

Computational Phenotypes for Patients with Opioid-Related Disorders Presenting to the Emergency Department

Richard Andrew Taylor, MD, MHS^{1,2,3*†}, Aidan Gilson, BS^{1†}, Wade Schulz, MD PhD^{3,4}, Kevin Lopez, PhD³, Patrick Young, PhD⁴, Sameer Pandya, MS⁴, Andreas Coppi, PhD⁵, David Chartash, PhD^{2,3,6}, David Fiellin, MD⁵, Gail D’Onofria, MD MS¹

¹Section for Biomedical Informatics and Data Science, Yale University School of Medicine.

²Department of Emergency Medicine, Yale School of Medicine.

³Program of Computational Biology and Bioinformatics, Yale University.

⁴Department of Laboratory Medicine, Yale School of Medicine.

⁵Department of Internal Medicine, Yale School of Medicine.

⁶School of Medicine, University College Dublin - National University of Ireland, Dublin.

†These authors contributed equally to this work.

Abstract

Objective: We aimed to discover computationally-derived phenotypes of opioid-related patient presentations to the emergency department (ED) via clinical notes and structured electronic health record (EHR) data.

Methods: This was a retrospective study of ED visits from 2013-2020 across ten sites within a regional healthcare network. We derived phenotypes from visits for patients 18 years of age with at least one prior or current documentation of an opioid-related diagnosis. Natural language processing was used to extract clinical entities from notes, which were combined with structured data within the EHR to create a set of features. We performed Latent Dirichlet allocation to identify topics within these features. Groups of patient presentations with similar attributes were identified by cluster analysis.

Results: In total 82,577 ED visits met inclusion criteria. The 30 topics discovered ranged from those related to substance use disorder, chronic conditions, mental health, and medical management. Clustering on these topics identified nine unique cohorts with one-year survivals ranging from 84.2-96.8%, rates of one-year ED returns from 9-34%, rates of one-year opioid event 10-17%, rates of medications for opioid use disorder from 17-43%, and a median Carlson comorbidity index of 2-8. Two cohorts of phenotypes were identified related to chronic substance use disorder, or acute overdose.

Conclusions: Our results indicate distinct phenotypic clusters with varying patient-oriented outcomes which provide future targets better allocation of resources and therapeutics. This highlights the heterogeneity of the overall population, and the need to develop targeted interventions for each population.

Keywords: Opioid Use Disorder, Computational Phenotypes, Machine Learning, Opioid Pandemic

Introduction

Background

Opioid overdose deaths have risen to alarming levels, worsening dramatically during the COVID-19 pandemic.[1] For the first time, the CDC reported over 100,000 U.S. drug related deaths in a one-year period, a staggering 28.7% increase from the year prior, with over three-quarters of the deaths attributed to opioids.[2] With approximately 130 million patient visits per year and a federal mandate to ensure emergency care for anyone who presents regardless of their ability to pay, emergency departments (EDs) are an essential healthcare setting for reducing the morbidity and mortality associated with the overdose crisis.[3]

EDs serve as the primary healthcare venue for individuals following an opioid overdose, where patients discharged with a primary diagnosis of opioid overdose have a 1-year all-cause mortality rate of 5.5%.^[4] Evidence-based interventions have offered a means to reduce the mortality rate, and are therefore crucial to implement. Despite the well-recognized importance of the ED in caring for patients with opioid-related illness, significant knowledge gaps persist concerning the characteristics and typology, or phenotypes, of opioid-related presentations to EDs and their associated outcomes.^[5] Better understanding of visit presentations and their association with future events can enhance the interpretation of risk, help ED providers address gaps in treatment and care paradigms, and uncover areas for further research.

Importance

Prior work on opioid related phenotypes in the ED has largely been confined to preconceived concepts (e.g., opioid use disorder (OUD), opioid overdose) using discrete structured data such as diagnostic codes or labs within electronic health records (EHRs).^[5–7] While essential for trial cohort identification and quality measurement, these methods often fail to address more nuanced and latent representations of patients with opioid related conditions who often have complex, multifactorial disease processes and psychosocial interactions.^[8, 9] Clinical notes within EHRs are an invaluable source of information about ED patients and visit presentations, often containing information not readily available in structured data fields (diagnostic codes, medication orders, etc.). Recent advances in natural language processing (NLP) concept extraction,^[10] topic modeling,^[11] and document level embedding^[12] offer the opportunity to process notes in innovative ways and model the aforementioned unstructured information to enhance insight into patients with opioid conditions presenting to the ED. In turn, derived novel patient cohorts from these techniques may have differential associations with downstream care processes including treatment decisions, disposition decisions, utilization of healthcare resources (care coordination, etc.), and expenditure. Using NLP, we can better examine the characteristics and typology of patients presenting with opioid related conditions, making possible the derivation of novel treatment paradigms by phenotype.

Goals of This Investigation

In this study, we aimed to discover phenotypes of ED patients with opioid conditions derived through NLP of ED clinical notes and structured data. We leveraged recent advances in NLP concept extraction to extract concepts from ED clinician and nursing notes and used Latent Dirichlet Allocation (LDA),^[13] a topic modeling method, to develop visit level topics of patient presentations within ED clinical notes. With these phenotypes in hand, we used advanced clustering methods to identify unique cohorts within the population of patients with opioid conditions. Furthermore, examining the downstream outcomes of these cohorts we aimed to demonstrate their clinimetric potential within the ED.

Methods

Study Design and Setting

This was a retrospective observational cohort study of ED visits between 2013 and 2020 across ten sites within a regional healthcare network in the northeastern United States. These EDs include two non-academic urban, two academic urban, one pediatric urban, and five non-academic suburban sites. The sites encompass a geographic area of approximately 650 square miles and closely resemble the overall national population.^[14] We included visits for all adult patients ≥ 18 years of age who had at least one opioid diagnosis (see Supplementary Table 1). This study followed STROBE reporting guidelines for observational studies.^[15] Our institutional review board approved this study and waived the need for informed consent (HIC# 1602017249).

Data Collection and Processing

Patient demographic and clinical data were extracted from the system-wide electronic health record (Epic, Verona, WI) using a centralized data warehouse (Helix). We initially identified a subset of ED notes representing patients with at least one current or prior opioid related diagnosis (the list of ICD10 codes used to identify patients is available in Supplement 1). Methodologic steps described below also necessitated the use of a representative subset of the entire ED population for use in the normalization of the opioid cohort concepts. For this step, an equal number of random ED encounters without an opioid diagnosis were identified along with associated ED notes.^[16] Diagnoses were grouped according to the AHRQ Clinical Classification Software coding system.^[17] Outpatient medications were mapped to the First DataBank Enhanced Therapeutic Classification System.^[18]

Clinical Concept Extraction

All subsequent steps of analysis are outlined in Figure 1. For clinical concept extraction, we use the Medical Concept Annotation Toolkit (MedCAT) that provides a novel self-supervised machine learning algorithm for extracting concepts to common data standards (UMLS, SNOMED-CT).[10] MedCAT overcomes many existing limitations noted during the comparison of tools such as MetaMap or cTAKES, including handling spelling mistakes and ambiguous concepts within a scalable solution.[19, 20] For the concept extraction task we use the full UMLS model trained on MIMIC III data. We further retrained this model on all 2.7 million ED physician notes and 6 million ED nursing notes written between 2013 and 2020 within the data warehouse. As the MIMIC II dataset is constructed from ICU encounters, this retraining allowed for the extraction of a broader set of clinical entities related to ED encounters and account for differences in documentation structure used in the MIMIC II dataset compared to our cohort. This model was then applied to both our opioid and random ED patient cohorts, and all text which was mapped to unique concept unique identifiers (CUIs) within each note were extracted. A knowledge source within the UMLS, the Semantic Network[21], was used to group all CUIs into larger categories (Semantic Types). We included all semantic type categories outlined in Supplemental Table 2. All other text in each note was removed. If a patient had multiple provider or nursing notes in a visit, all terms from each note were combined.

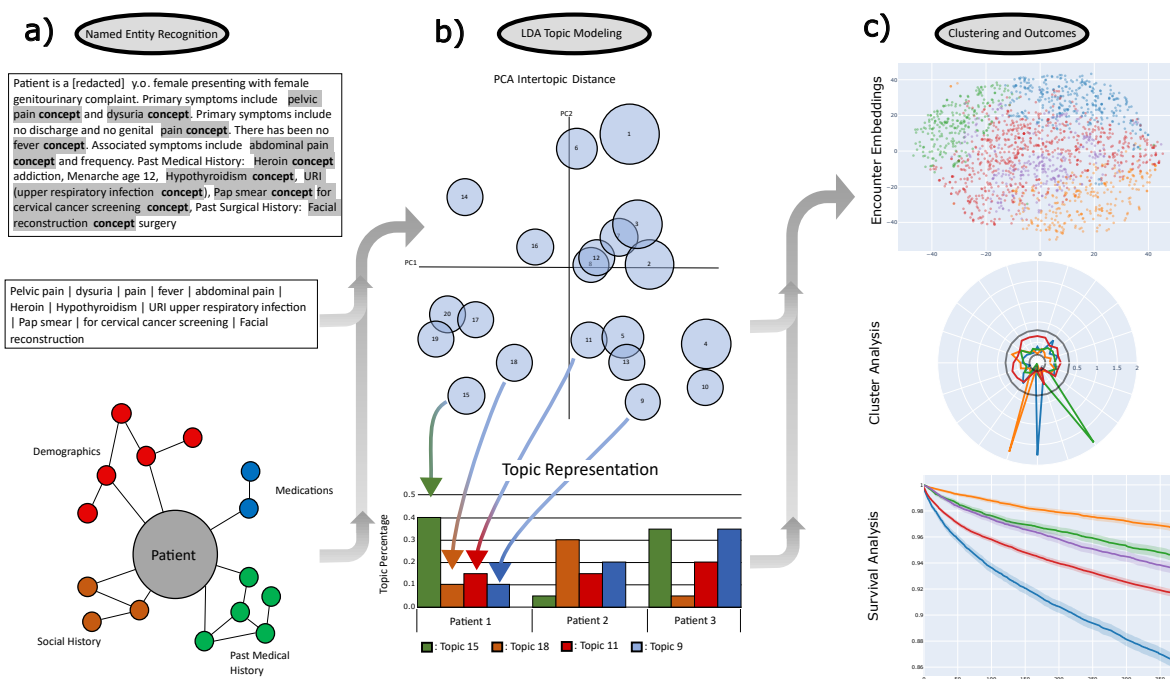


Fig. 1 Workflow a) Entity extraction with clinical concept extraction using the trained MedCat model. Structured data labels combined with identified entities to complete the feature set. b) LDA Topic modeling of the extracted features converted to patient presentation embeddings. c) Clustering of topics, and outcomes analysis.

Corpus Creation

Structured data was then added to the entity tags for each patient. The plaintext name of all structured information, for example “Low PaO2”, was added to each patient note if the code was present in the patient’s chart. For all continuous laboratory values, the plaintext name for the given laboratory value was added to the note along with the modifier “high” or “low” preceding the name if the patients value for the lab was greater than or less than two standard deviations from the mean of the matched random ED patient sample respectively.

Topic Generation

We then used a Latent Dirichlet Allocation (LDA) model to embed the combined notes from the electronic health record (EHR) into latent feature space as implemented by the software package Mallet (2.0.8).[22] LDA is a generative statistical model used to collect individual discrete data elements of a corpus (e.g. words) into topics and facilitate the classification of collected data elements (e.g. documents).[13] We first identified and removed data elements (words) that appear in more than five percent of both the opioid exposed and

random ED cohort datasets. This filtered out noise such as medical jargon related to common care processes and note templates. After removing these words, a contingency table was built from the remaining words to find keep words that have a p-value of less than 0.05 using a chi-square test for independence between cohorts. Keep words were then collected for each note within a patient encounter. Keep words from all notes within an encounter were then combined to create a term frequency corpus for each encounter. Mallet was used to generate the topics, each described by a set of keep words. The number of topics was chosen based on the model coherence plateau (see Supplementary Figure 1). For interpretability, topics were summarized with an overarching identifier through the analysis of keep words by expert consensus.[23] In Figure 2, word clouds were constructed for each topic with word size proportional to the probability of the data elements' presence within that topic.



Fig. 2 Wordclouds for the twenty topics identified by the LDA Mallet software. Word size is reflective of its importance within the topic. Topic titles were created by expert consensus to simplify discussion, but were not used in analysis.

We then used these topics to construct a n-dimensional embedding for each encounter within the opioid dataset, where n is the selected number of topics. The value of the ith dimension for a patient encounter is equal to the percentage that the ith topic is represented within the notes for that encounter, relative to all other topics. The sum of the embedding of all topics, and therefore all dimensions, will always equal one for a given patient encounter.

Discovery and Validation of Opioid Phenotype Groups

Finally we used the topic embeddings of encounters to generate novel computational phenotypes using unsupervised learning. We used K-means clustering, to allow for the evaluation of cluster quality and assessed

cluster cohesion, separation, and connectivity through silhouette coefficients, silhouette plots, and the Dunn metric respectively.[24] We analyzed the cluster performance through multiple external metrics: one-year survival following discharge, the average Charlson Comorbidity Index (CCI) of the cluster, 6-month and 12-month rate of all cause ED return, 6-month and 12-month rate of opioid related ED return, and the rates of outpatient prescription of naloxone, methadone and buprenorphine within each cluster.[25] K-means clustering was selected to allow for the evaluation of the clusters using these problem specific outcomes across a wide range of cluster number. Relative risk was used to qualify the presence of each structured variable within each cluster. For a cluster n and structured variable i , the relative rate (RR) was calculated using the formula:

$$RR = \frac{\text{patients with variable } i \text{ in cluster } n / \text{all patients in cluster } n}{\text{all patients with variable } i \text{ except those in cluster } n / \text{all patients except those in cluster } n}$$

All analyses were conducted in Python (version 3.6) and R statistical software (version 3.4.3). The code employed in the analysis is accessible via the designated GitHub repository (https://cchteam.med.yale.edu/gitlab/computational-emergency-medicine/computational-based-phenotyping/-/tree/Publication_Branch). The data used in this study can be made available upon request.

Results

In total, between March 1, 2013 and April 27, 2020, there were a total of 82,577 presentations of patients with prior or current documentation of an opioid related diagnosis which had associated structural related to ED presentation. All visit encounters contained a related provider note which was used in subsequent analysis regardless of presence of structured information. The characteristics of the patients with a history of opioid relative to the non-opioid cohort are shown in the Supplemental Table 3.

Discovery and Validation of Opioid Phenotype Groups

LDA model coherence plateaued when twenty topics were selected. Figure 2 shows word clouds for each of the 30-topics identified, as well as the overarching identifier. The top 30 keywords are shown, where word size is proportional to its importance in the topic. Topics, names here by expert consensus, ranged in areas from those directly related to substance use disorder (Opioid Withdrawal Symptoms, Acute Opioid Intoxication), past medical history (Congestive Heart Failure and Diseases of the Elderly, HIV and STIs, Chronic Back Pain and Disorders of the Spine, Cirrhosis and Liver Disease, etc.), mental health (Psychiatric Evaluation and Agitation), medical process (Laboratory and urine drug screen, Mental Status and Trauma exam), among others.

Clustering

Figure 3 shows clustering of the embeddings for each patient encounter formed from the representation of each LDA topic within the patient encounter. When optimized for each clustering metric, K-Means clustering identified 9 unique clusters. Cluster 1 (blue) contained patients defined by the topic Laboratory and urine drug screen. Cluster 2 (orange) was defined by mixed GI conditions and abdominal pain. The remaining clusters were 3 (green) mental status and trauma evaluation, 4 (red) mixed representation across all topics, 5 (purple) psychiatric evaluation, 6 (brown) joint replacement and kidney disease, 7 (pink) acute opioid intoxication, 8 (gray) general symptoms, 9 (yellow) social determinants of health.

Table 1 identifies the demographic data for each cluster as well as the rates of mental health disorders, substance use disorder, and sequelae of IV drug use. Figure 4 similarly shows the relative risk each cluster has for the structured data components, as well as the relative risk for different chief complaints. Figure 5 shows outcome metrics which can further separate clusters. Cluster 1 is a more white cohort with lower rates of mental health disorders. Cluster 2 is a female population with a high Emergency Severity Index (ESI). Cluster 3 is a male population with increased Medicaid usage. Cluster 4 is a slightly more African American population. Cluster 5 is younger male population with increased Medicaid usage and high rates of suicide and substance use disorder. Cluster 6 is an older, female cohort with high Medicare usage and high rates of admission as well as a high average CCI. Cluster 7 is a younger heavily male population with emergent ED presentations. Cluster 8 is a younger male population with decreased naloxone prescription. Finally, cluster 9 is equally male and female with very high rates of alcohol and other substance use disorder. Final cluster descriptions are shown in Table 2.

	Overall	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5	Cluster 6	Cluster 7	Cluster 8	Cluster 9
n	82577	6785	7770	4142	29989	7697	8634	4911	6143	6506
Age, median [Q1, Q3]	50.0 [36.0, 59.0]	51.0 [36.0, 62.0]	45.0 [34.0, 54.0]	45.0 [33.0, 55.0]	52.0 [40.0, 60.0]	41.0 [32.0, 51.0]	62.0 [55.0, 73.0]	40.0 [31.0, 52.0]	45.0 [33.0, 55.0]	45.0 [33.0, 53.0]
Sex, n (%)										
Male	44456 (53.8)	3269 (48.2)	3500 (45.0)	2592 (62.6)	15877 (52.9)	4864 (63.2)	3906 (45.2)	3444 (70.1)	3753 (61.1)	3251 (50.0)
Female	38103 (46.1)	3516 (51.8)	4270 (55.0)	1549 (37.4)	14098 (47.0)	2833 (36.8)	4726 (54.7)	1467 (29.9)	2389 (38.9)	3255 (50.0)
Unknown	18 (0.0)			1 (0.0)	14 (0.0)		2 (0.0)		1 (0.0)	
Insurance, n (%)										
Medicaid	45675 (55.3)	2504 (36.9)	4687 (60.3)	2743 (66.2)	15519 (51.7)	5697 (74.0)	2262 (26.2)	3571 (72.7)	4096 (66.7)	4596 (70.6)
Medicare	20813 (25.2)	2303 (33.9)	1585 (20.4)	704 (17.0)	8344 (27.8)	1026 (13.3)	4236 (49.1)	550 (11.2)	1023 (16.7)	1042 (16.0)
Medicare Managed Care	5921 (7.2)	390 (5.7)	460 (5.9)	194 (4.7)	2597 (8.7)	203 (2.6)	1399 (16.2)	119 (2.4)	278 (4.5)	281 (4.3)
Managed Care	3823 (4.6)	591 (8.7)	374 (4.8)	194 (4.7)	1426 (4.8)	273 (3.5)	309 (3.6)	204 (4.2)	250 (4.1)	202 (3.1)
BCBS	3399 (4.1)	421 (6.2)	385 (5.0)	161 (3.9)	1258 (4.2)	258 (3.4)	335 (3.9)	178 (3.6)	236 (3.8)	167 (2.6)
Other	2946 (3.6)	576 (8.5)	279 (3.6)	146 (3.5)	845 (2.8)	240 (3.1)	93 (1.1)	289 (5.9)	260 (4.2)	218 (3.4)
Race, n (%)										
White or Caucasian	53672 (65.0)	5184 (76.4)	4681 (60.2)	2790 (67.4)	18132 (60.5)	5497 (71.4)	5961 (69.0)	3340 (68.1)	3725 (60.6)	4362 (67.1)
Black or African American	17050 (20.6)	755 (11.1)	1829 (23.5)	762 (18.4)	7210 (24.0)	1122 (14.6)	1840 (21.3)	778 (15.9)	1367 (22.3)	1387 (21.3)
American Indian or Alaska Native	295 (0.4)	18 (0.3)	32 (0.4)	5 (0.1)	98 (0.3)	9 (0.1)	84 (1.0)	7 (0.1)	24 (0.4)	18 (0.3)
Asian	257 (0.3)	40 (0.6)	18 (0.2)	8 (0.2)	93 (0.3)	25 (0.3)	16 (0.2)	19 (0.4)	21 (0.3)	17 (0.3)
Native Hawaiian or Other Pacific Islander	151 (0.2)	10 (0.1)	19 (0.2)	13 (0.3)	54 (0.2)	18 (0.2)	1 (0.0)	12 (0.2)	12 (0.2)	12 (0.2)
Other	11143 (13.5)	778 (11.5)	1191 (15.3)	564 (13.6)	4398 (14.7)	1026 (13.3)	732 (8.5)	752 (15.3)	993 (16.2)	709 (10.9)
Ethnicity, n (%)										
Hispanic or Latino	14479 (17.5)	890 (13.1)	1617 (20.8)	751 (18.1)	5740 (19.1)	1324 (17.2)	901 (10.4)	956 (19.5)	1275 (20.8)	1025 (15.8)
Non-Hispanic	67782 (82.1)	5869 (86.5)	6129 (78.9)	3373 (81.4)	24128 (80.5)	6353 (82.5)	7702 (89.2)	3935 (80.1)	4843 (78.8)	5450 (83.8)
Unknown	316 (0.4)	26 (0.4)	24 (0.3)	18 (0.4)	121 (0.4)	20 (0.3)	31 (0.4)	20 (0.4)	25 (0.4)	31 (0.5)
ESI, n (%)										
Urgent	36663 (44.4)	4251 (62.7)	5534 (71.2)	1579 (38.1)	12866 (42.9)	2201 (28.6)	3953 (45.8)	1183 (24.1)	2858 (46.5)	2238 (34.4)
Emergent	32934 (39.9)	1592 (23.5)	1830 (23.6)	1795 (43.3)	11270 (37.6)	4365 (56.7)	3811 (44.1)	3290 (67.0)	1933 (31.5)	3048 (46.8)
Less Urgent	9936 (12.0)	740 (10.9)	335 (4.3)	600 (14.5)	4361 (14.5)	841 (10.9)	653 (7.6)	302 (6.1)	1130 (18.4)	974 (15.0)
Non-Urgent	1547 (1.9)	101 (1.5)	36 (0.5)	79 (1.9)	618 (2.1)	246 (3.2)	85 (1.0)	33 (0.7)	165 (2.7)	184 (2.8)
Immediate	1189 (1.4)	77 (1.1)	15 (0.2)	77 (1.9)	726 (2.4)	21 (0.3)	105 (1.2)	91 (1.9)	34 (0.6)	43 (0.7)
Unknown	308 (0.4)	24 (0.4)	20 (0.3)	12 (0.3)	148 (0.5)	23 (0.3)	27 (0.3)	12 (0.2)	23 (0.4)	19 (0.3)
ED Disposition, n (%)										
Discharge	44135 (53.4)	3253 (47.9)	3686 (47.4)	2550 (61.6)	14874 (49.6)	4983 (64.7)	2951 (34.2)	3374 (68.7)	4052 (66.0)	4412 (67.8)
Admit	34214 (41.4)	3197 (47.1)	3799 (48.9)	1373 (33.1)	13929 (46.4)	1792 (23.3)	5472 (63.4)	1243 (25.3)	1788 (29.1)	1621 (24.9)
Other	4228 (5.1)	335 (4.9)	285 (3.7)	219 (5.3)	1186 (4.0)	922 (12.0)	211 (2.4)	294 (6.0)	303 (4.9)	473 (7.3)
Alcohol Related Disorders, n (%)	28233 (34.2)	1702 (25.1)	2455 (31.6)	1602 (38.7)	9220 (30.7)	3360 (43.7)	2244 (26.0)	2144 (43.7)	1901 (30.9)	3605 (55.4)
Delirium/Dementia, n (%)	9487 (11.5)	508 (7.5)	595 (7.7)	349 (8.4)	3829 (12.8)	589 (7.7)	2069 (24.0)	354 (7.2)	440 (7.2)	754 (11.6)
HIV Infection, n (%)	4140 (5.0)	142 (2.1)	271 (3.5)	181 (4.4)	1892 (6.3)	261 (3.4)	386 (4.5)	191 (3.9)	359 (5.8)	457 (7.0)
Impulse Control Disorders, n (%)	1063 (1.3)	36 (0.5)	165 (2.1)	50 (1.2)	392 (1.3)	113 (1.5)	50 (0.6)	51 (1.0)	60 (1.0)	146 (2.2)
Miscellaneous Mental Health, n (%)	12911 (15.6)	763 (11.2)	1230 (15.8)	585 (14.1)	4651 (15.5)	1094 (14.2)	1667 (19.3)	609 (12.4)	712 (11.6)	1600 (24.6)
Mood Disorders, n (%)	48944 (59.3)	3306 (48.7)	4425 (56.9)	2322 (56.1)	17470 (58.3)	5269 (68.5)	5530 (64.0)	2586 (52.7)	3140 (51.1)	4896 (75.3)
Personality Disorders, n (%)	9023 (10.9)	514 (7.6)	750 (9.7)	500 (12.1)	2593 (8.6)	1267 (16.5)	845 (9.8)	608 (12.4)	511 (8.3)	1435 (22.1)
Schizophrenia, n (%)	15233 (18.4)	683 (10.1)	1143 (14.7)	801 (19.3)	4986 (16.6)	2108 (27.4)	1964 (22.7)	925 (18.8)	934 (15.2)	1689 (26.0)
Screening and History of Mental Health, n (%)	51081 (61.9)	2948 (43.4)	4787 (61.6)	2500 (60.4)	19106 (63.7)	4979 (64.7)	5847 (67.7)	2673 (54.4)	3501 (57.0)	4740 (72.9)
Suicide or Self Harm, n (%)	23061 (27.9)	1095 (16.1)	1838 (23.7)	1349 (32.6)	6672 (22.2)	3851 (50.0)	1604 (18.6)	1635 (33.3)	1533 (25.0)	3484 (53.6)
Substance Use Disorders, n (%)	63413 (76.8)	4024 (59.3)	6001 (77.2)	3284 (79.3)	22846 (76.2)	6679 (86.8)	6052 (70.1)	3848 (78.4)	4810 (78.3)	5869 (90.2)

Table 1. Feature Breakdown of Characteristics for Clusters Identified by K-Means

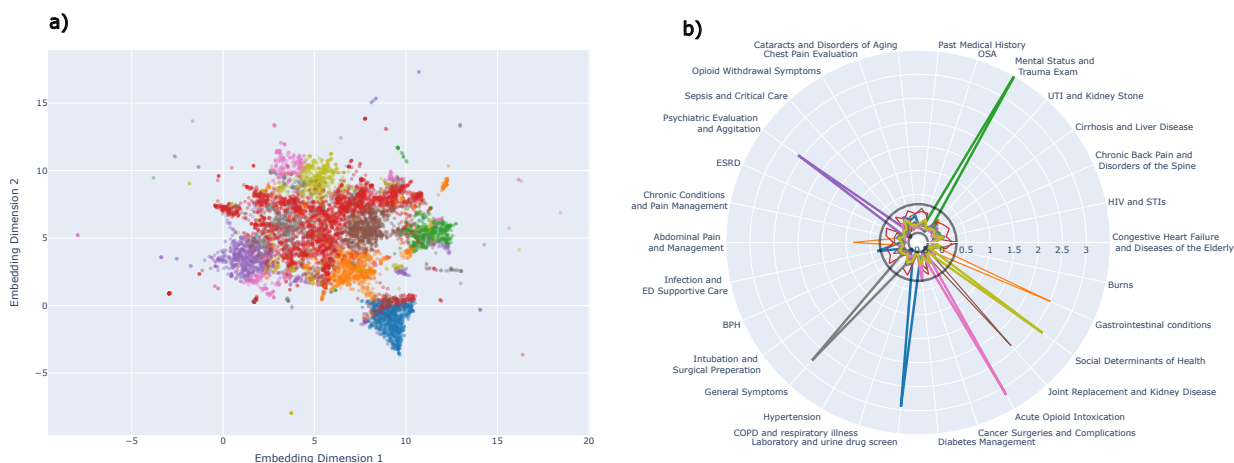


Fig. 3 (left) 2-D K-Means and Agglomerative clustering of LDA topic embeddings with tSNE projection. (right) Representation of the LDA topics within each of the clusters using absolute mean standardized deviation of each topic relative to the overall cohort of opioid patients for the structured data components.

Table 2 Cluster Descriptions

Cluster	Description
1 (blue)	White female population with increased Medicare use, lower rates of mental health disorders and substance use disorder, and moderate ED acuity.
2 (orange)	Younger female population with GI complaints, nausea and vomiting on presentation to ED, with a high ESI.
3 (green)	Male population with high Medicaid usage and trauma evaluation.
4 (red)	Black population with higher comorbidity burden, immediate ESI, high rates of HIV, presenting with constitutional complaints (shortness of breath, back pain, leg pain, leg swelling, fever, and cough).
5 (purple)	Young male population with high substance use disorder, and suicide rates, presenting as suicidal or in need of a psychiatric evaluation.
6 (brown)	Older female population with significant Medicare usage, high rates of dementia, a significant comorbidity burden, presenting with altered mental status, fall and/or weakness.
7 (pink)	Young male population with emergent presentation, high rates of ED discharge, presenting with acute intoxication or overdose.
8 (gray)	Younger male non-white population with low urgency ESI, presenting with altered mental status, dizziness, and/or headache.
9 (yellow)	A population with high rates of teenagers, with significant substance use disorder and mental health disorders and factors related to social determinants of health.

Outcomes Analysis

Figure 5 displays the results of the outcomes analysis. Cluster 5 had a significantly higher one-year survival (96.9%) and lower median CCI (2) than all other clusters ($p < 0.05$). Clusters 2, 3, 7, 8, and 9 had overlapping survival functions and were not separable (94.3%-95.7%). Clusters 1 and 4 had lower survival rates (90.2% and 89.3%), although cluster 4 has a significantly higher median CCI (5 vs 3). Finally, Cluster 6 has the lowest survival rate (84.2%) and highest median CCI (8).

Clusters 3,5,7, and 9 have the highest rates of all cause ED return in both the 6- and 12-month windows, however all clusters, with the exception of Cluster 6 have similar rates of ED return related to an opioid condition. Cluster 5 has the highest rates of outpatient methadone usage, (42.9%). Cluster 6 is the only cluster with similar rates of methadone and naloxone prescription (17.4% and 16.3% respectively), where no other cluster has naloxone prescriptions exceeding 60% of methadone prescriptions.

Discussion

The opioid epidemic presents a severe and multifaceted challenge that impacts patients in various ways, resulting in a considerable number of opioid-related fatalities, non-fatal overdoses, and secondary consequences.[26, 27] A crucial approach to addressing this issue involves identifying the diverse patient populations affected by the opioid epidemic and their unique characteristics.[5] In this study, we aimed to uncover these populations by developing opioid-related ED phenotypes through the integration of structured patient information in the EHR, unstructured data contained in notes, and cluster analysis. Our findings reveal nine distinct opioid-related cohorts, each characterized by unique profiles of demographic, diagnostic, pharmaceutical, and healthcare utilization. These cohorts exhibit larger trends, such as three younger male

populations with high acute severity and emergency department (ED) usage (Clusters 3, 5, and 7), two populations with significant comorbidity burden (Clusters 4 and 6), two predominantly female populations (Clusters 3 and 6), and two populations with notable substance use disorder (Clusters 5 and 9). This granular classification of patients facilitates targeted interventions for specific groups and offers a hierarchical approach for patient classification.

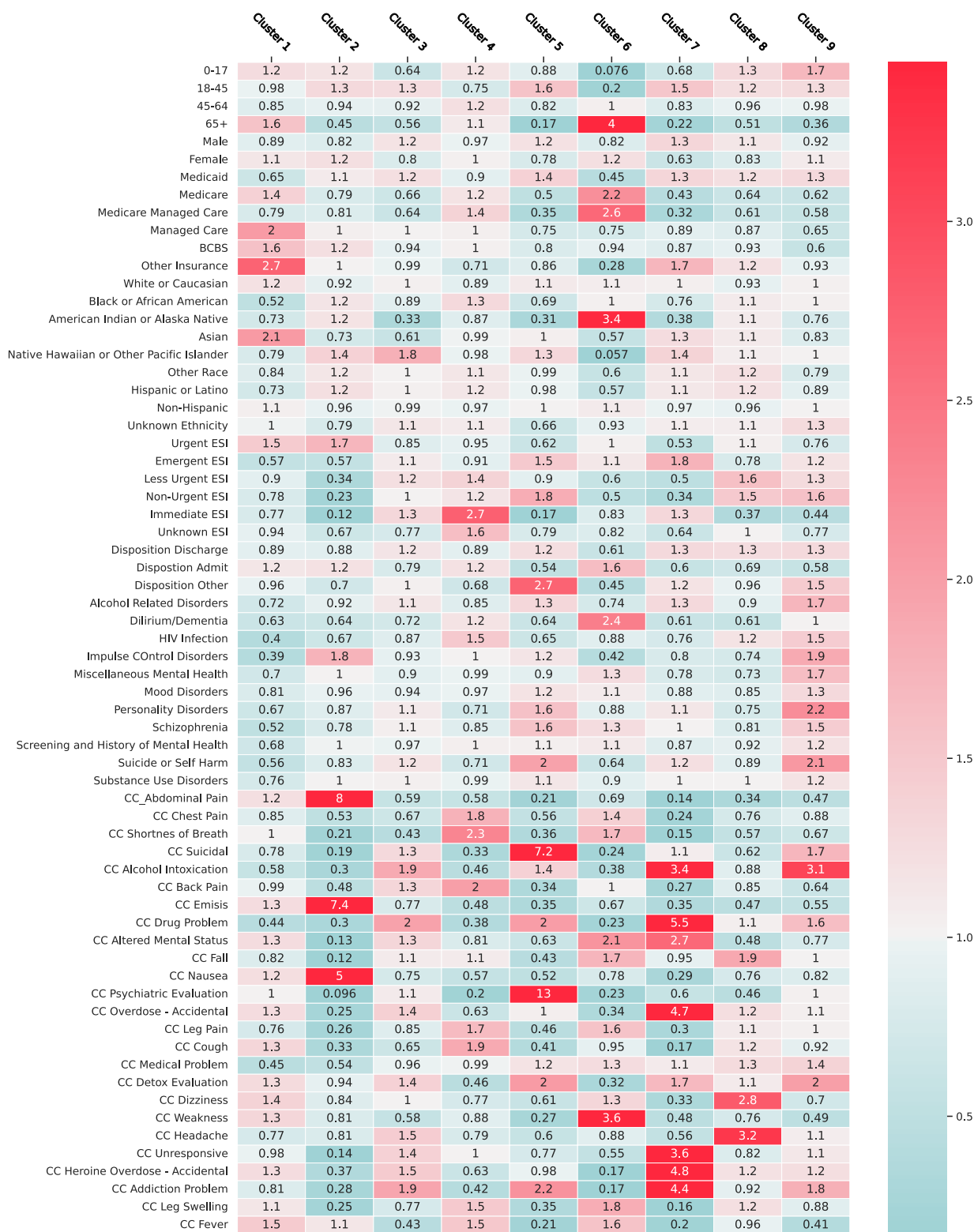


Fig. 4 Heatmap of the relative risk of the structured data component for each cluster. Relative to the entire population of opioid patients, red cells mark increased relative risk, while teal cells mark decreased relative risk. Each column therefore represents the information with a cluster that differs most significantly from the overall OUD population.

Previous studies on the phenotyping of ED patients presenting for opioid-related reasons primarily focused on differentiating opioid use disorder (OUD) from opioid overdose.[6, 7] The primary objective of these investigations centered on the precise identification of patient cohorts within Electronic Health Records (EHR) for the purposes of subject selection and pragmatic clinical trials. Methodologically, however, these studies faced limitations in their capacity to further discriminate between distinct subgroups. A few other studies have explored broader, computationally driven phenotypes of opioid related conditions. In 2019 Afshar et al. described four subgroups of opioid use disorder, patients with known disorder, patients with low socioeconomic status and psychosis, patients with alcohol use disorder, and patients with incidental opioid misuse.[28] Additionally, Schirle et al developed both a comorbidity and a NLP based scoring system for identifying the risk of OUD from EHR data.[29] These models identified structured elements with a correlation for OUD however they were not synthesized into complete phenotypes.

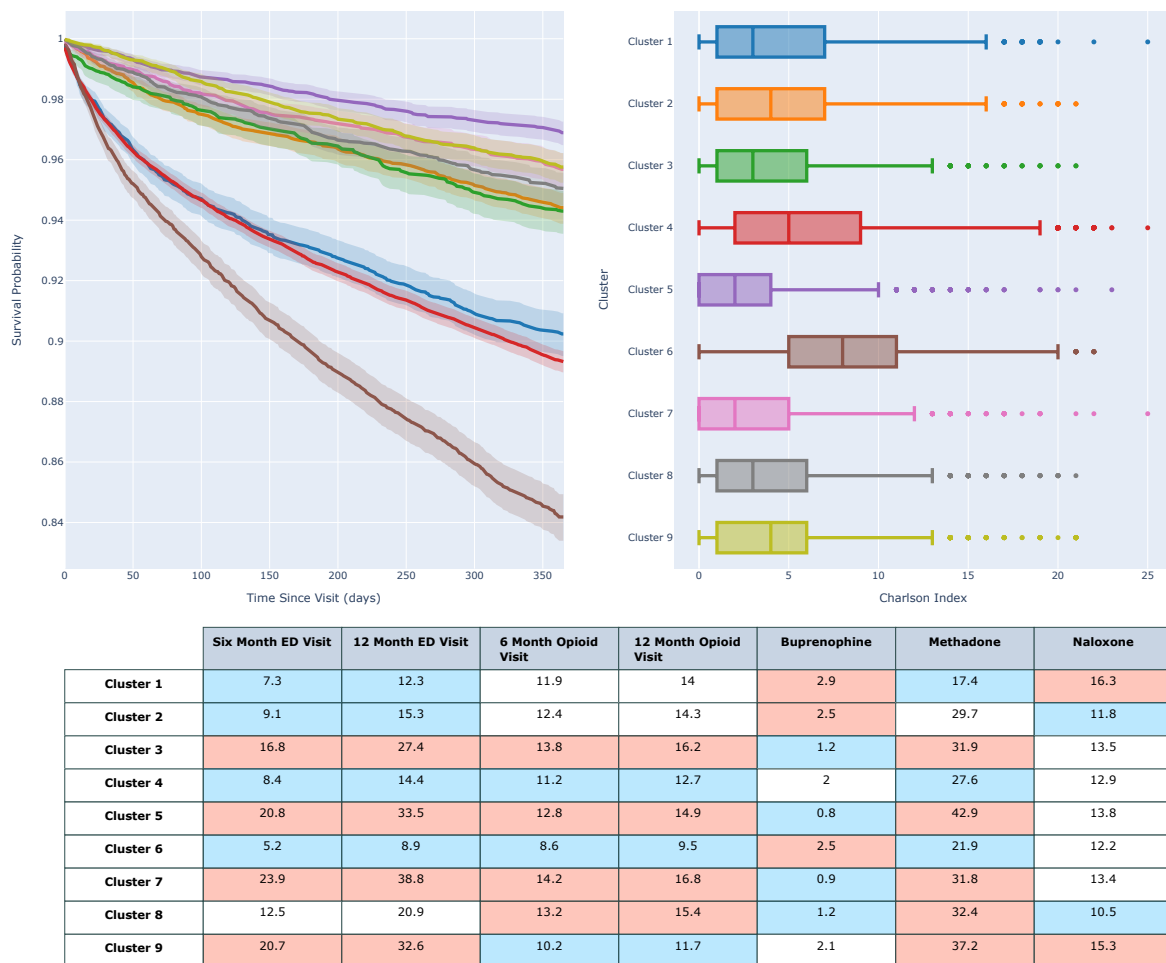


Fig. 5 a) Kaplan-Meier curve showing two-year survival following hospital discharge. b) Box and Whisker plot of the distribution of the CCI score for the patients within each cohort. c) Rates of 6 and 12 month ED return for all causes, rates of 6 and 12 month opioid related event, rates of prescription of Beuprenorphine, Methadone, and Naloxone. Relative to the entire dataset, red cells have statistically increased rates by chi-square test vs the remaining population, blue cells have statistically decreased rates by the same metric.

Our approach and findings diverge from those presented in previous literature by undertaking a more comprehensive search for opioid-related phenotypes focused on emergency department (ED) presentations. We accomplish this by employing a combination of structured and unstructured data sources, thereby capturing a more diverse and representative sample. Additionally, we extend our analysis by associating these phenotypes with crucial clinical and therapeutic outcomes, enabling an understanding of the relationship between opioid-related conditions and their respective consequences.

We identified four distinct groups within the chronic use disorder cohort (clusters 1, 4, 6, and 9), each exhibiting unique outcomes, characteristics, and prescribing patterns. For instance, Cluster 6, the older

female cohort, has a one-year survival rate 4% lower than the next phenotype within the chronic opioid use category (Cluster 4) and a significantly high comorbidity burden as measured by the average Charlson Comorbidity Index (CCI). However, this cluster also demonstrates the lowest ED return rates within the timeframes examined. Our analysis also identified three cohorts (Clusters 3, 5, and 7) more significantly associated with acute opioid overdose, with demographic breakdowns resembling those described in previous literature.[30] These cohorts exhibit the highest ED return rates among all identified clusters but also some of the lowest one-year mortalities, likely due to the lower average age compared to other cohorts.

The identification of these subgroups and their unique outcomes underscores the necessity of developing targeted interventions for each population. The substantial burden that the opioid epidemic has placed on both the general population and the healthcare system makes it challenging to provide patient-specific interventions when providers are overwhelmed with the number and variety of patients they encounter. The phenotypes discovered in this study can offer an initial framework for physicians to categorize patients under their care, leading to better resource allocation for patients with the highest mortality (Cluster 6), targeted integration of patients with the highest ED utilization with outpatient follow-up (Clusters 3, 5, and 7), distribution of naloxone and other life-saving medications to patients most at risk of acute overdose (Clusters 3, 5, and 7), and assessment of a patient's comorbidities at the initial prescription of an opioid to evaluate the risk of abuse related to underlying pain-causing conditions (Clusters 2 and 6).

In addition to the primary findings, the utilization of topic modeling techniques on provider notes has corroborated numerous intriguing opioid-related topics that span a diverse range of areas. This validation of previously established relationships further emphasizes the complex interplay between opioid use and various health conditions, strengthening the evidence base for these associations and reinforcing their significance in the context of opioid-related research and patient care. These areas include social determinants of health, psychiatric evaluations, and the management of chronic conditions such as diabetes, congestive heart failure, and chronic obstructive pulmonary disease.[31–35] Furthermore, several identified topics demonstrate a connection to pain-related conditions, including kidney stones, orthopedic procedures, surgeries, and burns. These discovered associations offer a deeper understanding of the intricate relationship between opioid use and various health conditions. For example, the association between diabetes management and opioid use may expose potential challenges in medication adherence and blood sugar regulation for patients dealing with simultaneous opioid dependencies. Likewise, the connection between chronic pain and opioid use underscores the necessity for alternative pain management strategies, emphasizing the importance of addressing the underlying causes of pain while minimizing opioid dependency.[36] The association between infections and opioid use further accentuates the well-documented susceptibility to infectious complications, particularly among intravenous drug users.[37] This relationship underscores the importance of vigilant monitoring and the implementation of preventive measures to mitigate the risk of infections in this vulnerable population. Finally, the association with opioid withdrawal symptoms highlights the demand for efficacious management strategies to mitigate the discomfort and obstacles encountered by patients during the withdrawal process.

The phenotypes identified in this study hold significant potential for a variety of applications and future use in both clinical and research settings. By providing a more granular classification of patients affected by the opioid epidemic, these phenotypes enable healthcare providers to tailor interventions and treatment strategies based on the unique characteristics and needs of each subgroup. This targeted approach can lead to more effective resource allocation, improved patient outcomes, and a better understanding of the underlying factors contributing to opioid-related conditions. Furthermore, the phenotypes serve as a valuable resource for future research, aiding in the development and evaluation of new therapeutic approaches, preventive measures, and public health initiatives. As our understanding of the multifaceted nature of opioid-related conditions continues to grow, the phenotypes can be refined and expanded upon to better capture the evolving landscape of the opioid epidemic, ultimately contributing to a more comprehensive approach to addressing this complex public health challenge.

Limitations

This study has several limitations. First, there is an inherent trade-off between specificity and interpretability of the identified phenotypes. Our topics represent broad level phenotypes of patients, but not all patients fall into one of these categories, as evidenced by cluster 4, which is not well defined. Increased specificity of the phenotypes, which may also necessitate an increased number, would allow for more precise stratification of patient outcomes, but would complicate grouping new patients into their phenotypes as well as complicate the ease of assigning new patients to a phenotype. As the assignment of phenotypes to a novel patient encounter is mutually exclusive, increased specificity of the phenotypes may result in the oversimplification of an encounter to a single phenotype when it shares elements from multiple. Second, there is a chance for the injection of preconceived biases when identifying LDA topics.[23] Although we found topic coherence plateaued at 30 topics, similar coherence levels were achieved at lower topic number and through variation of other hyperparameters inherent in the algorithm. Final selection of topics was based on clinical judgment

of topic coherence, which may be susceptible to biases in expected topics. We have included the topic descriptions for other topic numbers in the supplementary data. Third, the concept extraction technique used is limited by the ability of the natural language processing (NLP) algorithms to accurately identify and extract relevant information from the clinical notes. This may result in important information being overlooked or misclassified. Fourth, the use of Latent Dirichlet Allocation (LDA) for topic modeling is limited by the assumption that each document can be represented as a mixture of topics and that the topics themselves are generated from a Dirichlet distribution.[38] This may not accurately capture the complexity of clinical notes and EHR data and the actual distribution of topics within the encounters may not be completely represented by a Dirichlet distribution. Fifth, the limitations of labeling the LDA topics can lead to inaccuracies in the interpretation of the phenotypes. This may result from subjective judgments and the difficulty in accurately categorizing complex information into discrete concepts. Sixth, the limitations of unsupervised clustering and the optimization of clustering size can impact the validity of the phenotype groups identified. This includes issues such as the sensitivity of the clustering algorithm to the choice of initial conditions and the difficulty in determining the optimal number of clusters. Finally, the study was conducted within a single healthcare network in the northeastern United States, which may limit the ability to generalize the results to other populations or healthcare systems.

Conclusions

Our results indicate distinct phenotypic clusters with varying patient-oriented outcomes providing future targets better allocation of resources and therapeutics. Not only does this analysis separate the patients into acute and chronic populations, but also provides further stratification within each of these cohorts. Although all patients share the burden of opioid related conditions, our analysis shows that past this similarity there is significant heterogeneity within the population. This separability of phenotypes demonstrates the need for specific evidence-based interventions for each phenotype. The identified subgroups provide a starting point for physicians to organize and better understand the patients under their care. Future work is needed to determine the most effective interventions for each population.

References

- [1] Kuehn BM. Accelerated overdose deaths linked with COVID-19. *Jama*. 2021;325(6):523–523. Publisher: American Medical Association.
- [2] : U.S. Overdose Deaths In 2021 Increased Half as Much as in 2020 - But Are Still Up 15%. Available from: https://www.cdc.gov/nchs/pressroom/nchs_press_releases/2022/202205.htm.
- [3] Loren AW. Harder to treat than leukemia—opioid use disorder in survivors of cancer. *New England Journal of Medicine*. 2018;379(26):2485–2487. Publisher: Mass Medical Soc.
- [4] Weiner SG, Baker O, Bernson D, Schuur JD. One-year mortality of patients after emergency department treatment for nonfatal opioid overdose. *Annals of emergency medicine*. 2020;75(1):13–17. Publisher: Elsevier.
- [5] Taylor RA, Fiellin D, D’Onofrio G, Venkatesh AK. Identifying opioid-related electronic health record phenotypes for emergency care research and surveillance: An expert consensus driven concept mapping process. *Substance abuse*. 2022;43(1):841–847. Publisher: Taylor & Francis.
- [6] Chartash D, Paek H, Dziura JD, Ross BK, Noguee DP, Boccio E, et al. Identifying opioid use disorder in the emergency department: multi-system electronic health record-based computable phenotype derivation and validation study. *JMIR medical informatics*. 2019;7(4):e15794. Publisher: JMIR Publications Inc., Toronto, Canada.
- [7] Badger J, LaRose E, Mayer J, Bashiri F, Page D, Peissig P. Machine learning for phenotyping opioid overdose events. *Journal of biomedical informatics*. 2019;94:103185. Publisher: Elsevier.
- [8] Hawk K, McCormack R, Edelman EJ, Coupet E, Toledo N, Gauthier P, et al. Perspectives about emergency department care encounters among adults with opioid use disorder. *JAMA network open*. 2022;5(1):e2144955–e2144955. Publisher: American Medical Association.
- [9] Duber HC, Barata IA, Cioè-Peña E, Liang SY, Ketcham E, Macias-Konstantopoulos W, et al. Identification, management, and transition of care for patients with opioid use disorder in the emergency department. *Annals of emergency medicine*. 2018;72(4):420–431. Publisher: Elsevier.

- [10] Kraljevic Z, Searle T, Shek A, Roguski L, Noor K, Bean D, et al. Multi-domain clinical natural language processing with MedCAT: the medical concept annotation toolkit. *Artificial intelligence in medicine*. 2021;117:102083. Publisher: Elsevier.
- [11] Vayansky I, Kumar SA. A review of topic modeling methods. *Information Systems*. 2020;94:101582. Publisher: Elsevier.
- [12] Kim D, Seo D, Cho S, Kang P. Multi-co-training for document classification using various document representations: TF-IDF, LDA, and Doc2Vec. *Information Sciences*. 2019;477:15–29. Publisher: Elsevier.
- [13] Blei DM, Ng AY, Jordan MI. Latent dirichlet allocation. *Journal of machine Learning research*. 2003;3(Jan):993–1022.
- [14] Kolko J.: ‘Normal America’ Is Not A Small Town Of White People. Available from: <https://fivethirtyeight.com/features/normal-america-is-not-a-small-town-of-white-people/>.
- [15] Von Elm E, Altman DG, Egger M, Pocock SJ, Gøtzsche PC, Vandenbroucke JP, et al. The Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) statement: guidelines for reporting observational studies. *Annals of internal medicine*. 2007;147(8):573–577. Publisher: American College of Physicians.
- [16] Van Hoeven LR, Janssen MP, Roes KC, Koffijberg H. Aiming for a representative sample: Simulating random versus purposive strategies for hospital selection. *BMC medical research methodology*. 2015;15:1–9. Publisher: Springer.
- [17] : CLINICAL CLASSIFICATIONS SOFTWARE REFINED (CCSR). Hcup@ahrq.gov. Available from: https://www.hcup-us.ahrq.gov/toolssoftware/ccsr/ccs_refined.jsp#:~:text=The%20CCSR%20for%20ICD%2D10%2DCM%20diagnoses%20is%20intended%20to,risk%20adjust%20by%20clinical%20condition.
- [18] : Drug Vocabulary | FDB MedKnowledge Foundations. Available from: <https://www.fdbhealth.com/solutions/medknowledge-drug-database/medknowledge-foundations>.
- [19] Reátegui R, Ratté S. Comparison of MetaMap and cTAKES for entity extraction in clinical notes. *BMC medical informatics and decision making*. 2018;18(3):13–19. Publisher: BioMed Central.
- [20] Kraljevic Z, Bean D, Mascio A, Roguski L, Folarin A, Roberts A, et al. MedCAT–medical concept annotation tool. *arXiv preprint arXiv:191210166*. 2019;.
- [21] McCray A. The UMLS semantic network. In: *Proceedings. Symposium on Computer Applications in Medical Care*; 1989. p. 503–507.
- [22] McCallum AK. Mallet: A machine learning for languagetoolkit. <http://mallet.cs.umass.edu>. 2002;.
- [23] Chang J, Gerrish S, Wang C, Boyd-Graber J, Blei D. Reading tea leaves: How humans interpret topic models. *Advances in neural information processing systems*. 2009;22.
- [24] Gagolewski M, Bartoszek M, Cena A. Are cluster validity measures (in) valid? *Information Sciences*. 2021;581:620–636. Publisher: Elsevier.
- [25] Charlson ME, Charlson RE, Peterson JC, Marinopoulos SS, Briggs WM, Hollenberg JP. The Charlson comorbidity index is adapted to predict costs of chronic disease in primary care patients. *Journal of clinical epidemiology*. 2008;61(12):1234–1240. Publisher: Elsevier.
- [26] Linas BP, Savinkina A, Madushani R, Wang J, Yazdi GE, Chatterjee A, et al. Projected estimates of opioid mortality after community-level interventions. *JAMA network open*. 2021;4(2):e2037259–e2037259. Publisher: American Medical Association.
- [27] Ciccarone D. The rise of illicit fentanyl, stimulants and the fourth wave of the opioid overdose crisis. *Current opinion in psychiatry*. 2021;34(4):344. Publisher: NIH Public Access.

- [28] Afshar M, Joyce C, Dligach D, Sharma B, Kania R, Xie M, et al. Subtypes in patients with opioid misuse: A prognostic enrichment strategy using electronic health record data in hospitalized patients. *PLoS One*. 2019;14(7):e0219717.
- [29] Schirle L, Jeffery A, Yaqoob A, Sanchez-Roige S, Samuels DC. Two data-driven approaches to identifying the spectrum of problematic opioid use: A pilot study within a chronic pain cohort. *International Journal of Medical Informatics*. 2021;156:104621.
- [30] Visconti AJ, Santos GM, Lemos NP, Burke C, Coffin PO. Opioid overdose deaths in the city and county of San Francisco: prevalence, distribution, and disparities. *Journal of Urban Health*. 2015;92:758–772. Publisher: Springer.
- [31] Vozoris NT, Wang X, Fischer HD, Bell CM, O'Donnell DE, Austin PC, et al. Incident opioid drug use and adverse respiratory outcomes among older adults with COPD. *European Respiratory Journal*. 2016;48(3):683–693. Publisher: Eur Respiratory Soc.
- [32] Dawson NL, Roth V, Hodge DO, Vargas ER, Burton MC. Opioid use in patients with congestive heart failure. *Pain medicine*. 2018;19(3):485–490. Publisher: Oxford University Press.
- [33] Sharma P, Balhara YPS. Opioid use and diabetes: An overview. *Journal of Social Health and Diabetes*. 2016;4(01):006–010. Publisher: Thieme Medical and Scientific Publishers Private Ltd.
- [34] Barocas JA, Wang J, Marshall BD, LaRochelle MR, Bettano A, Bernson D, et al. Sociodemographic factors and social determinants associated with toxicology confirmed polysubstance opioid-related deaths. *Drug and alcohol dependence*. 2019;200:59–63. Publisher: Elsevier.
- [35] Strain EC. Assessment and treatment of comorbid psychiatric disorders in opioid-dependent patients. *The Clinical journal of pain*. 2002;18(4):S14–S27. Publisher: LWW.
- [36] Kumar K, Kirksey MA, Duong S, Wu CL. A review of opioid-sparing modalities in perioperative pain management: methods to decrease opioid use postoperatively. *Anesthesia & Analgesia*. 2017;125(5):1749–1760. Publisher: Wolters Kluwer.
- [37] Des Jarlais DC, Friedman SR, Stoneburner RL. HIV infection and intravenous drug use: critical issues in transmission dynamics, infection outcomes, and prevention. *Reviews of infectious diseases*. 1988;10(1):151–158. Publisher: The University of Chicago Press.
- [38] Chauhan U, Shah A. Topic modeling using latent Dirichlet allocation: A survey. *ACM Computing Surveys (CSUR)*. 2021;54(7):1–35. Publisher: ACM New York, NY, USA.