A Natural Language Processing Approach to Extract Detailed Cannabis Use Patterns from Unstructured EHRs

Kimia Zandbiglari, M.Sc.; Shobhan Kumar, Ph.D.; Sebastian Jugl, RPh, MS, BPharm; Masoud Rouhizadeh, PhD, MSc, MA.
University of Florida, Gainesville, FL, USA

Introduction

- There has been a notable increase in both medical and recreational cannabis use. This trend presents new challenges and opportunities for healthcare and policy making.
- Electronic Health Records (EHR) often lack detailed information about cannabis usage, such as the mode of use, frequency, and concurrent use of other substances.
- Capturing these details is crucial for designing effective healthcare interventions and informed public policies. Analyzing unstructured clinical notes through natural language processing can provide the needed insights into cannabis usage patterns.

Objective

- The goal is to use natural language processing to analyze clinical notes and identify detailed patterns of cannabis use, focusing on modes of administration, usage frequency, and co-use with other substances.

Methods

- Data of 500 unique patients at UF Health, analyzed over a 60-day period per patient starting from their first inpatient visit.
- A specialized Cannabis Use Lexicon (CULx) developed by our group to identify mentions of cannabis use within 1162 sentences from clinical notes.
- Manual review to categorize cannabis use mentions across:
  - Mode of Administration: By Mouth, Smoking/Vaping, Topical, Unknown.
  - Frequency of Use: Daily, Weekly, Occasionally, Unknown.
  - Co-use with Other Substances: Alcohol, Illicit Drugs, Opioids/Opiates, Other Medications, Tobacco/Nicotine, Unknown.
- NLP models employed for classification: BERT, RoBERTa, Clinical BERT, Decision Trees, and Logistic Regression.
- Performance assessment via accuracy, sensitivity, and positive predictive value (PPV). (Figure 1)

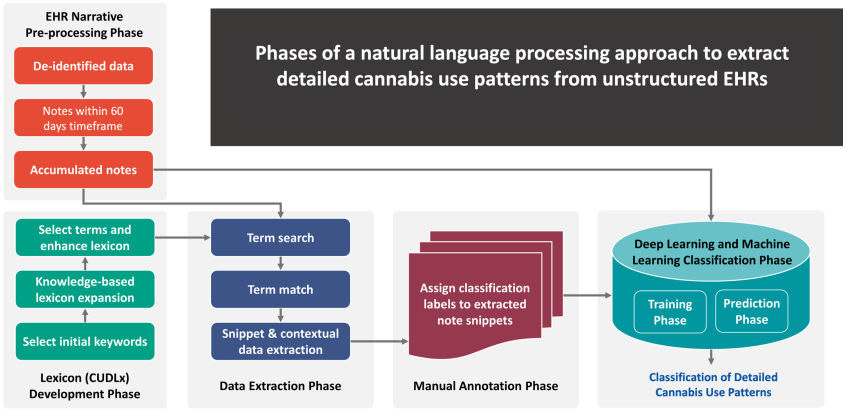

Figure 1: Phases of a natural language processing approach to extract detailed cannabis use patterns from unstructured EHRs

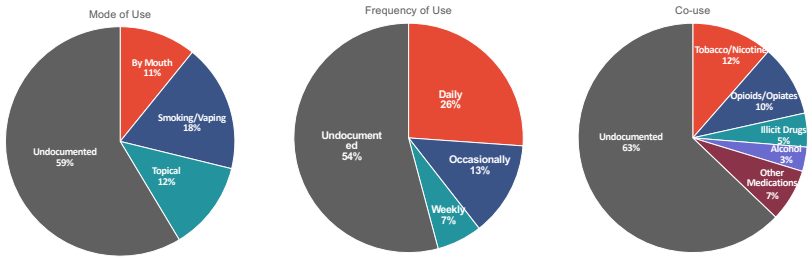

Figure 2: Detailed cannabis use patterns from unstructured EHRs

Table 1: Model performance in extracting detailed cannabis use patterns from unstructured EHRs

	Mode of Use			Frequency of Use			Co-use		
	PPV	accuracy	Sensitivity	PPV	accuracy	Sensitivity	PPV	accuracy	Sensitivity
BERT	88	88	88	88	88	88	94	94	94
RoBERTa	87	87	87	89	88	88	93	93	93
Clinical BERT	87	86	86	90	89	89	91	91	91
Decision Tree	94	93	93	90	88	88	94	94	94
Logistic Regression	93	93	93	89	88	87	93	93	93

Results

- Among 500 patients, already classified as cannabis users (Inter-Rater Reliability: 97%), 39.4% had documented cannabis use methods, 44.2% reported frequency, and 34% co-use with other substances, indicating significant integration in patient records.
- Smoking/vaping and daily use were the most common methods and frequencies, with notable co-use with tobacco/nicotine and alcohol among patients. (Figure 2)
- NLP classifiers achieved up to 93% accuracy, 94% PPV (Positive Predictive Value), and 93% sensitivity, closely matching human annotation.
- RoBERTa: Excelled in identifying mode of use and co-use with other substances, achieving over 93% PPV and sensitivity, and particularly strong in co-use metrics (95% PPV, 94% sensitivity).
- ClinicalBERT: Best at identifying frequency of use, with 89% PPV and 90% sensitivity. (Table 1)

Discussion

- The study revealed frequent mentions of detailed cannabis use patterns in unstructured EHR notes.
- Developed preliminary NLP system to replicate manual categorization of cannabis use details.
- Provides foundation for recognizing and categorizing cannabis use patterns within EHRs.
- Facilitates research into cannabis use trends within unstructured EHR notes.
- A significant portion of cannabis use instances remains undocumented in the EHR notes. While the system shows high sensitivity due to the inclusion of an "undocumented" category, the true nature of these cases is uncertain.
- The NLP system's performance may vary across different datasets and healthcare settings. Further validation is necessary to assess its generalizability and effectiveness in diverse EHR environments.

Contact: K.Zandbiglari@ufl.edu