ISDS 2019 Conference Abstracts

Development of Text-Based Algorithm for Opioid Overdose Identification in EMS Data

Andrew Patton[1, 2], Rochelle Ereman[2], Matt Willis[2], Haylea A. Hannah[2], Karina Arambula[2]

[1]Johns Hopkins Bloomberg School of Public Health, Baltimore, Maryland, United States

[2]Marin County Health and Human Services, San Rafael, California, United States

Objective

To develop and implement a classifcation algorithm to identify likely acute opioid overdoses from text fields in emergency medical services (EMS) records.

Introduction

Opioid overdoses have emerged within the last five to ten years to be a major public health concern. The high potential for fatal events, disease transmission, and addiction all contribute to negative outcomes. However, what is currently known about opioid use and overdose is generally gathered from emergency room data, public surveys, and mortality data. In addition, opioid overdoses are a non-reportable condition. As a result, state/national standardized procedures for surveillance or reporting have not been developed, and local government monitoring is frequently not specific enough to capture and track all opioid overdoses. Lastly, traditional means of data collection for conditions such as heart disease through hospital networks or insurance companies are not necessarily applicable to opioid overdoses, due to the often short disease course of addiction and lack of consistent health care visits. Overdose patients are also reluctant to follow-up or provide contact information due to law enforcement or personal reasons. Furthermore, collected data related to overdoses several months or years after the fact are useless in terms of short-term outreach. Therefore, given the potentially brief timeline of addiction or use to negative outcome, the current project set to create a near real-time surveillance and treatment/outreach system for opioid overdoses using an already existing EMS data collection framework.

Methods

Marin County Department of Health and Human Services EMS data (2015-2017) was used for development of the system. The pool of data for model development and evaluation consisted of 15,000 EMS records randomly selected from 2015, 2016, and 2017. Each record was manually classified in a binary manner with the criteria of “more likely than not opioid related”, using only selected text fields. The event did not need to be exclusively opioid related, nor did opioids have to be the primary cause for the EMS call. 2,000 records were selected for review by the medical director for Marin County EMS, with a Cohen’s kappa coefficient of approximately 0.94. Overall, the proportion of opioid overdoses was less than 0.01 amongst the 15,000 records. An enriched data set of 80 randomly selected overdoses and 320 randomly selected non-overdoses was created for the purposes of feature engineering. These 400 records were excluded for further use in model training and testing. Within the enriched set, the descriptive text fields were tokenized based on the hypothesis that opioid overdoses and non-overdoses are separable based on the content of the descriptive fields. Each field was tokenized as words, bigrams (pairs of consecutive words), and trigrams (triplets of consecutive words). The frequencies of each token as a percentage of overall words were calculated separately for opioid overdoses and non-overdoses. Structured fields used in the analysis were not tokenized prior to frequency calculations. The frequencies for each token/phrase were then compared across opioid overdoses status with a proportion test for equality at an alpha of 0.05 with a Bonferroni correction for multiple comparisons. The tokens/phrases that were statistically significantly more likely to be present in opioid overdoses were assigned to a quintile based on their p-value, with smallest p-values assigned five, and largest p- values assigned one. Tokens/phrases statistically significantly more likely to be present in non-overdoses were scored in the same manner, with the smallest p-value assigned negative five, and the largest p-value negative one. The tokens/phrases that were statistically different across opioid overdose status were stored along with their quintile scores in dictionaries that were kept for future modeling use. From the initial 15,000 classified records, excluding the 400 used for the enriched data set, 10,000 records were randomly selected for model training and development. Each record had their text fields tokenized into words, bigrams, and trigrams, and each was compared with the corresponding dictionary. If a token was present in the entry and also in the dictionary, that token’s quintile score was assigned to the record, with multiple tokens being summed to produce a score for each field-token option. The final created feature was the count of opioid specific terms such as “heroin”, “fentanyl”, “narcan”, etc. within the main narrative field. The intent was to create a variety of numerical features that were indicative of presence of tokens/phrases that were positively associated with opioid overdoses such that higher scores were more associated. Several models including support vector machines, neural nets, gradient boosted machines, and logistic regression were tested via 10-fold cross validation, with logistic regression yielding the best error rates and lowest computational costs. Although all models resulted in a sensitivity greater than 85 percent, logistic regression was by far the best in terms of false positive rate. The coefficients for the logistic regression model were selected from the eight created features along with patient sex and patient age by best subsets selection via Akaike information criterion (AIC), and the probability threshold for classification was selected via optimizing the receiver operating curve (ROC).

Results

Following the variable selection and threshold optimization for logistic regression, the sensitivity and specificity of the model were between 90 percent and 95 percent. However, given the large number of records fed through the algorithm either each week for 'real-time' surveillance and treatment/outreach, or for larger retrospective data sets, improving specificity is crucial to reduce the number of false positives. Additionally, given that a public health treatment/outreach staff has a finite amount of time and resources, limiting false positives will allow them to focus on the true cases. Further model improvements were made with a series of binary filters that allowed for overall sensitivity/specificity improvements as well as ensuring that the records sent for outreach are appropriate for outreach. The application of the filters pushed the classification sensitivity and specificity to greater than 99 percent. Further, the filters removed cases inappropriate for outreach at greater than 90 percent efficiency.

Conclusions

The algorithm was able to classify opioid overdoses in EMS data with a sensitivity and specificity greater than 99 percent. It was implemented into a viable public health treatment/outreach system through the Marin County Department of Health and Human Services in May 2018, and has identified approximately 50 overdoses for outreach as of September, 2018. It is possible, using minimal computational power and infrastructure to develop a fully realized surveillance system through EMS data for nearly any size public health entity. Additionally, the framework allows for flexibility such that the system can be tailored for specific clinical or surveillance needs - there is no 'black box' component. Lastly, the application of this methodology to other diseases/conditions is possible and has already been done using the same data for both sepsis and falls in older adults.

Acknowledgement

This project was supported by a grant from California Department of Public Health.

References

1. R Core Team. 2018. R: A Language and Environment for Statistical Computing. Available: https://www.r-project.org/.

2. RStudio Team. 2018. RStudio: Integrated Development Environment for R. Available: http://www.rstudio.com/.