C-TERM: Cannabis Terminology for EHRs and Medicine

Wei Wei, PhD, Matthew Woll, PhD, MBA, Antoinette DiCriscio, PhD, Vanessa Troiani, PhD
Geisinger College of Health Sciences, Danville, PA

Geisinger

BACKGROUND & OBJECTIVES

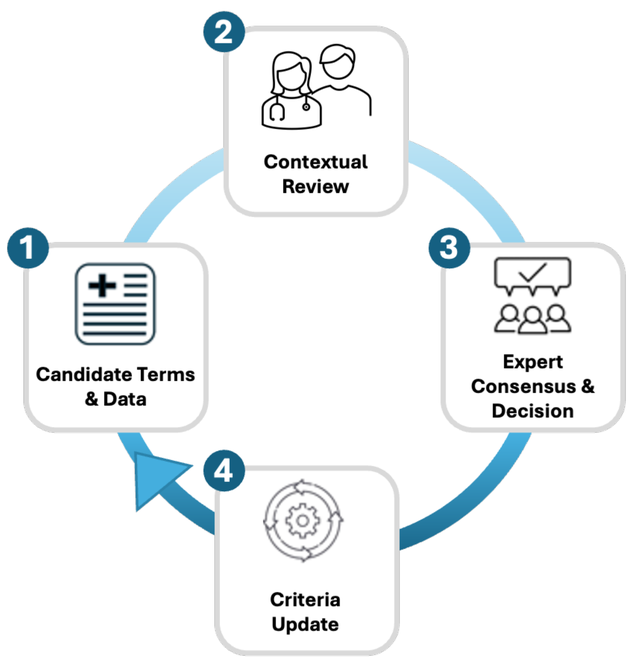
•Cannabis use is insufficiently documented in electronic health records (EHRs)[1], despite rising medical & recreational use in the U.S.

•We have previously found that cannabis use is predominantly captured in unstructured notes[2], thus developing a vocabulary is essential as a starting point for reliable data extraction.

•Prior vocabularies emphasize recreational cannabis use in primary care documentation[4], preoperative risk assessment[3], or web-based drug abuse surveillance[5] and have limited capture of therapeutic use.

•Our project bridges this gap by developing a standardized vocabulary that will be further developed into a formal ontology. The ontology will provide the semantic structure needed to represent relationships among products, chemical compounds, and observations to support interoperability, reasoning, and scalable NLP-based phenotyping.

FIGURE 1. Iterative workflow for refining the C-TERM vocabulary.



METHODS

We started with 284 candidate terms from prior studies and subject matter experts and used an iterative process for refinement. Inclusion and exclusion criteria *(see below)* were applied and term occurrences reviewed in 400 sampled Chronic Pain Cohort (**CPC**) notes. Contextual review and expert consensus guided term selection, with justifications documented and criteria updated throughout. *See Figure 1*.

Inclusion & Exclusion Criteria

1. Covering alternative names, chemical components, generic and brand names, dosage forms, routes of administration, reasons for use, and drug-drug interactions;
2. Aligning with existing terminologies (preferred terms, synonyms, and variants in SNOMED CT, RxNorm, ICD 10, Drug Abuse Ontology, etc.);
3. Minimizing ambiguous or broad terms, jargon, and redundancy.

Data Sets

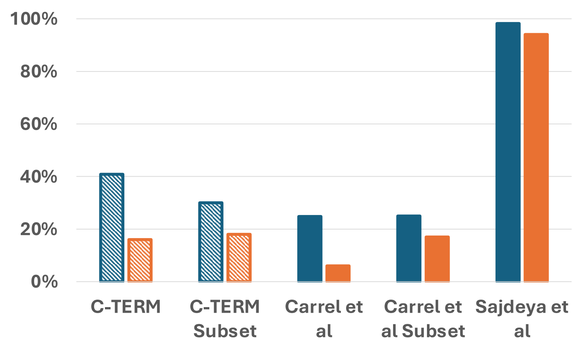
	# MRN	# Notes	Format	Granularity	Label Sampling Method
Chronic Pain Cohort (CPC)	100	174,768	Full notes	N/A	Unfiltered
Annotated Cannabis Snippets (ACS)	3,195	3,650	Text snippets	Snippet level	Keyword enriched
Cannabis Notes Dataset (CND)	32,580	544,818	Text snippets	N/A	Keyword enriched

RESULTS

The resulting vocabulary, Cannabis Terminology for Electronic Health Records and Medicine (**C-TERM**), contains 142 terms, including 26 primary terms that directly indicate cannabis use and 116 secondary terms that frequently co-occur with primary terms. C-TERM was evaluated on two datasets and compared with two baseline vocabularies[3,4]. Sajdeya's vocabulary includes 3,626 terms, while Carrel's vocabulary comprises 145 terms. For a focused comparison, a C-TERM subset (44 terms) and a Carrel subset (24 terms) were created by excluding condition- and symptom-related terms like injury.

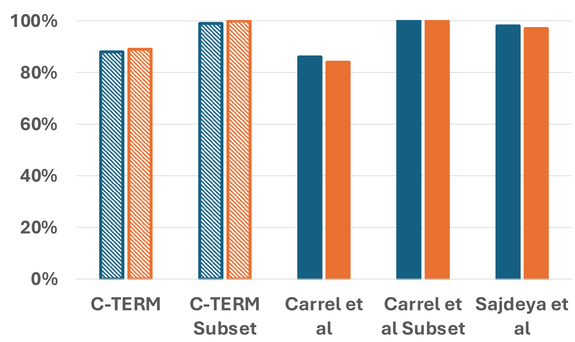
FIGURE 2. Evaluation of C-TERM and reference vocabularies. Blue bars represent metrics calculated on the ACS dataset and **orange** bars represent CND dataset. C-TERM results indicated with diagonal stripes.

A. Zero Hit Rate (lower is better)



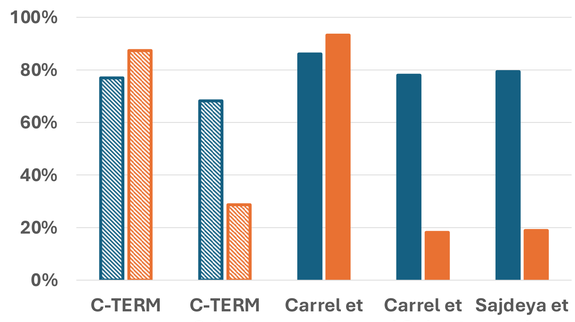
Zero Hit Rate (ZHR) = Number of zero-hit terms / Vocabulary size. C-TERM and Carrel exhibit lower ZHRs than Sajdeya, indicating stronger alignment with real-world EHR language and fewer unused terms.

B. Productivity@25 (higher is better)



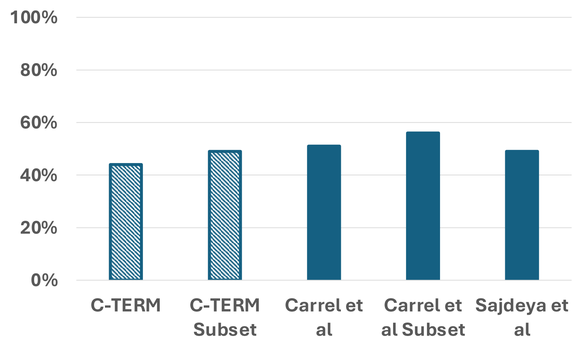
The 25 most frequent terms in the C-TERM subset captured 100% of all matches in the CND dataset, demonstrating that C-TERM is both compact and highly efficient.

C. Doc Hit Rate (higher is better)



Document Hit Rate (DHR) = Documents with hits / Dataset size. Carrel's higher DHR likely reflects inclusion of nonspecific terms (e.g., injury). After excluding these, C-TERM outperformed Carrel in the CND dataset.

D. Conditional PPV (higher is better)



Precision is measured by the proportion of retrieved snippets labeled positive. Higher values indicate better targeting of cannabis-use expressions and exclusion of vague or irrelevant terms.

DISCUSSION & NEXT STEPS

C-TERM is a streamlined vocabulary tailored for information extraction, avoiding bulky ontology or coding system terms like ICD-10. Compared to Sajdeya et al., it's more compact, and unlike Carrel's cannabis-adjacent vocabulary, C-TERM targets cannabis-specific expressions. In the CND dataset, C-TERM outperformed Carrel's vocabulary (29% vs. 18%), showing superior precision in identifying cannabis use in clinical text. We are expanding C-TERM into an ontology to support semantic reasoning and cross-source interoperability.

REFERENCES & ACKNOWLEDGEMENTS

1. Tavabi N, et al. npj Digit Med. 2023;6(1):138.
2. Pradhan AM, et al. J Patient Cent Res Rev. 2024;11:118-80.
3. Sajdeya R, et al. JAMIA. 2023;30(8):1418–28.
4.Carrell, D. S. et al. Subst. Abus. 43, 917–924 (2022).
5. Lokala U, et al. JMIR Public Health and Surveillance. 2022 Dec;8.

Geisinger is the recipient of research support from Ascend Wellness Holdings, an approved medical marijuana clinical registrant in Pennsylvania. The funder was not involved in the study.