Detecting Opioid-Related Aberrant Behavior using Natural Language Processing

Jesse M. Lingeman [1], Priscilla Wang [2], William Becker, M.D. [2,3], Hong Yu, PhD [4]
[1]University of Massachusetts: Amherst, Amherst, MA; [2]Yale Medical School, New Haven, CT; [3] West Haven VAMC, West Haven, CT; [4] University of Massachusetts Medical School, Worcester, MA

Abstract

The United States is in the midst of a prescription opioid epidemic, with the number of yearly opioid-related overdose deaths increasing almost fourfold since 2000[1]. To more effectively prevent unintentional opioid overdoses, the medical profession requires robust surveillance tools that can effectively identify at-risk patients. Drug-related aberrant behaviors observed in the clinical context may be important indicators of patients at risk for or actively abusing opioids. In this paper, we describe a natural language processing (NLP) method for automatic surveillance of aberrant behavior in medical notes relying only on the text of the notes. This allows for a robust and generalizable system that can be used for high volume analysis of electronic medical records for potential predictors of opioid abuse.

Introduction

Drug overdose deaths are the leading cause of accidental deaths in Americans, causing more fatalities than gunshot wounds or car accidents[2]. The majority of these deaths are due to prescription opioids, the third most commonly prescribed class of medications in the United States[3]. The most recent National Action Plan for Adverse Drug Event (ADE) Prevention identifies unintentional opioid overdose as the highest priority ADE for targeted prevention nationally[4]. Prevention efforts would be most successful via the use of high-quality surveillance tools, which are currently lacking in two ways: (1) they rely on clinician-coded ADEs, which often do not express a patient's full clinical presentation, and (2) they generally do not capture less severe ADEs (e.g. sedation, dizziness, confusion) or aberrant behaviors (e.g. requests for early refills) which may be prodromal signals for a catastrophic ADE, such as overdose.

As the American health care system moves toward a completely electronic health record system, the field of natural language processing has significant potential to aid clinical research, as it allows mining of free text in patient encounter notes in a high-volume and systematic manner[5]. This may prove particularly valuable in analyzing prodromal signals for severe opioid-related ADEs, which may be predominately described in the unstructured text of outpatient notes. Identification of phrases (and by proxy, patient symptoms or behaviors) predictive of catastrophic ADEs could aid clinicians to identify and intervene in the clinical course of high-risk patients.

Natural language processing techniques have been recently applied to identify problematic opioid usage or assess risk of inappropriate opioid-related behaviors[6,7]. However, these studies have predominately relied on explicit labels of opioid abuse or misuse, either via clinician-documented diagnosis or billing codes or statements in the medical record in which a clinician uses specific interpretative key phrases (e.g. "opioid abuse", "opioid dependence"). They do not include the full spectrum of opioid-related aberrant behaviors and related documentation language used by providers that have been described in literature[8]. Additionally, notes are complex documents. Detecting one or two sentences of aberrant behaviors in a relatively lengthy document creates a difficult classification task. In this study, we describe an annotation system that is not dependent on key words or billing codes, and a natural language processing system that can leverage these annotations to accurately identify notes that contain aberrant behaviors related to opioids.

Methods

Data

The medical notes used in this study were obtained from medical records from the University of Massachusetts Memorial Health Care. An initial dataset was collected by querying for opioid drug names and ICD-9 codes relating to opioid use and abuse (specifically 304.0 and V58.69). Specific notes were manually selected from the larger dataset based on the following inclusion criteria: the note had to be written by a provider documenting a primary care outpatient

encounter, and the patient had to have been taking a prescribed opioid medication other than methadone or buprenorphine. Notes were excluded if the patient was not on a prescribed opioid medication (i.e., if the patient was only using illicit opioids). A medically trained annotator then annotated the notes for any mentions of opioid-related aberrant behavior (defined below). In total, 112 notes were annotated, with 44 containing instances of aberrant behavior. Notes that were annotated to contain aberrant behavior are considered "true positives" in the classification task, and notes that contain a prescribed opioid medication but do not show signs of aberrant behavior are considered "true negatives". All notes were deidentified using DeID[9] prior to annotation or use in this system.

External Datasets

We use several external datasets for feature generation: SentiWordNet[10] and word embeddings trained on PubMed, Wikipedia, and a collection of medical notes (extended from[11] with medical notes). SentiWordNet is a database that indicates whether a word in a sentence has a positive, negative, or neutral connotation. For example, the word "happy" will have a positive sentiment, and will increase the overall sentiment of the document. On the other hand, the word "sad" has a negative connotation, and decreases the overall sentiment of a sentence. We use this dataset to construct features that measure the overall connotation of the medical note, described in Table 1.

Word embeddings[12] are a way of transforming a word into a numerical vector that embeds a word in some high dimensional space. Similar words should be close to each other in this vector space, and dissimilar words are far apart. We use this average vector of a sentence to represent its "meaning" in this space. These embeddings were trained using the word2vec algorithm[12] and are used to construct embedding features, also described in Table 1.

Annotation

Annotation was performed by a medically trained annotator using a coding manual to identify various types of documented opioid-related aberrant behaviors. Opioid-related aberrant patient behaviors were defined as the following:

- Patient behavior suggesting a loss of control of opioid use or compulsive/inappropriate opioid use, including behavior suggesting that the patient is seeking further doses of or other sources of opioids or consuming opioids in ways other than prescribed (e.g. "Patient claims that she has run out of her Percocet, but was just prescribed a 30-day supply one week ago", "Patient admits to crushing and snorting his oxycodone pills")
- Use of illicit substances or misuse of legal substances, not including their use of their prescribed opioid medication (e.g. "Patient has also been obtaining and using cocaine on the street")
- Emotions / strong opinions expressed by the patient in relation to opioids (e.g. "Patient loudly and angrily demanding that I agree to escalate his opioid dose")

We also chose to document clinician reactions in response to opioid-related aberrant patient behavior, specifically expressions of concern by the clinician or specific negative interpretative statements regarding patient's opioid use (e.g. "I am suspicious that the patient is abusing his Vicodin", "I am concerned that the patient is displaying drug-seeking behavior"). The annotator coded all instances of the types of behaviors described above, using the eHOST[13] annotation tool. For the purposes of this study, all four categories were grouped together into a single category of "aberrant behavior."

Support Vector Machines

Since the data were too small for multi-layer neural network models, we applied Support Vector Machines (SVM), a common and robust classification algorithm[14]. SVMs have been shown to perform very well across a wide array of classification tasks, from text classification to image classification. The basic idea behind an SVM is to first decompose an example, in our case a medical note, into a vector of relevant "features" that represent the contents of the note in some compressed form. The algorithm seeks to find a separating hyperplane that can effectively split the examples into two sides, where the examples matching class A are on one side of the hyperplane and examples matching class

B are on the other side. In practice, the examples are not actually perfectly separable in this way, so SVM employs what is called a "kernel trick" to project the examples into an infinite dimensional space, where finding this separating hyperplane is easier, and then projecting that plane back down into the dimensionality of the feature space.

We test using 3 different SVM models: Linear kernel with L1 regularization, linear kernel with L2 regularization, and the radial basis function kernel. L1 regularization is a technique for aggressive feature reduction: weights of uninformative features are quickly set to 0 to reduce noise in the outcome. L2 regularization aims to penalize non-informative features, but it is not as aggressive as L1. The idea behind both regularization functions is that we want to quickly down-weight uninformative features (such as an n-gram that only appears once in the dataset) while maintaining strong weights on those features that do help to differentiate the documents.

During training, we performed a sweep over the SVM hyper-parameter C. C controls the penalty for misclassification of an example: the greater the value of C the more we penalize misses. In practice C also controls the amount of over-fitting: for the model to be generalizable we expect to be incorrect for some examples, so a very high value for C may over-penalize misses and lead to poor test performance. We test values of C from 0.1 to 100 in increments of 0.5 to find the best fit.

All experiments were performed using 10-fold stratified cross-validation. Stratified cross-validation is the process of randomly splitting the dataset into 10 separate "folds", where each fold contains the same number of positive and negative examples. We then select one of the folds to be the test fold, and use the remaining 9 folds for training. Afterwards we select another fold to be the test fold, and so on, until we have cycled through all folds. The scores reported are the mean values of the predicted accuracies over these folds.

Features

Features are broken into two types: hand-crafted features that are specific for this dataset, and features that directly represent that data. Hand-crafted features include whether the note mentions the name of an illicit drug, whether the note includes words associated with drug-seeking behavior, whether the note contains words associated with withdrawal symptoms, and others. Features that directly represent the data include n-grams (of sizes 1 to 4), word-embeddings, and the positive, negative, and neutral sentiment values of the note as derived using SentiWordNet.

The sentiment features are broken down into 3 levels: document, sentence, and word. We measure the average positive and negative sentiment for all words in the record, the maximum positive and maximum negative sentiment over all sentences in the record, and the maximum positive and negative sentiments over all words in the record. Whether a word has positive or negative sentiment is done via a look up to the SentiWordNet database.

The n-gram features were selected by first tokenizing the document using the NLTK tokenizer[15], and then by grouping the tokens into unigrams, bigrams, trigrams, and 4-grams. To reduce the number of features and eliminate most of the n-grams that appear in only a single document, only the most common 1,000 n-grams from each group were used. All n-gram features are represented as binary features.

We also developed a set of handcrafted features specifically tuned to this dataset. The handcrafted features depend on words and phrases that appear in the document. For example, one feature checks for the presence of illicit drugs in the medical note (e.g., heroin). Another feature checks for words associated with anxiety, and checks to make sure that there isnt a negation preceding the mention. In total, we use 13 handcrafted features, described fully in Table 1.

Results

We find through 10-fold cross-validation that we can recover notes that contain aberrant behavior with 81.4% accuracy (std=0.12) using our hand-crafted feature set with sentiment information. However, adding word embeddings does not seem to help (78.5% accuracy, std=0.11). When the word embeddings and the n-gram features are added, our accuracy falls to 67.1% (std=0.10) in the best performing model. This is due to over-fitting: with only 121 notes to learn from, even aggressively pruning features using L1 regularization that are not useful still leads to over-fitting. Removing stop words from the document (using the NLTK English stopword list) did not improve performance. We hypothesize that with more notes the automatically generated features would be more useful. To reduce the dimensionality of the

automatically generated features, we can separate them by type. Hand-crafted features (HC), sentiment features (S), and word embeddings had an accuracy of 79.8%. HC, S, and unigrams had an accuracy of 78.5%. The full breakdown is reported in Figure 1. All sets of automatically generated features ended up hurting the overall accuracy during cross-validation due to over-fitting – there are simply too many features in each of these sets to be effectively fit given the small data size.

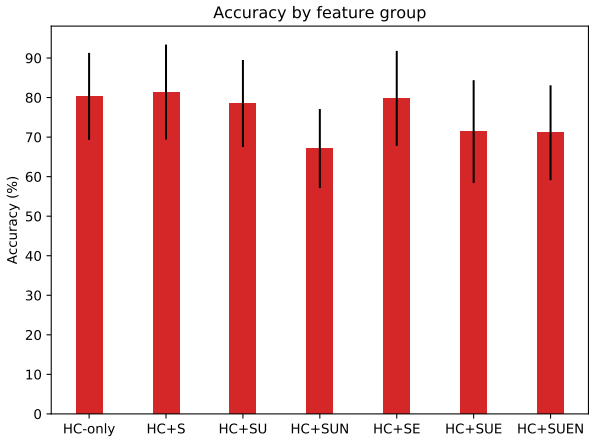

Figure 1: Accuracy over different sets of features. HC are hand-crafted features, S refers to sentiment features, U refers to unigram features, N refers to n-grams, E refers to embeddings.

We now break down the results by which features were most useful. We wanted to see if any single feature was more predictive than others. To test this, we performed a recursive feature elimination test to remove features that were not useful, and found that the best results occurred when all sentiment and hand-crafted features were in the model. As shown in Table 1, none of the features alone performed as well as the overall model, with the "patient anxiety" and "illicit drugs" features being the most predictive single features. Each of the hand-crafted features is dependent on the others to achieve high accuracy.

Next, we report results on classification of randomly selected unannotated notes. 50 random notes were pulled from the larger collection of medical notes that contain mentions of opioids and were not selected to only include outpatient notes. Of these 50 notes, the system using only hand-crafted features marked 15 notes as possibly containing aberrant behavior. Of these 15, the majority were either hospice care notes or notes following up on pain management treatments from surgery (11 in total). However, 4 of the notes selected contained aberrant behavior matching our criteria (all 4 included documentation of compulsive seeking of opioids, one included an opioid-related expression of anger by the patient). Within the 35 notes marked by the system as "negative" for aberrant behavior, two were incorrectly classified as negative (one included an opioid-related expression of concern by the physician, and one note indicated that the patient was misusing a legally prescribed medication).

Discussion and Conclusions

We find that we can differentiate clinical encounter notes that contain opioid-related aberrant behavior from those that do not with relatively high accuracy (81%), using only the free text of the note and without use of structured data, such as medical billing codes. We find that mentions of illicit drug use and patient anxiety are strong predictors of documented aberrant behavior, while n-grams and word embeddings are not particularly helpful for this task. However, we found that all of the hand-crafted features, when combined with sentiment information, had the best performance. It is likely that the ineffectiveness of n-gram and embedding based features is due to the small size of our annotated dataset.

Our future work will seek to enhance our tool's performance using a larger dataset of annotations. However, as manual annotations are challenging and costly due to the necessity of the annotator being a medical professional, we will also explore augmenting the dataset using a semi-supervised classification system. This will allow us to expand our dataset

beyond what would be possible with manual annotations, and help to ensure robustness. We also plan to analyze other aspects of medical notes such as prior diagnoses of PTSD, depression, or non-opioid drug use. Creating features that capture these patient-centric components may yield new insight into a system that could warn of potential abuse before it is mentioned in a medical note.

We have also demonstrated that our tool can be useful in a medical context to identify opioid-related aberrant behavior. Out of 50 medical notes randomly selected from diverse clinical settings for manual annotation, we recovered 4 out of 6 instances of documented aberrant behavior.

The majority of system-tagged false positives stemmed from non-primary care notes (excluded from the notes the system was trained on) and controlled settings in which pain medications are frequently prescribed (e.g. post-surgery and hospice settings). As we broaden the dataset on which the system is trained to include a larger array of clinical settings, we expect that the specificity of the system will increase. The majority of high volume analysis of electronic medical records relies on analysis of structured data, such as medical billing codes, which may exclude the richness of information included in the free text documentation of clinical encounters. This work represents a step forward in automated medical note analysis: we demonstrate that even with a small amount of annotated data, it is possible to automatically extract pertinent information about the notes without relying on external tagging systems. Our work suggests an approach by which free text-based natural language processing can be used to increase the effectiveness of public health surveillance systems, such as in the identification and prevention of opioid abuse.

References

1. V H Murthy. Ending the Opioid Epidemic—A Call to Action. *New England Journal of Medicine*, 2016.
2. Jiaquan Xu, Sherry L Murphy, Kenneth D Kochanek, and Brigham A Bastian. Deaths: Final Data for 2013. *National Vital Statistics Reports*, 2016.
3. N D Volkow, T A McLellan, J H Cotto, and M Karithanom. Characteristics of opioid prescriptions in 2009. *JAMA*, 2011.
4. U.S. Department of Health, Office of Disease Prevention Human Services, and Health Promotion. National Action Plan for Adverse Drug Event Prevention, 2014.
5. F Liu, C Weng, and H Yu. Natural language processing, electronic health records, and clinical research. *Health Informatics*, 2012.
6. D S Carrell, D Cronkite, R E Palmer, and K Saunders. Using natural language processing to identify problem usage of prescription opioids. *International Journal of Medical Informatics*, 2015.
7. I V Haller, C M Renier, and M Juusola. Enhancing Risk Assessment in Patients Receiving Chronic Opioid Analgesic Therapy Using Natural Language Processing. *Pain Medicine*, 2016.
8. J S Merlin, J M Turan, I Herbey, and A O Westfall. Aberrant DrugRelated Behaviors: A Qualitative Analysis of Medical Record Documentation in Patients Referred to an HIV/Chronic Pain Clinic. *Pain Medicine*, 2014.
9. Ishna Neamatullah, Margaret M Douglass, H Lehman Li-wei, Andrew Reisner, Mauricio Villarroel, William J Long, Peter Szolovits, George B Moody, Roger G Mark, and Gari D Clifford. Automated de-identification of free-text medical records. *BMC medical informatics and decision making*, 8(1):32, 2008.
10. Stefano Baccianella, Andrea Esuli, and Fabrizio Sebastiani 0001. SentiWordNet 3.0 - An Enhanced Lexical Resource for Sentiment Analysis and Opinion Mining. *2015 7th International Conference on Electronics, Computers and Artificial Intelligence (ECAI)*, 2010.
11. A N Jagannatha, J Chen, and H Yu. Mining and ranking biomedical synonym candidates from Wikipedia. *Proceedings of the Sixth International Workshop on Health Text Mining and Information Analysis*, 2015.
12. Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado, and Jeffrey Dean. Distributed Representations of Words and Phrases and their Compositionality. *arXiv.org*, 2013.

13. B R South, S Shen, J Leng, and T B Forbush. A prototype tool set to support machine-assisted annotation. *Proceedings ofthe 2012 Workshop on Biomedical Natural Language Processing*, pages 130–139, 2012.

14. C Cortes and V Vapnik. Support-vector networks. *Machine Learning*, 1995.

15. Edward Loper Bird, Steven and Ewan Klein. *Natural Language Processing with Python*. O'Reilly Media Inc., 2009.

Table 1: The names, description, and individual accuracy of all features used. Accuracy was determined by performing the same SVM experiment as with the full model, but with only that feature/feature set.

Feature Name	Feature Description	Accuracy
Declined	Note features word "declined".	65.2%
Anger	Note features anger words ("belligerent, "screaming", "angry", "anger", "violent", "frustration")	60.9%
Pain contract violation	Note contains "violate" and "pain contract" within 30 characters of each other.	63.6%
Patient depression	Note mentions depression and no negation words in sentence.	60.7%
Buying drugs	"Buying" or "bought" in close relationship to drug name.	60.7%
Drug abuse mentions	Note contains word "abuse" or "abused" or "abusing".	61.2%
Drug desire	Note contains request for specific drug.	62.9%
Drug addiction mentions	Note contains words "addict", "addiction", "drug seeking", "misuse", or "high risk".	67.1%
Patient anxiety	Note contains "anxiety" or "anxious" and is not negated.	70.7%
Patient complaints	Note contains "not helping" or "not working", these are common phrases in relation to treatment.	60.7%
Drug seeking	Note contains mention of needing drugs to "get through" or indications that they "ran out" early.	65.4%
Illicit drug use	Note contains mentions of illicit drugs (heroin, marijuana, cocaine).	72.1%
Sentiment	Overall positive and negative sentiments as measured by SentiWordNet.	63.3%
Unigrams	Note contains single words.	65.1%
N-grams	Note contains phrases of length 2-4.	64.4%
Word embeddings	Average word embedding vector for all words in note.	64.7%