A Knowledge Graph-based Approach for Exploring the U.S. Opioid Epidemic

Maulik R. Kamdar[1], **Tymor Hamamsy**[1], **Shea Shelton**[2], **Ayin Vala**[2,3], **Tome Eftimov**[2],
James Zou[4] and **Suzanne Tamang**[2,4]

[1]Biomedical Informatics Training Program, Stanford University, USA
[2]Center for Population Health Sciences, Stanford University, USA
[3]Foundation for Precision Medicine, San Francisco, USA
[4]Department of Biomedical Data Science, Stanford University, USA
{maulikrk,tymor,sheakat,ayinv,teftimov,jamesz,stamang}@stanford.edu

Abstract

The United States is in the midst of an opioid epidemic with recent estimates indicating that more than 130 people die every day due to drug overdose. The overprescription and addiction to opioid painkillers, heroin, and synthetic opioids, has led to a public health crisis and created a huge social and economic burden. Statistical learning methods that use data from multiple clinical centers across the US to detect opioid over-prescribing trends and predict possible opioid misuse are required. However, the semantic heterogeneity in the representation of clinical data across different centers makes the development and evaluation of such methods difficult and non-trivial. We create the Opioid Drug Knowledge Graph (ODKG) – a network of opioid-related drugs, active ingredients, formulations, combinations, and brand names. We use the ODKG to normalize drug strings in a clinical data warehouse consisting of patient data from over 400 healthcare facilities in 42 different states. We showcase the use of ODKG to generate summary statistics of opioid prescription trends across US regions. These methods and resources can aid the development of advanced and scalable models to monitor the opioid epidemic and to detect illicit opioid misuse behavior. Our work is relevant to policymakers and pain researchers who wish to systematically assess factors that contribute to opioid over-prescribing and iatrogenic opioid addiction in the US.

1 The Opioid Epidemic in the United States

The opioid abuse epidemic is one of the most challenging public health challenges that our nation has ever faced. The US FDA declared the over-prescribing of opioid painkillers to be a leading cause of the astronomical rise in opioid addiction, with 64,000 overdose deaths in 2016 and 2 million people currently addicted (McCance-Katzr et al., 2010). Under current conditions, the annual number of opioid overdose deaths in US is projected to reach nearly 82,000 by 2025, resulting in approximately 700,000 deaths from 2016 to 2025 (Chen et al., 2019). With eight states reporting more opioid prescriptions than residents (Reuben et al., 2015), modifying current prescribing practices is an important strategy for reducing surging rates of opioid overdoses in the US.

To gain a better understanding on opioid prescription trends across the entire United States and to predict possible opioid abuse behavior in patients, it is imperative to build statistical models using clinical data from multiple healthcare sites in the US. However, due to the vast semantic heterogeneity that still exists between different clinical systems (Sujansky, 2001), and the lack of use of standard terminologies to encode clinical features (e.g., patient medications), developing such statistical models is difficult. Raw patient data is often extracted from legacy databases across multiple clinical centers and transformed under a uniform representation format (e.g., Fast Healthcare Interoperability Resources format (Shickel et al., 2018) and OMOP Common Data Model (Hripcsak et al., 2015)) for use in these machine learning models. This is a burden on the side of the clinical centers and leads to the creation of multiple copies of private and secure patient data.

Table 1: First degree hops from the generic ingredient Morphine in the ODKG, listing few examples of different drug formulations, combinations, and tradenames.

Property Type	Example Classes
Part Of	Atropine / Morphine, Cyclizine / Morphine, Morphine / Naltrexone
Has Tradename	MS Contin, EMbeda, Avinza, Duramorph, Kadian
Has Form	Morphine Hydrochloride, Morphine Sulphate, Morphine Tartrate
Ingredient Of	Morphine Injectable Solution, Morphine Prefilled Syringe, Morphine / Naltrexone Extended Release Oral Tablet, Morphine Sulfate 20 MG/ML, Morphine hydrochloride 40 MG

Knowledge graphs can aid in the task of normalization of similar entities encoded using different identifiers and enable integration of data from multiple heterogeneous sources. Knowledge graphs are large directed networks of real-world entities and relations between those entities, with a fixed set of semantic classes and properties (Ehrlinger & Wöß, 2016). Knowledge graphs constructed from multiple, heterogeneous pharmacological data sources have been used to predict adverse side effects that manifest on the account of polypharmacy (Zitnik et al., 2018; Kamdar & Musen, 2017).

In this paper, we describe our efforts to create an Opioid Drug Knowledge Graph (ODKG). We use the ODKG to normalize drug strings from a data warehouse consisting of electronic medical record (EMR) data that was collected from one vendor with installations in 42 states in the US. We showcase how the ODKG can aid in generating summary statistics on the prescription of different opioids in the US. Finally, we will discuss potential applications of using the ODKG to develop a web-based US compendium that allows for exploring and visualizing opioid prescribing across the US. Although such tools exist to examine international opioid consumption trends (Sankaran et al., 2016; Gebert et al., 2018), there is no such resource for the greater US.

2 METHODS

2.1 GENERATION OF THE OPIOID DRUG KNOWLEDGE GRAPH

We use two terminologies to generate the Opioid Drug Knowledge Graph (ODKG): $\mathcal{T}_1$) **ATC** (WHO, 2003): Active ingredients of drugs classified according to their anatomical, therapeutic, and chemical properties, and $\mathcal{T}_2$) **RxNorm** (Liu et al., 2005): Standard names for clinical drugs and dosage forms, as well as relations between clinical drugs to their active ingredients, drug components, and related brand names. Both these terminologies are members of the Unified Medical Language System (UMLS) (Bodenreider, 2004) and are retrieved from the BioPortal repository of biomedical ontologies and terminologies (Whetzel et al., 2011). UMLS uses the notion of a Concept Unique Identifier (CUI) to map classes with similar meaning in different terminologies.

In the first step, we use hierarchical reasoning to retrieve all the descendants of five base opioid-related classes in the ATC terminology: (N02A) Opioid analgesics, (N01AH) Opioid anesthetics, (R05DA) Opium alkaloids and derivatives, (N07BC) Drugs used in opioid dependence, and (A06AH) Peripheral opioid receptor antagonists. That is, we retrieve active ingredients of opioid drugs. Using the UMLS CUI mappings we retrieve classes related to these opioid drug ingredients from the RxNorm terminology. RxNorm terminology has different classes pertaining to drug formulations (e.g., Morphine Sulfate 50 Mg), drug combinations (e.g., Atropine/Morphine), trade names, etc. Each RxNorm class may have a distinct CUI code and an RxCUI code. We retrieve these class relations through a fixed set of properties: *Ingredients Of*, *Has Form*, *Form Of*, *Part Of*, *Ingredient Of*, *Consists Of*, *Constitutes*, *Has Tradename*, and *Precise Ingredient Of*.

2.2 EMR DATA COLLECTION

The Electronic Medical Record (EMR) data is de-identified and is aggregated in structured form from more than 400 hospitals and healthcare facilities from across 42 states in US (59% South, 17% West, 13% Midwest, 12% Northeast) during 2009-2016. The mix of hospitals consist of large and small facilities in both urban (87%) and rural (13%) locations. Roughly 98% of providers submit both inpatient and outpatient data. The dataset includes hundreds of variables, such as each patient's demographics, diagnoses, procedures and prescribed medications; also, the type and location of facilities and utilization costs. Data from psychiatric admission facilities were excluded from the dataset due to HIPAA rules (Annas et al., 2003). The aggregated EMR data is stored in the Google BigQuery Analytics data warehouse (Tigani & Naidu, 2014).
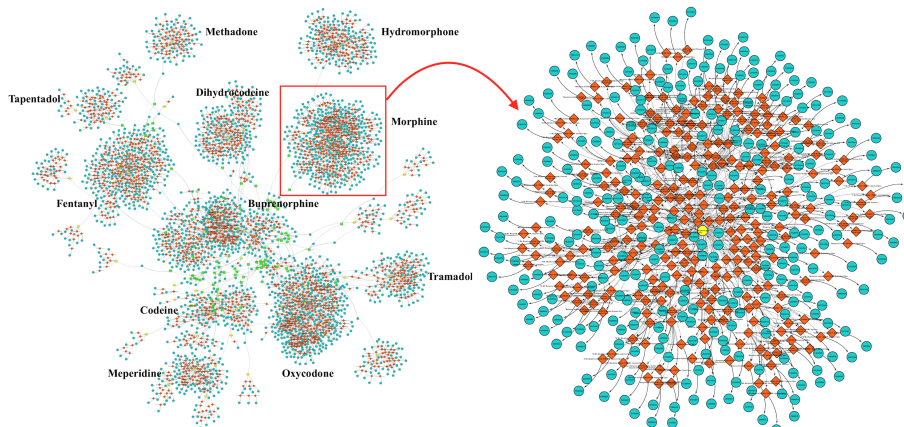
Figure 1: Opioid Drug Knowledge Graph: All the classes and their associated relations in the ODKG are visualized using a force-directed network layout. The different classes — RxNorm drugs, RxNorm generic ingredients, CUIs, and ATC drug classes, are shown as red diamond, yellow square, blue circle, and green square nodes respectively. The directed edges indicate a relation (e.g., *subClassOf*, *hasCUI*, or any property listed in **Section 2.1**) between the linked classes. The Morphine-specific network is shown in greater detail.

2.3 Normalization of Medication Information in the EMR Data Warehouse

Drugs administered to patients in different healthcare sites across the US are recorded and stored using site-specific identifiers with drug strings (e.g., *Duramorph 10mg/10ml EA*, *Morphine Sulfate (Concentrate) 10 mg/0.5ml OR Soln*, *Roxanol Liquid 120ml* all include the active ingredient Morphine). We extract a list of 425,059 unique drug strings from the aggregated EMR data warehouse. MedEx is a natural language processing system to extract medication information from clinical free text (Xu et al., 2010). We use MedEx to parse these drug strings and extract drug name (e.g., Morphine Sulfate), strength (e.g., 10 mg/0.5ml), dosage forms (e.g., OR Soln), etc. Drug names are mapped with corresponding CUIs and RxCUIs and ODKG classes are instantiated with these drug strings. It should be noted that one drug string can be mapped to multiple CUIs or RxCUIs.

3 Results

High resolution visualizations and detailed results for this research are made available online at `https://github.com/maulikkamdar/ODKG`.

3.1 Characteristics of the Opioid Drug Knowledge Graph

The Opioid Drug Knowledge Graph (ODKG), extracted from the ATC and RxNorm terminologies, has a hierarchical classification backbone with 97 ATC drug classes, 48 generic RxNorm drug ingredients, 4,960 other RxNorm classes (i.e., combinations, formulations, and ingredients), 5,051 CUI nodes and 5,188 RxCUI annotations, and 13,581 semantic relations of the type *subClassOf*, *hasCUI*, or any of the above-mentioned property types (see **Section 2.1**). The ODKG is visualized in **Figure 1**, with the Morphine-related community highlighted in more detail. The different classes — RxNorm drugs, RxNorm generic ingredients, CUIs and ATC drug classes are shown as red diamond, yellow square, blue circle, and green square nodes respectively. A directed edge between two ATC drug classes represents the *subClassOf* relation, whereas a directed edge between an active ingredient and an RxNorm node indicates a relation between those two classes subscribed under a property type. Each RxNorm node is associated with CUI node(s). Representation of clinical knowledge in graphical format enables ease of querying and abstraction of different drug strings. For example, as shown in **Table 1**, first degree hops from the generic ingredient Morphine enables the retrieval of different drug combinations, tradenames, formulations, and dosages.

3.2 Efficacy of MedEx for Normalizing Medication Information

There were 425,059 unique drug strings in EMR data warehouse. After extracting medication information using MedEx, 288,983 drug strings are mapped to at least 1 CUI (68% coverage), and 374,208 drug strings are mapped to at least 1 RxCUI (88% coverage). The opioid-related drug classes are instantiated with the normalized drug strings. There are 29 opioid-related active ingredients classes in the ODKG which have more than 10 drug strings from the EMR data warehouse
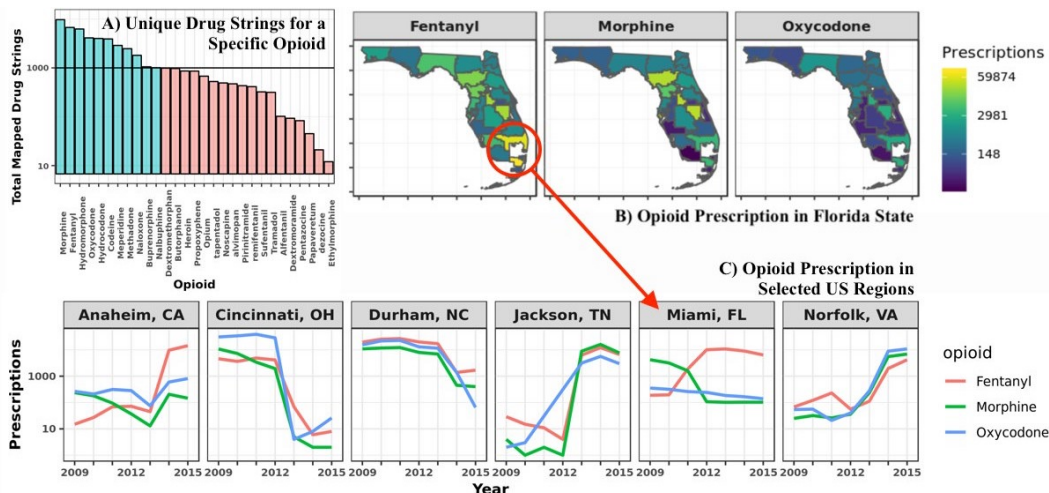
Figure 2: A) Total number of unique drug strings in the EMR data warehouse instantiated under the generic opioid classes in the ODKG. **B)** Florida region-wise visualization of unique prescription occurrences for three types of opioids: Fentanyl, Morphine, and Oxycodone. **C)** Increase or decrease in opioid prescriptions across selected US regions with more than 10,000 unique prescriptions for a particular opioid at a specific time period.

instantiated under them. It can be seen in **Figure 2A** that certain opioid painkillers, such as Morphine, Oxycodone, Hydromorphone, as well as synthetic opioids, such as Fentanyl, may have more than 1,000 drug strings instantiated under them. This demonstrates the efficacy of the ODKG toward drug string normalization in clinical data from across multiple centers.

3.3 SUMMARY STATISTICS ACROSS LOCATION AND TIME

We show a small application of using ODKG in conjunction with the aggregated EMR data warehouse to generate summary statistics of prescriptions of different opioids. Opioid prescriptions are categorized according to different US regions as well as different time periods (as determined through admission year of the patient). **Figure 2B** shows a visualization of the Florida state where it can be observed that the Miami region in Florida has $\approx$ 60,000 unique Fentanyl prescriptions. Moreover, as seen in **Figure 2C**, there is a distinct bump in the rate of Fentanyl prescriptions around 2012. **Figure 2C** also shows selected US regions that demonstrate more than 10,000 unique prescriptions for either Fentanyl, Morphine, or Oxycodone, in a specific year. Such visualizations may be used for hypothesis generation (e.g., increase in opioid prescription for a particular region).

4 DISCUSSION

Heterogeneous drug names and drug composition pose a significant challenge in performing data science and machine learning to study the opioid epidemic. In this work, we address this challenge by developing the Opioid Drug Knowledge Graph (ODKG), the first knowledge graph that captures how opioid drugs relate to each other. This knowledge graph makes it straightforward to translate medications from diverse electronic medical records into a common set of chemical-dosage features, which subsequently enables a large number of prediction and modeling tasks.

In order to identify the best strategies to reduce opioid over-prescription and misuse, a better understanding of country and regional consumption patterns, pharmaceutical industry influences, and sociopolitical factors that impact consumption, is needed. Our ODKG will be used to develop a web-based tool that can facilitate visualization of historical patterns and can enable comparisons across opioids, time, and US regions. Since our ODKG was data-driven, we hope to further refine it by consulting with a domain expert and tailor it for specific use cases and end users.

Additionally, we have identified several next steps for further research. We plan to compare our approach against the OMOP-based approach of transformation of clinical data (Hripcsak et al., 2015). Potential areas of application include the development of dynamic phenotyping methods to visually analyze individual pain medication use profiles, to identify potential risk factors for long-term use, and to detect adverse outcomes for incident user of prescription opioids, specific to the surgical setting, which some experts now consider the new 'gateway' to drug abuse.

REFERENCES

George J Annas et al. Hipaa regulations-a new era of medical-record privacy? *New England Journal of Medicine*, 348(15):1486–1490, 2003.

Olivier Bodenreider. The unified medical language system (umls): integrating biomedical terminology. *Nucleic acids research*, 32(suppl_1):D267–D270, 2004.

Qiushi Chen, Marc R Larochelle, Davis T Weaver, Anna P Lietz, Peter P Mueller, Sarah Mercaldo, Sarah E Wakeman, Kenneth A Freedberg, Tiana J Raphel, Amy B Knudsen, et al. Prevention of prescription opioid misuse and projected overdose deaths in the united states. *JAMA network open*, 2(2):e187621–e187621, 2019.

Lisa Ehrlinger and Wolfram Wöß. Towards a definition of knowledge graphs. *SEMANTiCS (Posters, Demos, SuCCESS)*, 48, 2016.

Theresa Gebert, Shuli Jiang, and Jiaxian Sheng. Characterizing allegheny county opioid overdoses with an interactive data explorer and synthetic prediction tool, 2018.

George Hripcsak, Jon D Duke, Nigam H Shah, Christian G Reich, Vojtech Huser, Martijn J Schuemie, Marc A Suchard, Rae Woong Park, Ian Chi Kei Wong, Peter R Rijnbeek, et al. Observational health data sciences and informatics (ohdsi): opportunities for observational researchers. *Studies in health technology and informatics*, 216:574, 2015.

Maulik R Kamdar and Mark A Musen. Phlegra: Graph analytics in pharmacology over the web of life sciences linked open data. In *Proceedings of the 26th International Conference on World Wide Web*, pp. 321–329. International World Wide Web Conferences Steering Committee, 2017.

Simon Liu, Wei Ma, Robin Moore, Vikraman Ganesan, and Stuart Nelson. Rxnorm: prescription for electronic drug information exchange. *IT professional*, 7(5):17–23, 2005.

Elinore McCance-Katzr, Deborah Houry, Francis Collins, and Scott Gottlieb. The Federal Response to the Opioid Crisis Written testimony on behalf of the following witnesses from the Department of Health and Human Services (HHS). Technical report, University of Zurich, Department of Informatics, 01 2010.

David B. Reuben, Anika A.H. Alvanzo, Takamaru Ashikaga, G. Anne Bogat, Christopher M. Callahan, Victoria Ruffing, and David C. Steffens. National Institutes of Health Pathways to Prevention Workshop: The Role of Opioids in the Treatment of Chronic PainThe Role of Opioids in the Treatment of Chronic Pain. *Annals of Internal Medicine*, 162(4):295–300, 02 2015. ISSN 0003-4819. doi: 10.7326/M14-2775. URL https://doi.org/10.7326/M14-2775.

Kris Sankaran, Suzanne Tamang, and Ami Bhatt. Opioid atlas: Mapping access to pain medication, 2016.

Benjamin Shickel, Patrick James Tighe, Azra Bihorac, and Parisa Rashidi. Deep ehr: A survey of recent advances in deep learning techniques for electronic health record (ehr) analysis. *IEEE journal of biomedical and health informatics*, 22(5):1589–1604, 2018.

Walter Sujansky. Heterogeneous database integration in biomedicine. *Journal of biomedical informatics*, 34(4):285–298, 2001.

Jordan Tigani and Siddartha Naidu. *Google BigQuery Analytics*. John Wiley & Sons, 2014.

Patricia L Whetzel, Natalya F Noy, Nigam H Shah, Paul R Alexander, Csongor Nyulas, Tania Tudorache, and Mark A Musen. Bioportal: enhanced functionality via new web services from the national center for biomedical ontology to access and use ontologies in software applications. *Nucleic acids research*, 39(suppl_2):W541–W545, 2011.

WHO. The anatomical therapeutic chemical classification system. *Oslo, Norway: WHO*, 2003.

Hua Xu, Shane P Stenner, Son Doan, Kevin B Johnson, Lemuel R Waitman, and Joshua C Denny. Medex: a medication information extraction system for clinical narratives. *Journal of the American Medical Informatics Association*, 17(1):19–24, 2010.

Marinka Zitnik, Monica Agrawal, and Jure Leskovec. Modeling polypharmacy side effects with graph convolutional networks. *Bioinformatics*, 34(13):i457–i466, 2018.