Text Fingerprinting and Topic Mining in the Prescription Opioid Use Literature

Huyen Le
National Center for Toxicological Research
Junxiu Zhou
Northern Kentucky University
Weizhong Zhao
Central China Normal University
Roger Perkins
National Center for Toxicological Research
Weigong Ge
National Center for Toxicological Research
Beverly Lyn-Cook
National Center for Toxicological Research
Henry Francis
Center for Drug Evaluation and Research
Huixiao Hong
National Center for Toxicological Research
Weida Tong
National Center for Toxicological Research
Wen Zou (wen.zou@fda.hhs.gov)
National Center for Toxicological Research

Research Article

Keywords: text mining, topic modeling, Latent Dirichlet Allocation, prescription opioid, codeine, morphine, hydrocodone, oxycodone, methadone.

Posted Date: March 18th, 2021

DOI: https://doi.org/10.21203/rs.3.rs-318083/v1

Abstract

Background Prescription opioids are powerful pain-reducing medications, but they may cause a variety of adverse effects. Long-term prescription opioid use (POU) is contributing to an opioid-related epidemic of addiction and death, and the scope of the opioid crisis continues to expand. As such, there is a need to identify the adverse effects associated with prescription opioid use (POU). Thousands of articles that focus on POU and its associated medical disorders have been published. However, it is time-consuming and labor-intensive to extract and understand the information of all POU-related published articles.

Methods In this study, we applied the well-adapted topic modeling method, Latent Dirichlet Allocation (LDA), to perform text mining on POU-related literature. We compiled six large academic abstract datasets by searching PubMed using the Medical Subject Headings (MeSH): prescription opioid, codeine, morphine, hydrocodone, oxycodone, and methadone. We then applied topic modeling to identify topics and analyze topic similarities/differences in these six datasets. Word clouds and histograms were used to depict the distribution of vocabularies over each topic in which the most prevalent words conveyed a topic's substance.

Results The LDA topics recaptured the search keywords in PubMed, and further revealed relevant themes, such as patients, drugs, side effects, and association links between different POU and risk factors, such as gender and age. Moreover, based on the topic modeling results, TreeMap was used to fingerprint abstracts, which revealed the possibility of constructing a visualized literature index by combining topic modeling and visualization tools such as TreeMap. Meanwhile, while performing trend analysis to explore the prevalent topic dynamics in the POU-related literature, we found that an increasing trend in opioid prescription and its associated health risks are assessed as the most central issues.

Conclusion The topic modeling results presented in this study not only convey an understandable and thematic structure of the POU literature, but also provide a means to discover which documents contain information about medical disorders associated with POU, thus, reducing the time and effort needed to review the literature for relevant articles. These results can be used as a preliminary study to systematically understand the risk factors related to increased POU-associated medical disorders.

Background

Prescription opioids are primarily used to treat pain resulting from surgery, injury, and cancer. Despite their effectiveness in treating pain, they sometimes exhibit serious side effects in the respiratory, gastrointestinal, musculoskeletal, cardiovascular, immune, endocrine, and central nervous systems. The side effects of opioids contribute to the cause of the unprecedented opioid epidemic in the United States, i.e., high and rapidly increasing numbers of abuse and overdose deaths. According to the 2015 national survey on drug use and health, 91.8 million (37.8%) U.S. adults used prescription opioids, including 11.5 million (4.7%) misusing opioids and 1.9 million (0.8%) with a use disorder [1]. The National Institute on Drug Abuse (NIDA) opioid overdose crisis report indicated that an average of 130 people in the United

States die after overdosing on opioids every day [2, 3]. One factor that contributes to this opioid crisis is the increasing number of opioids prescribed to patients since the 1990s. According to the U.S. Centers for Disease Control and Prevention (CDC), almost 450,000 people died in the United States from overdose-related prescription opioids from 1999 to 2018 [2, 3]. Furthermore, overdose deaths resulting from prescription opioids were six times higher in 2018 than in 1999. Therefore, the risk of prescription opioid use (POU) is a serious national health problem.

The risk associated with POU has attracted widespread interest among the medical research community. To date, a large number of articles that focus on POU and its associated medical disorders have been published. For example, Zedler and colleagues (2014) studied risk factors associated with prescription opioids in Veterans Health Administration patients [4]. The investigators revealed that even with the lower-than-maximum level of prescribed daily morphine equivalent dose (MED), there was a substantial risk of severe opioid-related toxicity and overdose. In addition, the study revealed that prescription opioids may cause liver disease [4]. Another study which focused on the risk factors of opioid-induced respiratory depression (OIRD) concluded that POU-caused-disorders had strong associations with serious OIRD [5]. This conclusion was consistent when the data were categorized according to demographics, clinical conditions, health care delivery systems, and clinical practices [5]. Serdarevic and colleagues (2017) studied both gender differences and risk factors in POU. The study found that women were more likely to use prescription opioids than men [6]. However, the risk factors analyzed were not well-presented with respect to gender differences, due to the lack of gender-related research work. An investigation was also made into the association between opioid analgesic and sedative risk factors with cardiopulmonary and respiratory arrest (CPRA) [7]. Analysis of millions of medical reports in the Premier database revealed that both opioid analgesic and sedative exposure could exacerbate CPRA risks [7]. A study into the opioid risk factors for severe respiratory depression (SRD) revealed that naloxone use was associated with a high probability of causing SRD [8]. It was also discovered that opioid misuse might worsen SRD clinical severity and that the severity of prescription opioid overdoses depended on the specific opioid medication [8]. A spatial epidemiological analysis was conducted to identify a relationship between opioid overdose deaths and potentially inappropriate opioid prescribing practices (PIP) in Massachusetts [9]. Stopka and colleagues (2019) found that the studied areas showed increasing rates of overdose over time. However, opioid overdose was not associated with PIP.

To fully understand the risk factors associated with POU, there is a need to utilize global POU-related research. However, it is a time-consuming and labor-intensive task to extract and understand all the information from POU-related publications. Even review articles, which tend to perform literature searches and summarize the findings, generally focus on a specific issue in each paper. Text mining is one tool that can be used to gain a broader understanding of the entire opioid-related dataset [10]. Therefore, to systematically identify the risk factors linked with increased POU-associated medical disorders, we applied the well-adapted topic modeling method, Latent Dirichlet Allocation (LDA) [11], for text mining of the POU-related literature.

Text mining aims to explore and analyze large amounts of structured or unstructured text data to extract meaningful information [12]. It plays an important role in identifying relationships and associations from a corpus of textual documents in the current trend of big data generation. Thus, text mining has been widely adopted in various fields, such as healthcare [13], bioinformatics [14, 15], chemistry [16], and marketing [17]. There also have been many studies on using text mining techniques to process medical data. For example, text mining has been applied to overcome challenges in the semantic analysis of published historical medical text [18]. Specifically, new resources were developed and established by mining diverse historical medical documents, which then further supported the text mining processing pipeline's ability to robustly detect semantic information. A proposed use of text mining was suggested for precision medicine as a method to uncover hidden information in the next-generation sequencing and electronic health records data for more effective clinical care [19]. Text mining was also applied to analyze literature on adolescent substance use and adolescent depression to reveal relevant issues [20]. Finally, to discover drug safety research trends, the use of a topic modeling-based text mining method found the top 3 most published topics as "benefit-risk assessment and communication", "diabetes", and "biologic therapy for autoimmune diseases" [21].

As one of the most popular text mining methods, topic modeling is a Bayesian statistical model which can identify salient patterns in text data. The basic assumption in topic modeling is that a document is a mixture of latent topics, where the expression of each is a distribution of words. LDA, a hierarchical Bayesian approach for modelling a corpus of unstructured data, is the most popular topic modeling method. LDA has been effectively applied to classify and understand scientific literature and databases [22, 23]. Here, we applied LDA topic modeling to perform text mining on POU-related literature. We assembled six large academic datasets consisting of abstracts dated through mid-February of 2020 using PubMed queries for the MeSH words: "prescription opioid", "codeine", "methadone", "hydrocodone", "morphine", "oxycodone". We then applied topic modeling for text mining to analyze the topics and their similarities/differences in six datasets. Word clouds and histograms were developed to depict the distribution of vocabularies over each topic in which the most prevalent words conveyed a topic's meaning. TreeMap [24] and trend analysis were performed to fingerprint abstracts and explore prevalent topic dynamics in POU-related literature.

Methods

Datasets

PubMed currently contains the largest number of citation records for biomedical literature. To collect articles on prescription opioid use, we tracked six categories of keywords in PubMed, including "prescription opioid", "codeine", "morphine", "hydrocodone", "oxycodone", and "methadone". Codeine and morphine are natural opioids, hydrocodone and oxycodone are semi-synthetic opioids, and methadone is a synthetic opioid. Data was then collected by searching for each keyword constrained by "Humans" species and "Abstract" article types. The date of publication was not limited because our goal was to discover general research status. Finally, the set of abstracts associated with each keyword was retrieved

on February 12, 2020 and stored in a separate text file. The number of abstracts for each keyword is shown in Table 1.

Table 1
Information of the collected datasets.

Keywords	Number of abstracts
Prescription opioid	5,556
Codeine	4,989
Morphine	20,739
Hydrocodone	823
Oxycodone	2,585
Methadone	9,793

Data Pre-processing

Datasets collected directly from PubMed may contain noise information that can compromise the efficiency of the results. Therefore, we performed the following pre-processing strategy to eliminate the effects of the noise information. First, Stanford CoreNLP software [25] was employed to conduct lemmatization pre-processing to reduce inflectional forms or derivationally related forms of a single word. Then, the general words, such as “background”, “aim”, “method”, “result”, “conclusion”, stop words, and numerical digits were also eliminated by the *remove-stopwords* function in MALLET (Machine Learning for Language Toolkit) [26], which is an open-source Java-based package for topic modeling and other applications for text manipulation.

Topic modeling

To extract meaningful information from the collected text datasets, we used the generative probabilistic topic model called Latent Dirichlet Allocation (LDA) [8]. LDA has been widely used to infer the hidden topic structure of documents by learning a set of thematic topics from words that tend to co-occur in documents. Specifically, the basic idea of LDA is to represent each document as a probability distribution over topics, $P(topic|document)$, and each topic as a probability distribution over words with a fixed vocabulary, $P(word|topic)$. LDA’s most important hyperparameters are α, β which specify Dirichlet distribution priors [27]. The α controls the mixture of topics for any given document. The higher α moves the topics away from the corners of the mixture of all the topics. The smaller α (normally $\alpha < 1$) leads to sparser choices within a few topics. However, as pointed out by Steyvers and colleagues (2004), different choices for the hyperparameters are task-dependent, i.e., depend on the number of topics and vocabulary size. In this work, we empirically adopted the $\alpha = 0.1$ and $\beta = 0.01$ setting for sparse topic modeling results. We used LDA in Mallet [26] to carry out Gibbs sampling [22] to obtain the posterior probability of assigning words to each topic as well as hidden topics to each document.

In LDA, different values of the topic number parameter show different topic results and it is very challenging to select the optimal number of topics. Previous research has proposed two methods to help determine the most appropriate number of topics. They are the perplexity-based method and the Rate of Perplexity Change (RPC)-based change point method (RPC method) [28]. Here, compared to the perplexity-based method, the RPC method proved to be more stable and accurate. Thus, the RPC method was repeated 10 times with different random seeds and five-fold cross-validation was utilized to select the most appropriate number of topics for each dataset. The topic numbers 5, 10, and increments of five up to 50 were used. In the RPC method, the optimal number of topics is defined as the first topic number *i* that satisfies $RPC(i) < RPC(i + 1)$. In our experiment, that condition becomes $RPC(i) < RPC(i + 5)$. Finally, the optimal number of topics for each dataset is shown in Table 2 and RPC values of LDA models with the optimal number of topics for each dataset are plotted in Fig. 1. We also note that in our runs, the stability result ranges from 60–85%, which is consistent with previous results [28].

Table 2
The most appropriate number of topics for each dataset.

Datasets	Number of topics
Prescription opioid	30
Codeine	35
Morphine	40
Hydrocodone	25
Oxycodone	35
Methadone	40

Visualization and Dynamics of Topics

WordCloud visualization tool (https://www.jasondavies.com/wordcloud/) was employed to display the discovered topics of datasets. WordCloud visualizes the distribution of words on each topic, i.e. the size of each word in WordCloud is proportional to the probability of the word within the topic, $P(word|topic)$. This provides a clear overview of the topic results.

Each abstract can be represented by the topic with the highest probability, $P(topic|document)$. Then, the frequency distributions of represented topics (e.g., histogram) were plotted for each dataset. In addition, we analyzed the dynamics of these topics. In each dataset, a trend analysis on each topic was conducted. A p-value smaller than Bonferroni-corrected alpha, which equals 0.05/the number of topics, was considered statistically significant.

In summary

An overview of the process is presented in Fig. 2. The collected datasets are pre-processed based on lemmatization and stop-words, and the LDA text mining method. The output of these processes is topics which come from each document. A topic consists of a cluster of words that frequently occur together.

Results And Discussion

Fingerprinting an Abstract with TreeMap

Figure 2 shows how each abstract in the dataset can be interpreted by topics analysis and their related weights. Generally, the first three to five words in each topic have the most important weights and contributions. Next, each topic can be interpreted by the identified words and their related weights. Thus, we can analyze each abstract from the perspective of topics and words with the help of TreeMap [24]. The TreeMap in Fig. 3 illustrates a randomly selected abstract from the prescription opioid dataset and displays the relationship between topics and words. In this example, we selected the top 3 keywords under each topic as an illustration. As shown in Fig. 3, topic T19, “pain” has the largest portion represented in the abstract. Additionally, the words “pain”, “dose”, and “patient” constituting the top 3 largest weights in this topic indicate that this opioid-related study is likely to focus on pain and dose for patients. The image in Fig. 3 may be used as a fingerprint of this abstract, which provides a visualized literature index by combining topic modeling and visualization tools such as TreeMap.

Topic Analysis of Datasets

Since each abstract can be represented by its topic with maximum weight, we analyzed the topic distribution in each dataset to provide an overview of opioid-related research aspects.

1) Prescription Opioid: Fig. 4 shows the distribution of assigned topics within the prescription opioid dataset. The five most prevalent topics were T6, T4, T21, T30, and T8. The corresponding word clouds for these topics are shown in Fig. 5. T4 shows the “patient”, “dose”, and “prescribe” with the general terms associated with opioid prescription. T21 was primarily concerned with “physician”, “prescribe”, and “practice”. T30 focuses on “chronic pain” and “patients”, and T8 may demonstrate topics associated with the increased risk for death from opioid overdose. The most dominant topic, T6, is assigned to 16.5% of the abstracts in the dataset which is nearly three times the number of abstracts assigned to the second most populous topic (T4). This trend indicates that many opioid-related research efforts have focused on the abuse and health risks of prescription opioids.

2) Codeine: According to the distribution of assigned topics for the codeine dataset (Fig. 6), the five most popular topics were T4, T33, T12, T35, and T13. The corresponding word clouds for these topics are shown in Fig. 7. T4 and T33 contain high-weighted terms such as: “extraction”, “detection”, “drug”, “clinical”, and “treatment”; T12, T13, and T35 show that the research on codeine has been closely related to pain-patient associations, such as the post-operative pain following surgery.

3) Morphine: According to the distribution of assigned topics for the morphine dataset (Fig. 8), the five most popular topics were T19, T26, T12, T13, and T6. The corresponding word clouds for these topics are

shown in Fig. 9. The most dominant topic, T19, was assigned approximately 9% of the abstracts in the dataset, with focus on “patients”, “post-operative pain”, and “treatment”; while the second top topic, T26, encompassed “cancer patient pain treatment”, and T12 primarily involved “epidural morphine” and “analgesics”. Topic T13 involved the “clinical drug treatment” and T6 contained aspects related to “opioid withdrawal”, including “naloxone”, “morphine” and “rat experiments”.

4) Hydrocodone: According to the distribution of assigned topics for the hydrocodone dataset (Fig. 10), the five most popular topics were T7, T9, T10, T12, and T2. The corresponding word clouds for these topics are shown in Fig. 11. They include not only the opioid-related common terms such as “prescription opioid: (T7), “drug” and “pain treatment” (T10), and “patient pain” (T12), but also focus on the “opioid metabolites” and “urine/specimen sample detection” (T9) and “comparison” of “hydrocodone” with “other analgesic medications” on “pain relief” (T2).

5) Oxycodone: According to the distribution of assigned topics for the oxycodone dataset (Fig. 12), the five most popular topics were T18, T25, T27, T14, and T23. The corresponding word clouds for these topics are shown in Fig. 13. T18, T25, and T14 were similar to some topics of other opioid datasets; T27 includes the comparison between “oxycodone” and “tapentadol” on the effect of “pain relief treatment”; and T23 contained the terms of “drug abuse” along with increased use of “prescription opioids”.

6) Methadone: According to the distribution of assigned topics for the oxycodone dataset (Fig. 14), the five most popular topics were T11, T32, T35, T3, and T15. The corresponding word clouds for these topics are shown in Fig. 15. The top three most popular topics, T11, T32, and T35, encompassed common topics such as “addiction”, “drug dose”, and “abuse in prescription opioid use”; T3 included a special concern regarding the possible effects of “methadone use” on “pregnant women” or “infants” such as “Neonatal Abstinence Syndrome (NAS)”. T15 shows that much research has been pursued on the relationships between “methadone maintenance treatment” with “HIV-related risk behaviors”.

Similarities

Although word clouds for the top 5 topics were analyzed separately for each dataset as above, checking word clouds for all topics (Supplemental Information Fig. 1) suggested that most topics for all six datasets could be grouped further into the following categories

- Opioid/Drug Prescription: topics containing general search terms (drug name) and opioid/drug prescription-related issues, such as increasing trend of dose.
- Patient/Pain: topics involved in opioid-use patients with pain resulting from conditions such as cancer, chronic, surgery post-operative pain.
- Misuse/Abuse: topics related to problems in using these opioids such as overdose leading to death.
- Adverse/Side Effects: topics associated with side effects from using these opioids such as RLS (Restless Legs Syndrome), NAS (Neonatal Abstinence Syndrome).

- Physician/Clinical Treatment: topics indicating that research on these opioids focused on practical experiments and treatments, especially those addressing opioid use disorder treatment.
- Gender/Age or Woman/Child: topics suggesting that these opioids may have gender and age disparities in pregnant women and infants.
- Genotype/CYP (Cytochrome P450): topics related to research on opioid metabolism and genotype.
- Review: topics related to current research reviews focused on prescription usage.
- *Differences*: We note that although most topics in all six datasets shared common categories/themes as described above, they were assigned different ratios of abstracts in each dataset. This explains why the top 5 topics were different for each dataset. For example, although the topic “pregnant women and infants” appeared in all datasets, this topic appeared in the top 5 topics only for the methadone dataset.

Furthermore, compared to other opioids, methadone was the only opioid which had several topics related to other factors (e.g., alcohol, smoking, heroin and cocaine use, HIV/AIDS, and Hepatitis C Virus (HCV)). Therefore, topics in the category of Physician/Clinical Treatment for methadone involved alcohol/smoking cessation/intervention, or cocaine treatment, or Opioid Substitution Therapy (OST) for HIV prevention. The association of these factors with methadone might partially be due to methadone being a synthetic opioid, which associates it with illegal drugs (e.g., heroin, cocaine). This is in contrast to codeine and morphine which are natural opioids or hydrocodone and oxycodone which are semi-synthetic opioids.

Another difference among the six datasets is in the category “Adverse/Side Effects”. The side effects in the codeine dataset contained more detailed symptom keywords such as cough, headache, bowel, renal, and side effects. The oxycodone dataset also contained more detailed symptom keywords such as bowel, renal, sleep, and RLS. Therefore, we may possibly conclude that different opioids may cause different side effects and/or the consequences vary between opioids and patients.

Trend Analysis

A trend analysis was applied to the proportion of topics by year for each dataset. The dynamic of the most prevalent topic, which was the most statistically significant increasing linear trend for each dataset is shown in Fig. 16. The word clouds in Fig. 17 show the meanings of these prevalent topics. The most prevalent topic in the prescription opioid dataset was related to prescription opioid dose prescribed for patients (Fig. 17(a)); while in the methadone dataset, continuing research has predominantly been conducted on methadone associated risk factors (Fig. 17(f)). The other datasets for codeine, morphine, hydrocodone, and oxycodone seem to share the same leading topic regarding the increasing trend in opioid prescription and associated health risks (Fig. 17(b,c,d,e)). We note that the leading topic is not necessarily the same as the most dominant topic in the topic distribution of each dataset. For example, the most dominant topic in the prescription opioid dataset was T6 (Figs. 4 and 5) while the leading topic was T4.

Conclusion

In this study we applied LDA for text mining of published articles dealing with prescription opioid use. Results showed the ability of topic modeling as a computational tool to segregate a vast quantity of articles into different themes that provide a systematic literature overview. Specifically, the LDA topics recaptured the search keywords in PubMed, and further revealed relevant themes such as pain treatment, opioid misuse/abuse, association links between opioid usage and pregnant women/infants. We also demonstrated that based on topic modeling's results, TreeMap can be used to fingerprint abstracts, suggesting the possibility of making visualized literature indices by combining topic modeling and visualization tools. Finally, using trend analysis to explore the dynamics of the proportion of topics categorized by year, we found that the increasing trend in opioid prescription and its associated health risks are assessed as the leading issues in the POU-related literature.

Declarations

Ethics approval and consent to participate

NA/Not Applicable

Consent for publication

NA/Not Applicable

Availability of data and materials

The datasets used and/or analysed during the current study available from the corresponding author on reasonable request.

Competing interests

The authors declare that they have no competing interests.

Funding

This work and the publication were funded by FDA.

Authors' contributions

HL and JZ performed all the calculations and data analysis and wrote the first draft of the manuscript. This work was established primarily by WZ and WZ (Zhao) in developing the methods, conceiving the original idea and guiding the data analysis and presentation of results. HL, JZ, RP, WZ (Zhao), and WZ participated in the dataset construction and the resulting figures. All authors contributed to data verification, approach evaluation, and assisted with writing the manuscript. All authors read and approved the final manuscript.

Acknowledgements

Huyen Le and Junxiu Zhou acknowledge the support of a fellowship from the Oak Ridge Institute for Science and Education, administered through an interagency agreement between the U.S. Department of Energy and the U.S. Food and Drug Administration.

References

1. Han B, Compton WM, Blanco C, Crane E, Lee J, Jones CM. **Prescription Opioid Use, Misuse, and Use Disorders in U.S. Adults: 2015 National Survey on Drug Use and Health**. *Ann Intern Med* 2017, 167(5):293-301.
2. National Institute on Drug Abuse. **Opioid Overdose Crisis**. Available at https://www.drugabuse.gov/drugs-abuse/opioids/opioid-overdose-crisis , 2020.
3. National Center for Health Statistics. **Wide-ranging online data for epidemiologic research (WONDER)**. Available at http://wonder.cdc.gov, 2020.
4. Zedler B, Lin X, Li W, Andrew J, Catherine V, Furaha K, Pradeep R, Onur B, and Lenn M. **Risk factors for serious prescription opioid-related toxicity or overdose among Veterans Health Administration patients**. *Pain Medicine* 2014, 15(11):1911-1929.
5. Nadpara P, Andrew J, Lenn M, Nathan C, Norman C, Marie B, and Zedler B. **Risk factors for serious prescription opioid-induced respiratory depression or overdose: Comparison of commercially insured and veterans health affairs populations**. *Pain Medicine* 2017, 19(1):79-96.
6. Serdarevic M, Striley WC, and Cottler LB. **Gender differences in prescription opioid use**. *Current Opinion in Psychiatry* 2017, 30(4):238.
7. Izrailtyan I, Qiu J, Overdyk FJ, Erslon M, and Gan TJ. **Risk factors for cardiopulmonary and respiratory arrest in medical and surgical hospital patients on opioid analgesics and sedatives**. *PloS one* 2018, 13(3): e0194553.
8. Fox LM, Hoffman RS, Vlahov D, Manini FA. **Risk factors for severe respiratory depression from prescription opioid overdose**. *Addiction* 2018, 113(1):59-66.
9. Stopka TJ, Amaravadi H, Kaplan AR, Hoh R, Bernson D, Chui KKH, Land T, Walley AY, LaRochelle MR, Rose AJ. **Opioid overdose deaths and potentially inappropriate opioid prescribing practices (PIP): A spatial epidemiological study**. *International Journal of Drug Policy* 2019, 68:37-45.
10. Ramage D, Rosen E, Chuang J, Manning CD, McFarland DA. **Topic modeling for the social sciences**. In: *NIPS 2009 Workshop on Applications for Topic Models: Text and Beyond*.
11. Blei DM, Ng AY, Jordan MI. **Latent Dirichlet Allocation**. *Journal of Machine Learning Research* 2003(3):993-1022.
12. Feldman R, Sanger J. **The text mining handbook: advanced approaches in analyzing unstructured data**. *Cambridge university press* 2007.

13. Koh HC, Tan G. **Data mining applications in healthcare**. *Journal of healthcare information management* 2011, 19(2):65.
14. Cohen AM, Hersh WR. **A survey of current work in biomedical text mining**. *Briefings in bioinformatics* 2005, 6(1):57-71.
15. Ananiadou S, McNaught J. **Text mining for biology and biomedicine**. *London: Artech Hous*e, 2006.
16. Krallinger M, Rabal O, Lourenco A, Oyarzabal J, Valencia A. **Information retrieval and text mining technologies for chemistry**. *Chemical reviews* 2017, 117(12):7673-761.
17. Guerreiro J, Rita P, Trigueiros D. **A text mining-based review of cause-related marketing literature**. *Journal of Business Ethics 2016*, 139(1):111-28.
18. Thompson P, Batista-Navarro RT, Kontonatsios G, Carter J, Toon E, McNaught J, Timmermann C, Worboys M, Ananiadou S. **Text mining the history of medicine**. *PloS one* 2016, 11(1):e0144717.
19. Simmons M, Singhal A, Lu Z. **Text mining for precision medicine: bringing structure to EHRs and biomedical literature to understand genes and health**. In: *Translational Biomedical Informatics* 2016, pp. 139-166.
20. Wang SH, Ding Y, Zhao W, Huang YH, Perkins R, Zou W, Chen JJ. **Text mining for identifying topics in the literatures about adolescent substance use and depression**. *BMC public health* 2016, 16(1):279.
21. Zou C. **Analyzing research trends on drug safety using topic modeling**. *Expert opinion on drug safety* 2018, 17(6):629-36.
22. Griffiths TL, Steyvers M. **Finding scientific topics**. *Proceedings of the National Academy of Sciences of the United States of America* 2004, 101(suppl. 1):5228-5235.
23. Blei DM. **Probabilistic Topic Models**. *Communications of the ACM* 2012, 55(4):77-84.
24. Ben S. **Tree visualization with tree-maps: 2-d space-filling approach.** *ACM Transactions on Graphics* 1992, 11:92-99.
25. Manning C, Surdeanu M, Bauer J, Finkel J, Bethard S, McClosky D. **The Stanford CoreNLP natural language processing toolkit**. In: *Proceedings of 52nd annual meeting of the association for computational linguistics: system demonstrations* 2014, pp. 55-60.
26. McCallum AK. **Mallet: A machine learning for language toolkit**. 2002.
27. Steyvers M, Smyth P, Rosen-Zvi M, Griffiths T. **Probabilistic author-topic models for information discovery**. In: *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, 2004, pp. 306-315.
28. Zhao W, Chen JJ, Perkins R, Liu Z, Ge W, Ding Y, Zou W. **A heuristic approach to determine an appropriate number of topics in topic modeling.** *BMC Bioinformatics* 2015, 16(Suppl 13):S8.
29. Chauvin M. **Prise en charge post-opératoire. La douleur après l'intervention chirurgicale [Postoperative patient management. Pain after surgical intervention]**. *Presse Med*. 1999 Jan 30;28(4):203-11. French. PMID: 10071636.

Figures

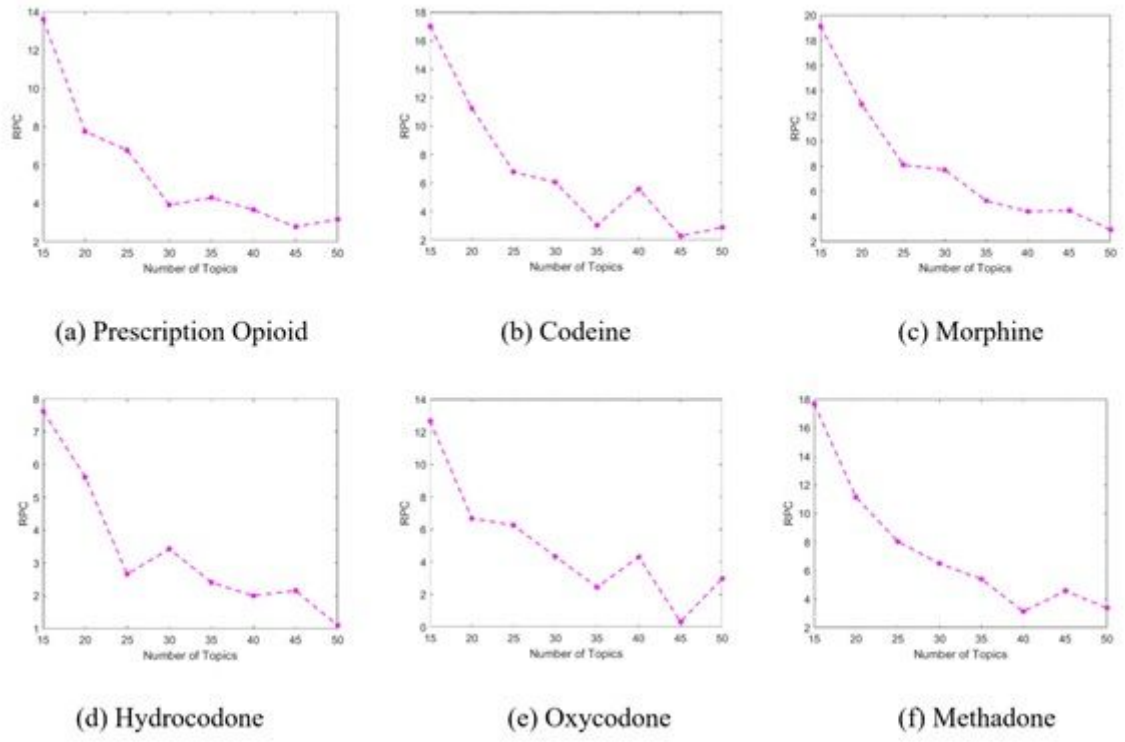

Figure 1

RPC values of LDA models with the optimal number of topics for each dataset.

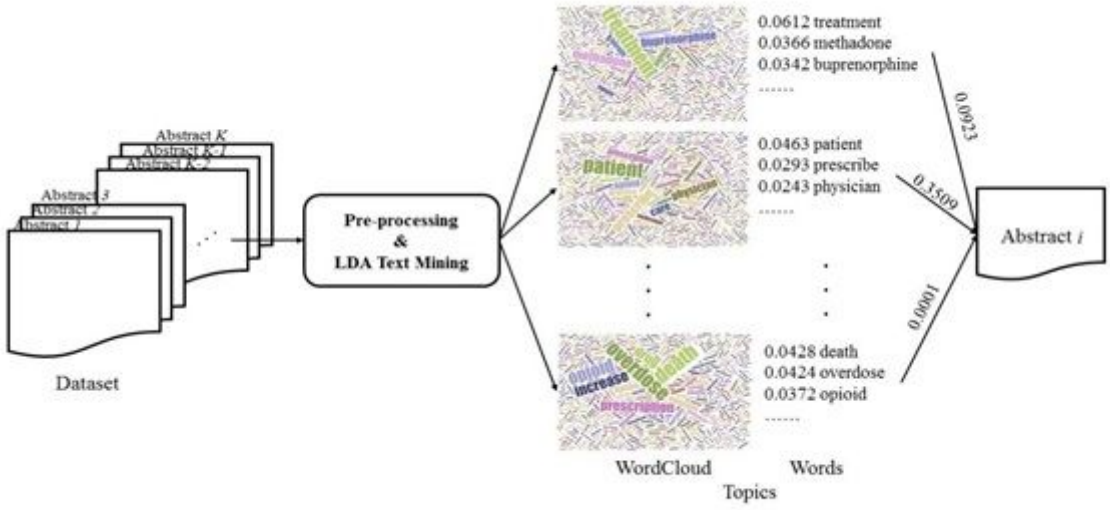

Figure 2

Workflow of proposed method with respect to dataset, topics, and abstract.

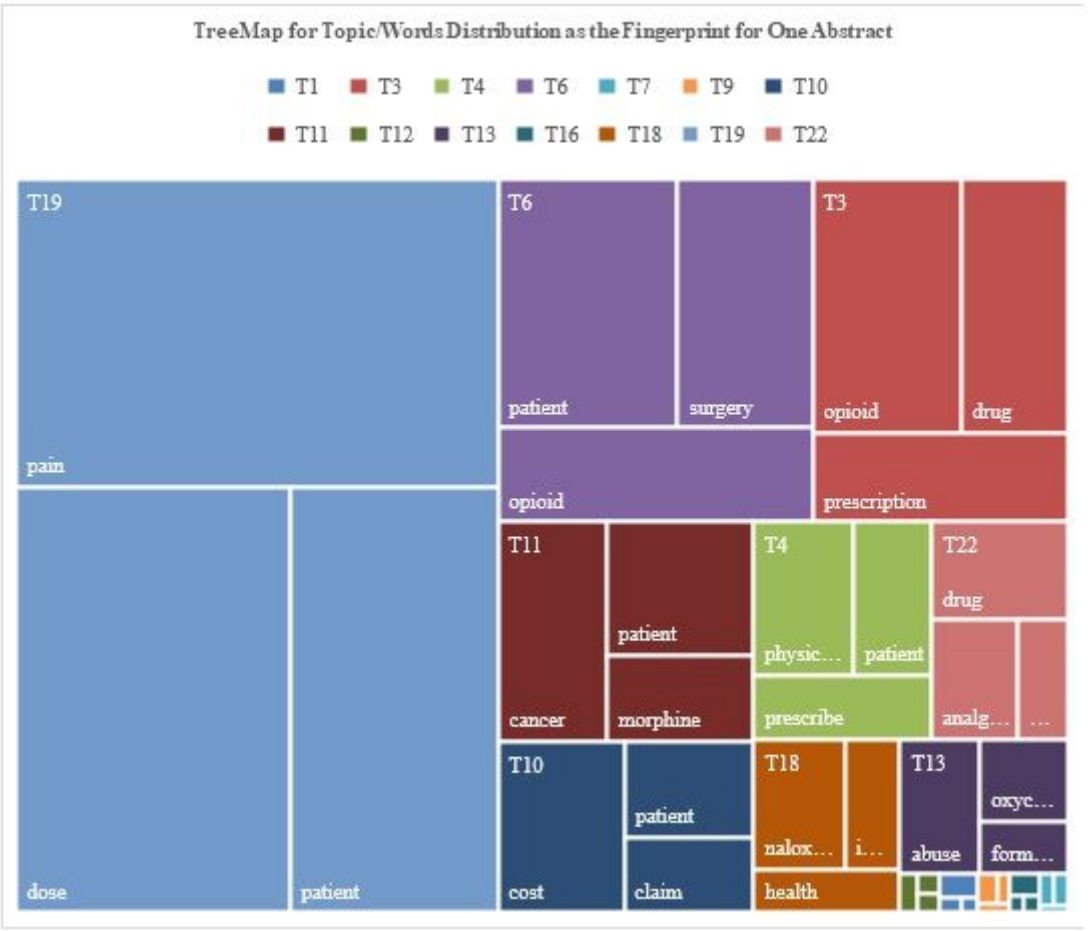

Figure 3

TreeMap visualization for one abstract [29] in the prescription opioid dataset.

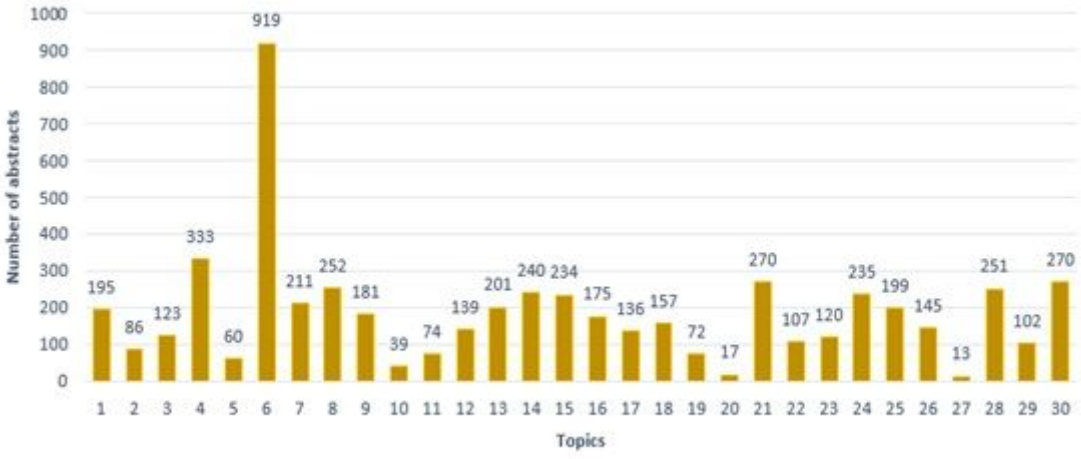

Figure 4

Distribution of assigned topics for the Prescription Opioid dataset.

Figure 5

Top-5 word clouds for the Prescription Opioid dataset (topic number = 30).

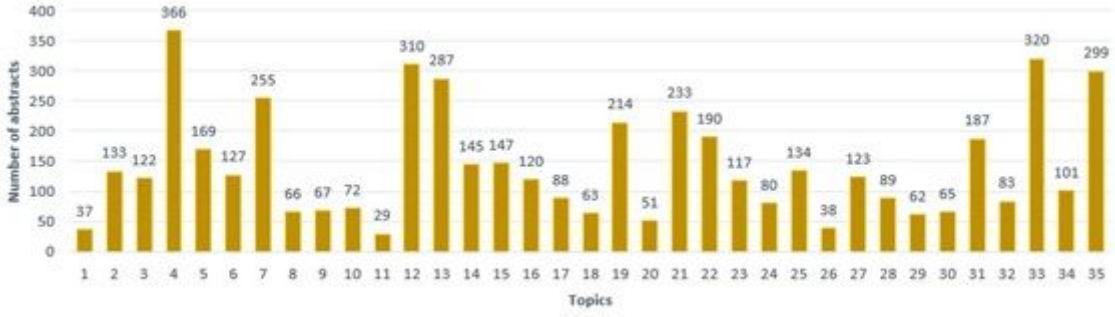

Figure 6

Distribution of assigned topics for the Codeine dataset.

Figure 7

Top-5 word clouds for the Codeine dataset (topic number = 35).

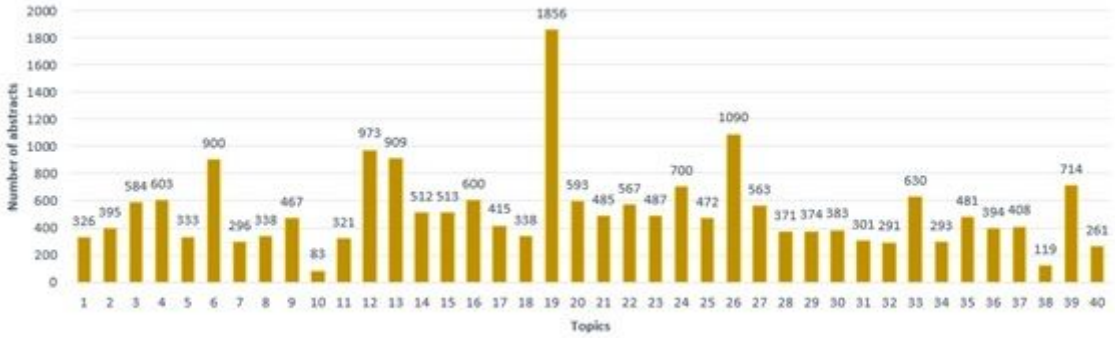

Figure 8

Distribution of assigned topics for the Morphine dataset.

Figure 9

Top-5 word clouds for the Morphine dataset (topic number = 40).

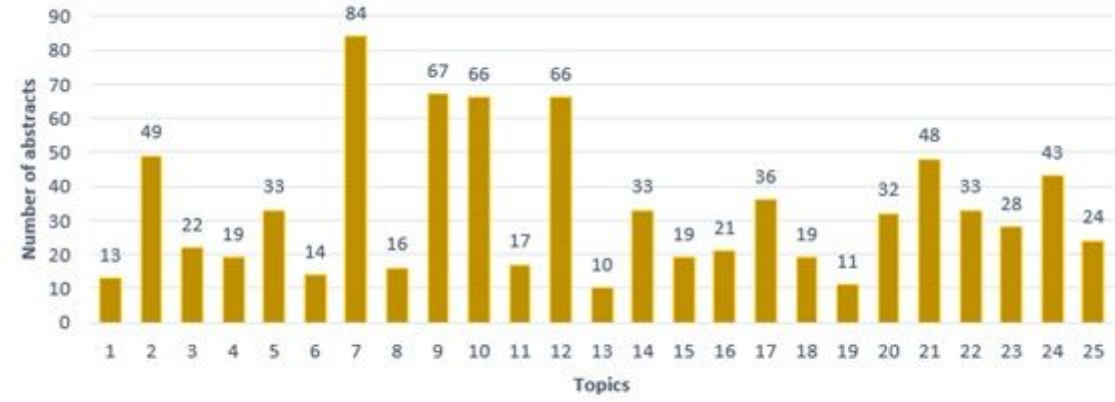

Figure 10

Distribution of assigned topics for the Hydrocodone dataset.

Figure 11

Top-5 word clouds of the Hydrocodone dataset (topic number = 25).

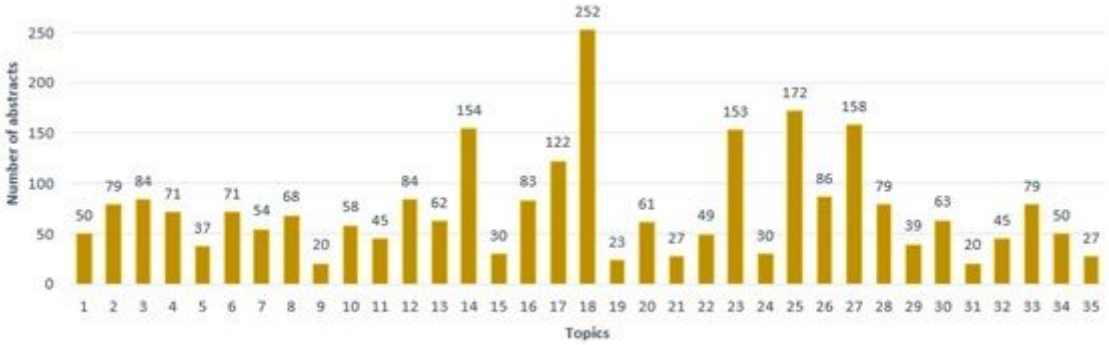

Figure 12

Distribution of assigned topics for the Oxycodone dataset.

Figure 13

Top-5 word clouds of the Oxycodone dataset (topic number = 35).

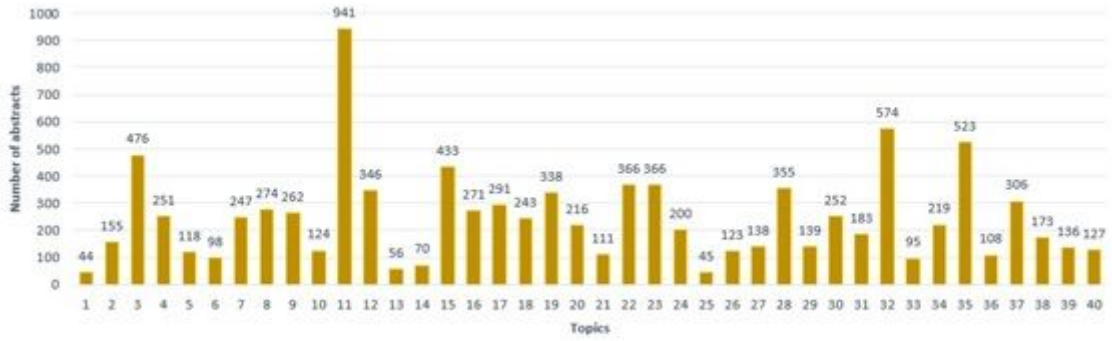

Figure 14

Distribution of assigned topics for the Methadone dataset.

Figure 15

Top-5 word clouds of the Methadone dataset (topic number = 40).

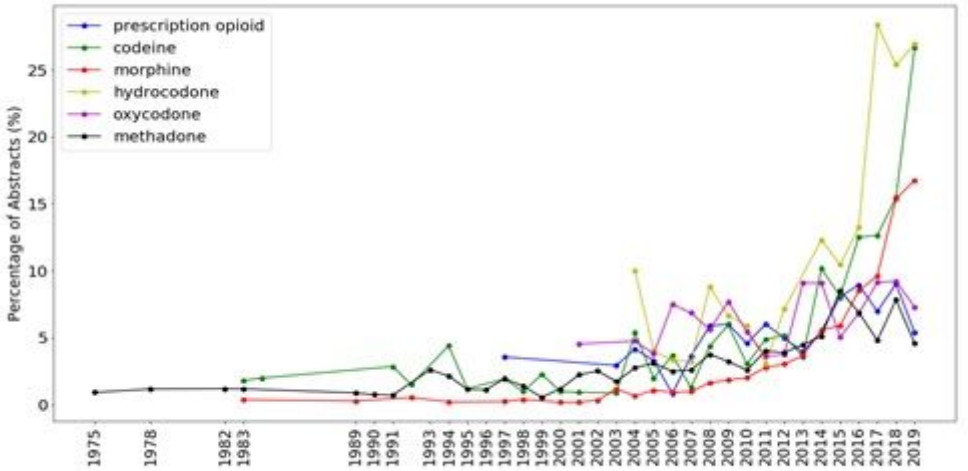

Figure 16

The dynamic of the leading topic for each dataset

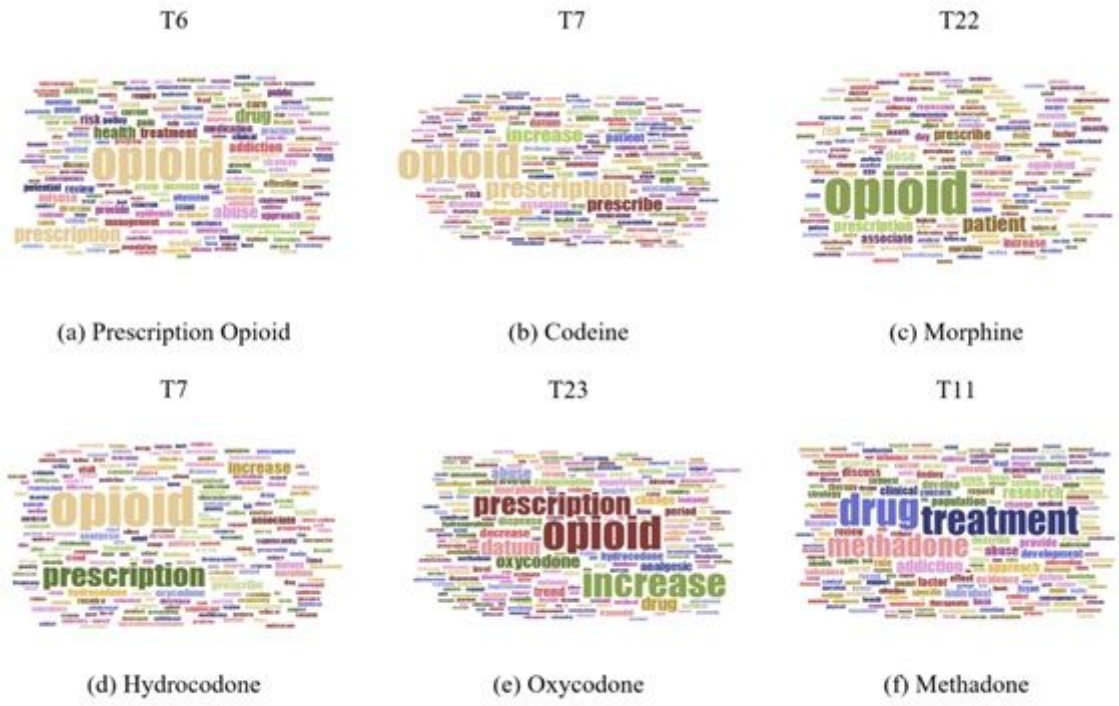

Figure 17

The leading topic's word cloud for each dataset.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- SupplementalInformation1.docx