A Natural Language Processing Pipeline to Study Disparities in Cannabis Use and Documentation Among Children and Young Adults
A Survey of 21 Years of Electronic Health Records

Nazgol Tavabi, PhD,[1,2] Marium Raza, BS,[1,2] Mallika Singh, MS,[1] Shahriar Golchin, MS,[1] Harsev Singh,[1] Grant D. Hogue, MD,[1,2]Ata M. Kiapour, PhD, MMSc[1,2#]

[1]Department of Orthopaedic Surgery and Sports Medicine, Boston Children's Hospital, Boston, MA

[2] Harvard Medical School, Boston, MA

Corresponding author.

The study was performed at Boston Children's Hospital, Harvard Medical School.

Corresponding Author:

Ata M. Kiapour PhD, MMSc
Department of Orthopedic Surgery and Sports Medicine
Boston Children's Hospital
Harvard Medical School
300 Longwood Ave, Boston, MA 02115
E: Ata.Kiapour@childrens.harvard.edu

Abstract

The legalizations of medical and recreational cannabis have generated a great deal of interest in studying the health impacts of cannabis products. Despite increases in cannabis use, its documentation during clinical visits is not yet mainstream. This lack of information hampers efforts to study cannabis effects on health outcomes. A clear and in-depth understanding of current trends in cannabis use documentation is necessary to develop proper guidelines to screen and document cannabis use. Here we have developed and used a hierarchical natural language processing pipeline (AUROC=0.94) to evaluate the trends and disparities in cannabis documentation on more than 23 million notes from a large cohort of 370,087 patients seen in a high-volume multi-site pediatric and young adult clinic over a period of 21 years. Our findings show a very low but growing rate of cannabis use documentation (<2%) in electronic health records with significant demographic and socioeconomic disparities in both documentation and use, which requires further attention.

Introduction

In the United States, cannabis is legal for medicinal use in 37 states, and legal for recreational use in 19 states. Up to 22-million Americans twelve years old or older use cannabis annually. This is part of an upwards trend. Daily reported usage of cannabis is increasing, from 2.1% in 2016 to 3.4% in 2019 according to the national survey on drug use. Within this population, there has also been increased usage among youth [1]. Over 11.8 million young adults report cannabis use. Daily use has increased from 5.9% to 6.9% for twelfth graders, 2.9 to 4.4% for tenth graders, and 0.8 to 1.1% for eighth graders from 2017 to 2020 [2].

Despite this increase in usage, proper documentation of cannabis use has not become mainstream. Such information is vital for a more accurate assessment of cannabis use rates and potential effects on treatment outcomes. For example, cannabis has shown to influence remodeling of a range of musculoskeletal tissues (e.g., bone) [3]. A recent study on pediatric patients with extremity fractures found significantly increased time to union in those who used cannabis [4]. There is also evidence suggesting cannabis use affects the outcomes of surgeries, including mortality, pain, comorbidities, and revision rates [5, 6]. One study found reduced mortality in cannabis users undergoing total hip arthroplasty, total knee arthroplasty, total shoulder arthroplasty, and traumatic femur fixation [7]. Other studies found that cannabis users had higher surgery revision rates [8] and cannabis users undergoing spine surgery had greater perioperative morbidity [9].

In terms of pain after surgery, results are contradictory as well. Cannabis users reported lower pain in the operative site in one study of 937 patients [10], while patients with preoperative cannabis use reported increased pain after major orthopedic surgery on an NRS scale in another study of 3793 patients [11]. A different study found that cannabis users had higher total prescribed opioids, and longer duration of use [12]. A recent review found that cannabis use in the form of combustive cigarettes represents perioperative risks for induction/anesthesia, post operative pain, and analgesia in teenagers [13].

Currently the most common way of cannabis use documentation in patients' health records is within unstructured clinical notes, in an unstandardized manner [14]. While studies have proposed the development of standardized screening protocols to streamline documentation [15], yet no such protocol has been adopted widely. Currently, cannabis itself is referred to in heterogeneous ways within clinical notes, alternatively referred to as MJ, cannabis, weed, CBD, and THC. Some of these terms, specifically weed and CBD, are nonspecific (i.e., weed also refers to other plants,

CBD also refers to common bile duct). These terms may also be misspelled. In addition, particularly within pediatric clinical notes, there are mixed mentions of cannabis use among patients and their family members. Additional analysis is needed to distinguish which subject is the cannabis user. Overall, the complexity of retrieving cannabis use screening information from this type of heterogeneous, unstructured data has made it difficult to monitor or study.

One approach to tackle this challenge is the use of natural language processing (NLP) to identify cannabis users. NLP uses linguistic knowledge to identify information (such as cannabis use) from human language input by identifying patterns in the data. Applying NLP techniques on clinical notes comes with its own challenges such as presence of noise, heterogeneity, different templates in the data, abbreviations, misspellings, incomplete sentences, etc. Hence, many of the state-of-the-art NLP models do not perform as well when applied directly to clinical notes and the data needs to be thoroughly preprocessed and cleaned before being fed into the models. Studies such as Tavabi et al. [16], Wang et al. [17], Ling et al [18], developed and evaluated such NLP pipelines and approaches on clinical notes for purposes like cohort identification and building registries. Some studies have also used NLP to identify substance use from clinical notes. One study used NLP on clinical notes to identify hospitalized trauma patients with alcohol misuse and demonstrated greater accuracy than EMR-based billing codes [19]. Another study detected alcohol, drug, or nicotine use from unstructured notes using NLP and achieved good performance over a wide breadth of notes [20]. When looking at cannabis use specifically, a separate study developed an NLP algorithm to identify cannabis related terms, historical mentions, and hypothetical mentions within EHR notes [21]. However, Carrell et al. [21] identified 54% of the notes with positive medical cannabis usage automatically and used an NLP-assisted manual review tool to identify the rest, which is a labor and time intensive process.

In this work, we developed an NLP pipeline to comprehensively assess the changes in cannabis use documentation over the past 21 years along with potential disparities in use and documentation in a multi-site high volume pediatric and young-adults orthopedic and sports medicine (OSM) academic practice, without the need for manual chart review. We looked at patients from the OSM considering the high prevalence of these injuries among children and young adults, and effects of cannabis in musculoskeletal tissue healing and remodeling. With a better understanding of how cannabis use is documented, care guidelines for a growing population of cannabis users can improve.

Methods

This study was performed in line with the principles of the Declaration of Helsinki. Approval was granted by the Institutional Review Board of Boston Children's Hospital. Considering the retrospective use of available data, the study was exempt from patient consent. A breakdown of the pipeline developed in this study is shown in Figure 1. First, through a process of physician-in-the-loop a dictionary of cannabis related keywords and their possible misspellings was generated. Based on the dictionary, a dataset of clinical notes was extracted, and after preprocessing and filtering out irrelevant notes, the sentences with cannabis related information were retrieved from the notes. A small percentage of the sentences were then annotated by a group of experts to generate training data for a classifier to confirm positive vs negative cannabis use. Afterwards, the structured Electronic Health Records (EHR) of the case and control group (patients with predicted positive and negative cannabis usage) such as demographics and list of diagnoses and procedures were extracted and used for analysis.

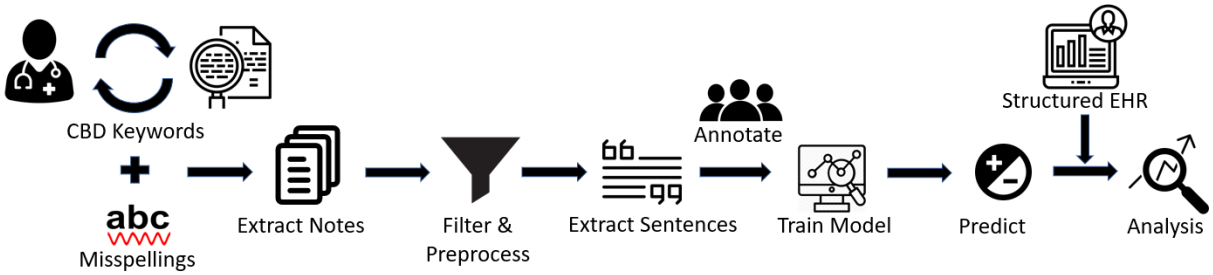

Figure 1 : Diagram of pipeline developed in this study to evaluate cannabis use documentation and to assess disparities in cannabis use among children and young adults with musculoskeletal injuries.

Data: Following IRB approval, all clinical notes, baseline demographics, and diagnostics and procedures of every patient who has visited the Boston Children's Hospital OSM clinics (6 locations across Massachusetts) between 2000-2021 were obtained (370,087 unique patients, 23,871,108 notes). The baseline demographics of the cohort are presented in Table 1. The Social Vulnerability Index (SVI) was calculated based on residence zip code. SVI is a composite index developed by the Centers for Disease Control and Prevention that characterizes community resilience and vulnerability relative to external stressors [22]. It calculates an overall index (0 – 1), with a higher index indicating greater social vulnerability. Additionally, current procedural terminology (CPT) billing codes were used to identify most common musculoskeletal procedures within the study population and diagnosis codes were used to identify patients with cannabis related diagnosis.

Table 1. Baseline characteristics of the studied cohort.

Characteristics	Patients
Total Sample, No.	370,087
Age at time of visit, mean (SD), y	11.6 (8.27)
Sex, No (%)	
Female	194,777 (52.6%)
Male	175,272 (47.3%)
Unknown	38 (<0.1%)
Race, No. (%)	
Asian	8,461 (2.2%)
Black	17,762 (4.8%)
Hispanic	24,766 (6.7%)
Other [a]	17,933 (4.8%)
White	207,098 (55.9%)
Unavailable	94,040 (25.4%)
SVI, mean (SD)	0.271 (0.244)

[a] Native American, Alaska native, native Hawaiian or other Pacific Islander, and multi-racial.

Diagnostic Codes for Cannabis Misuse: ICD (International Classification of Diseases) and SNOMED CT (Systematized Nomenclature of Medicine Clinical Terms) are 2 different coding systems used for diagnosing patients. ICD codes are mostly used after the care is completed for example for billing purposes, whereas SNOMED CT is used directly by healthcare providers during the process of care. There are different ICD codes and SNOMED CTs related to cannabis use. To identify them, we extracted all diagnoses with the keyword "cannabis" in their description, and after they were reviewed and confirmed, all cases with those codes and all the patients

diagnosed with them were extracted. Table 2 shows identified diagnostic codes, their descriptions, and their frequency. A patient might get diagnosed with the same code or a similar code more than once.

In the ICD coding system, there is a hierarchical structure. By adding more characters to the code, the diagnosis becomes more specific. For example, for F12 in Table 2 the actual diagnosis codes were F12.20, F12.10, F12.90, etc. Here they are aggregated to the most general cases (e.g., parent codes) related to cannabis use.

Table 2. Diagnostic codes related to cannabis use found in the dataset. Diagnostic codes with frequency less than 10 are not shown in the table.

Diagnosis Code	Source Vocabulary	Description	Count
F12	ICD-10-CM	Cannabis related disorders	10543
305.2	ICD-9-CM	Nondependent cannabis abuse	6990
304.3	ICD-9-CM	Cannabis dependence	4832
T40.7	ICD-10-CM	Poisoning by, adverse effect of and underdosing of cannabis (derivatives)	142
37344009	SNOMED CT	Cannabis abuse	136
428823006	SNOMED CT	Cannabis misuse	65
85005007	SNOMED CT	Cannabis dependence	53
191837001	SNOMED CT	Cannabis dependence, continuous	12

Natural Language Processing (NLP) Model: A hierarchical NLP approach was developed to first identify notes containing cannabis-related terms (screening step) and then classify them into positive (endorsed cannabis use) or negative (no cannabis use).

For the screening step, first a cannabis-related dictionary was generated with a physician-in-the-loop approach. Initially A few seed words (Marijuana, Cannabis, CBD, Weed, THC) and medical terms (Tetrahydrocannabinol, Epidiolex, Cannabidiol, …) from FDA approved cannabis drugs were chosen and notes containing them were collected. Afterwards, based on the notes collected, other keywords (such as MJ) were identified and added to the dictionary.

Additionally, possible misspellings of the keywords were identified and added to the dictionary. In order to add misspellings of dictionary keywords, words from all the notes were gathered and their N-grams (N= [1-3]) were computed. The cosine similarity of each word's Ngram and Ngram of the correct word was computed, and words with higher similarity than certain threshold were identified. The threshold was chosen manually for each keyword in the cannabis-related dictionary. Below is an example of how this is computed.

A = Marijuana N-grams (A:3, JU:1, AH:0…)

B = Marijauna N-grams (JA:1, JU:0, AU:1, JAU:1…)

C = Marijuahana N-grams (A:4, H:1, AH:1, AHA:1…)

Cosine Similarity (A, B) $= \frac{A \cdot B}{|A||B|} = 0.77$ *Cosine Similarity (A, C)* $= \frac{A \cdot C}{|A||C|} = 0.90$

When the dictionary was finalized, all the notes in the dataset (n=23,871,108) were screened for cannabis-related terms. . Figure 2 shows the number of notes identified with each keyword (some notes contained more than one keyword).

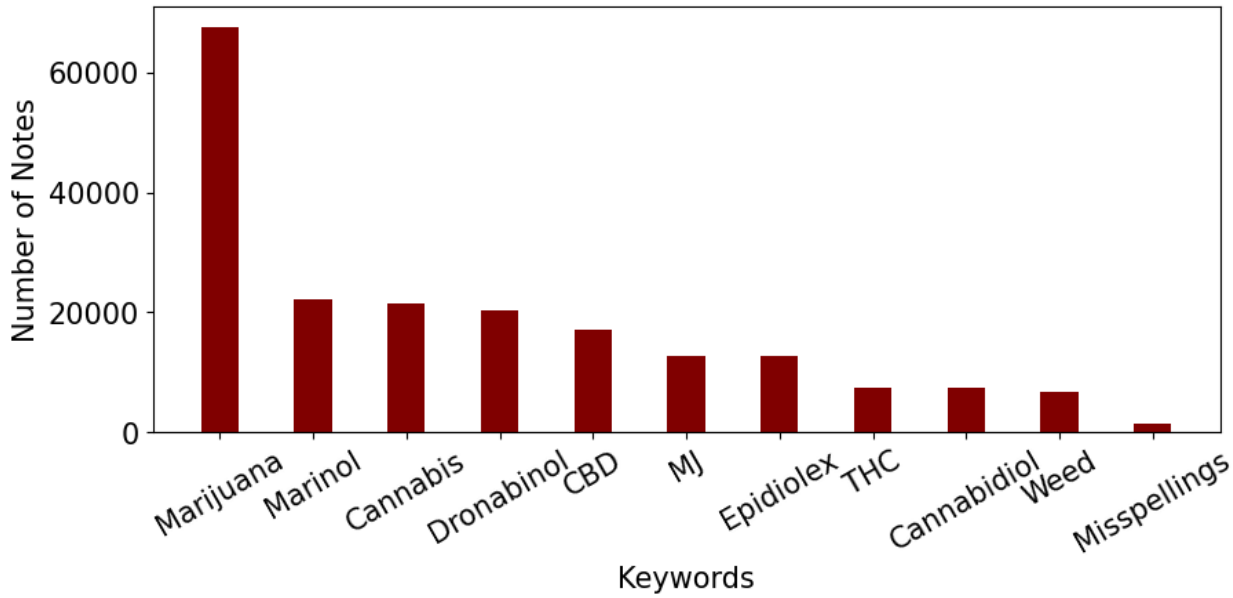

Figure 2. The keywords used to extract clinical notes and the number of notes containing each of them. Keywords like Syndros and Cesamet which appeared in less than 10 notes are not included in the graph.

After finding the notes with any of the keywords in the dictionary, the dataset was first cleaned and processed. The first step was to remove the notes that were incorrectly marked as cannabis related. As mentioned earlier words like CBD or Weed can also be used in other contexts. The acronym CBD is also used for Common Bile Duct and is used in notes related to gastrointestinal conditions. Also, the keyword Weed is used in notes related to allergy detection. To remove these irrelevant notes, 2 other sets of keywords were compiled. Words like abdominal, gallbladder, pancreas, common bile duct, … were used to identify and filter out notes related to common bile duct and words like pollen and allergy and its variations were used for allergy related notes. Additionally, cannabis related notes which correspond to very young patients (under 7) were primarily related to the cannabis use in their household and not by the patient, especially during pregnancy. Table 3 shows a few examples for patients under age 7. Since the purpose of this study

is to identify patients with direct cannabis use, patients younger than 7 years old were removed from the dataset.

Table 3. Examples of Cannabis Related Notes for Patients under 7.

Other Tobacco History: Mother smoked, but has quit, also use of MJ during pregnancy but has quit as well.
Birth Hx: Exposed to maternal THC in utero.
Pregnancy complicated by late to care, inconsistent prenatal care, marijuana use, and viral illness 5days prior to delivery.
He lives with his mother and occasionally with his grandmother in the housing projects where she says he is exposed to cigarette and marijuana smoke in the hallways.

After preprocessing the notes, the sentences containing keywords from the cannabis dictionary were extracted from each note. The types of extracted notes were very heterogeneous, including notes from "Social Work Clinic Note" to "General Inpatient Surgery", resulting in a heterogeneous collection of note contents and formats. Some notes such as toxicology reports were written as forms with different fields and without proper sentences and in some others like "Adolescent Substance Abuse Progress Visit", the note contained full sentences with a lot of detailed narratives. Distribution of the top 15 note types is shown in Figure 3. Since the cannabis use was only a small part of the whole notes and to minimize unwanted bias related to note types (i.e., template and content), only the sentences containing the keywords were extracted and used as input to the classifier.

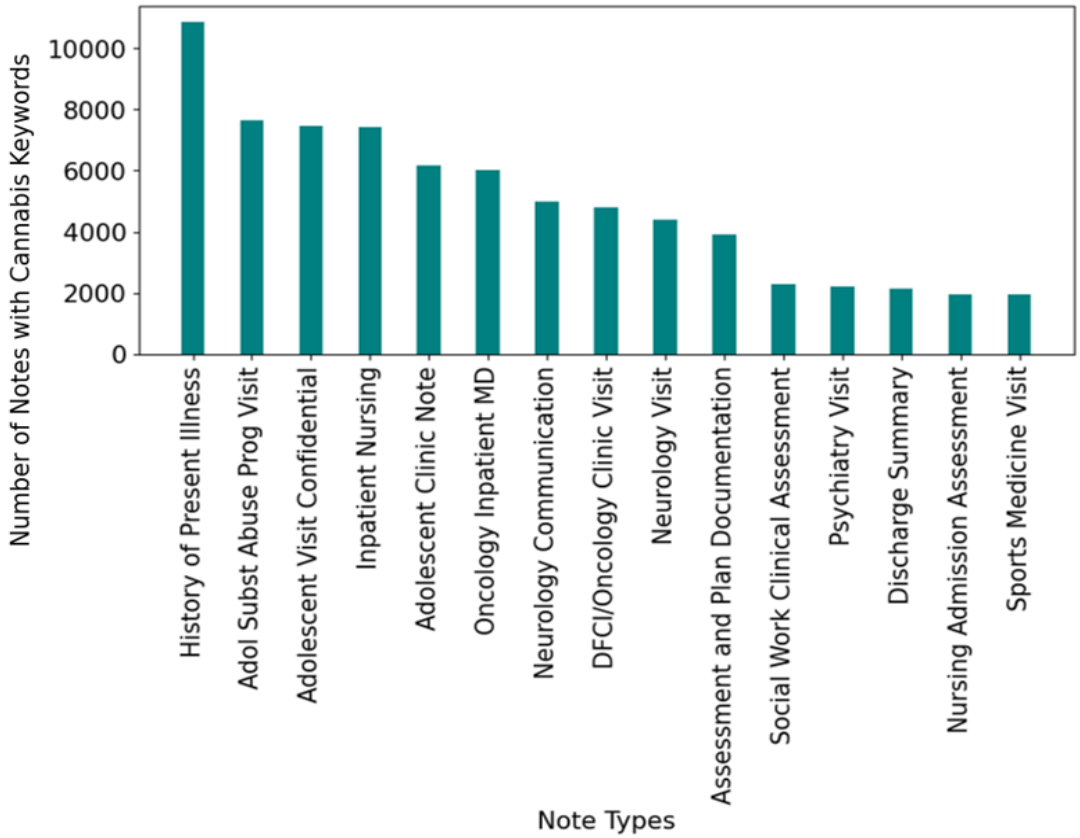

Figure 3. Distribution of the top 15 note types containing cannabis-related keywords.

After the screening step, the model Bidirectional Encoder Representations from Transformers (BERT) model was used to classify the extracted sentences. BERT is the state-of-the-art NLP model developed and trained by Google, which has been widely used across multiple domains, including healthcare [23]. BERT models provide contextual representations of words and sentences, which can then be used in classification. Domain specific BERT models yield significant improvement, hence for this project we used ClinicalBioBERT [24], which has been pretrained on medical literature (PubMed corpus) and publicly available MIMIC dataset [25]. We further pretrained the ClinicalBioBERT on all the available notes from Boston Children's Hospital (BCH_BERT; 23,871,108 notes for 1 epoch) so the model would be more familiar with the language used in the clinical notes. We then used the following deep learning model (Figure 4) to

classify the extracted sentences from the screening step into positive use (i.e., self-reported cannabis use by the patient or guardians currently or in the past, positive use reports by the clinical care team or toxicology reports) or negative use (i.e., patient or guardian reported no use, negative use confirmation by clinical care team or toxicology reports, discussions of pros and cons of cannabis use with no direct indication of positive use, reported abuse for someone other than the patient). Examples of different types of both positive and negative use are given in Table 4.

Table 4. Example notes of positive and negative cannabis use. The green color shows positive cannabis cases, and the red color shows negative.

Pt admits to smoking marijuana daily
MJ: +
agreed to try marinol with relief, continued scopolamine patch and all other antiemetics changed to prn. Be aware you may be light-headed or dizzy while using marinol RX: DRONABINOL (MARINOL(2.5MG/TAB
Denies drugs or alcohol; states used marijuana "one year ago". Urine tox screen: positive for amphetamines and cannabis.
He has not tried any other drugs and denies any use of marijuana, cocaine, heroine, or other drug abuse.
Labs - Urine tox sent - negative for amphetamines, barbiturates, benzos, cannabis, cocaine, opiates, PCP Wrist films and R hand film done - Negative for fracture
I discussed the interactions of stimulant medications with alcohol, which increases the concentration of the stimulant and marijuana, which increases the heart rate with use of a stimulant medication.
CRAFFT score is 1; he has been in the car when his brother has been smoking marijuana and driving, but his brothers will not allow him to do this.

The model was trained and tested on a diverse set of manually labeled notes (n=3,835, 73% positive, 27% negative, 80% train, 10% validation ,10% test). The training was done for 7 epochs on an NVIDIA Titan RTX GPU. The model achieved an accuracy of 0.95, area under the ROC curve of 0.94, sensitivity of 0.97 and specificity of 0.90 on the test set.

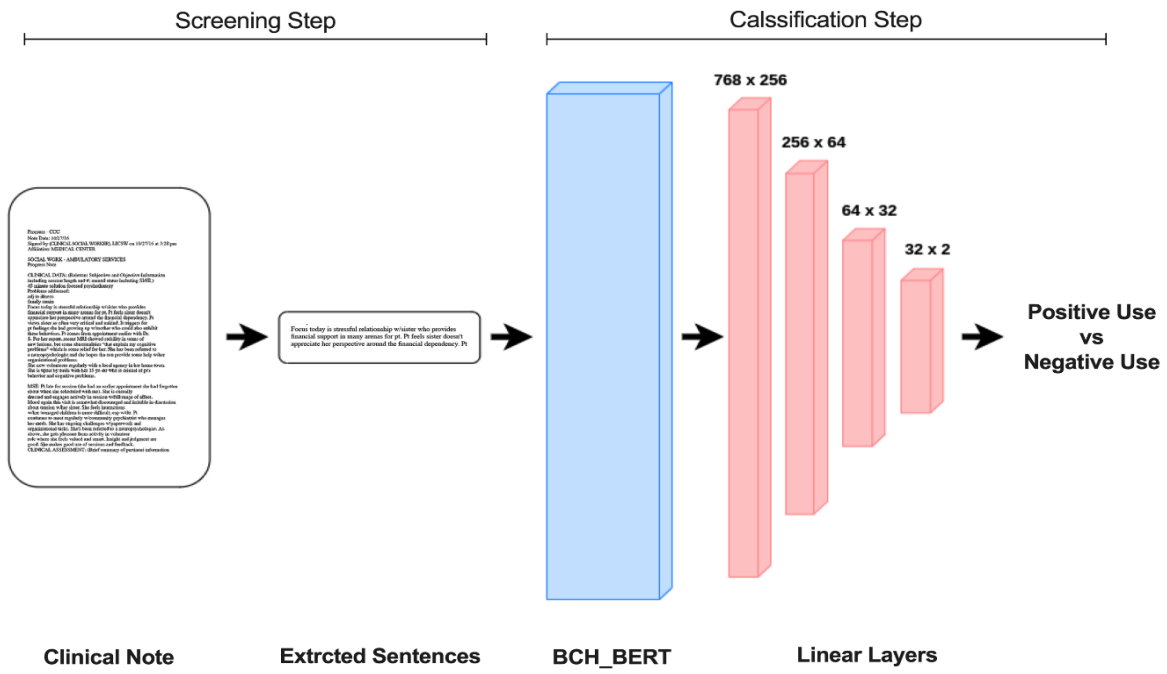

Figure 4. The deep learning model used to classify extracted sentences to identify positive cannabis use.

Evaluation of Disparities in Documented Cannabis Use: Disparities in the data were evaluated in 2 different settings. First, patients with documented cannabis use vs patients with no documents mentioning cannabis use. Second, patients with positive cannabis use vs patients with negative cannabis use. In the first setting the group with no documented cannabis use were never asked about cannabis usage and cannabis was not discussed in any of their notes whereas the other group had mentions of cannabis in at least one note which could be either a positive or a negative use.

A generalized linear mixed model was used to evaluate the associations between sex, race and SVI in rate of documented cannabis use (first setting). Adjusted odds ratios (aOR) were calculated and considered significant at $P<.05$ (SPSS v27) between those with documented cannabis use (positive or negative) and those with no cannabis use documentation. The aOR was calculated for female sex compared to male sex, and for asian, black, hispanic and other races compared to white race. For SVI, the aOR was calculated for every 0.01-point change in SVI. The analysis was repeated after excluding those with medical cannabis use. The same framework was applied to assess disparities in positive vs negative cannabis users with regards to sex, race and SVI (second setting).

Results

Of 23,871,108 notes (7.8% OSM notes), 166,530 (2.7% OSM notes) contained cannabis-related terms (23,974 patients), out of which 124,952 notes (2.3% OSM notes) deemed positive use (13,556 patients). The breakdown of patients with positive and negative cannabis use are presented in Table 5. Of all identified positive patients, only 1,971 (14.5%) had medical diagnosis of cannabis use disorder based on either ICD codes or SNOMED CT (Table 2). From 2000-2021, there were increases in overall and positive cannabis use documentation across all notes (Figure 5). There were stepwise increases in cannabis use documentation across all notes after legalization of medical (2012, black vertical dotted line) and recreational (2016, green vertical dotted line) cannabis (Figure 5). For the OSM notes, there were increases in overall and positive cannabis use documentation primarily after 2010 with marked increases in cannabis use documentation after legalization of recreational cannabis (2016, green vertical dotted line; Figure 5).

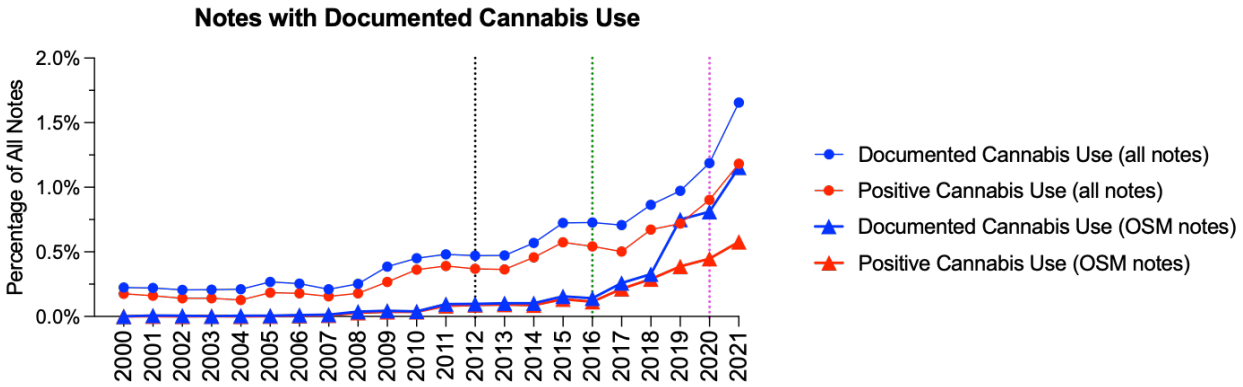

Figure 5. Cannabis use documentation as a percentage of all notes collected each year (circles) and cannabis use documentation in OSM notes as a percentage of all OSM notes collected each year (triangles). The vertical dotted lines indicate the legalization of medical (2012, black) and recreational (2016, green) cannabis in Massachusetts, and declaration of COVID-19 pandemic (2020, magenta).

From 2000 to 2021, there were increases in the total number of new patients with documented positive cannabis use (Figure 6A) with a substantial drop in 2020 corresponding to clinical care restrictions associated with the COVID-19 pandemic (magenta dotted vertical dotted line). Similar trends, with different magnitudes were observed for new patients with documented positive cannabis use within each race category (Figure 6B). While there were no consistent trends in the percentage of new female patients with documented positive cannabis use prior to 2012 (legalization of medical cannabis), the percentage of new female patients with documented positive cannabis use increased consistently from 2012-2021 (Figure 6C). Finally, the average age at first documented cannabis use increased from 2000-2021 (Figure 6D).

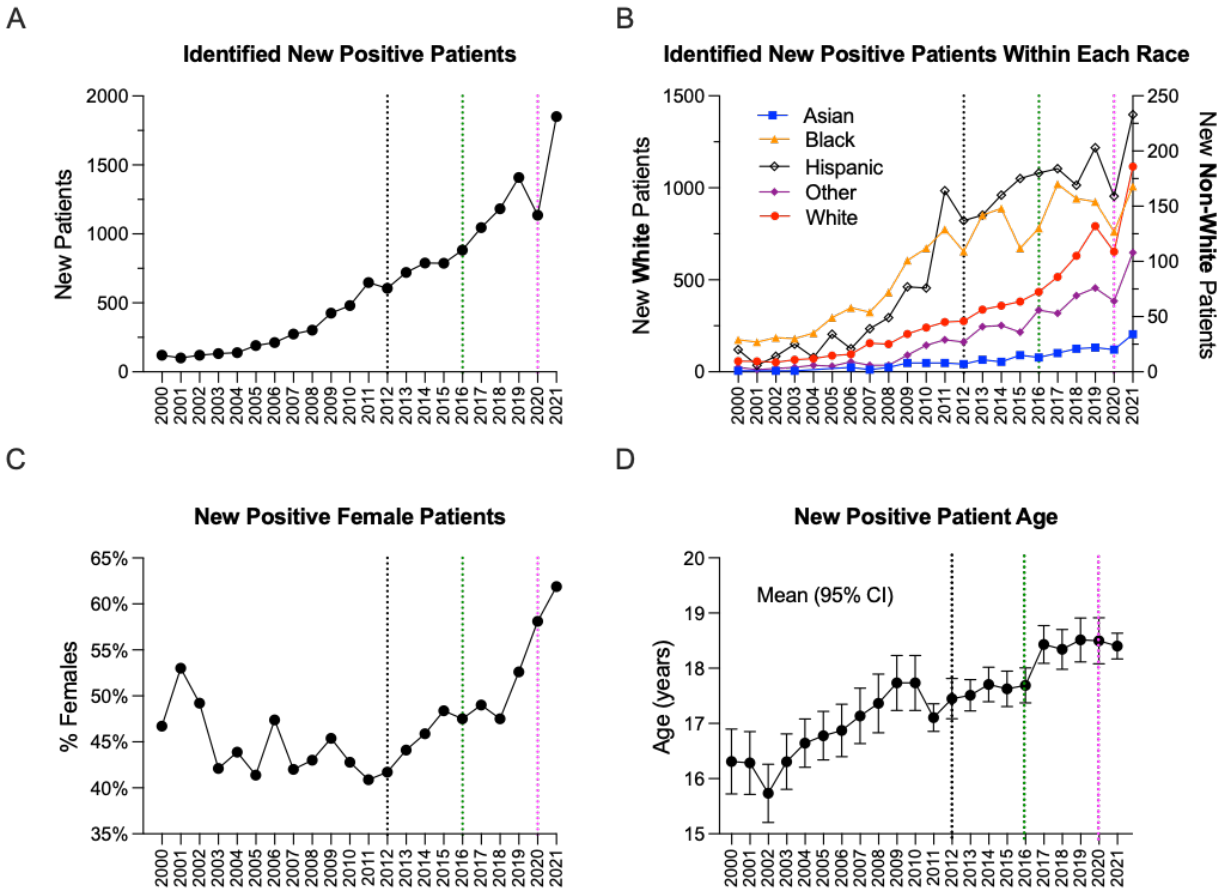

Figure 6. (A) Number of new identified cannabis positive patients per year. (B) Number of new identified cannabis positive patients by each race (other includes Native American, Alaska native, native Hawaiian or other Pacific Islander, and multi-racial). Due to a substantial imbalance in number of white and non-white patients, white patients are plotted on left y-axis and non-white patients are plotted on the right y-axis. (C) New identified female cannabis positive patients as a percentage of all new cannabis positive patients. (D) Average (95% CI) age of new identified cannabis positive patients. The vertical dotted lines indicate the legalization of medical (2012, black) and recreational (2016, green) cannabis in Massachusetts, and declaration of COVID-19 pandemic (2020, magenta).

The distribution of musculoskeletal procedures for both positive and negative cohorts were comparable with application of casting and splints for bone fractures as the most prevalent procedure (Figure 7).

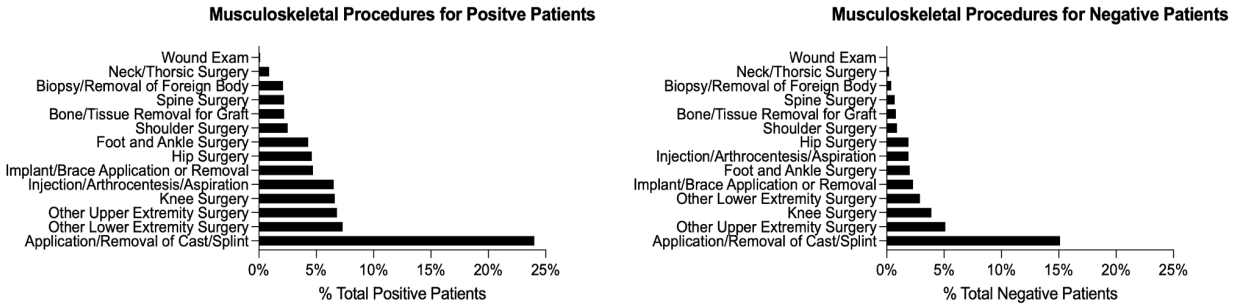

Figure 7. Distribution of common musculoskeletal procedure groups based on CPT billing codes.

Compared to males, females had higher odds of having cannabis use documented in their clinical notes. (aOR=1.074; P<.001; Table 5). Compared to white patients, Asian patients had lower odds (aOR=0.829, P<.001), while black patients (aOR=3.400, P<.001), Hispanic patients (aOR=2.245 P<.001) and those from other racial backgrounds (aOR=1.164, P<.001) had higher odds of cannabis use documentation within clinical notes (Table 5). Patients with higher SVI had higher odds of having cannabis use documented in their clinical notes (aOR=1.001, P<.001; Table 5). We observed similar trends when excluding the medical cannabis use from the patients with documented cannabis use (Supplementary Table S1).

Table 5. Characteristics of patients with and without cannabis use documentation.

Characteristics	Patients with Documented Cannabis Use	Patients without Documented Cannabis Use	aOR (95% CI)	P-Value
Total Sample, No.	21,839	348,248	NR	NR
Sex, No (%)				
Female	11,471 (52.5%)	183,311(52.6%)	1.074 [a] (1.039, 1.110)	**<.001**
Male	10,366 (47.5%)	164,901 (47.4%)	NR	NR
Unknown	2 (<0.1%)	36 (<0.1%)	NR	NR
Race, No. (%)				
Asian	404 (1.8%)	8,057 (2.3%)	0.829 [b] (0.742, 0.926)	**.001**
Black	3,359 (15.4%)	14,403 (4.1%)	3.400 [b] (3.220, 3.590)	**<.001**
Hispanic	3,578 (16.4%)	21,188 (6.1%)	2.245 [b] (2.129, 2.368)	**<.001**
Other c[c]	1,237 (5.7%)	16,723 (4.8%)	1.164 [b] (1.001, 1.002)	**<.001**
White	11,249 (51.5%)	195,849 (56.2%)	NR	NR
Unavailable	2,012 (9.2%)	92,028 (26.4%)	NR	NR
SVI, mean (SD)	0.306 (0.265)	0.269 (0.243)	1.001[d] (1.001, 1.002)	**<.001**

[a] aOR calculated relative to the male patients.
[b] aOR calculated relative to the white patients.
[c] Native American, Alaska native, native Hawaiian or other Pacific Islander, and multi-racial.
[d] aOR calculated per every 0.01-point change in SVI.

Compared to males, females had lower odds of positive cannabis use (aOR=0.957; P=.038; Table 5). Compared to white patients, Asian patients had lower odds (aOR=0.651, P<.001), while black (aOR=3.222, P<.001) and Hispanic (aOR=2.131, P<.001) patients had higher odds of positive

cannabis use (Table 5). Higher SVI was associated with greater odds of positive cannabis use (aOR=1.002, P<.001; Table 5).

Table 6. Characteristics of positive cannabis use patients compared to negative patients.

Characteristics	Positive Patients	Negative Patients	aOR (95% CI)	P-Value
Total Sample, No.	13,556	356,531	NR	NR
Age at the first documented use, mean (SD), y	17.9 (5.4)	NR	NR	NR
With diagnosed cannabis abuse disorder [a], No (%)	1,971 (14.5%)	NR	NR	NR
Sex, No (%)				
Female	6,728 (49.6%)	188,054 (53.7%)	0.957 [b] (0.918, 0.998)	**.038**
Male	6,827 (50.4%)	168,440 (47.2%)	NR	NR
Unknown	1 (<0.1%)	37 (<0.1%)	NR	NR
Race, No. (%)				
Asian	206 (1.5%)	8,255 (2.3%)	0.651 [c] (0.557, 0.761)	**<.001**
Black	2,144 (15.8%)	15,618 (4.4%)	3.222 [c] (3.014, 3.444)	**<.001**
Hispanic	2,280 (16.8%)	22,486 (6.3%)	2.131 [c] (1.994, 2.278)	**<.001**
Other [d]	685 (5.1%)	17,275 (4.8%)	1.007 [c] (0.918, 1.104)	.888
White	7,004 (51.7%)	200,094 (56.1%)	NR	NR
Unavailable	1,237 (9.1%)	92,803 (26.0%)	NR	NR
SVI, mean (SD)	0.309 (0.266)	0.269 (0.243)	1.002 [e] (1.001, 1.003)	**<.001**

[a] Determined based on diagnostic ICD codes.
[b] aOR calculated relative to the male patients.
[c] aOR calculated relative to the white patients.
[d] Native American, Alaska native, native Hawaiian or other Pacific Islander, and multi-racial.
[e] aOR calculated per every 0.01 point change in SVI.

Discussion

This study demonstrates the utility of NLP in evaluating cannabis use and its documentation from heterogeneous and unstructured clinical notes. The generated large-scale database enabled us to investigate the changes in cannabis use documentation and positivity rate, as well as potential disparities in cannabis use and documentation among children and young adults over a period of 21 years. We observed an increasing trend in cannabis use documentation, in particular over the past decade, with smaller proportions within the OSM clinical notes. These increases included both overall cannabis use documentation (e.g., positive or negative) and documented positive use. Multiple factors may have contributed to these increasing trends. In particular, the legalization of medical and recreational cannabis use, the development of new cannabis-based medications (e.g. Epidiolex, Marinol) as well as the cultural and policy changes may all have contributed to observed increases in cannabis use documentation. Cannabis was legalized for medical use in the state of Massachusetts in late 2012 and for recreational use later in 2016. These legalizations have possibly resulted in an increased cannabis usage [26, 27] and/or greater attention paid to the documentation of cannabis use during visits [28,29]. Despite increasing trends, a very small portion of notes overall documented cannabis use (<2%), which for the most part contained insufficient information (i.e., duration, frequencies, amount). These findings are in line with recent studies indicating discrepancies between patient reported cannabis (e.g., surveys) and cannabis use documentation in health records [30], while showing a much lower rate than those previously reported.

We saw significant disparities in cannabis use and its documentation. Female patients, non-white non-asian racial groups and those from a more vulnerable socioeconomic backgrounds (i.e., higher SVI) had higher rates of cannabis documentation (whether the patient was asked about cannabis usage or cannabis was discussed with the patient in any way) in their clinical notes. With regards to cannabis use (i.e., positive vs negative), we saw higher rates of cannabis use among male patients, black or Hispanic patients, and those from a more vulnerable socioeconomic background. While the observed discrepancies in use are in agreement with those reported in prior studies [31, 32], the demographic and socioeconomic differences in cannabis documentation highlight potential biases and implementation challenges in proper tracking of cannabis use among children and young adults. The observed racial and socioeconomic disparities in cannabis use

documentation may also be an underlying factor for the observed higher positive cannabis use rates among those patients. This highlights the need for a systematic and comprehensive approach to discuss and document cannabis use, regardless of patient demographic and socioeconomic background, which may in turn lead to a more accurate assessment of true disparities in cannabis use.

We saw a comparable distribution of musculoskeletal procedures between cannabis users and non-users. This may suggest that the use of cannabis is not primarily influenced by underlying injuries and health conditions. However, whether the use of cannabis has downstream effects on the treatment choices and outcomes requires further evaluations. The high prevalence of cast/splint procedures, often done for treatment of bone fractures, among cannabis users along with prior reports of changes in bone remodeling due to cannabis use [3, 4, 30, 33- 37] further highlight the importance of proper cannabis use documentation to improve treatment outcomes.

There are several limitations which should be considered when interpreting the current findings. The primary limitation of this study is limited generalizability. This study focused on a single institution, in a single state. Other states may have different policies regulating cannabis access, and patients in those states may have different cultural norms around cannabis use and disclosure. In addition, this study analyzed cannabis usage in a binary (positive/negative) manner since in most cases information on duration, frequency, and dosage were missing. Also, with the exception of toxicology reports and medical cannabis use (e.g., prescription drugs) , cannabis use was self-reported which may have biased our observations. Despite these limitations, this study is among the most comprehensive efforts to evaluate cannabis use documentation in a large corpus of clinical notes. Further, to our knowledge, our developed NLP pipeline has the highest performance metrics in capturing a diverse set of cannabis use patterns from clinical notes, which adds to the reliability of the observed findings.

Conclusions

In this study a hierarchical NLP pipeline was developed and used to conduct an in-depth analysis of cannabis use documentation in a large and heterogeneous corpus of clinical notes obtained over

a period of 21 years. The promising performance of the developed NLP pipeline (i.e. AUROC=0.94) supports the use of these approaches in large-scale screening of clinical notes for even complex tasks such as medical and recreational drug use. The findings also highlight the need for rigorous guidelines for proper screening and documentation of cannabis use in electronic health records. Such efforts would enable evidence-based research into the clinical impact of cannabis and support efforts to address observed disparities not only among cannabis users but also in whether patients were asked about cannabis use during their visits.

References

[1] , Substance Abuse and Mental Health Services Administration, "Key substance use and mental health indicators in the United States: Results from the 2019 National Survey on Drug Use and Health," (HHS Publication No., Rockville, MD: Center for Behavioral Health Statistics and Quality, Substance Abuse and Mental Health Services Administration, 2020.

[2] , G. Vargas, L. A. Shrier, N. Chadi and S. K. Harris, "High potency cannabis use in adolescence," The Journal of Pediatrics, 14 August 2022.

[3] , C. M. O'Connor, A. A. Anoushiravani, C. Adams, J. R. Young, K. Richardson and A. J. Rosenbaum, "Cannabinoid use in musculoskeletal illness: a review of the current evidence," Current Reviews in Musculoskeletal Medicine, vol. 13, pp. 379--384, 2020.

[4] , D. Heath, A. Ghali, D. Momtaz, L. Lee and . G. Hogue, "Marijuana use results in increased time to union in surgically treated pediatric fracture patients," Journal of Orthopaedic Trauma, vol. 36, pp. e106--e110, 2022.

[5] , D. G. Deckey, M. Doan, J. D. Hassebrock, K. A. Patel, K. Economopoulos, J. M. Tokish, J. S. Bingham and A. Chhabra, "Prevalence of Cannabinoid (CBD) Use in Orthopaedic Sports Medicine Patients," Orthopaedic Journal of Sports Medicine, vol. 10, no. 4, 2022.

[6] , B. J. Vivace, A. N. Sanders, S. D. Glassman, L. Y. Carreon, J. L. Laratta and J. L. Gum, "Cannabinoids and orthopedic surgery: a systematic review of therapeutic studies," Journal of Orthopaedic Surgery and Research, vol. 16, no. 1, pp. 1--15, 2021.

[7] , A. S. Moon, W. Smith, S. Mullen, B. A. Ponce, G. McGwin, A. Shah and S. M. Naranhe, "Marijuana use and mortality following orthopedic surgical procedures," Substance Abuse, 2018.

[8] , T. Y. Law, J. Kurowicki, S. Rosas, K. Sabeh, S. Summers, Z. Hubbard and M. Roche, "Cannabis use increases risk for revision after total knee arthroplasty," Journal of long-term effects of medical implants, vol. 28, 2018.

[9] , R. G. Chiu, S. Patel, N. Siddiqui, R. S. Nunna and A. I. Mehta, "Cannabis abuse and perioperative complications following inpatient spine surgery in the United States," Spine, vol. 46, no. 11, pp. 734--743, 2021.

[10] , S. H. Medina, V. Nadarajah, J. J. Jauregui, M. P. Smuda, M. Foster, S. J. Meredith, J. D. Packer and R. F. Henn, "Orthopaedic surgery patients who use recreational marijuana have less pre-operative pain," International Orthopaedics, vol. 43, no. 2, pp. 283-292, 2019.

[11] , A. Greis, E. Larsen, C. Liu, B. Renslo, A. Radakrishnan and A. R. Wilson-Poe, "Perceived efficacy, reduced prescription drug use, and minimal side effects of cannabis in patients with chronic orthopedic pain," Cannabis and Cannabinoid Research, 2021.

[12] , A. R. Bhashyam, M. Heng, M. B. Harris, M. S. Vrahas and M. J. Weaver, "Self-reported marijuana use is associated with increased use of prescription opioids following traumatic musculoskeletal injury," JBJS, vol. 100, no. 24, pp. 2095--2102, 2018.

[13] , B. J. Shore, R. Flaugh, B. A. Shannon, P. Curran and G. Hogue, "Preoperative considerations for teenagers undergoing orthopaedic surgery: VTE prevention, mental health assessment, vaping, and drug addiction," Journal of Pediatric Orthopaedics, vol. 41, pp. S64--S69, 2021.

[14] , Y. Ni, A. Bachtel, K. Nause and S. Beal, "Automated detection of substance use information from electronic health records for a pediatric population," Journal of the American Medical Informatics Association, vol. 28, no. 10, pp. 2116--2127, 2021.

[15] , S. A. Ryan, P. K. Gonzalez, S. W. Patrick, J. Quigley, L. Siqueria, L. R. Walker and a. others, "Medication-assisted treatment of adolescents with opioid use disorders," Pediatrics, vol. 138, no. 3, 2016.

[16] , N. Tavabi, M. Singh, J. Pruneski and A. M. Kiapour, "Systematic Evaluation of Common Natural Language Processing Techniques to Codify Clinical Notes," medRxiv, 2022.

[17] , L. Wang, L. Sha, J. R. Lakin, J. Bynum, D. W. Bates, P. Hong and L. Zhou, "Development and validation of a deep learning algorithm for mortality prediction in selecting patients with dementia for earlier palliative care interventions," JAMA network open, vol. 2, no. 7, 2019.

[18] , A. Y. Ling, A. W. Kurian, J. L. Caswell-Jin, G. W. Sledge Jr, N. H. Shah and S. R. Tamang, "Using natural language processing to construct a metastatic breast cancer cohort from linked cancer registry and electronic medical records data," JAMIA open, vol. 2, no. 4, 2019.

[19] , M. Afshar, A. Phillips, N. Karnik, J. Mueller, D. To, R. Gonzalez, R. Price, R. Cooper, C. Joyce and D. Dligach, "Natural language processing and machine learning to identify alcohol misuse from the electronic health record in trauma patients: development and internal validation," Journal of the American Medical Informatics Association, vol. 26, no. 3, pp. 254--261, 2019.

[20] , Y. Wang, E. S. Chen, S. Pakhomov, E. Arsoniadis, E. W. Carter, E. Lindemann, I. N. Sarkar and G. B. Melton, "Automated extraction of substance use information from clinical texts," AMIA Annual Symposium Proceedings, vol. 2015, p. 2121, 2015.

[21] , D. S. Carrell, D. J. Cronkite, M. Shea, M. Oliver, C. Luce, T. E. Matson, J. F. Bobb, C. Hsu, I. A. Binswanger, K. C. Browne and others, "Clinical documentation of patient-reported medical cannabis use in primary care: Toward scalable extraction using natural language processing methods," Substance Abuse, vol. 43, no. 1, pp. 917--924, 2022.

[22] , Agency for Toxic Substances and Disease Registry, "Agency for Toxic Substances and Disease Registry," Agency for Toxic Substances and Disease Registry, 15 March 2022. [Online]. Available: https://www.atsdr.cdc.gov/placeandhealth/svi/index.html.

[23] , J. Devlin, M.-W. Chang, K. Lee and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," arXiv preprint arXiv:1810.04805, 2018.

[24] , E. Alsentzer, J. R. Murphy, W. Boag, W.-H. Weng, D. Jin, T. Naumann and M. McDermott, "Publicly available clinical BERT embeddings," arXiv preprint arXiv:1904.03323, 2019.

[25] , A. E. Johnson, T. J. Pollard, L. Shen, L.-w. H. Lehman, M. Feng, M. Ghassemi, B. Moody, P. Szolovits, L. Anthony Celi and R. G. Mark, "MIMIC-III, a freely accessible critical care database," Scientific data, vol. 3, no. 1, pp. 1--9, 2016.

[26] , G. S. Wang, S. D. Davies, L. S. Halmo, A. Sass and R. D. Mistry, "Impact of marijuana legalization in Colorado on adolescent emergency and urgent care visits," Journal of Adolescent Health, vol. 63, no. 2, pp. 239--241, 2018.

[27] , A. M. Miller, R. Rosenman and B. W. Cowan, "Recreational marijuana legalization and college student use: Early evidence," SSM-Population Health, vol. 3, pp. 649--657, 2017.

[28] , J. M. Jennings, M. A. Williams, D. L. Levy, R. M. Johnson, C. L. Eschen and D. A. Dennis, "Has self-reported marijuana use changed in patients undergoing total joint arthroplasty after the legalization of marijuana?," Clinical Orthopaedics and Related Research, vol. 477, no. 1, p. 95, 2019.

[29] , D. I. Vigil, M. Van Dyke, K. E. Hall, A. E. Contreras, T. S. Ghosh and L. Wolk, "Marijuana use and related health care encounters in Colorado before and after retail legalization," International Journal of Mental Health and Addiction, vol. 16, no. 4, pp. 8106--812, 2018.

[30] , G. T. Lapham, T. E. Matson, D. S. Carrell, J. F. Bobb, C. Luce, M. M. Oliver, U. E. Ghitza, C. Hsu, K. C. Browne, I. A. Binswanger and others, "Comparison of Medical Cannabis Use Reported on a Confidential Survey vs Documented in the Electronic Health Record Among Primary Care Patients," JAMA Network Open, vol. 5, no. 5, 2022.

[31] , A. M. Jeffers, S. Glantz, A. Byers and S. Keyhani, "Sociodemographic characteristics associated with and prevalence and frequency of cannabis use among adults in the US," JAMA network open, vol. 4, no. 11, 2021.

[32] , S. S. Martins, L. E. Segura, N. S. Levy, P. M. Mauro, C. M. Mauro, M. M. Philbin and D. S. Hasin, "Racial and ethnic differences in cannabis use following legalization in US states with medical cannabis laws," JAMA Network open, vol. 4, no. 9, 2021.

[33] , D. S. Hasin, "US epidemiology of cannabis use and associated problems," Neuropsychopharmacology, vol. 43, no. 1, pp. 195--212, 2018.

[34] , N. M. Kogan, E. Melamed, E. Wasserman, B. Raphael, A. Breuer, K. S. Stok, R. Sondergaard, A. V. Escudero, S. Baraghithy, M. Attar-Namdar and others, "Cannabidiol, a major

non-psychotropic cannabis constituent enhances fracture healing and stimulates lysyl hydroxylase activity in osteoblasts," Journal of Bone and Mineral Research, vol. 30, no. 10, pp. 1905--1913, 2015.

[35] , H. Johal, C. Vannabouathong, Y. Chang, M. Zhu and M. Bhandari, "Medical cannabis for orthopaedic patients with chronic musculoskeletal pain: does evidence support its use?," Therapeutic advances in musculoskeletal disease, vol. 12, 2020.

[36] , A. M. Jakoi, G. J. Kirchner, Y. E. Kerbel, J. A. Iorio and A. S. Khalsa, "The effects of marijuana use on lumbar spinal fusion," Spine, vol. 45, no. 9, pp. 629-634, 2020.

[37] , J. M. Jennings, M. R. Angerame, C. L. Eschen, A. J. Phocas and D. A. Dennis, "Cannabis use does not affect outcomes after total knee arthroplasty," The Journal of Arthroplasty, vol. 34,

Competing Interest: The authors have no relevant financial or non-financial interests to disclose.

Author Contributions Statement: Study concept and design was done by AMK and GDH. Data collection was done by all authors. Analysis was done by NT and AMK. Interpretation of the findings was done by all authors. First draft of the manuscript was developed by NT, MR, MS, and AMK. All co-authors contributed to the reviewing and finalizing the manuscript.

Ethical Approval: This study was performed in line with the principles of the Declaration of Helsinki. Approval was granted by the Institutional Review Board of Boston Children's Hospital.

Consent to Participate: This a retrospective study and consent was not required as approved by the IRB.

Funding/Support: The study was funded by the Children's Orthopaedic Surgery Foundation (AMK) and Boston Children's Hospital Research Faculty Council (AMK). We also received hardware support from the NVIDIA Basic Research Accelerator Program (AMK).

Data Availability: The data used in this study cannot be publicly shared due to privacy concerns. Data can be access upon reasonable request pending approvals from Boston Children's Hospital.

Supplementary Materials

Supplementary Table S1. Characteristics of patients with and without cannabis use documentation excluding medical cannabis use.

Characteristics	Documented Patients	Not Documented Patients	aOR (95% CI)	P-Value
Total Sample, No.	20,312	348,248	NR	NR
Sex, No (%)				
Female	10,739 (52.9%)	183,311(52.6%)	1.098 [a] (1.061, 1.137)	**<.001**
Male	9,572 (47.1%)	164,901 (47.4%)	NR	NR
Unknown	1 (<0.1%)	36 (<0.1%)	NR	NR
Race, No. (%)				
Asian	365 (1.8%)	8,057 (2.3%)	0.826 [b] (0.735, 0.928)	**.001**
Black	3,303 (16.3%)	14,403 (4.1%)	3.713 [b] (3.514, 3.924)	**<.001**
Hispanic	3,427 (16.9%)	21,188 (6.1%)	2.362 [b] (2.236, 2.495)	**<.001**
Other c[c]	1,134 (5.6%)	16,723 (4.8%)	1.200 [b] (1.117, 1.290)	**<.001**
White	10,133 (49.9%)	195,849 (56.2%)	NR	NR
Unavailable	1,950 (9.6%)	92,028 (26.4%)	NR	NR
SVI, mean (SD)	0.308 (0.267)	0.269 (0.243)	1.001[d] (1.001, 1.002)	**<.001**

[a] aOR calculated relative to the male patients.
[b] aOR calculated relative to the white patients.
[c] Native American, Alaska native, native Hawaiian or other Pacific Islander, and multi-racial.
[d] aOR calculated per every 0.01 point change in SVI.