Title: Comparing computable structured phenotype- versus large language model-identification of opioid use disorder using electronic health record data

Authors: Melanie Molina, MD, MAS[1,2]; Cynthia Fenton[2]; Kathy T. LeSaint, MD[1]; Samuel D. Pimentel, PhD[3] Aaron E. Kornblith MD, MS[1,4]

Author Affiliations:
[1]Department of Emergency Medicine, University of California, San Francisco, California, USA
[2]Division of Clinical Informatics and Digital Transformation, Department of Medicine, University of California, San Francisco, California, USA
[3]Department of Statistics, University of California, Berkeley
[4]Department of Pediatrics, University of California, San Francisco, California, USA

Address manuscript correspondence to:
Melanie F. Molina, MD, MAS
UCSF Department of Emergency Medicine
Room L-126, Box 0209
505 Parnassus Avenue
San Francisco, CA 94143
Tel: 512-294-4366
Email: melanie.molina@ucsf.edu

Keywords: opioid use disorder, electronic health record data, large language models

Conflict of Interest: Dr. Melanie Molina is supported by the National Institute on Drug Abuse (K23DA060993) and the Harold Amos Medical Faculty Development Program. Dr. Aaron Kornblith is a consultant and co-founder for UCSF spinout, Capture Dx, and supported by the Eunice Kennedy Shriver National Institute of Child Health and Human Development (K23HD110716).

Word count: 1,975

Abstract

Introduction

Opioid use disorder (OUD) is common in emergency departments (EDs); identification via structured computable phenotypes may miss important clinical context.

Objective

Compare a computable structured OUD phenotype with a zero-shot large language model (LLM) using expert review as the reference.

Methods

We retrospectively analyzed 202 adult ED encounters. Two emergency physicians independently determined OUD status with consensus adjudication. The phenotype used ICD-10 codes, medications, toxicology, consult notes, and keyword rules. The LLM (GPT-4.1) classified OUD from concatenated ED notes. Test characteristics and McNemar's tests were computed.

Results

Experts classified 56 (28%) encounters as OUD (κ=0.77). The phenotype showed sensitivity 0.98 (95% CI 0.93–1.00) and specificity 0.54 (0.44–0.64). The LLM showed sensitivity 0.93 (0.87–0.96) and specificity 0.90 (0.77–0.97). Sensitivity (p=0.0117) and specificity (p<0.001) differed significantly.

Conclusion

A zero-shot LLM achieved balanced performance, outperforming the structured phenotype on specificity while maintaining high sensitivity, supporting tiered ED screening for OUD.

BACKGROUND

Opioid use disorder (OUD) affects over 2.5 million Americans[1] and contributed to more than 80,000 overdose deaths in 2022,[2] constituting a critical public health crisis. The emergency department (ED) serves as a vital access point for many patients with OUD,[3] whether presenting with overdose, withdrawal symptoms, or other complications. This presents a critical opportunity to connect patients with appropriate treatment and support services.

However, screening for OUD in the ED is often limited due to competing clinical priorities, time constraints, and resource limitations.[4] Failure to identify OUD during the ED visit is a missed opportunity, as recognition is the essential first step toward offering treatment, connecting patients to life-saving resources, and care coordination.[5] To address this challenge, prior work has attempted to define OUD using structured, computable electronic health record (EHR) phenotypes that could trigger downstream processes such as clinical decision support tools.[6] These approaches, validated across multiple health systems,[6] show promise by potentially improving workflow efficiency and bypassing the need for manual screening by busy clinicians.

Nevertheless, structured phenotypes may be insufficient to capture the full complexity of OUD presentation and often miss important contextual information embedded in clinical notes. Recent advances in artificial intelligence (AI) offer additional promise for improving OUD identification. While prior work using convolutional neural networks has demonstrated potential for efficient and reliable OUD identification,[7] studies have yet to analyze how large language models (LLMs) perform in this context, particularly in zero-shot scenarios without task-specific training.

Given that Epic and other EHR vendors have begun rolling out LLM-based chart summarization tools, there is immense potential for scalability of LLM-based OUD identification. However, the first step toward integrating novel LLM capabilities is understanding where they work well, where they fall short, and under what conditions their outputs are reliable.

OBJECTIVE

We sought to compare a computable structured OUD phenotype against an LLM for OUD identification, using expert human determination as a reference standard.

METHODS

Study Design and Data Source

This study analyzed EHR data extracted from Clarity at the University of California, San Francisco (UCSF), with IRB approval. We began with 70,880 ED encounters for patients ≥18 years between January 1, 2023 and October 17, 2024. We used a multi-step randomization process to generate a balanced sample of encounters for classification testing. We first stratified encounters by OUD status using a computable structured OUD phenotype[8] (described below). Within each stratum, we randomly selected one encounter per unique patient, eliminating duplicates. We then performed stratified random sampling to obtain 204 patient-encounters: 102 meeting the structured OUD phenotype and 102 not meeting the phenotype. All randomization and data processing were performed using standard SQL queries and window functions. To be eligible for inclusion, each ED encounter required associated ED notes for LLM analysis.

OUD Identification Methods

Computable Structured Phenotype

To ensure identification of vulnerable patients who could benefit from social work and care coordination, we employed a broad approach aimed at capturing patients with clinical evidence of OUD even without a formal diagnosis. We adapted the OUD phenotype originally developed

and validated by Chartash et al.[6] for our local implementation as part of a clinical decision support tool designed to identify high-risk OUD patients with unmet social needs. The original phenotype was enhanced to improve sensitivity and accommodate our institutional EHR structure, as OUD documentation is often incomplete due to stigma and variable clinical practices.

Our site-adapted phenotype[8] includes: ≥1 ICD-10 codes indicating OUD, withdrawal, or overdose in encounter (discharge and/or admission) diagnoses, problem lists, or past medical history; active prescriptions for buprenorphine or methadone preceding the index ED visit encounter; ≥1 positive urine toxicology for methadone, fentanyl, or heroin; ≥1 addiction care team consultation notes mentioning opioid-related terms; and text recognition of specific terms ("OUD," "opioid use disorder," "COWS," "heroin," "fentanyl") in ED clinical notes, including reason for visit free text.

Large Language Model
We used OpenAI ChatGPT 4.1 with all notes associated with the ED encounter (excluding student and procedure notes) concatenated into the prompt without additional labels other than note titles. The prompt asked the model to determine whether the patient had OUD based on the clinical documentation (**Supplementary Material Appendix A**).

Human Reference Standard
Two board-certified emergency medicine attending physicians, one with additional certification in addiction medicine and toxicology, independently reviewed each patient's entire chart to determine OUD status. Discrepancies were reconciled through consensus discussion.

Analysis
We used descriptive statistics to characterize patients within the ED encounters by OUD status. We assessed interrater agreement using Cohen's kappa with 95% confidence intervals. For each OUD identification method, we calculated sensitivity, specificity, and positive and negative likelihood ratios. We applied McNemar's test to assess whether differences in sensitivity and specificity across identification methods were statistically significant. All quantitative analyses were conducted in R version 4.4.2. We also performed qualitative analysis of discordant cases using rapid content analysis to identify themes.

RESULTS
Of the 204 selected encounters, two were excluded due to missing ED provider notes, consistent with the 0.8% frequency of missing notes in the overall population. The final analytic sample included 202 encounters, with 101 meeting the computable structured phenotype and 101 not meeting the phenotype (by design). Of the remaining encounters, 33% (67/202) had multiple ED note types. Patient demographic characteristics are shown in **Table 1**.

Table 1. Patient Demographic Characteristics of ED Encounters

	Total Encounters N (%)	OUD Encounters n (%)	Non-OUD Encounters n (%)
Overall	204 (100%)	102 (50%)	102 (50%)
Age, Mean (Std. Dev.)	50.9 (18.6)	48.6 (16.5)	53.3 (20.3)
Race and Ethnicity			
Asian	32 (15.7%)	6 (5.9%)	26 (25.5%)

Black	38 (18.6%)	20 (19.6%)	18 (17.6%)
Latinx	27 (13.2%)	11 (10.8%)	16 (15.7%)
Multi-racial	9 (4.4%)	6 (5.9%)	3 (2.9%)
Other[a]	11 (5.4%)	9 (8.8%)	2 (2%)
White	87 (42.6%)	50 (49%)	37 (36.3%)
Sex			
Male	124 (60.8%)	68 (66.7%)	56 (54.9%)
Female	76 (37.3%)	30 (29.4%)	46 (45.1%)
Other[b]	4 (2%)	4 (3.9%)	0 (0%)
Insurance			
Medicaid	57 (27.9%)	46 (45.1%)	11 (10.8%)
Medicare	68 (33.3%)	31 (30.4%)	37 (36.3%)
Commercial	63 (30.9%)	17 (16.7%)	46 (45.1%)
Other[c]	16 (7.8%)	8 (7.8%)	8 (7.8%)
Language: Interpreter Needed?	16 (7.8%)	4 (3.9%)	12 (11.8%)

Abbreviations: OUD, opioid use disorder
[a]Included “Other”, “Unknown/Declined”, “Southwest Asian and North African”
[b]Included “Choose not to disclose”, “Declines to answer”, “Something else”, “Unknown”
[c]Included “Other” and “Worker’s Comp”

There was strong interrater reliability ($\kappa = 0.77$, 95% CI 0.67-0.87) between the two physicians’ classifications. After consensus, 56 (28%) encounters were classified as having OUD and 146 (72%) as not having OUD. The test characteristics of the structured phenotype and LLM compared to the reference standard are shown in **Table 2**.

Table 2. Test Characteristics of Computable Structured Phenotype versus LLM

Identification Method	Sensitivity (95% CI)	Specificity (95% CI)	LR+ (95% CI)	LR- (95% CI)
Structured vs Reference	0.98 (0.93-1.0)	0.54 (0.44-0.64)	2.13 (1.72-2.63)	0.04 (0.01-0.15)
LLM vs Reference	0.93 (0.87-0.96)	0.90 (0.77-0.97)	8.9 (3.88-20.42)	0.08 (0.05-0.15)

Abbreviations: LLM, large language model

Compared to the reference standard, the computable structured phenotype demonstrated excellent sensitivity (0.98, 95% CI 0.93-1.0) but poor specificity (0.54, 95% CI 0.44-0.64). In contrast, the LLM showed slightly lower sensitivity (0.93, 95% CI 0.87-0.96) but substantially higher specificity (0.90, 95% CI 0.77-0.97), resulting in balanced performance across both metrics. Despite overlapping confidence intervals, McNemar’s test showed significant differences in sensitivity ($p = 0.0117$) and specificity ($p < 0.001$) between the methods.

Qualitative analysis of discordant cases (**Table 3**) revealed distinct error patterns for each approach. Computable structured phenotype false positives primarily resulted from single positive fentanyl urine toxicology results and chronic pain patients on methadone or

buprenorphine, while false negatives occurred when relevant OUD information was documented in notes external to the health system (i.e., Epic's "Care Everywhere"; see Discussion).

Table 3. Themes from Discordant Classifications

	False Positive	False Negative
Structured	- Urine toxicology resulted positive for fentanyl once - Chronic pain on opioid medication (including methadone or buprenorphine)	- Missed mention of overdose responsive to Narcan in Care Everywhere ED note - History of "opiate withdrawal" in Care Everywhere, cardiac arrest presumed to be related to "OUD"
LLM	- Suicide attempt by overdose of fentanyl - Opioid dependence in problem list counted as OUD - Counted withdrawal from kratom as opioid-like substance	- Missed mention of OUD on methadone as part of Care Everywhere - Missed mention of overdose responsive to Narcan in Care Everywhere ED note - Anesthesia Procedure Note mention of polysubstance abuse, including OUD - OUD documented in Care Everywhere notes
Expert 1	- Mention of buprenorphine prescription written "for OUD" - On buprenorphine for "substance abuse" but was kratom dependency	- On opioids for cancer pain, but also used to inject heroin in 1980s and last used in 2023 - Missed mention of regular fentanyl use in behavioral health ED note - Patient with delusions of being "fentanyled" by neighbors, but regular fentanyl positivity and documented use
Expert 2	- Mention of OUD in on ED provider note, but not elsewhere (severe opioid dependence) - Mention of OUD in a progress note, question of abusing prescriptions but on oxycodone for pain	- In the context of other substance use (e.g., methamphetamine), missed mention of occasional fentanyl use - Accidental opioid overdose with patient reporting that was only time he used but missed later urine tox positive for fentanyl - Patient denied fentanyl use but had admission concerning for fentanyl withdrawal

Abbreviations: ED, emergency department**;** OUD, opioid use disorder; LLM, large language model

LLM false positives included misclassifying suicide attempts involving fentanyl overdose as OUD, interpreting opioid dependence diagnoses in chronic pain contexts as OUD, and categorizing kratom withdrawal as opioid related. LLM false negatives predominantly occurred when OUD-related information was documented in Care Everywhere notes or anesthesia procedure notes that may have been less prominent in the concatenated text.

Human expert errors highlighted the complexity of OUD determination even among specialists. Expert 1 false positives included cases where buprenorphine was prescribed "for OUD" but the underlying condition was kratom dependency, while false negatives included missing current fentanyl use in patients with remote heroin use histories or overlooking fentanyl use documented in behavioral health notes. Expert 2 false positives involved interpreting mentions of "severe opioid dependence" in pain management contexts as OUD, while false negatives included missing fentanyl use in polysubstance users and cases where patients denied use despite positive toxicology or clinical signs of withdrawal.

DISCUSSION

To our knowledge, this is the first study directly comparing structured OUD identification to LLM identification. While the computable structured phenotype demonstrated slightly higher sensitivity, the LLM achieved consistently high performance across both sensitivity and specificity measures.

Our structured phenotype, modeled after previously developed algorithms based on ICD-10 codes and clinical keywords,[6] was enhanced with additional considerations including urine toxicology results, medications for opioid use disorder, and addiction care team consultation notes. This likely improved sensitivity compared to prior approaches but resulted in lower specificity, as expected when casting a wider net for case identification.

The complementary performance characteristics of these two approaches suggest a potentially synergistic implementation strategy with a tiered screening approach. Similar to HIV testing protocols that use highly sensitive screening tests followed by highly specific confirmatory tests, OUD identification could benefit from using the structured phenotype as an initial, computationally inexpensive screening tool, followed by LLM analysis for cases that screen positive. This sequential approach could optimize resource utilization by limiting expensive token consumption to the subset of encounters flagged by the structured phenotype, while maintaining the high specificity needed to minimize false positive alerts that contribute to alert fatigue. The structured phenotype's excellent sensitivity (0.98) makes it well-suited for this initial screening role, ensuring few true OUD cases are missed, while the LLM's superior specificity (0.90) could serve as an effective filter to reduce false positives before triggering clinical interventions or decision support tools.

Regarding LLM performance for OUD identification, prior work has compared encoder versus decoder LLMs for extracting OUD-related concepts from clinical notes, demonstrating generative LLMs' strength for structured data extraction.[9] Other studies have evaluated GPT-3.5 for identifying substance use in the MIMIC dataset[10] and quantifying substance use disorder severity.[11] Our study expands upon the literature by demonstrating that LLMs can be used as a screen for OUD altogether, with improved balanced performance compared to structured EHR phenotypes, which have traditionally been used to trigger clinical decision support.[12] Furthermore, we show strong LLM performance using only notes from a single ED encounter, suggesting that relatively limited clinical information can yield excellent diagnostic performance.

Our findings have several important implications. First, while expert interrater reliability was strong, it was not perfect, reflecting the inherent complexity of OUD diagnosis even among specialists. The cases where experts disagreed highlight scenarios where AI approaches face similar challenges: we cannot hold LLMs to a higher standard than human experts. The distinction between opioid dependence in chronic pain management and OUD represents a persistent challenge across all identification methods, emphasizing the need for careful prompt engineering and more nuanced classification approaches. Second, models must be able to access external health information exchange data, such as Epic's Care Everywhere, to comprehensively identify OUD, as critical diagnostic information is often documented across multiple health systems rather than contained within a single institution's records. Given neither the structured phenotype nor LLM had access to Care Everywhere, they were both disadvantaged compared to human reviewers who could review comprehensive cross-system documentation. Third, LLM performance may be particularly valuable in busy clinical environments, such as the ED, where comprehensive chart review is challenging. The ability to process multiple note types simultaneously represents a significant advantage over manual

screening approaches. The balanced performance characteristics of LLMs make them potentially suitable for both screening and diagnostic confirmation roles, unlike structured phenotypes, which may be better suited primarily for sensitive screening with subsequent human review.

While our study used emergency medicine physician reviewers to ensure ED applicability, future research should validate findings across different health systems, patient populations, and clinical settings, potentially including addiction medicine specialists. Our approach using only current ED visit notes, while limiting historical context, would be suitable for identifying OUD in admitted patients and subsequent ED visits where immediate clinical information is most relevant. Additionally, prospective studies examining the impact of LLM-based OUD identification on treatment, clinical outcomes, and healthcare utilization would provide valuable evidence for implementation decisions.

CONCLUSION

LLMs have strong potential for identifying conditions such as OUD, potentially saving time, resources, and costs by reducing the need for manual screening while maintaining high performance against human benchmarks. While the LLM in our study showed slightly lower sensitivity than the computable structured phenotype, it demonstrated substantially higher specificity, resulting in more balanced overall performance. Importantly, this LLM was not trained specifically for OUD identification, underscoring the advantage of general-purpose models that can perform multiple clinical tasks without additional tuning. As LLM summarization tools become more commercially available within EHR systems, automated screening using existing EHR data may become increasingly feasible and scalable across health systems.

REFERENCES

1. National Institute on Drug Abuse. Only 1 in 5 U.S. adults with opioid use disorder received medications to treat it in 2021 | National Institute on Drug Abuse (NIDA). August 7, 2023. Accessed October 30, 2025. https://nida.nih.gov/news-events/news-releases/2023/08/only-1-in-5-us-adults-with-opioid-use-disorder-received-medications-to-treat-it-in-2021

2. National Institute on Drug Abuse. Drug Overdose Deaths: Facts and Figures | National Institute on Drug Abuse (NIDA). August 21, 2024. Accessed October 30, 2025. https://nida.nih.gov/research-topics/trends-statistics/overdose-death-rates

3. Substance Abuse and Mental Health Services Administration. Use of Medication-Assisted Treatment in Emergency Departments. Published online 2021. Accessed May 27, 2023. https://store.samhsa.gov/sites/default/files/pep21-pl-guide-5.pdf

4. Hawk KF, D'Onofrio G, Chawarski MC, et al. Barriers and Facilitators to Clinician Readiness to Provide Emergency Department–Initiated Buprenorphine. *JAMA Network Open*. 2020;3(5):e204561. doi:10.1001/jamanetworkopen.2020.4561

5. Duber HC, Barata IA, Cioè-Peña E, et al. Identification, Management, and Transition of Care for Patients With Opioid Use Disorder in the Emergency Department. *Ann Emerg Med*. 2018;72(4):420-431. doi:10.1016/j.annemergmed.2018.04.007

6. Chartash D, Paek H, Dziura JD, et al. Identifying Opioid Use Disorder in the Emergency Department: Multi-System Electronic Health Record-Based Computable Phenotype Derivation and Validation Study. *JMIR Med Inform*. 2019;7(4):e15794. doi:10.2196/15794

7. Afshar M, Resnik F, Joyce C, et al. Clinical implementation of AI-based screening for risk for opioid use disorder in hospitalized adults. *Nat Med*. 2025;31(6):1863-1872. doi:10.1038/s41591-025-03603-z

8. Molina MF, Pimentel SD, Fenton C, Adler-Milstein J, Gottlieb LM. Characterizing emergency clinician engagement with social drivers of health data among patients with opioid use disorder. *medRxiv*. Preprint posted online December 17, 2025:2025.12.15.25342249. doi:10.64898/2025.12.15.25342249

9. Paredes D, Talankar S, Peng C, et al. Identifying Opioid Overdose and Opioid Use Disorder and Related Information from Clinical Narratives Using Large Language Models. *AMIA Jt Summits Transl Sci Proc*. 2025;2025:414-421.

10. Shah-Mohammadi F, Finkelstein J. Extraction of Substance Use Information From Clinical Notes: Generative Pretrained Transformer–Based Investigation. *JMIR Medical Informatics*. 2024;12(1):e56243. doi:10.2196/56243

11. Mahbub M, Dams GM, Srinivasan S, et al. Decoding substance use disorder severity from clinical notes using a large language model. *npj Mental Health Res*. 2025;4(1):5. doi:10.1038/s44184-024-00114-6

12. Taylor RA, Fiellin D, D'Onofrio G, Venkatesh A. Identifying opioid-related electronic health record phenotypes for emergency care research and surveillance: An expert consensus driven concept mapping process. *Subst Abus*. 2022;43(1):841-847. doi:10.1080/08897077.2021.1975864