposes of phenotyping, their primary concern was positive predictive value (PPV), which addresses the question: "of all patients identified by algorithm as having OUD, what proportion truly have OUD, relative to a gold standard?" This is different from sensitivity, which is defined as the proportion of all true OUD cases in the sample that are identified by the algorithm. The authors reported a PPV of 0.95 (95 % CI 0.85–0.99), indicating that most patients identified as having OUD did, in fact, have OUD. Sensitivity and specificity could not be assessed given their study design.

In addition to ICD codes, EHR contain substantial information in the form of unstructured, narrative text entered by healthcare providers in the course of treatment [24]. These clinical notes include but are not limited to information on patient symptoms, conditions, behaviors, as well as healthcare advice and plans [26]. Generally, information in unstructured notes may include demographics, medical encounters, developmental history, obstetric history, medications and medical allergies, family history, social history, habits, and immunization records [6].

Natural language processing (NLP), a branch of artificial intelligence that is concerned with computer understanding of human languages [1], holds great promise as a tool for extracting information from unstructured textual data in many domains [5,10,17]. Typically, information extraction involves splitting text into units called tokens, which comprise individual words and punctuation marks, etc. [5]. Rule-based NLP approaches to information extraction attempt to identify matches to pre-specified sequences of tokens [12].

Carrell et al investigated the potential to apply NLP to EHR records to increase the identification of problem use of prescription opioids (POU) among patients undergoing chronic opioid therapy [3]. POU was defined as indications of addiction, abuse, misuse or overuse, and is thus more broadly defined and less specific than OUD. The study documented POU between 2006 and 2012 in a sample of 22,142 patients who received chronic opioid therapy, defined as at least a 70-day supply of prescription opioid medications dispensed in a calendar quarter, within a large health plan serving the state of Washington. They used a rule-based approach to identify "mentions" of POU in clinical notes, such as phrases of the form "opioid addiction", "dependence on methadone", or "no evidence of drug abuse". Candidate mentions were then manually validated by trained reviewers. POU prevalence increased under the NLP approach: POU prevalence was 10.1 % based on ICD-9-CM codes alone, and 13.4 % including patients identified by NLP – an increase of nearly 33 %.

In a more recent investigation, Zhu et al studied a patient sample similar to that by Carrell et al (N = 13,654 adult non-cancer patients receiving chronic opioid therapy at Medical University of South Carolina from 2013 to 2018) [28]. Using Linguamatics I2E software, they implemented a similar rule-based approach to identify patients with OUD (rather than POU). They reported that NLP achieved high performance for identifying OUD versus gold standard expert review of test cases (98.5 % PPV), and that including NLP-identified cases increased OUD identification by 40 % over ICD codes alone. Agreement between cases identified by NLP and ICD was modest (Kappa = 0.63).

These previous studies in patients undergoing COT provide valuable insights into the ability of NLP methods to enhance ascertainment of OUD in populations were prevalence of OUD or POU is assumed to be relatively high. However, a majority of patients with OUD who present to emergency departments (ED) and hospitals will not necessarily be undergoing chronic opioid treatment (COT). Therefore, the effectiveness of NLP approaches in broader patient populations where prevalence of OUD is lower remains unknown [7]. Further, in 2015 the United States transitioned to ICD-10-CM for medical coding [15]. Thus, we aimed to conduct a study of the application of NLP to OUD ascertainment in a general patient population where ICD-10-CM codes are used. Drawing on inpatient and ED EHR records (spanning 2017–2019) from the University of Kentucky HealthCare (UKHC) Albert B. Chandler Hospital, we developed and investigated the performance of a rule-based NLP algorithm in the identifications of OUD cases among hospital inpatients.

## 2. Methods

### 2.1. Study population

Data were drawn from all adults (age 18 years and older) inpatient and ED visits occurring at the UKHC Albert B. Chandler Hospital between January 1st, 2017 and December 31st, 2019. Due to high prevalence of opioid use for the treatment of cancer-related pain [27], we excluded visits for patients with active cancer (ICD-10-CM code: C00-C27, C30-C42, C43-C59, C60-C81, C7A.*, C7B.*, C81-C97, D37-D50) [13]. Additionally, we required that patient visits had at least one of the following five types of notes, which we considered most likely to include information pertaining to opioid use disorder: ED triage, ED general, History and Physical, Addiction Medicine Consult, and Discharge Summary notes. Addiction Medicine Consult notes were included because our broadly-defined cohort included patients who presented for reasons completely unrelated to OUD. When OUD is not a primary reason for seeking care, the patient visit may be less likely to receive an ICD code. However, if OUD is suspected, a consult may be ordered. In addition to unstructured provider notes, structured EHR data on patient demographics and diagnosis codes were extracted. This study was approved by both the UK Institutional Review Board (IRB# 20548) and the UKHC Data Management Committee.

### 2.2. ICD-10 definition of OUD

The ICD-10-CM definition for OUD included the codes for opioid abuse (F11.10, F11.11, F11.12, F11.14, F11.18, F11.19), opioid dependence (F11.20, F11.21, F11.22, F11.23, F11.24, F11.25, F11.28, F11.29), and unspecified opioid use (F11.90, F11.92, F11.93, F11.94, F11.95, F11.98, F11.99). For each patient visit in the study sample, the encounter was classified as positive for OUD if any of these diagnosis codes were present.

### 2.3. NLP-based definition of OUD

#### 2.3.1. Overview of algorithm development process

The NLP algorithm was developed in phases: dictionary development (Phase 1), parsing rule and algorithm development (Phase 2), and final classification (Phase 3). In Phase 1, we used information from extant literature to create dictionaries of OUD-related terms. The dictionaries were refined based on advice from an expert in medical toxicology and emergency medicine (author PDA), as well as manual review of notes from 50 randomly selected patient visits occurring in 2017: 25 with OUD identified by ICD-10-CM and 25 without OUD identification by ICD-10-

CM. The developed dictionaries were used to create parsing rules to identify evidence that the patient visit did or did not indicate that a classification of OUD was supported.

In Phase 2, we used data from patient visits occurring in 2018 to develop the algorithm. We applied the initial version of the algorithm to these data to classify each encounter as OUD or non-OUD. Next, to evaluate algorithm performance, we selected a 1 % random sample for review each from the visits NLP classified as OUD and from the visits classified as non-OUD. An expert clinician (PDA) independently reviewed the EHR records for these 300 cases and classified them as OUD or non-OUD, without knowledge the algorithmic classification. The conditions that were taken as evidence of OUD when manually reviewing the cases were refined from a list reported in Carrell [3], and are listed in Table 1. Based on findings from the manual review, we updated the dictionaries and revised the protocol pipeline to optimize performance. In Phase 3, the finalized algorithm was applied to the data set consisting of patient visits occurring in 2019.

### 2.4. Dictionaries

Five dictionaries were constructed (Table 2). To facilitate analysis, our dictionary terms were specified in all lower case (see Table 2), and we transformed the EHR text data into all lower case, because matching text strings is case-dependent. Based on published literature (Carrell 2015) and expert knowledge, we created dictionaries for opioid types (dictionary 1; e.g., "morphine", "narcotic", "oxycodone") and terms suggestive of use disorder (dictionary 2; e.g., "abuse", "dependence", "use disorder"). Next, we queried our training database (2018 patient visits) to refine these dictionaries. For example, we discovered spelling errors (e.g., "depandance" for "dependence"), commonly used abbreviations (e.g., "sub" for "subutex", "od" for "overdose", "vico" for "vicodin"), and other types of nonstandard text. We created a third dictionary consisting of terms (e.g., "denies", "without", "negative") that are used in clinical notes to negate a mention of OUD. The fourth dictionary was created to capture specialized terms used related to drug use. Those terms were found primarily in the social history and related to the negation of opioid drug use, e.g. "IVDA/intranasal: Denies". The fifth dictionary included the name of the UKHC opioid treatment clinic where patients may be referred to following discharge (First Bridge Clinic). We developed six parsing rules representing combinations of dictionary terms (Table 3).

### 2.5. Parsing rules

In the EHR system, each type of clinical note (i.e., ED triage, ED general, History and Physical, Addiction Medicine Consult, and Discharge Summary) has a consistent set of labeled sections. For example, a Discharge Summary note is divided into the following sections: Reason for Hospitalization, Significant Findings, Procedures and Treatment, Patient's Discharge Condition, Patient and Family Instructions, and Attending Physician's Signature. In order to reduce computing time, the algorithm first scanned each Note section for any mention of terms contained in dictionaries 1 and 5. If no mention was found, no further parsing was done. If a mention was found, those sections were processed sequentially by first separating the section into sentences, and then breaking down each sentence into individual tokens. Parsing rules were then used to identify sequences of individual tokens that provide evidence for (or against, in the case of negation rules) the presence of OUD. As an example, Parsing Rule 1 consists of an opioid term (dictionary 1), followed by zero to three other valid tokens, followed by a use disorder term (dictionary 2). Sequences of tokens that would be identified by this rule include, for example, "opioid use disorder", "oxycodone dependence", or "addiction to heroin." The complete list of parsing rules that we used is summarized in Table 3. We allowed for up to 3 intervening, valid tokens between opioid terms and use disorder terms, in order to capture more complex OUD mentions. Piotrkowitz et al provided an example of an application of NLP to oncology, where up to 5 intervening tokens were allowed [18]. The value of the token parameter is application-specific can be tuned by inspecting test cases.

**Table 1**
Conditions indicating appropriate classification as OUD.

| |
|---|
| Admits to opioid use disorder |
| Recent inpatient admission for detox |
| Referral for opioid addiction treatment at the First Bridge clinic |
| Currently receiving methadone or suboxone treatment for opioid addiction |
| Loss of control of opioid, craving |
| Family member reported opioid addiction to clinician |
| Current or recent opioid overdose |
| Obtained opioids from multiple MDs surreptitiously |
| Opioid taper/wean due to problems (not due to expected pain improvement) |
| Unsuccessful taper attempt |
| Physician or patient wants immediate taper |
| Positive response to Narcan treatment |

**Table 2**
Dictionaries of opioid use disorder and negation terms, and additional specialized terms, which were combined via parsing rules to form search phrases.

| Dictionary | Key words |
|---|---|
| 1. Opioid term | fentanyl, heroin, hydromorphone, dilaudid, oxymorphone, opanum, opana, methadone, oxycodone, oxycotin, roxicodone, percocet, morphine, hydrocodone, vicodin, vico, lortab, codeine, meperidine, demerol, tramadol, ultram, meloxicam, kratom, carfentanil, buprenorphine, meperidine, narcotic, dihydrocodeine, levorphanol, naloxone, naltrexone, pentazocine, suboxone, subutex, sub, tapentadol, vivitrol, opiate, opioid, opium |
| Use disorder terms | abuse, abuses, abused, abusive, abusing, addict, addicts, addicting, addicted, addiction, dependence, dependant, dependance, dependency, misuse, misuses, misused, misusing, overdose, overdoses, overdoes, over dose, over dosed, od, over use, over used, overuse, use disorder, use-disorder, inject, injected, injects, injection, injecting, ivda, intravenous drug abuse, iv drug use, intravenous drug user, iv drug user, ivdu, intravenous drug abuse, iv drug abuse, iv drug abuse, iv drug abuser, withdrawal, withdraw, withdrew, withdrawling |
| Negation terms | absence, absent, deny, denies, denied, denying, do not, don't, donnot, exclude, excluded, excludes, excluding, lack, lacked, lacks, lacking, negative, negation, never, no, no evidence, did not have, no history, no hx, no sign, no signs, not observed, not present, without, without evidence, suspect, suspected |
| 2. Specialized terms | See Supplement Table 2 for specialized term lists |
| Specific clinic | first bridge clinic, the bridge |

**Table 3**
Parsing rules defining the combinations of dictionary terms used in the identification of opioid use disorder.

| Parsing rules | Rule contracture | Example |
|---|---|---|
| Rule 1 | Opioid term + <= 3 valid tokens + use disorder term | Opioid use disorder<br>Opiate dependence |
| Rule 2 | Use disorder term + <= 3 valid tokens + opioid terms | Addicted to suboxone |
| Rule 3 | Negation term + <=3 valid tokens + opioid terms + use disorder term | Denies opioid addiction |
| Rule 4 | Opioid term + use disorder term + <=3 valid tokens + negation term | Opioid dependence is denied |
| Rule 5 | Specialized terms (use dictionary 4) | IVDA/intranasal: Denies |
| Rule 6 | Specific clinic (use dictionary 5) | First Bridge Clinic |

### 2.6. Implementation

The finalized algorithm was applied to the 2019 patient data. The

NLP algorithm was coded using the Python programming language and the Natural Language Toolkit (NLTK) and Spacy modules; the logic and pipeline is shown in Fig. 1. The search rules were implemented as regular expression searches. The text in each Note section was converted to lowercase and scanned for terms in dictionaries 1 and 5. If a match was found, the section was further parsed into sentences. Each sentence was then scanned for mentions of OUD matching one of the negation rules (parsing rules 3, 4 and 5). If there was a match in a Note section, an OUD mention record was created, which included the matching text, the parsing rule to which it matched, and the Note and section name. If no matching terms were found, the next Note section was processed. Otherwise, the algorithm continued scanning the same sentence for matches to parsing rules 1, 2 and 6 (positive OUD mentions). If a match was found, an OUD mention record was created and the algorithm continued to search the whole document. If no mention was found, the next sentence was processed. In this way, all mentions of OUD appearing in a clinical note associated with a particular visit were extracted and classified as either positive or negative mentions. One or multiple OUD mentions can be identified from one Note.

### 2.7. Classification

To classify each visit as "OUD" or "non-OUD", we applied the following logic: if all mentions of OUD for a visit were positive, we classified it as OUD. If a visit had no mentions of OUD, or if all identified OUD mentions were negative, the visit was classified as "non-OUD." A

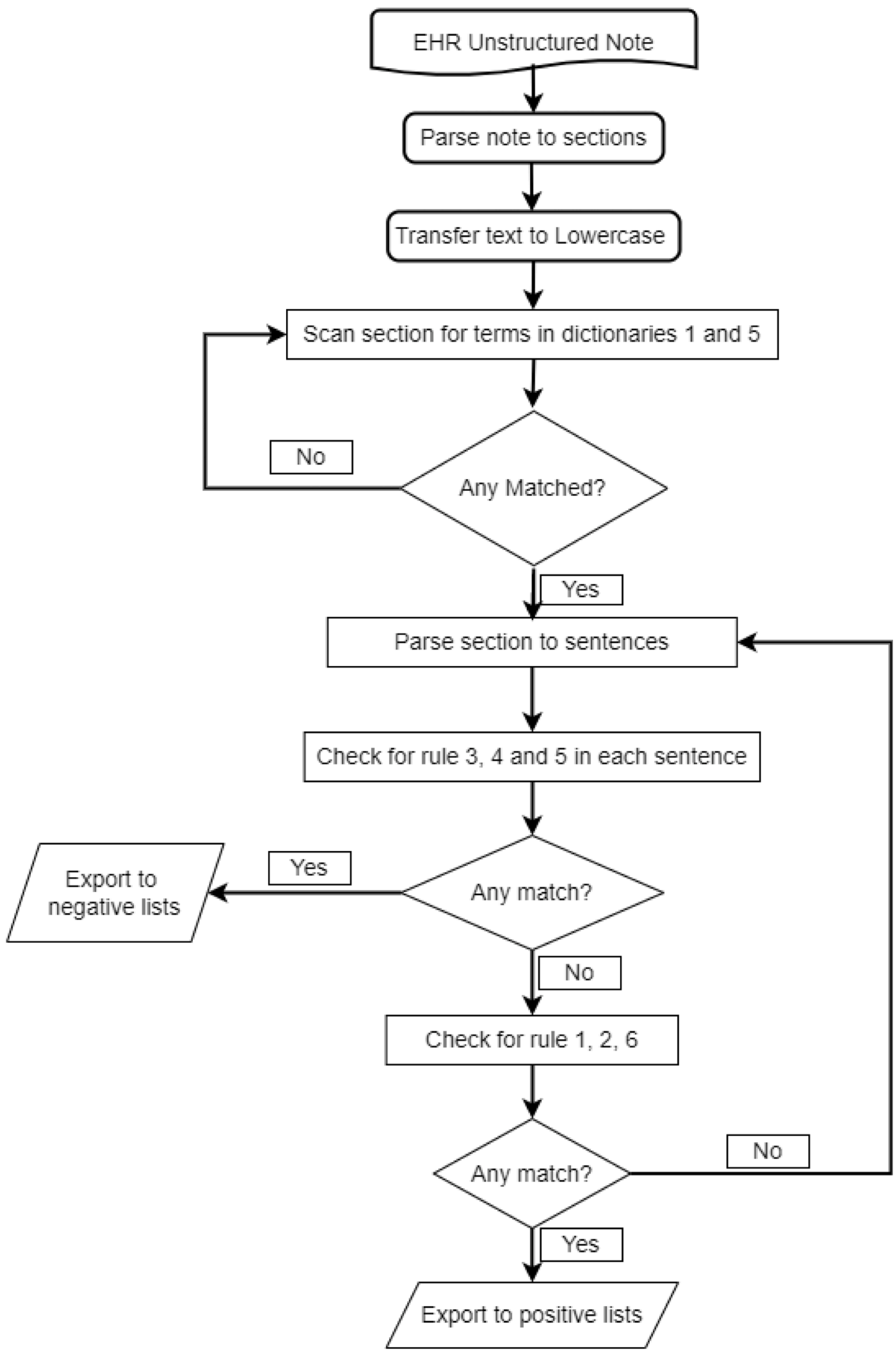


**Fig. 1.** Rule-based natural language processing parsing and classification process for individual mentions of opioid use disorder in electronic health record clinical notes.

small proportion of visits (162 out 29,212) had both positive and negative mentions of OUD. We manually reviewed these cases and classified them as OUD or non-OUD.

### 2.8. Statistical methods

Sensitivity and specificity of the NLP algorithm were computed with reference to 300 manually reviewed cases (the 1 % sample) as the gold standard diagnosis. OUD cases were compared between the classifications by the NLP algorithm and by ICD-10 codes. Demographic characteristics of patients classified as positive for OUD by ICD-10 only, NLP only, and both ICD-10 and NLP were compared using means and proportions. All statistical analyses were carried out using SAS 9.4® M6 (SAS Inc.; Cary, NC).

## 3. Results

A 1 % random sample of cases was manually reviewed to establish a ground truth data set for validation. The final sensitivity and specificity for the NLP algorithm, applied to the 2018 training data, were estimated against the ground truth sample as 81.8 % and 97.5 %, respectively. The positive predictive value (PPV) was 72.0 %.

We identified 29,212 hospital inpatient and ED visits occurring in 2019, of which 28,079 (95.1 %) met inclusion criteria. Those 28,079 visits generated 116,974 unstructured clinical notes, an average of 3.96 of the 5 notes of interest per patient visit. These notes comprised 59,780 Discharge Summaries (51.1 % of all notes), 22,080 History and Physical notes (18.9 %), 18,679 ED General notes (16.0 %), 14,927 ED Triage notes (12.8 %), and 1,508 Addiction Medicine Consult notes (1.3 %). Nearly all (98.7 %) patient visits in 2019 had at least one of these five clinical notes, which suggests minimal selection bias was introduced into the cohort by this requirement. Additionally, every included patient visit had an associated Discharge Summary.

About 67.0 % of the 2019 study patients were between 35 and 74 years old, 20.4 % were between 18 and 34 years old, and 12.9 % were 75 years or older at the time of visit. The majority of visits were among male patients (52.0 %) and with patient race reported as European American for 88.5 % of the visits. Ten percent of patient visits were among African American patients (9.9 %), and patients with reported other or unknown races accounted for 1.6 %. Similarly, the majority of patients were reported as non-Hispanic ethnicity (86.4 %). Hispanic ethnicity was reported for 2.8 % patients, with the remainder having no ethnicity information reported.

### 3.1. OUD case ascertainment by ICD-10-CM

We identified 1,811 patient visits having any ICD-10-CM code for OUD. Of these, 57 (3.1 %) were ED visits and 1,754 (96.9 %) were inpatient hospital visits. The distribution of opioid use disorder ICD-10-CM codes for these visits was as follows: opioid dependence accounted for 48.9 %, opioid abuse was 42 %, and opioid use was 9.1 %. Additional details are available in Supplemental Table 1.

### 3.2. OUD case ascertainment by NLP

The NLP algorithm identified 1,902 patient visits as having evidence of OUD in the clinical notes. Of these, 1,844 (97.0 %) were identified from inpatient hospital visit data and 58 (3.0 %) from ED visits. The NLP algorithm identified 24,822 total mentions of OUD across the 29,212 visits and the five selected notes (Table 4). ED General Notes and History and Physical Notes contained the majority of OUD mentions, and these tended to be negative (for example, “denial of opioid misuse”). The vast majority of the positive mentions of OUD were identified in Discharge Summary Notes (43.8 %) or Addiction Medicine Consult Notes (29.8 %), with another 18.3 % identified in History and Physical Notes. Most of the negative mentions of OUD (94 %) came from ED General Notes or History and Physical Notes.

**Table 4**
Distribution of opioid use disorder (OUD) mentions by note type.

| Note Group | Mentions of OUD (n = 22,715) | Positive mentions[1] (n = 6,186) | Negative mentions[2](n = 16,529) |
|---|---|---|---|
| ED General | 10,676 (47.0) | 446 (7.2) | 10,230 (61.9) |
| History and Physical | 6, 494 (30.5) | 1,134 (18.3) | 5,360 (32.4) |
| Discharge Summary | 3,948 (17.3) | 2,710 (43.8) | 788 (4.8) |
| Addiction Medicine Consult | 1,972 (8.7) | 1,841 (29.8) | 131 (0.8) |
| ED Triage | 75 (0.3) | 55 (0.9) | 20 (0.1) |
| **Total** | **22,715** | **6,186** | **16,529** |

[1] A positive mention of OUD is one which indicates the presence of OUD for the present visit.
[2] A negative mention of OUD is one which indicates the absence of OUD (e.g. “patient denies opioid abuse,” refers to a historical condition (e.g. “history of opioid abuse”), etc.

### 3.3. Comparison of NLP and ICD-10-CM results

The number of OUD cases identified by ICD-10-CM and NLP is summarized in Table 5. The absolute number of patient visits with evidence of OUD identified by each method was similar, with NLP identifying 91 more cases. While there was substantial overlap in the identified cases (1,381 [59.2 %]), overall 2,332 unique visits were identified. Of the total unique visits, 430 (18.4 %) were identified only by ICD-10-CM codes, and 521 (22.3 %) were identified only by NLP. The prevalence of visits with evidence of an OUD diagnosis in this sample, ascertained using only ICD-10-CM codes, was 1,811/29,212 (6.1 %). Including the additional 521 visits identified only by NLP, the estimated prevalence of OUD is 2,332/29,212 (7.9 %), an increase of 29.5 %.

Demographic characteristics of patient visits by OUD classification method are presented in Table 5. Compared to ICD codes alone, NLP codes alone identified a greater proportion of males (54.7 % vs 49.1 %), patients aged 55 or older (29 % vs 17.7 %), African American patients (10 % vs 5.1 %) and Hispanic patients (1.3 % vs 0.5 %), and married patients (23.2 % vs 17.2 %).

### 3.4. Analysis of ambiguous cases

A small proportion of visits (162 out 29,212) had both positive and negative mentions of OUD. An expert clinician (PDA) manually reviewed these cases and classified them as OUD or non-OUD. Only 22 of

**Table 5**
Patient characteristics by opioid use disorder (OUD) ascertainment method.

| Characteristics | OUD-ICD only (n = 430) | OUD-NLP only (n = 521) | Common (n = 1,381) |
|---|---|---|---|
| **Gender** | | | |
| Male | 211 (49.1 %) | 285 (54.7 %) | 672 (48.7 %) |
| Female | 219 (50.9 %) | 236 (45.3 %) | 722 (52.3 %) |
| **Age** | | | |
| 18–34 | 127 (29.5 %) | 118 (22.6 %) | 507 (36.7 %) |
| 35–54 | 227 (52.8 %) | 252 (48.4 %) | 692 (50.1 %) |
| 55–74 | 71 (16.5 %) | 135 (25.9 %) | 169 (12.2 %) |
| 75+ | 5 (1.2 %) | 16 (3.1 %) | 13 (0.9 %) |
| **Race** | | | |
| Black | 22 (5.1 %) | 52 (10.0 %) | 54 (3.9 %) |
| White | 405 (94.2 %) | 461 (88.5 %) | 1318 (95.4 %) |
| Other | 3 (0.7 %) | 8 (1.5 %) | 9 (0.7 %) |
| **Ethnicity** | | | |
| Hispanic | 2 (0.5 %) | 7 (1.3 %) | 6 (0.4 %) |
| Non-Hispanic | 379 (88.1 %) | 448 (86.0 %) | 1211 (87.7 %) |
| Other | 49 (11.4 %) | 66 (12.7 %) | 164 (11.9 %) |

these cases (13.5 %) were determined to be non-OUD. The other 140 (86.5 %) were confirmed to be OUD cases. For all 140 OUD cases, the source of the negated mentions of OUD was the social history in the ED general notes, ED Triage notes or Discharge Summary notes.

## 4. Discussion

Systematic reviews have demonstrated the potential for using NLP to extract clinical information from EHR notes, for a wide range of disease conditions. The most commonly studied diseases have been cancer, venous thromboembolism, peripheral arterial disease, and diabetes mellitus [21,26]. Previous investigations of NLP-based identification of OUD have been limited to patients undergoing chronic opioid therapy. Our objective was to extend this work to a general population of hospital inpatients.

We measured the prevalence of OUD in UKHC hospital and ED patient visits using ICD-10-CM codes and a rule-based NLP algorithm. Our study identified 1,902 OUD (6.4 %) visits out of 29,212 total visits by NLP algorithm, while a search of ICD-10-CM codes identified 1,811 (6.1 %) OUD cases from the same population. Combining ICD-10-CM codes and NLP-identified OUD visits gives 2,332 (7.9 %) OUD cases in our sample, an increase of nearly 30 % over the prevalence estimated using ICD codes alone. This increase—in our general adult patient population–is similar to previous reports of 33 % and 40 % increases in estimated prevalence, respectively, in Carrell et al and Zhu et al [3,28], which were limited to patients undergoing COT. For studies seeking to identify all cases of OUD in EHR data, we recommend that approaches like NLP be considered to supplement the structured data elements.

Eighteen percent of the identified OUD cases were identified by ICD codes but not by NLP. The most common reason for this discordance was the use of ambiguous language in the clinical notes. Visits with terms that indicated a substance use disorder but did not specify the substance – such as “polysubstance abuse”, “substance abuse” or “intravenous drug use/abuse” were not classified as OUD by the NLP algorithm. Varying terminology hindered the efficiency of NLP. This can occur in clinical documentation because an adequate history is not obtained, or the specific substance is not believed to be germane to the patient’s treatment.

Additionally, we did not search all clinical notes. Instead, we selected only the 5 types of notes that we believed were most likely to contain mentions of OUD. In several cases, on follow-up review of cases where NLP did not identify OUD but ICD codes did, OUD mentions were found in other clinical notes such as behavioral health notes, pharmacy consultation notes, physical therapy assessments, and others. We identified the clinical notes that were most likely to be present for our target population. We primarily relied on notes authored by the physician provider who would assign the ICD codes for the patient. This is not a foolproof method as collateral information may be gathered from other healthcare team members; however, this information should have filtered into the physician note. Future NLP studies of OUD in EHR data should consider including all possible notes, with the caveat that this may substantially increase the needed computing resources. Third, in some cases the ICD-10-CM OUD diagnosis conflicts with the textual evidence. For example, one case specified that the patient “denied drug use,” and no other opioid use information was given, but there was an ICD-10-CM code for OUD for this patient visit.

On the other hand, 22 % of OUD cases were identified by NLP but not by ICD codes. This may occur when OUD was documented as a secondary condition to the main reason for presenting for care, but either was not treated or was not considered relevant to reimbursement for treatment of the primary condition [8]. Unfortunately, errors may occur during documentation, where the correct answer is not recalled by the documenter. Errors may also occur if the history given by the patient changes.

Moreover, there were several important differences in the demographic characteristics of patient visits with OUD identified by only NLP vs only by ICD-10-CM. The group of visits identified only by NLP included slightly higher percentages of men, older patients, African American patients, and Hispanic patients. Prior studies have yielded evidence that such discrepancies are not uncommon in EHR data. In a study comparing patient race and ethnicity in structured and unstructured EHR data using NLP methods, Sholle et al reported that NLP increased the identification of black patients by 26 %, and Hispanic patients by 20 %, over structured variables alone [22]. Moreover, they found that patients who were identified as black or Hispanic only by NLP tended to be older and male and have higher comorbidity and were less likely to have commercial insurance. Vest et al compared structured and unstructured data for identifying patients in need of services to address social determinants of health (SDoH) [25]. They reported that unstructured data identified different patients than structured data, and that the patients documented as having a need for SDoH services in unstructured data tended to be medically more complex. These studies, and our own findings, suggest that biases may exist in the EHR structured data quality based on types of patients and medical conditions. NLP shows promise for addressing such biases.

### *4.1. Limitations*

Our sample was limited to hospital and ED visits in a single academic medical center in a geographic area with higher levels of opioid use disorder. The findings should not be generalized to all healthcare settings. To conserve computing resources, we limited the text mining search to five types of notes that were considered most likely to include mentions of OUD. In particular, we did not include psychiatry notes that include behavioral and mental health information. In our rule-based algorithm, the development of keyword dictionaries and parsing rules relied on literature reviews and expert opinion. Although we included over 1,000 entries that healthcare workers might use to describe OUD in text notes, it is still possible that terms were missed by our dictionaries and parsing rules. In particular, OUD mentions that contain abbreviations or spelling errors may result in false negatives. Also, negation rules are imperfect, and may result in misclassification of cases. For example, one of our negation rules allowed 3 or fewer tokens between the negation term and the opioid use term. This rule would fail to correctly negate the following mention of OUD: “Patient denies fevers or chills, vomiting or diarrhea, and IV drug abuse” and “patient denies any history of smoking, drug use.” Finally, a small proportion (0.5 %) of cases were ambiguous and had to be manually reviewed. Although this percentage is small, it could present an obstacle to scaling up the approach to very large samples. A possible solution may be to omit sections of clinical notes that are likely to contain negated mentioned of OUD that are not relevant to the present case – such as social history – from NLP processing. This is a point for future investigation.

## 5. Conclusion

The findings in this study support using rule-based NLP algorithms to identify potential OUD cases in EHR data and to improve surveillance of opioid use disorder compared to methods that only rely on ICD-10-CM codes. NLP is advantageous both in terms of improving the completeness of ascertainment, as well as mitigating biases in EHR structured data quality with regard to the age, gender, and race/ethnicity of the patient.

**Ethics and Consent statement**

This study received ethics approval from the Institutional Review Board at University of Kentucky (Study No. 60548). The IRB granted the study a waiver of authorization for participation due to the retrospective design using data already collected.

**Availability of data and materials**

The data was obtained from University of Kentucky HealthCare. The data was obtained through a data use agreement. These data cannot be shared with other scientists by the authors as they contain protected health information, and sharing the data would constitute a breach of the agreement. The NLP algorithm Python code is available upon reasonable request.

## Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

## Acknowledgments

This study involved a research collaboration among University of Kentucky HealthCare IT department and College of Public Health at the University of Kentucky. The authors thank Roshan A. Hussain, Christopher Burton, Daniel Steinberg and Yosef Wirian for their support.

## References


[1] O. Baclic, M. Tunis, K. Young, C. Doan, H. Swerdfeger, J. Schonfeld, Challenges and opportunities for public health made possible by advances in natural language processing, Can. Commun. Dis. Rep. 46 (2020) 161–168.

[2] K.S. Beam, M. Lee, K. Hirst, A. Beam, R.B. Parad, Specificity of International Classification of Diseases codes for bronchopulmonary dysplasia: an investigation using electronic health record data and a large insurance database, J. Perinatol. 41 (4) (2021) 764–771.

[3] D.S. Carrell, D. Cronkite, R.E. Palmer, K. Saunders, D.E. Gross, E.T. Masters, T. R. Hylan, M. Von Korff, Using natural language processing to identify problem usage of prescription opioids, Int. J. Med. Inf. 84 (12) (2015) 1057–1064.

[4] D. Chartash, H. Paek, J.D. Dziura, B.K. Ross, D.P. Nogee, E. Boccio, C. Hines, A. M. Schott, M.M. Jeffery, M.D. Patel, T.F. Platts-Mills, O. Ahmed, C. Brandt, K. Couturier, E. Melnick, Identifying opioid use disorder in the emergency department: multi-system electronic health record-based computable phenotype derivation and validation study, JMIR Med. Inform. 7 (2019) 1–10.

[5] E. Ford, J.A. Carroll, H.E. Smith, D. Scott, J.A. Cassell, Extracting information from the text of electronic medical records to improve case detection: a systematic review, J. Am. Med. Inform. Assoc. 23 (2016) 1007–1015.

[6] R.E, Gliklich, M.B, Leavy, N.A, Dreyer. . Tools and Technologies for Registry Interoperability, Registries for Evaluating Patient Outcomes A User's Guide, 3rd Edition, Addendum 2. *L&M Policy Research* 2019 DOI: https://doi.org/10.23970/AHRQEPCREGISTRIES3ADDENDUM2.

[7] A.D. Kaye, M.R. Jones, A.M. Kaye, J.G. Ripoll, V. Galan, B.D. Beakley, F. Calixto, J. L. Bolden, R.D. Urman, L. Manchikanti, Prescription opioid abuse in chronic pain an updated review of opioid abuse predictors and strategies to curb opioid abuse Part1, Pain Physician 20 (2017) 111–133.

[8] N.Y. Kirson, A. Shei, J.B. Rice, C.J. Enloe, K. Bodnar, H.G. Birnbaum, P. Holly, R. Ben-Joseph, The Burden of Undiagnosed Opioid Abuse Among commercially Insured Individuals, Pain Med. 16 (7) (2015) 1325–1332.

[9] I. Kurbasic, H. Pandza, I. Masic, S. Huseinagic, S. Tandir, F. Alicajic, S. Toromanovic, The advantages and limitations of international classification of diseases, injuries and causes of death from aspect of existing health care system of bosnia and herzegovina, Acta. Inform Med. 16 (2008) 159–161.

[10] S.M. Meystre, G.K. Savova, K.C. Kipper-Schuler, J.F. Hurdle, Extracting Information from Textual Documents in the electronic health record A review of recent research, IMIA Yearbook of Med. Informatics. 173 (2008) 128–145.

[11] J.E. Mezzich, International Surveys on the Use of ICD-10 and Related Diagnostic Systems, Psychopathology 35 (2002) 72–75.

[12] P.M. Nadkarni, L. Ohno-Machado, W.W. Chapman, Natural language processing: an introduction, J. Am. Med. Inform. Assoc. 18 (5) (2011) 544–551.

[13] Neoplasms. 2021. C00-D49: https://www.icd10data.com/ICD10CM/Codes/C00-D49.

[14] K.J. O'Malley, K.F. Cook, M.D. Price, K.R. Wildes, J.F. Hurdle, C.M. Ashton, Measuring diagnoses: ICD code accuracy, Health Serv. Res. 40 (5p2) (2005) 1620–1639.

[15] Office of the Secretary, H. 2014. Administrative Simplification: Change to the Compliance Date for the International Classification of Diseases, 10th Revision (ICD-10-CM and ICD-10-PCS) Medical Data Code Sets. Final rule. *Fed Regist;* 79: 45128-34.

[16] S.A. Palumbo, K.M. Adamson, S. Krishnamurthy, S. Manoharan, D. Beiler, A. Seiwell, C. Young, R. Metpally, R.C. Crist, G.A. Doyle, T.N. Ferraro, M. Li, W. H. Berrettini, J.D. Robishaw, V. Troiani, Assessment of probable opioid use disorder using electronic health record documentation, JAMA Netw. Open 3 (2020) 1–12.

[17] S.A. Pendergrass, D.C. Crawford, Using electronic health records to generate phenotypes for research, Curr. Protoc. Hum. Genet. 100 (2019) 1–28.

[18] A. Piotrkowicz, O. Johnson, G. Hall, Finding relevant free-text radiology reports at scale with IBM Watson Content Analytics: a feasibility study in the UK NHS, J. Biomed. Semantics. 10 (2019) 21–29.

[19] H. Quan, B. Li, L.D. Saunders, G.A. Parsons, C.I. Nilsson, A. Alibhai, W.A. Ghali, I. Investigators, Assessing validity of ICD-9-CM and ICD-10 administrative data in recording clinical conditions in a unique dually coded database, Health Serv. Res. 43 (2008) 1424–1441.

[20] S.I, Ranapurwala, I, Alam, M, Clark, T, Carey, P.R, Chelminski, B, Pence, J, Korte, W, L-T, M, Wolfson, S, Christensen, M, Capata, H, Douglas, L, Greenblat, L, Bowlby, J, Spangler, S, Marshall. LIMITATIONS OF OPIOID USE DISORDER (OUD) ICD CODES: DEVELOPMENT AND VALIDATION OF A NEW OUD IDENTIFICATION ALGORITHM IN ELECTRONIC MEDICAL RECORDS. *Inj Prev*. 10 2021 2-3.

[21] S. Sheikhalishahi, R. Miotto, J.T. Dudley, A. Lavelli, F. Rinaldi, V. Osmani, Natural language processing of clinical notes on chronic diseases: systematic review, JMIR Med. Inform. 7 (2019) 1–8.

[22] E.T, Sholle, L.C, Pinheiro, P, Adekkanattu, M.A, Davila, S.B, Johnson, J, Pathak, S, Sinha, C, Li, S.A, Lubansky, M.M, Safford, T.R, Campion Underserved populations with missing race ethnicity data differ significantly from those with structured race/ethnicity documentation 2019. *J Am Med Inform Assoc*. 26:722-729.

[23] R. Smart, C.A. Kase, E.A. Taylor, S. Lumsden, S.R. Smith, B.D. Stein, Strengths and weaknesses of existing data sources to support research to address the opioids crisis, Prev. Med. Rep. 17 (2020) 1–14.

[24] I. Spasic, G. Nenadic, Clinical Text Data in Machine Learning: Systematic Review, JMIR Med. Inform. 8 (2020) 1–19.

[25] J.R. Vest, S.J. Grannis, D.P. Haut, P.K. Halverson, N. Menachemi, Using structured and unstructured data to identify patients' need for services that address the social determinants of health, Int. J. Med. Inf. 107 (2017) 101–106.

[26] Y. Wang, L. Wang, M. Rastegar-Mojarad, S. Moon, F. Shen, N. Afzal, S. Liu, Y. Zeng, S. Mehrabi, S. Sohn, H. Liu, Clinical information extraction applications: A literature review, J. Biomed. Inform. 77 (2018) 34–49.

[27] S.-C. Wong, C.-W. Cheung, Optimization of opioid utility in cancer pain populations, Ann. Palliat. Med. 9 (2) (2020) 558–570.

[28] V.J. Zhu, L.A. Lenert, K.S. Barth, K.N. Simpson, H. Li, M. Kopscik, K.T. Brady, Automatically identifying opioid use disorder in non-cancer patients on chronic opioid therapy, Health Informatics J. 28 (2022) 1–17.