2021

Enhancing Drug Overdose Mortality Surveillance through Natural Language Processing and Machine Learning

Patrick J. Ward
University of Kentucky, pward5656@gmail.com
Author ORCID Identifier:
https://orcid.org/0000-0001-7331-1772
Digital Object Identifier: https://doi.org/10.13023/etd.2021.131

STUDENT AGREEMENT:

I represent that my thesis or dissertation and abstract are my original work. Proper attribution has been given to all outside sources. I understand that I am solely responsible for obtaining any needed copyright permissions. I have obtained needed written permission statement(s) from the owner(s) of each third-party copyrighted matter to be included in my work, allowing electronic distribution (if such use is not permitted by the fair use doctrine) which will be submitted to UKnowledge as Additional File.

I hereby grant to The University of Kentucky and its agents the irrevocable, non-exclusive, and royalty-free license to archive and make accessible my work in whole or in part in all forms of media, now or hereafter known. I agree that the document mentioned above may be made available immediately for worldwide access unless an embargo applies.

I retain all other ownership rights to the copyright of my work. I also retain the right to use in future works (such as articles or books) all or part of my work. I understand that I am free to register the copyright to my work.

REVIEW, APPROVAL AND ACCEPTANCE

The document mentioned above has been reviewed and accepted by the student's advisor, on behalf of the advisory committee, and by the Director of Graduate Studies (DGS), on behalf of the program; we verify that this is the final, approved version of the student's thesis including all changes required by the advisory committee. The undersigned agree to abide by the statements above.

Patrick J. Ward, Student

Dr. April M. Young, Major Professor

Dr. Heather Bush, Director of Graduate Studies

Enhancing Drug Overdose Mortality Surveillance through Natural Language Processing and Machine Learning

---

DISSERTATION

A dissertation submitted in partial fulfillment of the
requirements for the degree of Doctor of Philosophy in the
College of Public Health
at the University of Kentucky

By
Patrick J. Ward
Lexington, Kentucky
Director: Dr. April M. Young, Associate Professor of Epidemiology
Lexington, Kentucky
2021

https://orcid.org/0000-0001-7331-1772

ABSTRACT OF DISSERTATION

ENHANCING DRUG OVERDOSE MORTALITY SURVEILLANCE THROUGH NATURAL LANGUAGE PROCESSING AND MACHINE LEARNING

Epidemiological surveillance is key to monitoring and assessing the health of populations. Drug overdose surveillance has become an increasingly important part of public health practice as overdose morbidity and mortality has increased due in large part to the opioid crisis. Monitoring drug overdose mortality relies on death certificate data, which has several limitations including timeliness and the coding structure used to identify specific substances that caused death. These limitations stem from the need to analyze the free-text cause-of-death sections of the death certificate that are completed by the medical certifier during death investigation. Other fields, including clinical sciences, have utilized natural language processing (NLP) methods to gain insight from free-text data, but thus far, adoption of NLP methods in epidemiological surveillance has been limited. Through a narrative review of NLP methods currently used in public health surveillance and the integration of two NLP tasks, classification and named entity recognition, this dissertation enhances the capabilities of public health practitioners and researchers to perform drug overdose mortality surveillance. This dissertation advances both surveillance science and public health practice by integrating methods from bioinformatics into the surveillance pipeline which provides more timely and increased quality overdose mortality surveillance, which is essential to guiding effective public health response to the continuing drug overdose epidemic.

KEYWORDS: Machine learning; substance use; drug overdose; surveillance; deep learning

Patrick James Ward

04/14/2021

Date

ENHANCING DRUG OVERDOSE MORTALITY SURVEILLANCE THROUGH NATURAL LANGUAGE PROCESSING AND MACHINE LEARNING

By
Patrick J. Ward

April M. Young, PhD
Director of Dissertation

Heather Bush, PhD
Director of Graduate Studies

04/14/2021
Date

ACKNOWLEDGMENTS

There are too many people who have helped me along my academic career to thank everyone. I want to first thank my committee; my chair, Dr. April Young, for being in my corner and guiding me through the PhD process from beginning to end; Dr. Svetla Slavova, for being a supportive co-worker, supervisor, and advisor, and for always telling me how you felt, good or bad; Dr. Heather Bush, for encouraging me to be creative in how I think about my work and always providing thoughtful, constructive criticism; and Dr. Jennifer Havens, whose subject area expertise has helped immensely in grounding this methodological work to always be focused on having practical applications for improving public health.

I want to thank my wife, who has been always supportive of my dreams, for being willing to move several states for me to pursue further education; this accomplishment is not possible without you being a steadying hand for me throughout. I would also like to thank the rest of my family, particularly my parents, who set an example for me that I can do anything I put my mind to, while also instilling in me the value of hard work. My academic accomplishments would not be possible without you.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1

Introduction

In 2019, there was a total of 70,630 overdose deaths in the United States,[1] increasing from 67,367 overdose deaths 2018.[2] Drug overdoses continue to be a major cause of morbidity and mortality in the United States,[2] in large part due to the opioid epidemic.[3-5] Evidence from morbidity data indicates that the novel coronavirus pandemic has led to increased drug overdoses[6] in 2020, suggesting that drug overdose outcomes are still rising. Identifying and characterizing the burden of the epidemic is critical to developing interventions and guiding policy and funding initiatives to combat this public health crisis. Epidemiologists and data scientists working at the state, local, and national levels use surveillance data to identify regions with high burdens of overdoses, demographic subgroups at disproportionate risk of overdose, and substances involved[7] in causing morbidity and mortality. Mortality surveillance data has highlighted increases in deaths caused by drugs such as fentanyl[8,9] and psychostimulants.[10]

Evidence from surveillance data is used to guide public health interventions and public safety initiatives. To effectively guide practice, surveillance data must be both accurate and timely. With the ever-changing landscape of the drug overdose epidemic,[4] timely data is essential. Public health interventions should be targeted to the specific substances causing morbidity and mortality in near-real time, not substances that were affecting the community in a year prior. Similarly, accurately identifying what substances are causing morbidity and mortality is crucial. Public health interventions for prescription opioids may consist of increased analysis of prescription drug monitoring program data[11] to identify and stop “doctor shopping”, while an initiative aimed at illicit opioid mortality

may focus more on supply-side interventions and harm reduction. Similarly, interventions for preventing opioid overdoses may include expanding access to medication for opioid use disorder,[12,13] while interventions for misuse of methamphetamine, a substance causing increased morbidity and mortality during the opioid crisis,[10] would include increasing access to cognitive behavioral therapy.[14] Without timely and accurate surveillance data, actionable evidence for the development of public health initiatives will be limited.

A current barrier in surveillance of drug overdoses is the reliance on death certificate data for mortality surveillance. While death certificates contain detailed information about the decedent's demographics and cause- and manner-of-death, death certificates often lag behind other surveillance sources due to both the time a death investigation takes[15] and the process of adding International Classification of Diseases, 10th Revision (ICD-10)[16] codes to the death certificates. Additionally, medical certifiers who complete death certificates for drug overdose deaths in some states are not required to have a medical background.[17] The medical certifier is charged with completing the free-text cause-of-death and description-of-injury sections of the death certificate, from which information on what substances caused an overdose death is obtained. Data quality stemming from errors or misspellings in these sections present additional challenges in using death certificate data. For example, the medical certifier may incorrectly spell the name of a specific substance, causing the record to be missed when identifying the number of overdoses involving that substance.

This dissertation aims to address the existing limitations in epidemiological analyses relying on death certificate data by enhancing the timeliness and completeness

of the information derived from death certificates to advance drug overdose surveillance science. The limitations in timeliness and quality both stem from the need to analyze free-text data, specifically, the free-text cause-of-death section of the death certificate that is completed by the medical certifier during the death investigation. Free-text data can be analyzed through the use of natural language processing (NLP) and machine learning (ML) methods. NLP and ML have seen extensive use in clinical sciences[18,19] but only limited use in public health.[20] An example of NLP use in other fields include identifying drugs from clinical texts,[21-23] which could be similarly adopted to identify drugs from death certificate text.

With the current limitations of death certificate data stemming from the need to analyze free-text, leveraging NLP methods is the next step in advancing public health surveillance for drug overdose. NLP provides a suite of methods for transforming unstructured, free-text data into "features" which can be fed as variables into classical statistical methods as well as ML algorithms. NLP typically leads to extremely high dimensional data, necessitating ML models capable of handling high dimensionality to analyze. In a typical NLP and ML pipeline, data is split into training and testing sections. Free-text training data is processed via NLP methods and the created features are fed into an ML model that, using these features, are trained to predict some outcome (y). This trained algorithm is then deployed onto the test data, predicting the outcome for each record in the test data set. The model is then scored based on how well it performed on the test data against a gold standard. The final, trained algorithm can then be used on new data to predict the outcome when it is unknown or unavailable.

The following chapters address three specific aims for developing and integrating NLP and ML models in the public health surveillance pipeline for drug overdose mortality to advance surveillance science and improve public health practice through more timely and accurate surveillance. These aims seek to advance surveillance science through the tenets of integration (applying techniques that emerged from a different field allows us to uncover new phenomena, linked to previous work in clinical sciences) and development (the methods developed in the dissertation are the first of their kind for applying NLP to overdose mortality surveillance, and provide a base for researchers to build further NLP models for surveillance), defined as two of the five ways in which scientific progress is achieved according to the National Research Council.[24]

- *Aim 1/Chapter 2*: Review the existing literature on NLP and ML use in public health surveillance to identify current uses and gaps in the literature,
- *Aim 2/Chapter 3*: Develop a machine learning model that can identify drug overdose deaths through the analysis of free-text to improve the timeliness of surveillance data, and
- *Aim 3/Chapter 4*: Advance surveillance science by improving the detection and identification of novel drugs and drug misspellings through a state-of-the-art deep learning model.

CHAPTER 2

Natural language processing and machine learning methods in public health surveillance: a narrative review

Introduction

Public health surveillance (PHS) is the systematic, ongoing collection of health-related data and its use to assess the health of populations.[25] PHS is critical for detecting epidemics,[26,27] observing long-term health trends,[4,28,29] and detecting emerging threats to public health.[30,31] For example, surveillance of the substance use epidemic has helped reveal the rise in fentanyl and its contribution to overdoses in the United States.[8,9] Traditionally, PHS has relied on structured data sources including death certificates,[5,7,9,28,29,32-35] hospitalization discharge billing data,[36,37] disease registries,[38,39] and systems that track specific conditions, such as mandatory reporting databases.[40] Each has limitations, including timeliness of death certificate data[15,41] and lack of specificity in the coding structure in hospital billing data.[42,43] To address these limitations, PHS has begun to increasingly rely on unstructured data sources for surveillance such as free-text chief complaint fields from electronic health record (EHR) and emergency department data[44,45] and narrative text from emergency medical service reports.[46-50]

Natural language processing (NLP) and machine learning (ML) methods have been developed and applied in clinical sciences[51-53] to analyze free-text data. Broadly, NLP consists of methods that enable computers to process and analyze human language. ML algorithms can be used to develop predictive models using variables ("features") derived from NLP output. In clinical science, NLP and ML tasks typically focus on classification (e.g., identifying reportable cancer cases[52]) or information extraction (e.g.,

identification of cardiac risk factors[54]). Using NLP and ML allows researchers to extract information without performing lengthy, manual reviews. Relatively recent advances in ML, particularly deep learning, the use of neural networks with multiple hidden layers for use on complex, feature-rich data,[55] have improved the accuracy of NLP methods.[56-59]

Compared to relatively common use in clinical science,[18,19] NLP and ML have been used more rarely in PHS. Increased integration of NLP and ML into PHS activities could advance both the science and practice of surveillance by improving capabilities for processing bigger, unstructured data.[20] Studies examining NLP and ML in PHS have been published in a variety of sources, given the cross-disciplinary nature of these methods. This narrative review sought to collect peer-reviewed literature that applied NLP and ML in PHS and describes the current state of these techniques' use in the field and their current and potential future contribution of NLP to the assessment and science of population health.

Methods

To identify existing literature for this narrative review, PubMed was searched for articles containing all of the following terms: ("public health" AND "surveillance" AND "machine learning" AND "natural language processing"). When performed in November 2019, this search resulted in a total of 82 articles. Inclusion criteria consisted of being peer-reviewed and describing either the development or application of an NLP method that utilizes ML for PHS purposes. No date cutoff was used to exclude articles.

Of the 82 articles identified, 14 met inclusion criteria. Most of the 68 articles that did not meet inclusion criteria were articles that developed NLP and ML methods, but not explicitly for PHS; rather, they mentioned that such a method could be used for PHS in

their discussion. Additional reasons for exclusion included articles that applied text-matching methods (such as regular expressions) to analyze text, but did not apply ML algorithms. The included articles' reference lists were searched for additional candidates, resulting in the addition of 8 articles for a total of 22 articles included. An abstraction form programmed in REDCap was used to guide information abstraction from the articles. Abstracted information included year published, purpose and rationale, ML task (the formal purpose of the ML model, in NLP typically classification or information extraction), training and testing sample size (the number of records used to develop the ML algorithm and the number of records used to validate the model, respectively), evaluation metrics, score on final evaluation metric, strengths, limitations, and future directions noted by the authors.

Results

Of the 22 articles included in this narrative review, 12 involved traditional data sources[41,52,60-69] and 10 involved online media data,[70-79] summarized in Tables 1 and 2 respectively. Studies that applied NLP to traditional PHS data covered several health conditions, including autism spectrum disorder (ASD),[61,69,70,77] drug overdose,[41] cancer,[52,60,62,65,66] and infectious diseases.[67] Applying NLP and ML addressed different challenges in traditional surveillance, including the absence of a defined surveillance definition,[63] lengthy reviews required to identify cases,[61,69] and enhancing the timeliness of reporting.[41] Most articles that used online media analyzed Twitter (www.twitter.com) data; Twitter is a microblogging website where users post "tweets," messages containing up to 280 characters. Challenges addressed by utilizing NLP on online media included

monitoring disease activity in near-real time[71,72,75,76,78,79] and monitoring conditions that are difficult to track typically.[70,73,77]

Discussion

Applying Natural Language Processing to Traditional Public Health Data Sources

Traditional PHS data that were enhanced with NLP and ML included death certificate data and emergency department data, including EHRs and pathology reports. Several studies utilized NLP and ML to address time lags present in mortality surveillance, which relies on International Classification of Diseases, Tenth Revision (ICD-10) codes[16] for identification of conditions. To eliminate time lag of ICD-10 coding for drug overdose mortality surveillance, Ward et al developed a support vector machine (SVM) classifier that utilized features from free-text on an individual's death certificate to classify deaths as drug overdose deaths, achieving an F-score[80] of 0.97.[41] NLP and ML methods have been similarly used on death certificates for cancer surveillance.[62,65,66] Butt and colleagues[66] used previously developed named-entity-recognition tools to extract features for ML classifiers to identify death certificates as cancer for reporting cancer cases, resulting in an F-score of 0.99. This methodology was then extended[65] to categorize identified cancers' ICD-10 codes to determine the type of cancer. While their analysis achieved high accuracy on common cancers (which made up 85% of all cancers), results were lower on rare cancers. This limitation was further addressed[62] by developing a hybrid ML and rule-based approach, which achieved an overall F-score of 0.80.

NLP and ML have also been utilized on mortality data for general surveillance.[67,68] Koopman and colleagues[67] utilized NLP to extract features for an ML

classifier that predicts if a death was caused by diabetes, influenza, pneumonia, or HIV, achieving an F-score of 0.96, allowing for faster mortality surveillance. Another study developed a deep learning architecture to perform ICD-10 coding of death certificates from free-text.[68] The final model had an accuracy of 0.76, with infrequently occurring ICD-10 codes having poor prediction scores. The authors note that sparse modeling methods could improve the prediction on these rarer ICD-10 codes. While this analysis had relatively low accuracy, the task of coding every ICD-10 code is ambitious.

PHS utilizing NLP and ML has also been performed for morbidity surveillance using EHRs[63,64] and pathology reports.[52,60] Surveillance of suicide attempts is difficult with structured emergency department data. To address this limitation, Metzger et al[64] utilized text from EHRs to develop features for ML classifiers to identify suicide attempt visits. This method had an F-score of 0.953, and authors intend to integrate the model into national surveillance. Another analysis of EHR data identified adverse drug events utilizing named-entity-recognition and relation extraction.[63] This analysis compared classical ML methods to a more state-of-the-art, long-short-term memory (LSTM) model, with the developed SVM outperforming the LSTM (F-score 0.89), providing evidence that traditional ML models can still be utilized for PHS over state-of-the-art methods. Finally, text from pathology reports were used in conjunction with NLP and ML for reporting to cancer registries. Osborne et al[52] developed a ML method to classify reportable cancer cases among pathology reports, with a maximum entropy model attaining an F-score of 0.85, while Alawad and colleagues[60] developed a multi-task convolutional neural network to extract case information (histological type, cancer cite, etc.) required by cancer registries. While the overall scores for this method were not

particularly strong, the study displayed that multi-task convolutional neural networks have improved classification performance compared to single-task models, particularly for imbalanced data.

Another application for morbidity surveillance utilized free-text data from the Autism and Developmental Disabilities Monitoring network, which are reviewed by clinicians to identify ASD cases.[69] This is a lengthy process that has led to lagging surveillance. In order to address this time lag, Maenner and colleagues[69] developed a random forest classifier utilizing features from words and phrases from the network's data. Their random forest was trained on 1,162 records from 2008 and tested on 1,450 records from 2010. Despite this limited training data, their algorithm resulted in an area under receiver-operating characteristic curve[80] of 0.932. The gold standard method of estimating prevalence for ASD results in an estimate of 1.55%, while their ML method estimates a prevalence of 1.46%, showing promise for utilizing the method for ASD surveillance. Leroy et al[61] extended this work by using NLP and ML to extract diagnostic criteria for ASD from EHRs; while the evaluation scores were not very high (0.76 and 0.43 precision and recall scores, respectively) it is clear there is promise for NLP in ASD surveillance with additional work needed.

Applying Natural Language Processing to Online Media

While most analyses of online media utilized social media data,[70,72-79] Feldman and colleagues[71] used news reports to develop a ML classifier to track global infectious disease epidemics. Recognizing that media reports often discuss infectious disease outbreaks prior to the signal in surveillance systems, their developed algorithm takes

news articles that discuss infectious diseases and identifies those that discuss disease activity levels. Their ensemble ML model had a final F-score of 0.87, showing potential for real-time monitoring of disease activity levels with media reports. Limitations of this include potential error in media reports and data is limited by media interest.

Another study sought to perform PHS on a range of health conditions utilizing Twitter data.[72] A system called "Crowdbreaks"[81] was developed that identifies tweets discussing health conditions, then relying on crowdsourcing for annotation. Several built-in ML models perform surveillance. One task that was explored was performing sentiment analysis (classifying a tweet as positive, negative, or neutral) on tweets that discussed vaccines, which resulted in precision and recall scores of 0.77. While Crowdbreaks is an innovative approach, relying on crowdsourced labels may lead to errors in labeling (typically, annotating in studies is performed by multiple trained annotators) and a lack of labeled data to train the built-in models.

Most of the literature that discussed NLP and ML for PHS using social media data focused on a specific condition, rather than a range of conditions. To address challenges in substance use surveillance, one analysis identified tweets that discuss prescription drugs and developed a ML model that identified if the tweet is discussing abuse of the prescription.[77] This study achieved an F-score of only 0.45 in the highest performing SVM, with a relatively small sample size (6,400 tweets). With a larger sample size, state-of-the-art deep learning models would likely improve classification. Another analysis focused on substance use sought to use Twitter and Reddit data to identify individuals to target for substance use disorder interventions.[70] Reddit (www.reddit.com) is a compilation of forums (called "subreddits") where users can share messages, videos, and

other content and have discussions. The classifier developed relied primarily on features from lexicons and lookup tables, and as a short report detailed information about the modeling strategy is limited. Their final algorithm had an accuracy of 0.90 for identifying individuals open to recovery interventions.

Two other studies utilized Twitter data for maternal and child health surveillance. A 2018 analysis identified and characterized tweets discussing birth defects to estimate the prevalence of specific birth defects, which are difficult to measure.[73] A lexicon and rules were developed and bootstrapping methods were used to classify tweets that discussed a child with a birth defect and not just discussing birth defects in general. This approach resulted in a recall score of 0.95, indicating that expert rules can be utilized effectively for text classification. Sarker et al[74] sought to identify women who are pregnant through user tweets to develop cohort data for drug safety surveillance. Basic pattern matching filtered tweets to identify if a user is potentially pregnant, and then several ML models were tested utilizing features derived from tweets to determine which of the filtered tweets indicated the user is pregnant. Their best performing SVM model resulted in an F-score of 0.88, outperforming a neural network.

One topic that has been extensively explored was the use of Twitter for influenza surveillance. A common issue in using Twitter for PHS is filtering tweets that discuss a disease but are not talking about an individual having the condition. One analysis explored this issue utilizing a two-staged approach of determining if a tweet 1) discusses influenza and 2) determining if it is discussing an individual who is infected.[78] Their method was applied at local and national levels, with counts of ML-model identified tweets highly correlated at the week level with counts of influenza-like-illness (ILI) from

CDC national surveillance data, but local-level counts not as correlated. Allen et al[79] performed a similar analysis, utilizing an SVM to identify tweets that discussed an influenza infection. Their model's F-score was 0.786, with significant correlation between the national ML-model identified rate and the national ILI rate. Similarly to Broniatowski et al's study[78] however, regional-level surveillance was less accurate.

An additional study[76] identified if Twitter users misdiagnose themselves with influenza to determine if Twitter is a feasible option for influenza surveillance. While their study was limited by a small sample size (1,274), the analysis indicated that Twitter data incorrectly reports a flu season occurring from late 2011 to early 2012, despite no increases in WHO positive influenza counts. A final influenza study[75] utilized Twitter data to extract topics from tweets that mentioned drugs used to treat influenza. Tweets mentioning influenza drugs of interest were identified and then words from the tweet were used as features to train ML classifiers that predict if a tweet is indicating consumption of the mentioned drug. Their highest performing SVM had an F-score of 0.82, indicating promise for tracking drug consumption with Twitter.

A common limitation in studies that use online media is that terminology individuals use to indicate health conditions are likely different from medical terminology common in PHS. Tweets often contain abbreviations/slang. Additionally, a tweet indicating a user has the "flu" may not be discussing influenza at all, as people use "flu" to indicate a common cold. These limitations should be addressed as social media is used for disease surveillance. Other limitations include private Twitter accounts, whose tweets are not available for analysis, and the large samples required for ML models which first require annotation.

Public Health Implications

While NLP and ML methods have been used extensively in clinical science to analyze free-text data,[18,19] they have been used less frequently for PHS despite unstructured data becoming more common.[20] NLP and ML methods have shown promise in studies to enhance surveillance activities, particularly when they focus on a specific condition opposed to a variety of conditions,[41,62,65,66] but their potential has been largely untapped. Further integrating NLP methods into surveillance workflows will advance surveillance science and public health practice, with several avenues for advancement identified through the review. While Ward et al[41] classified deaths as drug overdoses, an extension to this method that identifies the drug(s) that caused the death would further improve surveillance. Similarly, while detection of cancer-caused deaths had high scores,[66] there is still opportunity to further improve the classification of rare cancers.[62] New surveillance systems, such as emergency medical services data,[44,45] should be explored and NLP methods applied to extract information and classify cases.

There are also gaps in the literature on utilizing social media data for PHS. While influenza has been studied relatively extensively,[75,76,78,79] there is little literature on other infectious diseases. Additionally, given the substance use epidemic occurring in the United States, there is opportunity to use social media to track trends related to substance use and misuse. Sentiment analysis on tweets that mention specific drugs could create an "early warning system" to identify drugs that are rising in use before they appear in traditional PHS.

Overall, most of the literature reviewed utilized traditional ML models opposed to deep learning techniques while still achieving high scores. While many of the reviewed

articles discussed the development of a process to improve surveillance, some studies lacked key information on how these methods were developed. As public health researchers begin to use and develop ML models more extensively, care should be taken to ensure a thorough description of methods is present in articles, including a step-by-step breakdown of the model's development. For example, common missing details included how the model was tuned to select hyperparameter values (cross validation is common for traditional methods to avoid overfitting[82]) and clear indication of whether the training and testing data were split to ensure the model is validated on an out-of-sample dataset, which could lead to look-ahead bias if it did not occur.[83] Table 3 displays facets of ML algorithm development that should be included in future research. Another limitation that many studies had was sample size. ML models require relatively large amounts of data for accurate predictions,[84] especially deep learning methods. Finally, most studies required manual annotation, which can be a costly, time-consuming process. Contributing more resources to manual annotation on PHS data and social media data will be required for NLP and ML methods to reach their full potential as a surveillance tool.

Limitations

A limitation of the narrative review format compared to a structured systematic review is that it does not capture detailed information on the number of articles included and excluded due to various review criteria, nor does it follow a strict information extraction protocol. However, the narrative review format was appropriate for the purpose of providing a general overview of the current state of NLP and ML use in PHS and identifying future opportunities to use these methods in the field. Another limitation

of this review was its narrow search criteria; as a result, some studies that fit the overall purpose of this review may have been missed due to the intentionally narrow inclusion criteria. Additionally, this review did not examine other text analysis methods (such as regular expressions and key word searches) that are more frequently used in PHS for analyzing free-text data. A final limitation is that PubMed was the only database searched; most if not all articles applying NLP and ML to PHS are likely indexed in PubMed, but some candidates may have been missed. As NLP and ML become more widely used in PHS, a systematic review of this topic may become warranted.

Conclusions

While several studies have used NLP and ML methods for PHS, there are limitations and gaps that these methods can address. With the increasing size of data and unstructured data existing in public health, utilizing NLP and ML methods is essential for gaining insights. The overall promise that many studies in this review display indicate that NLP and ML have an important place in future PHS. These methods have already improved the timeliness of PHS, extracted additional information from data sources, and positioned social media as an emerging data source for monitoring disease trends in near-real time. The continued development and evaluation of these methods will be key to PHS moving forward.

Table 1: Articles applying NLP to traditional data sources

Study	Condition(s)	Description of method	ML Task	NLP Task	Key Findings
Butt et al (2013)	Cancer death	Classify cancer deaths from death certificate free-text for cancer registry reporting	Binary classification	Extracting stems, bigrams, and concept-based features from free-text	Traditional ML method was highly accurate for classifying cancer deaths
Koopman et al (2015)	Deaths caused by diabetes, HIV, pneumonia, influenza	Develop a ML model that predicts if a death is caused by these conditions using free-text	Multiclass classification	Extracting term and concept-based features from free-text	These relatively common conditions can have classification high accuracy with a single model
Koopman et al (2015)	Cancer death	Classify the ICD-10 codes of cancer deaths from free-text death certificates	Multiclass classification	Extracting term and concept-based features from free-text	Common cancers had high accuracy, rare cancers had low accuracy
Osborne et al (2016)	Cancer cases	Identify cancer cases that are mandatory reportable from pathology reports	Binary classification	Extract concept unique identifiers from pathology reports	High F-score highlights potential for automating reporting processes
Maenner et al (2016)	Autism spectrum disorder	Classify case reports to diagnose autism spectrum disorder	Binary classification	Extracting and stemming tokens, creating a term-frequency—inverse document frequency matrix	NLP can be used to classify autism in reports to estimate prevalence in the population
Metzger et al (2017)	Suicide attempts	Identify suicide attempts from emergency department EHR free-text	Binary classification	Extracting concept unique identifiers from free-text	Pilot study had high accuracy, authors seek to deploy as part of a national surveillance system

Duarte et al (2018)	All-causes of death	Develop a deep learning model that classifies the ICD-10 cause of death from free-text	Multiclass classification	Creating word embeddings for the neural network	Poor prediction scores for infrequently occurring ICD-10 codes; models for specific conditions are more accurate
Munkhdalai et al (2018)	Adverse drug events	Using named-entity-recognition and relation extraction, identify adverse drug events within EHR free-text	Information extraction, relationship extraction	Extract token distance between entities, creating word embeddings	Traditional model outperformed deep learning model, indicating that traditional ML can be used for PHS successfully
Koopman et al (2018)	Cancer death	Develop a hybrid rule-ML model for increased classification accuracy of rare cancers	Multiclass classification	Extracting term and concept-based features from free-text	Adding rule-based model to ML increases the accuracy of rare cancer classification
Leroy et al (2018)	Autism spectrum disorder	Extract diagnostic criteria for autism spectrum disorder from EHRs	Information extraction	Extract expressions of DSM criteria using a lexicon	Extracting specific criteria for autism is a more difficult task than classifying the disease
Ward et al (2019)	Drug overdose death	Classify deaths caused by drug overdoses using free-text on death certificates	Binary classification	Extract tokens, bigrams, and trigrams from free-text	Traditional ML methods can be utilized for accurate death certificate classification of drug overdose
Alawad et al (2020)	Cancer	Extract required case information that cancer registries require from pathology reports	Information extraction	Text pre-processing for use in several neural networks	Low accuracy scores overall, but study demonstrated promise for multi-task convolutional neural networks for imbalanced data

Table 2: Articles applying NLP to online media data sources

Study	Condition(s)	Description of method	ML Task	NLP Task	Key Findings
Broniatowski et al (2013)	Influenza	Two-staged approach to identify Tweets that mention influenza and then identify if the tweet is discussing an infected individual	Binary classification	Extract words, bigrams, trigrams from tweets	Trends in classified tweets followed trends in influenza-like illness nationally, with less correlation at the local level
Sarker et al (2016)	Prescription medication abuse	Given a tweet discussing a prescription drug, classify whether or not the tweet is discussing abuse of the drug	Binary classification	Stemming, extracting n-grams, part-of-speech tagging	Developed model has a relatively low F-score, but more state-of-the-art methods and additional training data could improve results
Mowery (2016)	Influenza	Determine if Twitter is a feasible source for influenza surveillance by analyzing if users misdiagnose themselves in tweets	Binary classification	Extract words and bigrams from tweets	Authors identified a spike in reported influenza on Twitter during a period without a real spike, suggesting caution should be taken when using social media for surveillance
Allen et al (2016)	Influenza	Classify tweets that discuss an individual infected with influenza and utilize geographic	Binary classification	Extract words, bigrams, and trigrams from tweets	Identified tweet trends were correlated with national and local data, but with less correlation at the local level

		information systems to normalize tweets			
Kagashe et al (2017)	Influenza	Develop a classifier that can track influenza drug consumption from tweets	Binary classification	Parsing tweets for dependency features to identify words with relations to an annotated drug	Study showed promise for tracking user drug consumption through tweets, which can be used as a proxy for influenza surveillance
Sarker et al (2017)	Pregnancy	Using tweets, classify if a Twitter user is pregnant for safety surveillance	Binary classification	Extract n-grams and create word embeddings for vector representation of words, extract tweet sentiments	Traditional ML method outperformed deep learning, providing evidence traditional ML can still be used for accurate classification
Klein et al (2018)	Birth defects	Given a tweet discussing a birth defect, classify whether or not the tweet is discussing a child with a birth defect	Binary classification	Filtering tweets using lexicon developed from Unified Medical Language System	High evaluation scores indicate incorporating rule-based methods can lead to successful classification for surveillance
Müller et al (2019)	Various	Using crowdsourced labeling of tweets, monitor changes in disease trends and beliefs through social media	Several including classification, sentiment analysis	Create word embeddings for tweets	An innovative approach, but using crowdsourced labels as opposed to train annotators may lead to errors
Feldman et al (2019)	Infectious diseases	Classify news articles that discuss infectious diseases to identify those that	Binary classification	Extract features from news articles for supervised ML	Model has potential for real time identification of outbreaks from media articles, but is limited by what conditions the news media is interested in

		discuss disease activity			
Jha, Singh (2019)	Substance use disorder	From Reddit posts and tweets, identify social media users to target for substance use recovery and interventions	Binary classification	Sentiment analysis, drug term identification	Model developed for Twitter data had high accuracy, indicating promise for using social media for this task

Table 3: Suggested key elements to be reported in publications reporting results of machine learning models applied in public health surveillance

Topic	Checklist items
Dataset description	✓ Time period of data collection ✓ Total sample size ✓ Training sample size ✓ Test sample size ✓ How train/test split was performed
Feature engineering	✓ Detailed description of how features included in the model were developed, including published code if privacy restrictions allow
Algorithms tested	✓ Machine learning models used ✓ Describe why these specific models were selected
Training process	✓ Describe how model hyperparameters were selected and any other tuning steps
Evaluation criteria	✓ List the metrics used for evaluation ✓ Describe why these are important for the application

CHAPTER 3

Enhancing timeliness of drug overdose mortality surveillance: a machine learning approach

Introduction

Death certificates (DCs) are the primary source for state and local drug overdose (OD) mortality surveillance and are currently the only nationwide source.[85] DCs provide information about decedents (including demographic information, residence, and place of death), cause and manner of death, and substance(s) involved in an OD that are important to developing drug OD prevention programs and policies.[86,87] In order to design and implement effective public health interventions, this information must be available to public health practitioners in a timely manner.

In the case of a suspected drug OD death, a coroner or a medical examiner serving the jurisdiction where the death occurred determines the cause-of-death and completes a DC.[88,89] The DC is then filed (electronically or as a paper copy) with the state office of vital statistics (OVS). An electronic record with selected DC fields, including the free text information for the cause-of-death,[90] is transmitted to the National Center for Health Statistics (NCHS) and coded according to the guidelines of the International Classification of Diseases, Tenth Revision (ICD-10) to allow standardized classification of the causes of death.[91-94] A copy of the ICD-10-coded record, containing one underlying cause-of-death (UCOD) and up to 20 supplementary causes of death, is sent back to the state OVS to be used for epidemiological analysis. There is a significant time lag between the day of death and the day when an ICD-10-coded DC record is available for identification of a drug OD death (the consensus definition for drug OD mortality surveillance is based on the UCOD code in the range X40-X44, X60-X64, X85, or Y10-

Y14[95,96]). Spencer et al. reported that only 37.8% of the drug OD death certificates are available to NCHS by 13 weeks (vs. 83.9% for overall deaths), mostly due to delays in DC completion related to required forensic toxicology analysis.[97] Additional time lag is acquired at the NCHS as about two-thirds of the deaths with an UCOD of drug OD are coded manually, compared to one-fifth of all-cause deaths.[7]

Motivation

Understanding the critical role of surveillance data to inform prevention and response to the opioid epidemic, the Centers for Disease Control and Prevention (CDC) provided dedicated funding to states to build capacity for more timely and comprehensive opioid OD surveillance data.[98] This paper examines the feasibility of using natural language processing (NLP) and machine-learning (ML) methods to identify potential OD deaths from free-text DC fields, allowing the identification of potential drug OD deaths (and the initiation of gathering of additional medicolegal data for these cases) before the DC records are sent to the NCHS for ICD-10 coding. Figure 1 displays the overall workflow of the death investigation and public health surveillance approach; the proposed method eliminates the time lag associated with steps 3a-7a, replacing these with steps 3b and 4b. We chose NLP and ML as the methodological base given they (a) provide an intuitive mapping from free text fields to categories using classification techniques and (b) are generally more accurate than rule-based systems in data rich settings. Our main goal is to build a practical computational solution that can be employed by epidemiologists in public health agencies for near real time OD mortality

surveillance with reasonably high accuracy. Hence, all the code used in this effort is made publicly available: https://github.com/pjward5656/dcnlp.

NLP and ML Background

Text is ubiquitous in healthcare and biomedicine coming from different sources including biomedical literature (journal articles, conference abstracts), clinical notes (e.g., discharge summaries and pathology reports), and social media text (e.g., Twitter, Reddit, and specialized forums such as the Cancer Survivors Network). Text classification methods from NLP and ML have been shown to play a critical role in health and biomedical applications especially when structured sources do not fully capture all the information necessary. The NLP component deals with extracting interesting "features" (independent variables) based on simple n-grams (typically words and two-to-three word phrases) and more involved syntactic constructs such as parts-of-speech for each word (e.g. noun, adjective) and constituency and dependency parse trees that represent inter-word grammatical relations within each sentence. The ML component then learns a model with these features as independent variables and the category as the outcome variable using, typically, hand-coded training data.

A basic text classification problem typically deals with an outcome variable that is **binary**. For example, based on a pathology report can we identify whether a cancer case is reportable or not?[52] There could also be cases where the problem is **multiclass**, where one of more than two categories ought to be chosen. For instance, in cancer registries, certified tumor registrars read pathology reports to code major sites from a list of dozens

of such sites and this task can be expedited using automated methods.[51] A variant of the multiclass problem is **ordinal** classification, where the categories are ordered in some manner specific to the task. Although one class is chosen in ordinal models, errors are counted differently based on how far away the prediction is from the true category. As an example, in psychiatry a recent shared application[53] dealt with assigning symptom severity categories (*absent*, *mild*, *moderate*, and *severe*) based on content in a psychiatric evaluation note. Finally, **multilabel** classification handles scenarios where more than one category is typically assigned to each input instance. A use-case is when coders assign multiple diagnosis codes[99] to electronic medical records for every patient visit. For elaborate details of specific applications of NLP in biomedicine, please refer to broad reviews.[100,101] The problem at hand in this current effort is binary classification to identify deaths due to OD based on DC text.

Methods

Data

The Kentucky Injury Prevention and Research Center (KIPRC), as bona fide agent of the Kentucky Department for Public Health, receives weekly extracts of DCs to perform injury surveillance. Coded DCs for years 2017-2018 as of November 1st, 2018 (n=84,142), were used for this analysis. The ML process requires that the data is split into training and testing sets; the coded 2017 DCs (n=48,016) were used as training data and the coded 2018 DCs (n=36,126) were used as testing data. In total, 2,478 (2.9%) of the DCs were coded as OD deaths based on the ICD-10 code assigned by NCHS in the UCOD field.[102] These cases were treated as the “true positive” cases when training and testing the ML algorithm.

Task

The task for the ML algorithm in this paper is to take an un-coded DC and classify it as an "OD death" or "not OD death" using free-text fields. To automate this task free-text fields on the DC were used to create features for a classifier. ML algorithms require feature vectors to train the model. In this analysis, a feature is a binary [0,1] variable created from an aspect of the free-text present on the DC, and a feature vector is *z*-dimensional vector of features where *z* is the total number of features.

Feature engineering

DCs in Kentucky are certified by county coroners or physicians depending on the manner of death. The certifier completes the cause-of-death section, consisting of an "immediate cause" field ("line a") followed by three sequential fields indicating conditions that were "leading to the cause listed in line a" [7]. In addition, the certifier may also complete two other sections: 1) "other significant conditions contributing (SCC) to death but not resulting in the underlying cause" and 2) "describe how the injury occurred." The latter is only completed if the death resulted from an injury.

These free-text fields (used by the NCHS to assign ICD-10 codes for underlying and contributing causes-of-death) were used to create features for the ML algorithm. Two different field combinations were examined for this task: 1) all three free-text sections of the death certificate and 2) the cause-of-death section and the description of injury section. The latter option was considered as ODs and substance use may not cause death but may contribute to other types of morbidity that later cause death. Excluding the SCC

section may result in better separation between OD deaths and deaths that involved substance use or previous ODs.

Certifiers in Kentucky typically do not write in full sentences using grammatically or syntactically correct language when completing the DC. The text is often concise. Table 4 shows an example of a typical OD DC. Due to these considerations, to simplify the task the fields in the training data were combined into one free-text field. All punctuation was removed from this field. Using the scikit-learn[103] library in Python, this field was tokenized into individual words, bigrams (adjacent two-word sequences), and trigrams (adjacent three-word sequences), excluding stop words ("the", "an", "and", etc.). All free-text in the data exists in fully capitalized form, and the capitalized tokens were used, unaltered, for the analysis. Any token appearing less than five times was discarded. The 2017 DC data including the SCC field contained 2,184 unique words and 11,261 bi/tri-grams. When excluding the SCC field, the word and bi/tri-gram list decreased to 1,820 and 8,029, respectively. Features for each of these words and phrases were then created for the model; a feature was given the value of 1 if the word/phrase it represents appeared in the text and a value of 0 otherwise. To illustrate, a feature representing the bigram "acute cocaine" would be 1 for the DC in Table 1 while a feature for the word "poisoning" would be 0.

Classifiers

Several classification methods were considered and examined for this task. Linear support vector machines (SVM), random forests (RF), and multilayer perceptrons (MLP)

were tested for this classification task. The SVM approach was selected due to both the low computational requirements of this method as well as its use in previous, similar classification tasks,[65-67] while RFs and MLPs were examined as more complex nonlinear methods that may identify additional interaction features compared to the linear SVM. Figure 2 shows the overall process used to predict if a DC is an OD using only the free-text fields. While this classification task was implemented in the Python environment, the algorithms tested have well documented and easily programmable methods within the e1071 and CARET[104] packages in R, making this an accessible method.

Training

The algorithms were trained on all coded 2017 DCs using 3 times repeated, stratified 10-fold cross validation within the scikit-learn Python library. Repeated cross validation was selected as it is recommended over other methods for general classification use.[82] Stratified cross validation was used so that each fold created during the procedure would have the same makeup of OD deaths and non-OD deaths that the entire dataset has (~97% non-OD, ~3% OD) [105]. The cost (regularization) hyperparameter of the linear SVM, the maximum depth, number of trees, and maximum number of features hyperparameters of the RF, and the hidden layer sizes and alpha (regularization) hyperparameters of the MLP were tuned based on F-score, the harmonic mean of positive predictive value (PPV) and sensitivity.[80] F-score was selected as the tuning metric due to the class imbalance in the data; tuning based on accuracy would bias the algorithm to correctly classifying non-OD cases, as ~97% of the data are not OD deaths. To tune the hyperparameters, potential values were initially selected.. After training on these values,

an additional search was performed around the value(s) that the previous step indicated was the best value(s) for the hyperparameter(s). This process was repeated subsequent times until the ideal hyperparameter value(s) was identified. After this, the entire training data was re-trained using the ideal value(s). This tuning approach was implemented as it is a straightforward grid-search method commonly used for tuning ML algorithms.[106]

Testing

After the learner was trained on the 2017 DC data it was deployed on the 2018 DC data. The same word/phrase features used in the 2017 DC data were created for the 2018 DC data. Both free-text field combination algorithms were tested.

Rule-based method

A final rule-based classification method was also tested for this analysis on the 2018 DC data. This rule-based method scanned the free text of the DCs for 37 words or bigrams that were indicative of an OD death. These words/bigrams were selected from a review of OD DCs by epidemiologists with a combined 10 years of experience in OD surveillance. For this review, 2017 OD deaths were examined and common tokens identified. These words and bigrams are available in Appendix 1. Any DC that contained one or more of the words/bigrams in this list were automatically classified as an OD death, and any death that did not contain a word/bigram in this list were classified as a non-OD death. This rule-based method is a simpler version of previously proposed text-matching methods,[107] and acts as a baseline to compare to the more computationally costly ML models.

Evaluation

Methods were evaluated based on their performance on the test data. To compare the results of the methods the F-score was calculated along with sensitivity and PPV. For this task PPV was considered the most important metric, as large numbers of false positives could be problematic when attempting to develop near real-time interventions for an OD outbreak. Two-proportion *z*-tests were performed to test for statistically significant differences between the sensitivity and PPV of the best performing ML model and rule-based method.

Results

The model including all three sections had a total of 13,445 features while the model excluding the SCC section had a total of 9,849 features. Computing time for training the ML models was not significant; even when using stratified 10-fold, 3 times repeated cross validation the longest any model took to train was roughly ~3 hours on a Windows machine with 32 GB of RAM. Deploying the models on the 2018 data took seconds. The rule-based method does not involve any training and produced predictions instantly.

Table 5 displays the F-score, sensitivity, and PPV from the final methods when training the models with features constructed from all three sections and Table 6 displays the results when training the models with features from the cause-of-death and description of injury sections. All of the models were highly specific. Only ~3% of all deaths in the data were OD deaths and all of the ML models had high performance in identifying deaths that were not OD deaths. The ML models without the SCC section,

however, all performed equally or higher than their counterpart ML models with the SCC section with regards to sensitivity, PPV, and F-score. The SVM model without the SCC section was the best performing model overall, with an F-score of 0.9695, and also achieved the highest PPV (0.9622), while the RF model without the SCC section achieved the highest sensitivity (0.9803). Table 7 displays the confusion matrix for the SVM excluding the SCC section while Table 8 displays the confusion matrix for the rule-based model excluding the SCC section.

The rule-based models did not perform as well as their counterpart ML models. The rule-based model excluding the SCC section had similar PPV to the ML models (0.9504), however this model had lower sensitivity than the ML models (0.9243). The rule-based model excluding the SCC section had a higher F-score than the rule based model including this section, similar to the results of the ML models. Comparing the best performing ML model (SVM, Table 6) to the best performing rule-based model (Rule-based, Table 6), two-proportion single-tailed *z*-tests show that the ML model has a significantly higher sensitivity ($p<0.001$) but no statistical difference for PPV ($p=0.13$).

Discussion

Model performance

This paper presents an accessible method to quickly identify OD deaths from free-text DCs for rapid surveillance purposes. The performance of the ML algorithms developed were very high. The F-score of 0.9695 for the best performing model is comparable, and in some cases superior, to that of other models for cause-of-death classification in the literature,[65-68,108] many of which were trained on larger datasets with more common causes of death. The ML models performed higher than their counterpart

rule-based models, including significantly higher results for sensitivity, providing evidence for further increasing the use of ML in public health surveillance over more traditional methods.

Of the methods compared, the ML models that excluded the SCC section produced the best performing models, as the ML models that included all three sections had more false positives than their counterpart models. This is likely because features that identify an OD death when appearing in the description of injury section or the cause-of-death section have different meaning in the SCC section. For example, the bigram "drug overdose" clearly indicates the death was caused by an OD when present in the cause-of-death section, but when present in the SCC section may indicate that the individual had a previous OD event, but it did not directly lead to death. Including the SCC section results in a feature space that contains information that is not directly involved in causing death, leading to a slightly biased model. Further research should examine including the SCC section in a classifier to identify drug-related deaths, which are important for drug-related surveillance.

Public health implications

DC data is the only national source for OD mortality surveillance[85] and as such is extremely important for understanding the opioid epidemic and developing responses to it. These data, however, comes at a non-trivial time lag[7,97] that prohibit availability of actionable data until several weeks following the event of an OD death. The method proposed in the present study eliminates the time lag from transferring DCs between the state OVS and the NCHS and the time for ICD-10 coding by the NCHS, thus allowing

state or local jurisdictions to process their own mortality data as soon as the DC is available in their database with free-text cause-of-death information. This provides more timely provisional counts for OD deaths, allowing for detection of overdose death spikes and identification of new patterns in contributing drugs in a shorter time window, which could be key as novel designer drugs emerge.[3,8,109] This could lead to faster mobilization of community stakeholders to implement harm reduction strategies, such as targeted naloxone distribution to communities.

An additional application of the ML method developed is its potential use as a data quality instrument. While many OD deaths are manually coded at NCHS,[7] there is still the potential for coding errors to occur. Records that the classifier identifies as OD deaths but NCHS codes as non-OD deaths (false positives) can be reviewed by a medical examiner to determine if the case is truly a false positive or if the UCOD was coded incorrectly. This data quality process will capture OD deaths that previously were not identified and lead to more accurate, complete surveillance of the OD epidemic.

Future directions

To further improve this model, additional feature engineering could be explored. This could include adding regular expression features to the model to identify patterns that frequently appear on OD DCs. Additionally, more n-gram features (such as non-adjacent bigrams) could be added to the model. Features that do not arise from NLP that use other fields on the DC (demographics, manner of death, place of death, etc.) could also be exploited in a future classifier. Part-of-speech tagging and dependency parsing, which creates a tree-like structure that explains the grammatical relationships between

words in sentences,[110] was considered for generating additional features for this analysis. Likely because the free-text present on DCs is often not grammatically correct, these features did not significantly increase the predictive power of the models. To build a simpler model they were excluded from the final analysis.

Deep neural networks, specifically convolutional neural networks, were considered for use in this task as they have been used for classifying free-text DCs.[68,108] Due to the small amount of text present on DCs (some contain only two words) a traditional classification method was selected in our effort. Future research should examine deep learning methods (particularly the use of pre-trained word embeddings) and determine if they have higher performance for this task. In general, however, improving performance scores that are already in the high nineties (our F-score was 0.9695) will be a challenging endeavor that is worth further exploration. Deep learning should also be explored in the future for more complicated DC classification tasks, particularly multilabel classification of the substances that caused a drug OD death. An OD DC may contain no information about the substances causing the drug OD or list several substances involved in the death. A deep learning framework that can classify drug OD deaths and the substances causing the death from DC free-text would be a noteworthy extension to the methods developed in the present study. Since the completion of postmortem toxicology alone can delay the completion of the DC for weeks, the next step for improved timeliness in OD death surveillance is to expand the proposed machine learning models to work directly on medicolegal death investigation data, utilizing unstructured data from coroner and medical examiner case management systems (e.g., death scene investigation notes, autopsy reports, police reports, coroner

notes) thus identifying likely drug overdose cases before the DC is officially filed. As these reports typically contain much more free-text than DCs, deep learning models should be explored for this process as well.

Strengths and limitations

The present study has several strengths. A thorough literature review indicates that this is the first study of its kind to our knowledge to describe a method for classifying OD deaths from free-text DCs using NLP and ML. The design of this experiment mimics how the model would be applied in a real-life use case scenario, with features engineered from previous year(s) data exploited to classify new data, giving the metrics on the test data validity for future use. This setup also adds to the difficulty of the task, as new substances that may be important in drug OD classification in a future year may not be present in the past year that the algorithm was trained on. Despite this, the ML models excluding the SCC field all achieved F-scores above 0.96.

Another strength of this study is the availability of data for our models to be tested on, further research to be performed, and similar algorithms to be developed. A major limitation of most ML applications is the large amounts of training data needed which coincides with the long, tedious process of labeling training data. Previous years coded DCs, however, are readily available for training at most state's health departments. These DCs are ready to be used for training ML models without the typical initial requirement of labeling. This labeled data, while not publically available, can be requested from states' OVS for research purposes. In addition, the NCHS operates the Research Data Center (RDC) to allow researchers access to restricted-use data. In 2019,

the NCHS made available to researchers a Redacted Death Certificate Literal Text File (LTF)[111] that includes the cause-of-death text record for every U.S. resident death. The access to LTF will allow researchers to test our GitHub source code on DC records that come from all U.S. jurisdictions.

The study has some limitations. First, the models were trained on only Kentucky DCs, meaning that the model may not perform as well on data from other jurisdictions if certifiers use different words or phrases on their DCs. Kentucky, however, has a hybrid-coroner/medical examiner system with 120 county coroners and deputy coroners who certify deaths (per KRS 72.025) along with medical examiners who assist coroners in determining the cause and manner of deaths.[112] Physicians also certify natural deaths. Therefore, there is a diverse group of individuals certifying deaths in Kentucky, meaning the language on Kentucky DCs represents a range of medical backgrounds. Another limitation, inherent in ML, is the difficulty in diagnosing errors. The models described here have features numbering near the 10,000s—determining exactly what features are causing the model to incorrectly predict is inherently more difficult than for a simpler model with a small number of variables. A detailed error analysis of the SVM model excluding the SCC field is available in Appendix 2.

Conclusion

The present study compares three methods for identifying OD deaths from free-text DCs using NLP and several ML models as well as a simple rule-based method. Classifying OD deaths using free-text would substantially reduce the time it currently takes a surveillance systems to identify OD deaths, from several months to a few weeks. The described ML methods performed better than the rule-based methods on testing data,

providing evidence that ML methods should be implemented in public health surveillance tasks, particularly for OD mortality surveillance. The programming code used to develop the model is publically available, which can facilitate further testing and development in other jurisdictions. Further research is needed to explore the potential for other causes of death to be classified using ML methods as well as additional exploration of ML, including deep learning, to improve drug-related surveillance

Figure 1: Certification, registration, and analysis of drug overdose deaths

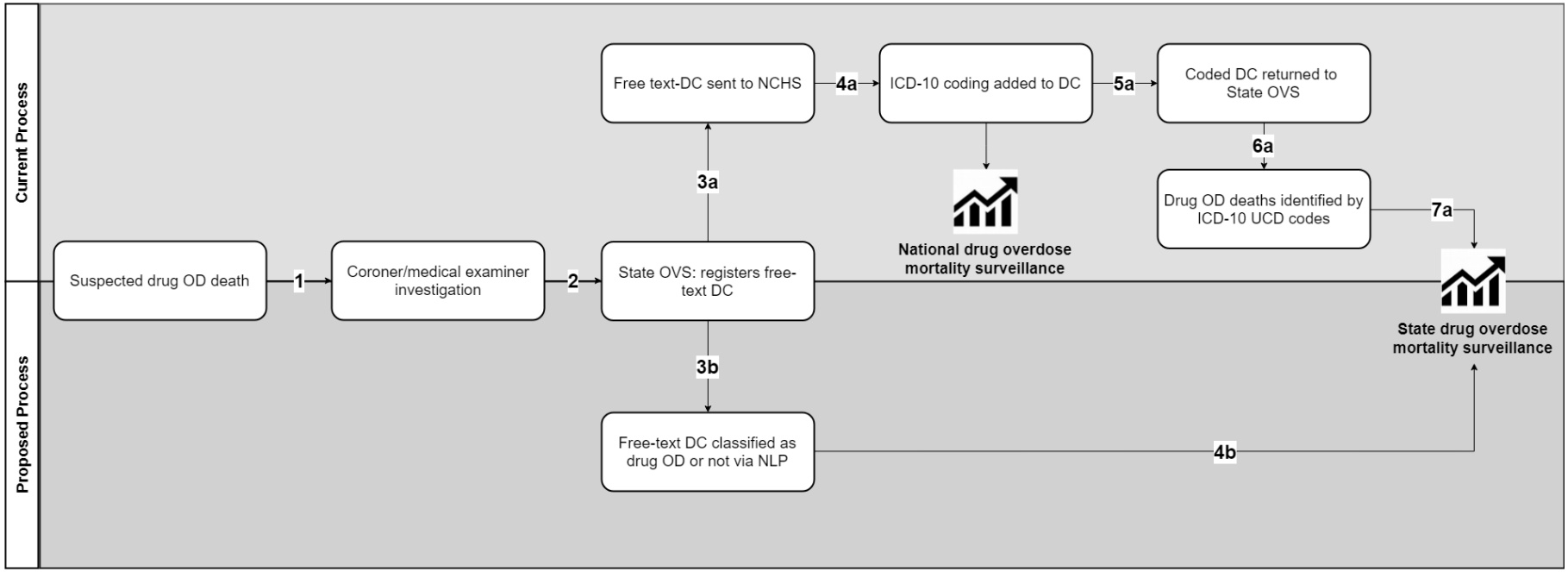

Figure 2: Example Analytic Pipeline

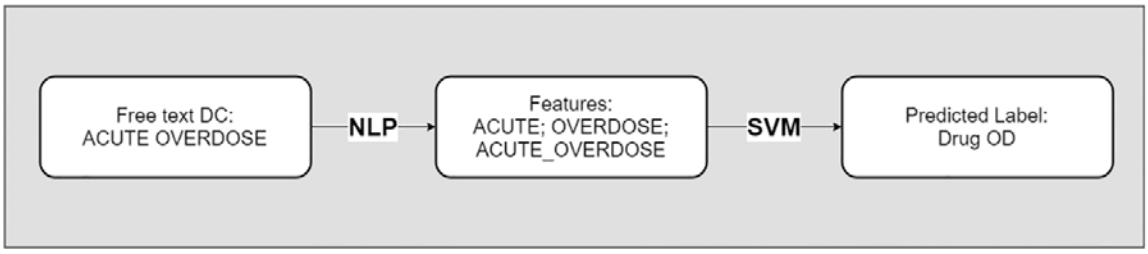

Table 4: Example Death Certificate Free Text

Field	Text
Immediate cause-of-death	ACUTE COCAINE TOXICITY
Due to (or as a consequence of)	
Due to (or as a consequence of)	
Due to (or as a consequence of)	
Significant conditions contributing	HYPERTENSIVE CARDIOVASCULAR DISEASE, OBESITY
Description of injury	ACCIDENTAL OVERDOSE

Table 5: Final model results on test data, features from all death certificate sections

Method	PPV	Sensitivity	F-score
SVM	0.9549	0.9748	0.9647
RF	0.9328	0.9748	0.9533
MLP	0.9518	0.9737	0.9626
Rule-based	0.9215	0.9265	0.9240

Table 6: Final model results on test data, features from cause-of-death and description of injury fields

Method	PPV	Sensitivity	F-score
SVM	0.9622	0.9770	0.9695
RF	0.9531	0.9803	0.9665
MLP	0.9621	0.9737	0.9678
Rule-based	0.9504	0.9243	0.9372

Table 7: Confusion matrix, SVM model, features from cause-of-death and description of injury fields

		UCOD Label		
		Drug OD	Not Drug OD	Totals
Predicted Label	Drug OD	891	35	926
	Not Drug OD	21	35,179	35,200
	Totals	912	35,214	36,126

Table 8: Confusion matrix, rule-base model, features from cause-of-death and description of injury fields

		UCOD Label		
		Drug OD	Not Drug OD	Totals
Predicted Label	Drug OD	843	44	887
	Not Drug OD	69	35,170	35,239
	Totals	912	35,214	36,126

CHAPTER 4

Deep neural networks for fine-grained surveillance of overdose mortality

Introduction

Death certificates (DCs) are the primary data source for drug overdose (OD) mortality surveillance, providing information about the cause- and manner-of-death.[2,5] Medical certifiers write the cause-of-death in free-text fields on the DC,[7,113] which are then used to assign International Classification of Diseases, 10th Revision[16] (ICD-10) codes by the National Center for Health Statistics (NCHS) to each death certificate. ICD-10 codes are used at the national and local levels to calculate OD mortality statistics, to monitor trends in the drugs involved in overdoses, and for epidemiological analyses. Specifically, ICD-10 codes assigned as an underlying cause-of-death are used to identify drug OD deaths and those assigned as supplemental cause-of-death codes (up to 20) are used to identify the drug(s) involved in the OD.[2,7,102] This process is central to the surveillance of drug OD mortality and is the primary way in which OD mortality information is reported to communities.

The ICD-10 classification process of drug overdose deaths (also called drug poisoning in the ICD-10 terminology) has limitations, however, including timeliness of data,[7,15] and the lack of fine-grained granularity in the coding structure of ICD-10 for identification of specific drugs contributing to an overdose death.[114] Some drugs have specific codes (e.g., T40.1 indicates heroin involvement, T40.3 indicates methadone involvement), but other drugs of interest do not. For example, ICD-10 code T43.6 specifies poisoning by "psychostimulants with abuse potential[10]", which includes illicit substances such as methamphetamine and MDMA in addition to licit substances such as

prescription amphetamine but excludes the psychostimulant cocaine, which has its own ICD-10 code (T40.5).[10] ICD-10 code T40.4 specifies "synthetic narcotics[5]", which includes fentanyl (prescribed or illicitly manufactured) and fentanyl analogs that have been a major cause of drug OD deaths in the United States,[4,8,9] but also prescription drugs like tramadol.

This inexhaustive coding structure of ICD-10 has led epidemiologists at the national and local levels to rely on text analyses to identify specific drugs involved in OD deaths when the codes present are not indicative of a specific drug. In 2015, the Council of State and Territorial Epidemiologists (CSTE) released a tool to identify specific drug mentions from the free-text cause-of-death section of DCs.[115] This tool consisted of a SAS program[115] that looped through individual words in the cause-of-death section and matched the words to a lookup table containing search terms (dug names, metabolites, and common misspellings) and a crosswalk to a referent drug. The tool produced a list of drugs found in each record. In 2016, the NCHS developed a methodology for identifying specific drug involvement from the cause-of-death section that again relied on a lookup-table.[7] This methodology additionally considered contextual information and had a more extensive table than the CSTE tool. While these tools presented improvements over using ICD-10 codes alone, they were limited by their respective lookup tables, requiring frequent updates with novel drugs. Additionally, maintaining every possible misspelling or metabolite for substances is a difficult, ongoing, and resource-intensive task.

Modern approaches within the field of natural language processing (NLP), including named entity recognition (NER),[116] could be applied in the area of drug overdose surveillance to 1) expedite the process for identifying potential drug overdose

deaths before the ICD-10 coding of the death certificates is completed; and 2) improve the identification of specific drug involvement by identifying drug search terms that are missing in the lookup tables.

NER involves tagging particular words as "named entities" (e.g., relevant to overdoses, tagging substance names, misspellings, metabolites, and generics as "drug entities") and then training a machine learning (ML) algorithm to identify these entities from free-text. This allows for the trained algorithm to predict what word(s) on new DC records are drug entities.[22,23,117]

Modern NER techniques leverage deep learning[21,117-119] methods that involve artificial neural networks with several hidden layers. This allows for the model to automatically learn complex and robust features that are predictive of outcomes based on textual inputs.[120] Manual feature engineering that was the hallmark of traditional NLP methods gave way to dense neural representations that have been shown to be more powerful in information extraction applications, including our current task. An advantage of this approach is that NER does not rely on a lookup table to determine which word(s) are indicative of a substance, rather, the ML algorithm through training learns which words (and surrounding contexts) are indicative of substances. NER has been extensively used in clinical science to identify substances in free-text,[21-23,117,121] but has not been utilized for DCs to our knowledge.

The present study sought to advance the science and practice of drug OD mortality surveillance through the development of a modern NER tool for identifying substances on drug OD DCs. Integrating NER methods into the drug OD mortality surveillance pipeline will advance surveillance science,[24] and in turn enhance public

health practice by improving the granularity and timeliness of OD mortality surveillance. This study demonstrates how the adaptation of the NER methodology enhances the current public health surveillance practice of identification of specific drug involvement in ODs.

Methods

Data

Data for this study was extracted from the Kentucky Death Certificate Database, Kentucky Office of Vital Statistics, Department of Public Health on March 9th, 2020. All records coded with the consensus definition for drug OD death[2] for years 2014-2019 were identified and pulled from the database, resulting in a total of n=8,146 OD DC records used to develop the NER drug identification tool. Using R,[122] the free-text data was parsed and reshaped so that every row was one individual element of the free-text from the cause-of-death section. To demonstrate, a DC that said "HEROIN, FENTANYL, AND OXYCODONE OVERDOSE" was transformed into 7 rows (one for each token), corresponding to the elements "HEROIN", ",", "FENTANYL", ",", "AND", "OXYCODONE", "OVERDOSE". This process resulted in a dataset with 95,566 total tokens. The text on the death certificates is fully capitalized, and these tokens were used, unaltered.

Annotation

To tag the individual words as drug entities, the dataset was split in half. Two trained annotators labeled each split according to the beginning, inside, last, outside, unit (BILOU) tagging scheme.[123] In the BILOU scheme, a single word that represents an

entity is tagged as "U-*entity*" (unit). An entity that spans numerous tokens is tagged with "B-*entity*" (beginning), "I-*entity*" (inside), and "L-*entity*" (last), for the first token in the entity, any tokens between the first and last tokens in the entity, and the last token in the entity, respectively. Any words that do not represent an entity of interest are tagged with "O" (outside). Table 9 shows how a record annotated with the BILOU scheme looks for drug entities. Kappa statistics were calculated for each split and for the entire dataset. Any disagreements between annotators were identified and resolved at meetings of all four annotators to produce the final, annotated dataset.

NER Drug Identification Tool and Modeling strategy

Development of the NER tool utilized the Flair[124] library in Python. Flair is a flexible, powerful NLP library that provides a suite of word embeddings[125] in addition to an easy to navigate modeling framework for programming deep neural networks for NLP tasks, including NER. Word embeddings are semantic vector-space representations of words. In a typical embedding setting, words that are closely related have similar vector representations, which provides a more desirable quality for word-representation in NLP tasks compared to simple dictionary-key representations. The Flair library provides both previously developed embeddings (such as GloVe[125]) as well as Flair embeddings,[124] which were developed using a character-level recurrent neural model that is contextualized by surrounding text, meaning an individual word can have multiple embeddings depending on the context. More recently, Flair embeddings were extended to "pooled Flair embeddings,[126]" which addresses the limitation of producing useful embeddings for rarely used words.

In general, development of a model in the Flair framework follows the following steps: 1) labeled corpus creation, consisting of creating a train-test-validation split and pre-processing data to fit into Flair's framework; 2) choosing word embeddings; 3) selecting a model; 4) training the chosen model using the selected embeddings. For the present study, two models were developed using this framework, only differing in step 2. The n=8,146 DCs were split into training (n=6,108), validation (n=816), and test (n=1,222 DCs) sets, roughly a 75%-10%-15% split, respectively. This split was performed chronologically; DC records were sorted by date of death, and the first 75% were assigned to the training data, the next 10% were assigned to the validation data, and the final 15% were assigned to the test data. For word embeddings, a model with GloVe embeddings was developed as well as a model using forwards- and backwards-trained pooled Flair embeddings. Hypothetically, the contextual Flair embeddings should out-perform GloVe. Testing the pooled embeddings was of interest as DC text contains frequent misspellings as well as rare words such as metabolites of drugs and novel substances. The model development pipeline is displayed in Figure 3.

The modeling framework selected was a bidirectional long-short-term memory (BiLSTM) conditional random field (CRF) model. BiLSTM-CRF was selected as it achieved state-of-the-art performance in NER tasks previously.[119] Flair has a BiLSTM-CRF model built into the library, which utilizes the PyTorch framework.[127] For training, the model was set to run for 150 epochs. After each epoch, the model was tested on the validation set and the model's accuracy was calculated. The highest-scoring model on the validation set was saved during training and deployed on the test data. Documented code for this process is available in Appendix 3. Positive predictive value (PPV), sensitivity,

and F-score (harmonic mean of PPV and sensitivity)[80] were calculated for each models' performance on the test set for evaluation. High PPV indicates that there are few false positives; a PPV of 90% means that 9 out of every 10 entities the model identifies as drug entities are truly drug entities. In this example, the false positive rate is 10%. High sensitivity indicates that there are few false negatives; a sensitivity of 90% means that 9 out of every 10 tokens that are truly drug entities are correctly identified as drug entities. In this example, the false negative rate is 10%. F-score provides one score to directly compare competing models. To demonstrate the advantage of the deep learning approach over traditional machine learning methods, a Naïve Bayes model was also developed using a dictionary-key representation for words and evaluated on the test data for a baseline comparison.

Comparison to lookup table approach

A widely used methodology for identification of specific drug involvement is based on literal text search for drug names, metabolites, and misspellings, cross-walked to a "referent drug" and included in a lookup table of search terms. The methodology was described and implemented in a CSTE tool[115] in 2015. The Kentucky Drug Overdose Fatality Surveillance System[32] has been updating the initial CSTE dictionary, and the current table includes more than 250 referent drugs (Appendix 4). To determine if the NER model provides an improvement over current methods, the lookup table approach was used to identify drug entities on the test data. Unlike the neural NER model developed, the lookup table approach has no false positives (as any word identified from the lookup table is already confirmed to be a drug). The number of entities that the best

performing BiLSTM-CRF identified that the lookup table did not was calculated to display how adding entities identified from the NER model could improve existing dictionaries.

Results

Annotation

The dataset had a total of n=95,566 tokens that required annotation and each token needed to be assigned a BILOU tag as discussed in the Methods. The kappa statistic between the first two annotators (first half of the dataset) was 0.996 and between the second two annotators (second half of the dataset) 0.973. Overall, the entire dataset after annotation had a kappa statistic of 0.983. The kappa values indicate the annotation task resulted in "perfect" inter-rater agreement level as per suggested rule of thumb[129] making it a very high-quality dataset.

Modeling

Table 10 displays the results from the respective BiLSTM-CRF models with the basic GloVe embeddings and Flair embeddings as well as the results from the Naïve Bayes model. While both deep learning models achieved considerably high performance for this task, the model utilizing pooled Flair embeddings performed better, with both fewer false positives and false negatives, resulting in higher scores for Flair vs GloVe for PPV (99.16% vs 98.63%), sensitivity (99.10% vs 98.08%) and F-score (99.13% vs 98.35%). Both deep learning models achieved substantially higher scores than the baseline Naïve Bayes model (F-score=15.58%). Figure 4 displays the validation loss during training for the pooled Flair model.

The BiLSTM-CRF model using pooled Flair embeddings was able to recognize drug entities that the lookup table was not. In total, the deep learning approach identified 168 more entities than the lookup table did. Appendix 5 displays the 130 unique drug search terms that the lookup table did not contain that were present in the test data. Many of these entities were terms that spanned multiple tokens, such as “DESIGNER OPIOIDS”. While the lookup table had the term “OPIOIDS”, the deep learning method identified the entity “DESIGNER OPIOIDS”, which provides more information on the specific drug than the term “OPIOIDS” alone. Other missed entities included misspellings, such as “OXMORPHONE”, “ALPRAZOLM”, and “METHAMPHETTAMINE”, which all have the correctly spelled terms in the lookup table but not these specific misspellings. Table 11 displays 3 example death certificate records and shows the entities identified by each approach. By contrast, there 13 instances where the opposite occurred, in which the lookup table identified a drug entity when the NER model did not.

Discussion

Model Performance

This study presents a highly accurate method for identifying drug entities on free-text drug OD DCs utilizing a modern NLP model. The F-score of 99.13% for identifying drug entities achieved in the present study show significant promise for using deep learning methods in public health surveillance. The BiLSTM-CRF model leveraging pooled Flair embeddings also identified substantially more drug entities than the current lookup table for identifying drugs on free-text DCs, clearly displaying that this method is

an improvement over currently available surveillance tools. In total, the best deep learning approach identified 130 new unique drug entities compared to the lookup table.

During training, the model achieved lower (better) loss on the validation data than the training data, as displayed in Figure 4. This is an unexpected result, as typically models will have higher scores on the in-sample training data than the out-of-sample validation set. Potential reasons for this include the relatively small sample size of the validation set (n=816) compared to the training set (n=6,108). Additionally, it is possible that the training data contained more difficult examples for the model to categorize than the validation data. A practical reason for this is how the data was split; the training data includes the oldest (chronologically) DC records. In Kentucky, public health initiatives have worked with medical certifiers to improve DC completion and accuracy for drug overdose deaths, so DCs completed earlier in the epidemic may be of lower quality and thus more difficult for the model to learn than more recent DCs.

Implications for public health practice

While knowledge of deep learning methods and NER is not typically part of a public health practitioner's toolkit, epidemiologists working in OD surveillance possess knowledge of programming and text analytic tools as they have become required for identifying drugs that cause morbidity and mortality in communities.[7] The high performance achieved by the developed model shows that deep learning should be leveraged moving forward in public health practice as a tool to solve challenges that appear in free-text data. Epidemiologists working at the state, local, and national levels should develop and implement these methods into surveillance pipelines to both improve

public health practice and improve surveillance data quality. Neural NLP methods will become essential as public health surveillance continues using more data sources that contain free-text, including electronic health records,[40,49,61,64] emergency medical services run data,[44,45,130,131] and syndromic surveillance,[27,46,48,132] which are used for both drug ODs and other health conditions.

As many jurisdictions performing OD mortality surveillance will not have the expertise nor the computational power to develop and test NLP models, the developed model can advance surveillance efforts through improvement of the current surveillance pipeline. In analysis of Kentucky data alone, the deep learning model identified over 100 unique drug entities that a lookup table did not contain. Identified entities can be extracted from the model's results and added to the lookup tables of current tools, which can be disseminated to jurisdictions to improve identification of specific drugs. This will allow for the increased specificity of drug entities that the model provides without needing to run a complex model in situations where computational power or expertise is limited. With periodic runs on a fresh set of DCs, our model can thus surface new drug terms and improve operations in jurisdictions that do not necessarily have the resources to train and deploy neural models.

Integration of additional entities recognized by the model into existing tools in the drug OD mortality surveillance workflow will increase the specificity of drugs identified on OD DCs through fine-grained spotting, compared to both ICD-10 coding and other methods.[7] Importantly, ongoing application of the NLP approach as part of the routine drug OD surveillance analysis will allow for the detection of novel substances as soon as they appear on a death certificate, before the lookup tables are updated with the new

substance names. Addition of a new substance to a lookup table depends on circumstances within a rapidly changing drug market, which can be demanding, as evidenced by the difference in performance between the two lookup table methods tested. Since the developed algorithm can utilize context, it has the ability to recognize novel drug entities without constant modification of the underlying model. The model's ability to recognize novel entities will provide the opportunity for early warning signals for novel substances, and thus faster public health and public safety response. Additionally, the increased granularity will provide an overall improvement in surveillance data quality, which will lead to more accurate reporting of information to communities and stakeholders. The proposed neural NER methods expand our previous work on DC free-text[41] where we developed a NLP algorithm to capture drug OD death cases prior to ICD-10 coding. In combination, the two NLP algorithms can be added to the routine drug OD mortality surveillance tools to improve the early identification of drug OD deaths and emerging new drugs of concern, providing opportunities for timelier public health and safety response.

Future research

Future studies should apply the developed model to data from other jurisdictions and assess the model's performance to explore generalizability of the model. This generalizability would indicate that the model can be directly integrated into the surveillance work of other states and localities without the need for annotation and training. Additionally, future research should utilize a similar workflow for identifying substances on other surveillance data sources. While it is unlikely that the developed

model, due to the nature of DC free-text, would achieve high scores on other free-text OD surveillance sources such as emergency medical services run data, future studies should explore the use of the Flair library and particularly pooled Flair embeddings for drug NER tasks. One limitation of the lookup table approach for identification of drugs involved in fatal ODs is the lack of context analysis. For example, if "history of heroin abuse" was mentioned on a drug OD death certificate, the lookup table approach as well as our currently proposed NLP algorithm would identify the OD death as heroin-involved. Since NLP methods can be trained to recognize context, a future improvement of our NLP algorithm would be the filtering out of the drug entities in situations where the drugs were not mentioned as contributory to the OD death. This specific scenario did not occur frequently in Kentucky data used to develop the model, but will be an important improvement to avoid false positive cases for drug OD involvement of specific drugs.

Strengths and limitations

The developed method has several strengths. First, the developed dataset consisted of a large sample (n=8,146) of OD DCs spanning multiple years of data. The dataset was annotated by 4 trained annotators, ensuring accurate labels were produced which was verified by a high overall kappa statistic (0.983). Another strength of the method was the way in which the train-validation-test split was performed. By using earlier records for training and validation, and testing on later DCs, the model demonstrated that it can use older data to produce accurate predictions on new data, mimicking the real-world use-case of the model. The final strength of the model is the

high F-score (99.13%) achieved, displaying high accuracy for identifying drug entities on free-text DCs.

The study does have a few limitations. A primary limitation is that the developed NER model cannot be used as a stand-alone surveillance tool; rather, it should be implemented as an enhancement in current OD mortality surveillance work to improve drug identification on death certificates. The lookup table-based methods contain crosswalks of drug search terms to their parent drug (for example, misspelling "CLONAZPAM" is cross-walked to parent drug "CLONAZEPAM"). Since the NER method, by design, recognizes novel entities, these entities are not present in dictionaries and their respective crosswalks, and therefore the NER model cannot be used as a surveillance tool alone. This limitation is addressed, however, by utilizing the model periodically on new data to recognize novel entities and adding these entities to existing lookup tables, so that novel entities can be identified by surveillance tools. Further, the inclusion of an entity in a lookup table guarantees that it will be recognized on every record it appears in, regardless of context—this addresses the rarely occurring scenario when a drug entity was not identified by the NER model due to it appearing in different context than the model learned.

An additional limitation is that the entirety of the data used for training the model came from Kentucky DCs. The performance of the model on data from other jurisdictions should be evaluated. Additionally, the comparison of model performance to performance of the lookup table method is conditional on how up-to-date a given lookup table is; other jurisdictions may have more complete tables, so the performance improvement from the NER model may not be as high as in the present analysis. Another limitation is the

complexity of the method and the need for computationally powerful hardware when applied to large datasets. Finally, the complexity of the model makes diagnosing errors difficult, which is an inherent limitation in most ML applications.

Conclusion

To our knowledge, this study is the first of its kind to use deep neural networks for drug NER on DCs. The highest performing model developed achieved an F-score of 99.13%, indicating that the method is highly accurate at this task. The high performance of the developed model clearly shows that deep learning models should be integrated into public health surveillance workflows. Particularly for drug OD mortality surveillance, the method could improve surveillance data quality and timeliness, enabling public health practitioners to more quickly recognize novel substances and more accurately report data to communities. The developed method advances the science of public health surveillance by integrating NLP models not currently used in the field into surveillance workflows and advance public health practice through enhancing both data quality and timeliness of reporting. These surveillance improvements are key in monitoring the continuing drug OD epidemic and informing interventions to address this national crisis.

Table 9: Example Annotated Death Certificate Free-text

Word	Tag
7	B-drug
-	I-drug
AMINOCLONAZEPAM	L-drug
AND	O
HEROIN	U-drug
OVERDOSE	O

Figure 3: Flair NLP model development and evaluation for overdose mortality surveillance

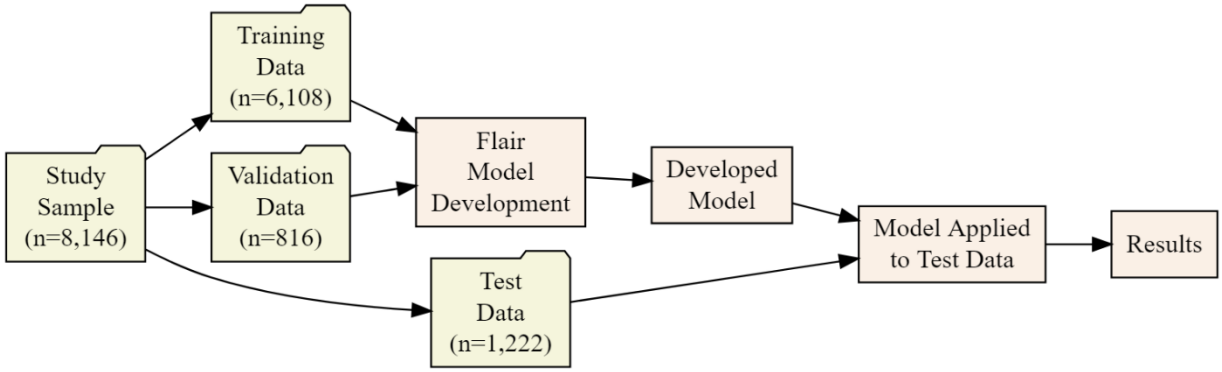

Table 10: Performance of machine learning models on test set

Method Tested	True Positives	False positives	False negatives	PPV	Sensitivity	F-score
GloVe	3168	44	62	98.63%	98.08%	98.35%
Pooled Flair	3201	27	29	99.16%	99.10%	99.13%
Naïve Bayes	383	1686	2847	18.51%	13.45%	15.58%

Figure 4: Validation and Training Loss, Pooled Flair Model

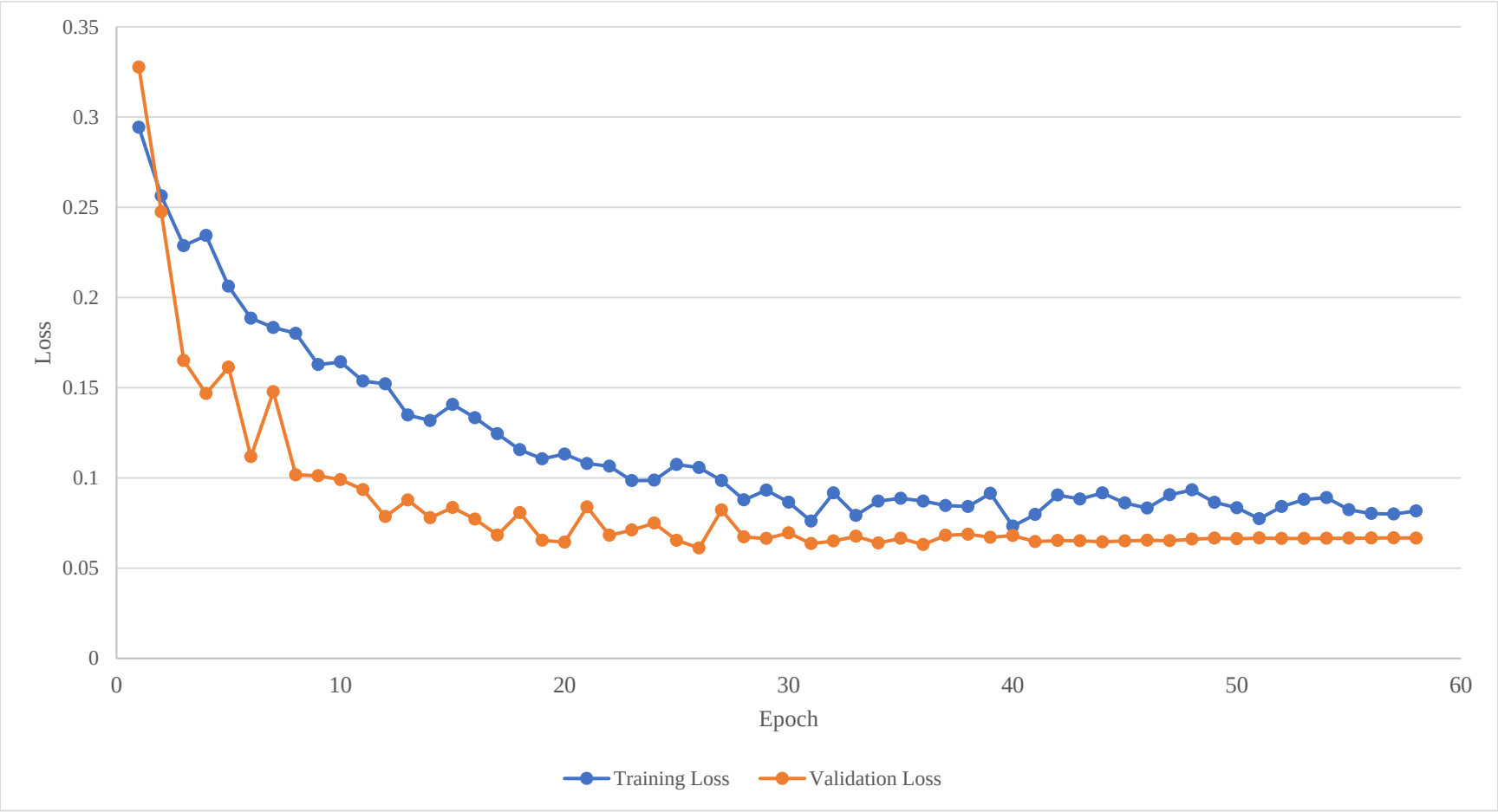

Table 11: Example Classified Death Certificate Records

Death certificate tokens[1]	Recognized by the lookup table	Recognized by NER model[2]
ACUTE INTOXICATION BY THE COMBINED EFFECTS OF FENTANYL , ACETYLFENTANYL , METHAMPHETTAMINE , TRAMADOL , AND GABAPENTIN SUBSTANCE ABUSE	✓	✓
MULTIPLE DRUG INTOXICATION (METHAMPHETIMINE , CLONAZAPAM , OXYCODONE , GABAPENTIN) MULTIPLE DRUG INTOXICATION	✗	✓
ACUTE COMBINED TOXIC EFFECTS OF		

[METHAMPHETAMINE , FENTANYL , HEROIN , GABAPENTIN , & PARAFLUOROBUTYRYLFENTANYL] SELF ADMINISTRATION OF ILLICIT DRUGS	✓ ✓ ✓ ✓ ✗	✓ ✓ ✓ ✓ ✓

1. Highlighted tokens are drug entities; a horizontal line indicates the start of a new death certificate record.

CHAPTER 5

Discussion and Conclusions

This dissertation sought to advance surveillance science and public health practice through examining and testing the integration of natural language processing (NLP) and machine learning (ML) methods into the drug overdose mortality surveillance workflow. A thorough review of the literature found that NLP and ML methods, particularly deep learning models, are used scarcely in public health surveillance. However, evidence from clinical sciences[18,19] indicates that there is untapped potential for NLP and ML methods to enhance public health surveillance activities. Applying NLP methods into surveillance advances science through the tenets of integration and development[24], by leveraging techniques used primarily in other disciplines to improve understanding and stimulating additional research in the field, respectively. The dissertation focused on advancing surveillance practice in drug overdose mortality surveillance in particular as overdose mortality is one of the most pressing public health crises the United States has faced in recent history,[4] with the age-adjusted drug overdose death rate increasing from 20.7 per 100,000 people in 2018 to 21.6 per 100,000 people in 2019[1]. This dissertation presents the development and application of two NLP models that improve drug overdose mortality surveillance timeliness and data quality.

Enhancing timeliness

A primary limitation of drug overdose mortality surveillance is the timeliness of the data.[97] Drug overdose mortality surveillance relies on death certificates, which become available to state and local epidemiologists after a substantial time lag attributable to the death investigation and the process of death certificate coding at the

National Center for Health Statistics (NCHS). Delays in obtaining data limit the ability of practitioners to disseminate surveillance data to stakeholders as well as prohibiting timely responses to increases in drug overdose mortality. To improve drug overdose mortality data timeliness, Chapters 3 and 4 developed classification and named entity recognition (NER) models, respectively, that can analyze free-text death certificate data. The developed methods eliminate the time required to transfer death certificates between the state Office of Vital Statistics (OVS) and the NCHS and the time for NCHS to add International Classification of Diseases, 10th Edition[16] (ICD-10) codes to the death certificate. Utilization of the developed models in tandem to enhance currently existing surveillance tools allows for full-scale drug overdose mortality surveillance—both case identification and identification of drugs that caused the death—prior to ICD-10 coding.

Pre-processing of death certificates using the model developed in Chapter 3 while records undergo ICD-10 coding allows public health practitioners to obtain provisional counts of overdose deaths more quickly. Faster data can lead to quicker implementation of harm reduction strategies, such as targeted naloxone distribution, thereby advancing public health practice. Enhancing the timeliness of data through NLP also reflects an improvement in the science of surveillance as well. According to the National Research Council, two of the five ways in which science can be advanced is by integration (e.g., linking insights and methodologies from other fields and levels of analysis) and development (e.g., stimulating additional research in the field)[24]. This dissertation accomplishes both by integrating methodologies from other disciplines into public health surveillance and by developing a base on which future scientists can build to add further NLP methods in surveillance as described in more detail below.

Enhancing data quality

In addition to timeliness, drug overdose mortality surveillance is also limited by weaknesses in data quality. Death certificate data is the only national source for overdose mortality surveillance,[5] and is the primary source for state and local mortality surveillance, making the quality of the data essential so that jurisdictions can provide accurate information to communities, promote appropriate evidence-based practices and policies, and monitor the impact of interventions and changes in policy. A primary data quality limitation in drug overdose mortality surveillance is the coarse granularity of information on what substances caused death (e.g., ICD-10 code indicates T43.6, "psychostimulants with abuse potential" as the substance, when the specific substance involved was methamphetamine). To address this challenge in data quality, the method presented in Chapter 4 leverages NLP to identify novel drugs and/or drug misspellings currently absent from lookup tables used to determine drugs involved in overdose deaths.

Chapter 4 focused on developing a state-of-the art method for identifying drug mentions on free-text death certificate data. The developed deep learning model was able to identify novel drugs and misspellings that look-up tables traditionally used for performing overdose mortality surveillance were unable to identify. The developed method improves surveillance data quality by identifying novel drugs and misspellings on free-text death certificates that are not present in current lookup tables. These new entities can be added to the lookup tables of existing surveillance tools.

Adding novel entities identified by the model to the lookup tables of current surveillance tools will improve the identification of drugs on the death certificate. The ability of the model to learn context enables it to identify novel drugs and misspellings

that are not present in the tables currently used. This improved identification ensures that jurisdictions are responding to the specific drugs that are causing morbidity and mortality in a community. Many jurisdictions will not have the expertise nor the computational power to develop and apply NLP models. Therefore, it will be necessary for centers with the expertise to periodical run the developed deep learning model on new death certificates, add novel entities to lookup tables, and disseminate the updated lookup tables to jurisdictions. This process will provide increased specificity of detecting drug entities without the need for each jurisdiction to run a complex model.

As public health initiatives rely on surveillance data, ensuring that the data is as accurate as possible is essential. The enhancement of surveillance data quality through the implementation of NLP methods advances public health practice by ensuring that decisions are made on accurate, high quality data. Additionally, surveillance science is advanced by integration (by applying techniques that emerged from a different field, linked to previous work in clinical sciences).[24] Applying NLP methods into surveillance science are a necessity as unstructured, free-text data sources have become common in epidemiological surveillance[44,45] and the methods developed for death certificates in the present dissertation show promise for utilization of NLP on other sources, further advancing surveillance science through the tenet of development.[24]

Limitations

The contribution of this dissertation to practice and science should be interpreted in light of its limitations. Despite the promise for NLP methods described in the narrative review, adoption of these methods into surveillance workflows may not be possible in all jurisdictions. Epidemiologists working in state and local public health departments may

not have the expertise or computing infrastructure needed to implement NLP and ML models, including those discussed in Chapters 3 and 4, in their work. However, even in these contexts, the NER model from Chapter 4 has value. NER can be used to identify new drug entities that can be added to look-up tables that those working in public health currently use, allowing them to detect novel substances and identify misspellings that were previously missing. Integrating the model presented in Chapter 4 as a tool to improve existing drug dictionaries addresses another limitation of this method, as its use as a stand-alone surveillance tool is limited since the entities the tool identifies must be manually linked to a parent drug. Adding novel entities to a lookup table, which typically contain crosswalks to parent drugs, addresses this limitation.

Another limitation is that the methods developed were trained using only data from Kentucky death certificates. While Kentucky has a robust drug overdose mortality surveillance system,[32] thereby boosting the rigor of the analyses, it remains unknown whether models will be generalizable outside of Kentucky. Death certificate free-text is typically short and direct, so it is unlikely that the language used on other jurisdictions' death certificates would differ substantially. A final limitation is the inability to diagnose errors in the models' performance. Complex ML models suffer from the "black box" problem,[133] which makes identifying why the model incorrectly predicted a response difficult due to the large number of inputs and interactions present in the underlying model.

Future research

This dissertation highlights the ability of NLP and ML models to advance public health practice and surveillance science specifically in the context of drug overdose

mortality surveillance. The narrative review showed that despite widespread use in clinical sciences there was limited use of NLP methods in public health surveillance, despite data sources including free-text elements.[20] An additional gap in the use of NLP methods in public health was the lack of state-of-the-art models used, indicating that current advances in NLP science had not yet been translated to public health practice. Implementing deep learning approaches in surveillance, as Chapter 4 of the dissertation demonstrates, will improve surveillance data quality and increase case identification.

Future research in the field of drug overdose mortality surveillance should test the developed models and develop new models on data from additional jurisdictions. Demonstrating that the current models can accurately predict labels on data from other jurisdictions or training a new model on data from several jurisdictions would eliminate the current limitation of having data from only one state present. The latter would also increase the sample size used to train the model, which typically will improve model performance and lead to a more accurate model overall. Additionally, further improvements can be made to the methods developed in Chapters 3 and 4. Additional feature engineering, such as non-sequential tokens, could be explored to improve the classification model developed in Chapter 3. The model developed in Chapter 4 could be improved to learn additional context so that it does not identify drug entities appearing in phrases such as “history of heroin abuse” which are not indicative that the drug was involved in the overdose death.

Finally, there are other surveillance data sources currently used for drug overdose mortality surveillance that contain free-text data, including emergency medical services data[44,45,131] and emergency department syndromic surveillance data.[132] Studies should

investigate the use of NLP, specifically the Flair[124] library used in Chapter 4, as a tool for information extraction and classification of these data sources. Adding NLP and ML to the surveillance workflow for these sources will improve the quality of drug overdose morbidity surveillance in a similar manner to how this dissertation improves the quality of drug overdose mortality surveillance.

Conclusion

The work of the present dissertation advances the current state of surveillance science and public health practice through the integration of NLP and ML methods that are not currently used in the field. Integrating NLP and ML methods into the surveillance pipeline enhances both the timeliness and quality of drug overdose mortality surveillance data. Timely, accurate data is essential to monitoring the ongoing drug overdose epidemic, as it ensures public health and public safety resources are directed at the substances currently causing morbidity and mortality in communities.

The methods developed had high scores for positive predictive value, sensitivity, and F-score. In particular, the sensitivity of the NLP methods was substantially higher than those of currently used surveillance tools. This improved sensitivity leads to a low false negative rate, meaning more cases that meet case criteria are identified than the currently used text-matching methods and look-up tables. Increased case identification allows for the collection of additional data sources in a timelier manner which can improve the public health response to increases in overdose mortality.

NLP and ML methods are under-utilized in public health surveillance overall. Implementation of these tools are necessary as the data collected for surveillance purposes continues to contain free-text elements. Making accurate insights from free-text

data is important to ensure that the public health response is based on the best available evidence. This dissertation shows that integrating NLP and ML methods into a surveillance workflow can lead to enhancements in both data timeliness and data quality and provides an overall advancement to public health practice and surveillance science through the development of two highly accurate models that ensure timely, high quality data for drug overdose mortality surveillance.

APPENDICES

Appendix 1: Phrase list for rule-based method

1. Overdose
2. Polypharmacy
3. Drug intoxication
4. Multiple drug
5. Combined drug
6. Acute combined
7. Intoxication drug
8. Drug toxicity
9. Acute fentanyl
10. Fentanyl intoxication
11. Fentanyl toxicity
12. Heroin intoxication
13. Multidrug intoxication
14. Heroin toxicity
15. Acute intoxication
16. Combined effects
17. Toxic effects
18. Acute heroin
19. Multi drug
20. Multiple drugs
21. Illicit drugs
22. Abused fatal
23. Heroin fentanyl
24. Intoxication fentanyl
25. Fentanyl morphine
26. Intoxication methamphetamine
27. Acute multidrug
28. Intoxication heroin
29. Illicit drug
30. Acute methamphetamine
31. Drug fentanyl
32. Drugs including
33. Gabapentin drug
34. Methamphetamine intoxication
35. Intoxication overdose
36. Drug heroin
37. Including heroin

Appendix 2: Error analysis

The false positives and false negatives of the SVM model excluding the SCC section were examined manually to determine if a post-processing step after the classifier is deployed could be used to improve classification. As Table 4 of the manuscript shows, this model had a total of 35 false positives and 21 false negatives on the test data. Of the 35 false positives, more than 10 appear to be data quality errors in the UCOD field on the DC records. These cases have free-text that points to the death being an OD death, but the text listed in the UCOD field is not an ICD-10 code.

Another category of the false positive cases are cases that appear to be OD deaths that were wrongly coded at NCHS. Of these cases, several of them mention the term "overdose" which likely lead the classifier to predict that these cases are ODs. Other cases in this category mention the decedent dying as a result of "drug intoxication" (listing a specific drug or indicating multiple drugs) or mention elevated levels of a substance in the decedent's system. Interestingly, most of the cases in these categories list other complications, such as "asphyxiation due to drug overdose", with NCHS then coding the cases as an asphyxiation death. However, if the OD was an event that directly caused the individual to asphyxiate, these deaths should perhaps be coded as OD deaths. This presents another use for this classifier—potentially identifying additional OD deaths that were miscoded at NCHS.

The majority of the other false positives are cases that involve deaths caused by chronic drug abuse (not an acute OD event) or an OD exacerbating some sort of condition or causing an injury that then leads to death. These cases include examples of

an individual overdosing and then falling into a river and drowning, or an individual ingesting drugs and aggravating their existing chronic obstructive pulmonary disease. The presence of specific words, such as substances and terms like “intoxication”, likely lead to the classifier mistaking these for OD deaths.

Many of the 21 false negatives were cases with a small amount of text that contained substances that are rarely seen in OD deaths (such as acetaminophen) or contain misspellings. For example, one false negative misspells “toxicity” as “tocicity” and others misspell “drug” as “drue.” Employing an automated spell checker when preprocessing the text may fix some of these errors. The other false negatives are the opposite—well written text that is much longer than what typically appears on a DC. Many of these list that the decedent died from a combination of drugs and alcohol, so the presence of the word alcohol (or related terms such as ethanol) may signal the classifier that the death may be an alcohol related death and not a drug OD death, resulting in misclassification.

Appendix 3: Flair Modeling Code

In [1]:

```
# Reading in the corpus
from flair.data import Corpus
from flair.datasets import ColumnCorpus

columns = {0: "text", 1: "ner"}
data_folder = "/home/pjwa227/Corpus2/"

corpus: Corpus = ColumnCorpus(data_folder, columns,
                              train_file = "train.txt",
                              test_file = "test.txt",
                              dev_file = "val.txt")
2020-12-11 09:17:50,035 Reading data from /home/pjwa227/Corpus2
2020-12-11 09:17:50,036 Train: /home/pjwa227/Corpus2/train.txt
2020-12-11 09:17:50,036 Dev: /home/pjwa227/Corpus2/val.txt
2020-12-11 09:17:50,036 Test: /home/pjwa227/Corpus2/test.txt
```

In [2]:

```
# Quick check
print(len(corpus.train))

print(corpus.train[1].to_tagged_string("ner"))
6108
INTRAVENTRICULAR HEMORRHAGE COUMADIN <S-DRUG> TOXICITY
```

In [3]:

```
# Tell flair what tag we want to predict
tag_type = "ner"
tag_dictionary = corpus.make_tag_dictionary(tag_type=tag_type)
```

In [4]:

```
# Load embeddings
from flair.embeddings import PooledFlairEmbeddings, StackedEmbeddings

embeddings : StackedEmbeddings =
StackedEmbeddings([PooledFlairEmbeddings("news-forward"),

PooledFlairEmbeddings("news-backward"),])
```

In [5]:

```
# Initialize the sequence tagger
from flair.models import SequenceTagger

tagger : SequenceTagger = SequenceTagger(hidden_size =256,
                                         embeddings = embeddings,
                                         tag_dictionary =
tag_dictionary,
                                         tag_type = tag_type,
                                         use_crf = True)
print(tagger)
SequenceTagger(
```

```
  (embeddings): StackedEmbeddings(
    (list_embedding_0): PooledFlairEmbeddings(
      (context_embeddings): FlairEmbeddings(
        (lm): LanguageModel(
          (drop): Dropout(p=0.05, inplace=False)
          (encoder): Embedding(300, 100)
          (rnn): LSTM(100, 2048)
          (decoder): Linear(in_features=2048, out_features=300,
bias=True)
        )
      )
    )
    (list_embedding_1): PooledFlairEmbeddings(
      (context_embeddings): FlairEmbeddings(
        (lm): LanguageModel(
          (drop): Dropout(p=0.05, inplace=False)
          (encoder): Embedding(300, 100)
          (rnn): LSTM(100, 2048)
          (decoder): Linear(in_features=2048, out_features=300,
bias=True)
        )
      )
    )
  )
  (word_dropout): WordDropout(p=0.05)
  (locked_dropout): LockedDropout(p=0.5)
  (embedding2nn): Linear(in_features=8192, out_features=8192,
bias=True)
  (rnn): LSTM(8192, 256, batch_first=True, bidirectional=True)
  (linear): Linear(in_features=512, out_features=8, bias=True)
  (beta): 1.0
  (weights): None
  (weight_tensor) None
)
```

In [7]:

```
# Train
from flair.trainers import ModelTrainer
trainer : ModelTrainer = ModelTrainer(tagger, corpus)

trainer.train('resources/taggers/pooled_flair_ner',
              learning_rate=0.1,
              mini_batch_size=32,
              max_epochs=150)
Results:
- F1-score (micro) 0.9913
- F1-score (macro) 0.9913

By class:
DRUG       tp: 3201 - fp: 27 - fn: 29 - precision: 0.9916 - recall:
0.9910 - f1-score: 0.9913
2020-12-11 09:42:20,443 ----------------------------------------------
------------------------------------------------------
```

Out[7]:

Appendix 4: Drug Overdose Fatality Surveillance System Drug Entities

25B-NBOME
25C-NBOME
25D-NBOME
25H-NBOME
25I-NBOME
3METHYLFENTANYL
3-METHYLFENTANYL
3METHYLMORPHINE
3-METHYLMORPHINE
3-METHYLTHIOFENTANYL
4_ANPP
4ANPP
4-ANPP
4-METHOXYBUTYRYLFENTANYL
5F-AB-PINACA
5F-ADB
5F-ADB-PINACA
5F-AMB
5F-NNEI
5F-PB-22
5F-THJ
6-AM
6MAM
6-MAM
6-MOMOACETYLMORPHINE
6-MONACETYLMORPHINE
6-MONOACETYLMOPRHINE
6MONOACETYLMORPHINE
6-MONOACETYLMORPHINE
6-MONOACETYMORPHINE
6-MONOACETYTMORPHINE
6-MONOACEYTLYMORPHIN
6-MONOACEYTLYMORPHINE
6-MONOACEYTLYMORPHONE
6-MONOACTEYLMORPHINE
7-AMINOCLONAZEAPM
7AMINOCLONAZEPAM
JWH-073
JWH-122
JWH-210
JWH-250
LAMOTRIGINE
LAZANDA
LEVAMISOLE
LEVETIRACETAM
LEVOFLOXACIN
LIDOCAINE
LOPERAMIDE
LORAZEPAM
LORAZEPAN
LORCET
LORTAB
LYRICA
M-144
MAB-CHMINACA
MA-CHMINACA
MAM-2201
MARIJUANA
MATRIFEN
MAXIDONE
MDMA
MDMB-CHMICA
MDMB-FUBINACA
MDPV
MDPV
MEDICATION
MEPERIDINE
MEPHEDRONE
MEPHEDRONE
MEPROBAMATE
METAHDONE
META-METHYMETHOXYACETYL FENTANYL
METAXALONE
METFORMIN

7-AMINOCLONAZEPAM
AB-CHMINACA
AB-FUBINACA
AB-PINACA
ABSTRAL
ACETAMINOPHEN
ACETYL-ALPHA-METHYLFENTANYL
ACETYLFENTANYL
ACID
ACRYLFENTANYL
ACTIQ
ADB-FUBINACA
ADBICA
ADB-PINACA
ADDERALL
AH-7921
ALCOHOL
ALFENTANIL
ALPHA-METHYLFENTANYL
ALPHA-METHYLTHIOFENTANYL
ALPHA-PBP
ALPHA-PPP
ALPHA-PVP
ALPRAOZOLAM
ALPRAZOLAM
ALPRAZOLAN
ALPRAZOLEM
ALRPAZOLAM
AMB
AMBIEN
AMIDONE
AMINOCLONAZEPAM
AMIODARONE
AMITRIPTYLINE
AMLODIPINE
AMPHETAMINE
AMPHETAMINES
ANEXSIA
ANPP
ANTICONVULSANT
ANTICONVULSANTS
ANTIDEPRESSANT
METH
METHADONE
METHADOSE
METHAMHPETAMINE
METHAMPHETAMINE
METHANOL
METHEDRONE
METHEDRONE
METHOCARBAMOL
METHORPHAN
METHOTREXATE
METHOXETAMINE
METHOXYACETYLFENTANYL
METHOXYBUTYRYLFENTANYL
METHYLENEDIOXYMETHAM
METHYLETHCATHINONE
METHYLETHCATHINONE
METHYLFENTANYL
METHYLONE
METHYLONE
METOCLOPRAMIDE
METONITAZENE
METOPROLOL
MIRTAZAPINE
MIRTAZEPINE
MITRAGYNINE
MN18
MO-CHMINACA
MOLLY
MOMOACETYLMORPHINE
MONACETYLMORPHINE
MONOACETYLMOPRHINE
MONOACETYLMORPHINE
MONOACETYMORPHINE
MONOACETYTMORPHINE
MONOACEYTLYMORPHIN
MONOACEYTLYMORPHINE
MONOACEYTLYMORPHONE
MONOACTEYLMORPHINE
MONOHYDROXYOXCARBAZEPINE
MONOXIDE
MOPRHINE

ANTIDEPRESSANTS
ANTIPSYCHOTICS
A-OH-ALPRAOZOLAM
A-OH-ALPRAZOLAM
A-OH-ALPRAZOLAN
A-OH-ALPRAZOLEM
A-OH-ALRPAZOLAM
A-OH-APRAZOLAM
APP-FUBINACA
APRAZOLAM
A-PVP
ATENOLOL
BACLOFEN
BARBITURATE
BARBITURATES
BENADRYL
BENZODIAEPINE
BENZODIAEPINES
BENZODIAZEPINE
BENZODIAZEPINES
BENZOS
BENZOYLECGONINE
BENZOYLECGONINE
BENZTROPINE
BETA-HYDROXY-3-METHYLFENTANYL
BETA-HYDROXYFENTANYL
BETA-HYDROXYTHIOFENTANYL
BLEOMYCIN
BRORPHINE
BUPHEDRONE
BUPHEDRONE
BUPRENORFINE
BUPRENORPHINE
BUPROPION
BUPROPRION
BUSPIRONE
BUTALBITAL
BUTYLONE
BUTYLONE
BUTYRYLFENTANYL
CAFFEINE
CANNABINOIDS
MORPHINE
MORPHONE
MT-45
NALOXONE
NALTREXONE
NARCOTIC
NARCOTICS
NARCOTISM
N-DESMETHYL-TRAMADOL
NEURONTIN
NEUROSTIL
NICOTINE
NNEI
NORBUPRENORPHINE
NORCO
NORDIAZEPAM
NORDOXEPIN
NORFENTANYL
NORFLUOXETINE
NORPROPOXYPHENE
NORSERTRALINE
NORTRAMADOL
NORTRIPTYLINE
NORVENLAFAXINE
NUPENTIN
OCFENTANIL
O-DESMETHYL-TRAMADOL
OLANZAPINE
ONSOLIS
OPIATE
OPIATES
OPIOID
OPIOIDS
ORPHENADRINE
ORTHO-FLUORO FENTANYL
OXAZEAPM
OXAZEPAM
OXCARBAZEPINE
OXISET
OXYCODONE
OXYCODONE
OXYCONTIN

CANNABIS
CARBAMAZEPINE
CARBAZEPINE
CARBON
CARFENTANIL
CARFENTANIL
CARFENTANYL
CARISOPRODOL
CARISPRODOL
CHARCOAL
CHLORDIAZEPOXIDE
CHLORIDE
CHLOROPHENYLPIPERAZINE
CHLORPHENIRAMINE
CHLORPROMAZINE
CITALOPRAM
CLOMIPRAMINE
CLONAZEPAM
CLOZAPINE
COCAETHYLENE
COCAINE
CODEINE
CODEINE
COTININE
COUMADIN
CRYSTAL
CYANIDE
CYCLOBENZAPRINE
CYCLOPROPYLFENTANYL
DAMASON-P
DARVOCET
DARVOCET
DARVON
DEMEROL
DEPRESSANT
DESIPRAMINE
DESOMORPHINE
DEXTROMETHORPHAN
DEXTROPROPOXYPHENE
DIACETYLMORPHINE
DIAMORPHINE
DIAZDEPAM
OXYCONTIN
OXYCOTIN
OXYGEN
OXYMOPHONE
OXYMORPHONE
OXYNORM
OZAZEPAM
PALLADONE
PANACET
PARA-FLUOROBUTYRYLFENTANYL
PARA-FLUOROFENTANYL
PARA-FLUOROISOBUTYRYLFENTANYL
PARA-METHYMETHOXYACETYL FENTANYL
PAROXETINE
PB-22
PENRAL
PENTEDRONE
PENTEDRONE
PENTYLONE
PERCOCET
PERCODAN
PETNYLONE
PHARMACEUTICAL
PHARMACOLOGIC
PHENCYCLIDINE
PHENOBARBITAL
PHENTERMINE
PHENYTOIN
PIPERACILLIN
POLYDRUG
POLYPHARMACY
POLYSUBSTANCE
POTASSIUM
PREGABALIN
PRESCRIPTIONS
PROMETHAZINE
PROOXYPHENE
PROPAFENONE
PROPANE
PROPANOLOL
PROPOFOL
PROPOFOL

DIAZEPAM
DICYCLOMINE
DIFLUOROETHANE
DIGOXIN
DIHYDROCODEINE
DIHYDROCODEINONE
DIHYDROMORPHINONE
DILAUDID
DILTIAZEM
DIPHENHYDRAMINE
DISKETS
DOLOPHINE
DOXEPIN
DOXYLAMINE
DULOXETINE
DURAGESIC
DURAGESIC
DUROGESIC
EAM-2201
ECSTASY
EDDP
ENDOCET
ENDODAN
EPHEDRINE
EPHEDRINE
ESCITALOPRAM
ETHABOL
ETHANOL
ETHYLENE
ETHYLMETHCATHINONE
ETHYLMETHOCATHINONE
ETHYLONE
ETHYLONE
ETIZOLAN
ETOH
ETOMIDATE
EXALGO
FAB-144
FANATREX
FDU-PB-22
FEENTANYL
FENATNYL
PROPOSYPHENE
PROPOXIPHENE
PROPOXITENE
PROPOXPHENE
PROPOXTYPHENE
PROPOXY
PROPOXYCODONE
PROPOXYPHEN
PROPOXYPHENA
PROPOXYPHENE
PROPOXYPHERE
PROPOXYPHINE
PROPOXZPHENE
PROPRANOLOL
PROPXYPHENE
PSEUDOEPHEDRINE
PSYCHIATRIC
PX1
PX2
PX3
QUETIAPINE
REMIFENTANIL
ROXICET
ROXICODONE
ROXISET
SALICYLATE
SALICYLATES
SDB-006
SEROQUEL
SERTRALINE
SUBLIMAZE
SUBOXONE
SUBSTANCE
SUDAFED
SUFENTANIL
TEMAZEPAM
TETRAHYDROCANNABINOL
TETRAHYDROFURANFENTANYL
THC
THC-COOH
THIOFENTANYL
THJ

FENTANIL
FENTANLY
FENTANOL
FENTANY
FENTANYL
FENTANYL
FENTATYL
FENTAYNL
FENTNAYL
FENTORA
FETANYL
FLEPHEDRONE
FLEPHEDRONE
FLUOROBUTYRYLFENTANYL
FLUOROISOBUTYRYLFENTANYL
FLUOXETINE
FLURAZEPAM
FURANYLFENTANYL
GABAPENTIN
GABAPIN
GABARONE
GABRION
GLYCOL
GRALISE
GUAIFENESIN
HALDID
HALOPERIDOL
HERION
HEROIN
HYCODAN
HYDROCADONE
HYDROCHLOROTHIAZIDE
HYDROCOCONE
HYDROCODINE
HYDROCODONE
HYDROCODONE
HYDROMORPHINE
HYDROMORPHONE
HYDROXYCHLOROQUINE
HYDROXYTHIOFENTANYL
HYDROXYZINE
HYSINGLA ER
THJ-018
THJ-2201
TIZANADINE
TOBACCO
TOLUENE
TOPIRAMATE
TRAMADAL
TRAMADOL
TRAMADONE
TRAMDOL
TRAMEDOL
TRAMELL
TRAMIDOL
TRAMODOL
TRAZADONE
TRAZODONE
TREMEDEL
TREMEDOL
TUSSINEX
TYLENOL
TYLOX
U47700
U-47700
U-49900
U-51754
UR-144
VALERYL FENTANYL
VALIUM
VALPROIC
VENLAFAXINE
VERAPAMIL
VICIDAN
VICODIN
VICODIN
WARFARIN
XANAX
XANAX
XLR-11
XLR11
XLR12
XLR-12
XYLAZINE

IBUPROFEN	ZOHYDRO ER
INSTANYL	ZOLPIDEM
ISOPROPANOL	ZYDONE
ISOPROPYL	NALMEFENE
ISOTONITAZENE	QUININE
JWH-015	SUBLOCADE
JWH-018	SUBUTEX
JWH-019	NALBUPHINE

Appendix 5: Drug entities present in test data not in Drug Overdose Fatality Surveillance System Table

ALPRAZOLM	FLOUXETINE
HYDROCONE	NORFLUOXITINE
METHAMPHETAMINES	DEMOXEPAM
ACETALFENTANYL	DESIGNER FENTANYLS
METHAMPHETTAMINE	CANNABOIDS
IMODIUM	4-AANP
OYCODONE	ACETYLFENTAYL
TETRAHYDROFURANFENTANYL	DESIGNER OPOIDS
FENANTYL	ALPRAZPOAM
NORIDIAZEPAM	BENZO
ACETYFENTANYL	FENTANL
METHAMPETAMINE	3,4 METHYLENEDIOXY-METHAMPHETAMINE
INULIN	ALPROZOLAM
INSULIN	KETAMINE
ANDACETYLFENTANY	METHOXYCETYFENTANYL
HYDORCODONE	GAHAPENTIN
BENZO-DIAZEPINES	ARIPIPRAZOLE
INDOMETHACIN	DIPHENHYDRA
FENTANYLN	SERTALINE
BENZODIAZAPENIES	VENAFLAXZINE
4-NAPP	BENZOYLECGONINE.QUANT
THC COOH	NITROGLYCERIN
OPATES	METHAMPHRTAMINE
CODIENE	CLONZEPAN
4-ANNP	BENZODIAZIPINE
AMPHETAIMES	BENZODIAPINE
METHAMPHETIMINE	ZANAFLEX
CLONAZAPAM	METHOXYACETYLFENTANY
CANNABINOID	GGABAPENTIN
7-AMINOCIONAZEPAM	HYDROXYAINE
HYDROOXYZINECAUSING	CONTININE

ACETYFENTANTANYL
BUTYRYFENTANYL
TOPIRMATE
TCC-COOH
MORPINE
METHAMPH
EFFEXOR
PRISTIQ
FLEXERIL
OXMORPHONE
$-ANPP
OXYCODE
METHAPHETAMINE
1,1-DIFLUOROETHANE
TCH-COOH
DESIGNER OPIOIDS
NIFEDIPINE
ACETYL FENTANYL
AMHPETAMINE
4-AMPP
PARAFLUROBUTYRYLFENTANYL
DESPROPIONYLFENTANYL
DESIGNER OPIATE
4-ANP
OXYCODEONE
METAMPHETAMINE
4ANNP
CRACK COCAINE
OXYCODONC
GABAPCNTIN
LOPERAMITE
IMMODIUM
OVERDOSE:FENTANYL
CALCIUM CHANNEL BLOCKER
O-DESMETHYLVENLAFAXINE
MEXILETINE
NAPROXEN
FANTANYL
ACETYFENTANY
FLUXETINE
ACETLFENTANYL
ACETYFENTAN
ACETYLFENTANYL1
SODIUM NITRATE
BUSPRENORPHINE
WITHACETYLFENTANYL
COCAETHYENE
TRAZODON
HYDROX
METHANPHETAMINE
AMPHETATMINE
BENZODIAZOPINE
METHAMPHETEMINE
METHANPHETAMINES
ACETTYLFENTANYL
SODIUM NITRITE
4NAPP
ACEYLFENTANYL
METHAMPPHETAMINE
BUBRENORPHINE
METHAMPHATAMINE
METHAMPHETAMINR
DESIGNER DRUGS
GABAPENTIN_
DIIIIIAZEPAM
PARAFLUOROBUTYRYLFENTANYL
ACEYTLFENTANYL
PARA-FLUOROISOBUTYRYL FENTANYL

REFERENCES

1. Hedegaard H, Minino AM, Warner M. *Drug Overdose Deaths in the United States, 1999-2019.* National Center for Health Statistics;2020.
2. Hedegaard H, Miniño AM, Warner M. Drug Overdose Deaths in the United States, 1999-2018. *NCHS Data Brief.* 2020;356.
3. Concheiro-Guisan M, Chesser R, Pardi J, Cooper G. Postmortem toxicology of new synthetic opioids. *Frontiers in Pharmacology.* 2018;9:1210.
4. Jalal H, Buchanich JM, Roberts MS, Balmert LC, Zhang K, Burke DS. Changing dynamics of the drug overdose epidemic in the United States from 1979 through 2016. *Science (New York, NY).* 2018;361(6408).
5. Warner M, Hedegaard H. Identifying Opioid Overdose Deaths Using Vital Statistics Data. *American Journal of Public Health.* 2018;108(12):1587-1589.
6. Slavova S, Rock P, Bush HM, Quesinberry D, Walsh SL. Signal of increased opioid overdose during COVID-19 from emergency medical services data. *Drug and alcohol dependence.* 2020;214:108176.
7. Trinidad JP, Warner M, Bastian BA, Miniño AM, Hedegaard H. Using literal text from the death certificate to enhance mortality statistics: Characterizing drug involvement in deaths. *National vital statistics reports : from the Centers for*

Disease Control and Prevention, National Center for Health Statistics, National Vital Statistics System. 2016;65(9):1-15.

8. O'Donnell JK, Halpin J, Mattson CL, Goldberger BA, Gladden RM. Deaths Involving Fentanyl, Fentanyl Analogs, and U-47700 - 10 States, July-December 2016. *MMWR Morb Mortal Wkly Rep.* 2017;66(43):1197-1202. doi:10.15585/mmwr.mm6643e1.
9. Slavova S, Costich JF, Bunn TL, et al. Heroin and fentanyl overdoses in Kentucky: Epidemiology and surveillance. *Int J Drug Policy.* 2017;46:120-129.
10. Kariisa M, Scholl L, Wilson N, Seth P, Hoots B. Drug Overdose Deaths Involving Cocaine and Psychostimulants with Abuse Potential—United States, 2003–2017. *Morbidity and Mortality Weekly Report.* 2019;68(17):388-395.
11. Patrick SW, Fry CE, Jones TF, Buntin MB. Implementation Of Prescription Drug Monitoring Programs Associated With Reductions In Opioid-Related Death Rates. *Health Aff (Millwood).* 2016;35(7):1324-1332.
12. Blanco C, Volkow ND. Management of opioid use disorder in the USA: present status and future directions. *The Lancet.* 2019;393(10182):1760-1772.
13. Connery HS. Medication-assisted treatment of opioid use disorder: review of the evidence and future directions. *Harvard review of psychiatry.* 2015;23(2):63-75.
14. Ciccarone D. Stimulant abuse: pharmacology, cocaine, methamphetamine, treatment, attempts at pharmacotherapy. *Journal of Primary Care.* 2011;38(1):41-58.

15. Spencer M, Ahmad F. Timeliness of Death Certificate Data for Mortality Surveillance and Provisional Estimates. In: National Center for Health Statistics; 2017.

16. *World Health Organization. International Statistical Classification of Diseases and Related Health Problems: Tenth Revision Version for 2007 (ICD-10).* 2007.

17. NCHS. *A reference guide for completing the death certificate for drug toxicity deaths.* Hyattsville, MD: National Center for Health Statistics;2019.

18. Ford E, Carroll JA, Smith HE, Scott D, Cassell JA. Extracting information from the text of electronic medical records to improve case detection: a systematic review. *J Am Med Inform Assoc.* 2016;23(5):1007-1015.

19. Wang Y, Wang L, Rastegar-Mojarad M, et al. Clinical information extraction applications: A literature review. *Journal of biomedical informatics.* 2018;77:34-49.

20. Mooney SJ, Pejaver V. Big Data in Public Health: Terminology, Machine Learning, and Privacy. *Annu Rev Public Health.* 2018;39:95-112.

21. Chen Y, Zhou C, Li T, et al. Named entity recognition from Chinese adverse drug event reports with lexical feature based BiLSTM-CRF and tri-training. *Journal of biomedical informatics.* 2019;96:103252.

22. Korkontzelos I, Piliouras D, Dowsey AW, Ananiadou S. Boosting drug named entity recognition using an aggregate classifier. *Artificial intelligence in medicine.* 2015;65(2):145-153.

23. Björne J, Kaewphan S, Salakoski T. UTurku: drug named entity recognition and drug-drug interaction extraction using SVM classification and domain knowledge.

Paper presented at: Second Joint Conference on Lexical and Computational Semantics (* SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013) 2013.

24. *A strategy for assessing science: Behavioral and social research on aging.* National Research Council; 2007.

25. Teutsch SM, Churchill RE. *Principles and practice of public health surveillance.* Oxford University Press, USA; 2000.

26. Allen DJ, Trainor E, Callaghan A, O'Brien SJ, Cunliffe NA, Iturriza-Gomara M. Early Detection of Epidemic GII-4 Norovirus Strains in UK and Malawi: Role of Surveillance of Sporadic Acute Gastroenteritis in Anticipating Global Epidemics. *PLoS One.* 2016;11(4):e0146972.

27. Sir-Ondo-Enguier PN, Ngoungou EB, Nghomo YN, et al. Syndromic surveillance of potentially epidemic infectious diseases: Detection of a measles epidemic in two health centers in Gabon, Central Africa. *Infect Dis Rep.* 2019;11(1):7701.

28. Hedegaard H, Miniño AM, Warner M. Drug Overdose Deaths in the United States, 1999-2017. *NCHS Data Brief.* 2018;329.

29. Warner M, Trinidad JP, Bastian BA, Miniño AM, Hedegaard H. Drugs most frequently involved in drug overdose deaths: United States, 2010–2014. *National vital statistics reports : from the Centers for Disease Control and Prevention, National Center for Health Statistics, National Vital Statistics System.* 2016;65(10):1-15.

30. Frey MT, Meaney-Delman D, Bowen V, et al. Surveillance for Emerging Threats to Pregnant Women and Infants. *J Womens Health.* 2019;28(8):1031-1036.

31. Stoto MA, Kang M, Song T, Bouey J, Boyce MR, Katz R. At the frontier of the global battle against emerging infections: surveillance and management of avian influenza A(H7N9) in Guangdong Province, China. *Journal of Global Health Reports.* 2019;3.

32. Hargrove SL, Bunn TL, Slavova S, et al. Establishment of a comprehensive drug overdose fatality surveillance system in Kentucky to inform drug overdose prevention policies, interventions and best practices. *Injury Prevention.* 2018;24(1):60-67.

33. Slavova S, Bunn TL, Hargrove SL, Corey T, Ingram V. Linking Death Certificates, Postmortem Toxicology, and Prescription History Data for Better Identification of Populations at Increased Risk for Drug Intoxication Deaths. *Pharm Med.* 2017;31(3):155-165.

34. Slavova S, Miller A, Bunn TL, et al. Prevalence of gabapentin in drug overdose postmortem toxicology testing results. *Drug Alcohol Depend.* 2018;186:80-85.

35. Slavova S, O'Brien DB, Creppage K, et al. Drug Overdose Deaths: Let's Get Specific. *Public Health Reports.* 2015;130(4):339-342.

36. Smith MG, Royer J, Mann JR, McDermott S. Using Administrative Data to Ascertain True Cases of Muscular Dystrophy: Rare Disease Surveillance. *JMIR Public Health Surveill.* 2017;3(1):e2.

37. Jafarzadeh SR, Thomas BS, Gill J, Fraser VJ, Marschall J, Warren DK. Sepsis surveillance from administrative data in the absence of a perfect verification. *Ann Epidemiol.* 2016;26(10):717-722 e711.

38. Noone AM, Cronin KA, Altekruse SF, et al. Cancer Incidence and Survival Trends by Subtype Using Data from the Surveillance Epidemiology and End Results Program, 1992-2013. *Cancer Epidemiol Biomarkers Prev.* 2017;26(4):632-641.

39. Hallowell BD, Saraiya M, Thompson TD, et al. Population-Based Assessment of HPV Genotype-Specific Cervical Cancer Survival: CDC Cancer Registry Sentinel Surveillance System. *JNCI Cancer Spectr.* 2018;2(3).

40. Frank CJ, Kushner SE, Doran DA, Stehr-Green J. Mandatory Reporting of Fatal and Nonfatal Opioid Overdoses in a Rural Public Health Department. *Am J Public Health.* 2018;108(12):1646-1648.

41. Ward PJ, Rock PJ, Slavova S, Young AM, Bunn TL, Kavuluru R. Enhancing timeliness of drug overdose mortality surveillance: A machine learning approach. *PLoS One.* 2019;14(10):e0223318.

42. Burnham JP, Kwon JH, Babcock HM, Olsen MA, Kollef MH. ICD-9-CM coding for multidrug resistant infection correlates poorly with microbiologically confirmed multidrug resistant infection. *Infection Control & Hospital Epidemiology.* 2017;38(11):1381-1383.

43. Slavova S, Costich JF, Luu H, et al. Interrupted time series design to evaluate the effect of the ICD-9-CM to ICD-10-CM coding transition on injury hospitalization trends. *Inj Epidemiol.* 2018;5(1):36.

44. Ray BR, Lowder EM, Kivisto AJ, Phalen P, Gil H. EMS naloxone administration as non-fatal opioid overdose surveillance: 6-year outcomes in Marion County, Indiana. *Addiction.* 2018;113(12):2271-2279.

45. Moore PQ, Weber J, Cina S, Aks S. Syndrome surveillance of fentanyl-laced heroin outbreaks: Utilization of EMS, Medical Examiner and Poison Center databases. *Am J Emerg Med.* 2017;35(11):1706-1708.

46. Nolan ML, Ehntholt A, Merrill T, Weiss D, Lall R, Paone D. Novel use of syndromic surveillance to monitor the impact of synthetic cannabinoid control measures on morbidity. *Inj Epidemiol.* 2019;6:33.

47. Hartnett KP, Kite-Powell A, Patel MT, et al. Syndromic Surveillance for E-Cigarette, or Vaping, Product Use–Associated Lung Injury. *New England Journal of Medicine.* 2019.

48. Abat C, Chaudet H, Rolain JM, Colson P, Raoult D. Traditional and syndromic surveillance of infectious diseases and pathogens. *Int J Infect Dis.* 2016;48:22-28.

49. Perlman SE, McVeigh KH, Thorpe LE, Jacobson L, Greene CM, Gwynn RC. Innovations in Population Health Surveillance: Using Electronic Health Records for Chronic Disease Surveillance. *American Journal of Public Health.* 2017;107(6):853-857.

50. Klompas M, Cocoros NM, Menchaca JT, et al. State and Local Chronic Disease Surveillance Using Electronic Health Record Systems. *American Journal of Public Health.* 2017;107(9):1406-1412.

51. Kavuluru R, Hands I, Durbin EB, Witt L. Automatic extraction of ICD-O-3 primary sites from cancer pathology reports. *AMIA Summits on Translational Science Proceedings.* 2013;2013:112.

52. Osborne JD, Wyatt M, Westfall AO, Willig J, Bethard S, Gordon G. Efficient identification of nationally mandated reportable cancer cases using natural

language processing and machine learning. *Journal of the American Medical Informatics Association.* 2016;23:1077-1084.

53. Rios A, Kavuluru R. Ordinal convolutional neural networks for predicting RDoC positive valence psychiatric symptom severity scores. *Journal of biomedical informatics.* 2017;75:S85-S93.

54. Moon S, Liu S, Scott CG, et al. Automated extraction of sudden cardiac death risk factors in hypertrophic cardiomyopathy patients by natural language processing. *International journal of medical informatics.* 2019;128:32-38.

55. Hinton G. Deep Learning-A Technology With the Potential to Transform Health Care. *JAMA.* 2018;320(11):1101-1102.

56. Mikolov T, Sutskever I, Chen K, Corrado GS, Dean J. Distributed representations of words and phrases and their compositionality. Paper presented at: Advances in neural information processing systems2013.

57. Huang Z, Xu W, Yu K. Bidirectional LSTM-CRF models for sequence tagging. *arXiv.* 2015;abs/1508.01991.

58. Bengio Y, Ducharme R, Vincent P, Jauvin C. A neural probabilistic language model. *Journal of machine learning research.* 2003;3(Feb):1137-1155.

59. Collobert R, Weston J. A unified architecture for natural language processing: Deep neural networks with multitask learning. Paper presented at: Proceedings of the 25th International Conference on Machine Learning2008.

60. Alawad M, Gao S, Qiu JX, et al. Automatic extraction of cancer registry reportable information from free-text pathology reports using multitask convolutional neural networks. *J Am Med Inform Assoc.* 2020;27(1):89-98.

61. Leroy G, Gu Y, Pettygrove S, Galindo MK, Arora A, Kurzius-Spencer M. Automated Extraction of Diagnostic Criteria From Electronic Health Records for Autism Spectrum Disorders: Development, Evaluation, and Application. *J Med Internet Res.* 2018;20(11):e10497.

62. Koopman B, Zuccon G, Nguyen A, Bergheim A, Grayson N. Extracting cancer mortality statistics from death certificates: A hybrid machine learning and rule-based approach for common and rare cancers. *Artif Intell Med.* 2018;89:1-9.

63. Munkhdalai T, Liu F, Yu H. Clinical Relation Extraction Toward Drug Safety Surveillance Using Electronic Health Record Narratives: Classical Learning Versus Deep Learning. *JMIR Public Health Surveill.* 2018;4(2):e29.

64. Metzger MH, Tvardik N, Gicquel Q, Bouvry C, Poulet E, Potinet-Pagliaroli V. Use of emergency department electronic medical records for automated epidemiological surveillance of suicide attempts: a French pilot study. *Int J Methods Psychiatr Res.* 2017;26(2).

65. Koopman B, Zuccon G, Nguyen A, Bergheim A, Grayson N. Automatic ICD-10 classification of cancers from free-text death certificates. *International journal of medical informatics.* 2015;84(11):956-965.

66. Butt L, Zuccon G, Nguyen A, Bergheim A, Grayson N. Classification of cancer-related death certificates using machine learning. *The Australasian medical journal.* 2013;6(5):292-299.

67. Koopman B, Karimi S, Nguyen A, et al. Automatic classification of diseases from free-text death certificates for real-time surveillance. *BMC medical informatics and decision making.* 2015;15:53.

68. Duarte F, Martins B, Pinto CS, Silva MJ. Deep neural models for ICD-10 coding of death certificates and autopsy reports in free-text. *Journal of biomedical informatics.* 2018;80:64-77.

69. Maenner MJ, Yeargin-Allsopp M, Van Naarden Braun K, Christensen DL, Schieve LA. Development of a Machine Learning Algorithm for the Surveillance of Autism Spectrum Disorder. *PLoS One.* 2016;11(12):e0168224.

70. Jha D, Singh R. SMARTS: the social media-based addiction recovery and intervention targeting server. *Bioinformatics.* 2019.

71. Feldman J, Thomas-Bachli A, Forsyth J, Patel ZH, Khan K. Development of a global infectious disease activity database using natural language processing, machine learning, and human expertise. *J Am Med Inform Assoc.* 2019;26(11):1355-1359.

72. Müller MM, Salathé M. Crowdbreaks: Tracking health trends using public social media data and crowdsourcing. *Frontiers in public health.* 2019;7(81).

73. Klein AZ, Sarker A, Cai H, Weissenbacher D, Gonzalez-Hernandez G. Social media mining for birth defects research: A rule-based, bootstrapping approach to collecting data for rare health-related events on Twitter. *Journal of biomedical informatics.* 2018;87:68-78.

74. Sarker A, Chandrashekar P, Magge A, Cai H, Klein A, Gonzalez G. Discovering Cohorts of Pregnant Women From Social Media for Safety Surveillance and Analysis. *J Med Internet Res.* 2017;19(10):e361.

75. Kagashe I, Yan Z, Suheryani I. Enhancing Seasonal Influenza Surveillance: Topic Analysis of Widely Used Medicinal Drugs Using Twitter Data. *J Med Internet Res.* 2017;19(9):e315.

76. Mowery J. Twitter Influenza Surveillance: Quantifying Seasonal Misdiagnosis Patterns and their Impact on Surveillance Estimates. *Online J Public Health Inform.* 2016;8(3):e198.

77. Sarker A, O'Connor K, Ginn R, et al. Social Media Mining for Toxicovigilance: Automatic Monitoring of Prescription Medication Abuse from Twitter. *Drug Saf.* 2016;39(3):231-240.

78. Broniatowski DA, Paul MJ, Dredze M. National and local influenza surveillance through Twitter: an analysis of the 2012-2013 influenza epidemic. *PLoS One.* 2013;8(12):e83672.

79. Allen C, Tsou MH, Aslam A, Nagel A, Gawron JM. Applying GIS and Machine Learning Methods to Twitter Data for Multiscale Surveillance of Influenza. *PLoS One.* 2016;11(7):e0157734.

80. Sokolova M, Japkowicz N, Szpakowicz S. Beyond accuracy, F-score and ROC: a family of discriminant measures for performance evaluation. Paper presented at: Australasian Joint Conference on Artificial Intelligence2006.

81. Müller MM. Crowdbreaks. Published 2019. Accessed 03/03/2020, 2020.

82. Kim J-H. Estimating classification error rate: Repeated cross-validation, repeated hold-out and bootstrap. *Computational Statistics & Data Analysis.* 2009;53(11):3735-3745.

83. Choi J, Liu C, Woo J. An efficient approach for removing lookahead bias in the least square monte carlo algorithm: Leave-one-out. *arXiv.* 2018.

84. Juba B, Le HS. Precision-Recall versus Accuracy and the Role of Large Data Sets. Paper presented at: The Thirty Third AAAI Conference on Artificial Intelligence2019; Honolulu, Hawaii.

85. Warner M, Hedegaard H. Identifying Opioid Overdose Deaths Using Vital Statistics Data. *American journal of public health.* 2018;108(12):1587-1589.

86. Williams KE, Freeman MD, Mirigian L. Drug Overdose Surveillance and Information Sharing via a Public Database: The Role of the Medical Examiner/Coroner. *Academic Forensic Pathology.* 2017;7(1):60-72.

87. ASTHO. *Improving Drug Specificity and Completeness on Death Certificates for Overdose Deaths: Opportunities and Challenges for States. Stakeholder Meeting Report.* Association of State and Territorial Health Officials; Feb. 23, 2018 2018.

88. NCHS. *Medical examiners' and coroners' handbook on death registration and fetal death reporting. 2003 revision.* National Center for Health Statistics2003.

89. Ruiz L, Posey BM, Neuilly MA, Stohr MK, Hemmens C. Certifying Death in the United States. *J Forensic Sci.* 2017.

90. NCHS. *U.S. Standard Certificate of Death. 2003 revision.* National Center for Health Statistics2003.

91. NCHS. *Instructions for Classifying the Underlying Cause-of-Death, ICD-10, 2017.* National Center for Health Statistics2017.

92. NCHS. *Instructions for Classifying the Multiple Causes of Death, ICD-10, 2017.* National Center for Health Statistics2017.

93. WHO. International classification of diseases, tenth revision: version 2016. World Health Organization. http://apps.who.int/classifications/icd10/browse/2016/en. Published 2016. Accessed.

94. WHO. *International statistical classification of diseases and related health problems. 10th revision. Volume 2. Instruction manual.* 2011.

95. ISW7. *Injury Surveillance Workgroup 7. Consensus recommendations for national and state poisoning surveillance. .* http://www.safestates.org/?page=ISWReports The Safe States Alliance; April 2012 2012.

96. Hedegaard H, Miniño A, Warner M. *Drug Overdose Deaths in the United States, 1999–2017.* November 2018 2018.

97. Spencer M, Ahmad F. *Timeliness of Death Certificate Data for Mortality Surveillance and Provisional Estimates.* National Center for Health Statistics; December 2016 2016.

98. CDC. Enhanced State Opioid Overdose Surveillance. Centers for Disease Control and Prevention. https://www.cdc.gov/drugoverdose/foa/state-opioid-mm.html. Published 2018. Accessed.

99. Rios A, Kavuluru R. Few-Shot and Zero-Shot Multi-Label Learning for Structured Label Spaces. Paper presented at: Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing. 2018.

100. Simpson MS, Demner-Fushman D. Biomedical text mining: a survey of recent progress. In: *Mining text data.* Springer; 2012:465-517.

101. Demner-Fushman D, Chapman WW, McDonald CJ. What can natural language processing do for clinical decision support? *Journal of biomedical informatics.* 2009;42(5):760-772.

102. Warner M, Paulozzi LJ, Nolte KB, Davis GG, Nelson LS. State Variation in Certifying Manner of Death and Drugs Involved in Drug Intoxication Deaths. *Academic Forensic Pathology.* 2013;3(2):231-237.

103. Pedregosa F, Varoquaux G, Gramfort A, et al. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research.* 2011;12(Oct):2825-2830.

104. Kuhn M. Caret package. *Journal of Statistical Software.* 2008;28(5):1-26.

105. Witten IH, Frank E. Data Mining: Practical Learning Tools and Techniques with Java Implementations. In. San Diego, CA: Academic Press; 2000.

106. Hsu C-W, Chang C-C, Lin C-J. A practical guide to support vector classification. 2010.

107. Riedl B, Than N, Hogarth M. Using the UMLS and Simple Statistical Methods to Semantically Categorize Causes of Death on Death Certificates. *AMIA Annual Symposium proceedings AMIA Symposium.* 2010;2010:677-681.

108. Duarte F, Martins B, Pinto CS, Silva MJ. A Deep Learning Method for ICD-10 Coding of Free-Text Death Certificates. 2017; Cham.

109. Gerace E, Salomone A, Vincenti M. Analytical approaches in fatal intoxication cases involving new synthetic opioids. *Current Pharmaceutical Biotechnology.* 2018;19(2):113-123.

110. *spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing* [computer program]. 2017.

111. NCHS. Redacted Death Certificate Literal Text File. In: Prevention CfDCa, ed2019.

112. Centers for Disease Control and Prevention. Death investigation--United States, 1987. *MMWR.* 1989;38(1):1.

113. Davis GG. National Association of Medical Examiners Position Paper: Recommendations for the Investiation, Diagnosis, and Certification of Deaths Related to Opioid Drugs. *Academic Forensic Pathology.* 2013;3(1):77-83.

114. Slavova S, Delcher C, Buchanich JM, Bunn TL, Goldberger BA, Costich JF. Methodological complexities in quantifying rates of fatal opioid-related overdose. *Current epidemiology reports.* 2019;6(2):263-274.

115. Davis J, Sabel J, Wright D, Slavova S. Epi Tool to Analyze Overdose Death Data. Council of State and Territorial Epidemiologists. CSTE Features Web site. https://www.cste.org/blogpost/1084057/211072/Epi-Tool-to-Analyze-Overdose-Death-Data. Published 2015. Accessed 8/25/2020.

116. Yadav V, Bethard S. A survey on recent advances in named entity recognition from deep learning models. *arXiv preprint arXiv:191011470.* 2019.

117. Zeng D, Sun C, Lin L, Liu B. LSTM-CRF for drug-named entity recognition. *Entropy.* 2017;19(6):283.

118. Xu K, Yang Z, Kang P, Wang Q, Liu W. Document-level attention-based BiLSTM-CRF incorporating disease dictionary for disease named entity recognition. *Computers in biology and medicine.* 2019;108:122-132.

119. Luo L, Yang Z, Yang P, et al. An attention-based BiLSTM-CRF approach to document-level chemical named entity recognition. *Bioinformatics.* 2018;34(8):1381-1388.

120. Lopez MM, Kalita J. Deep Learning applied to NLP. *arXiv preprint arXiv:170303091*. 2017.

121. Alshaikhdeeb B, Ahmad K. Biomedical named entity recognition: a review. *International Journal on Advanced Science, Engineering and Information Technology*. 2016;6(6):889-895.

122. *R: A language and environment for statistical computing.* [computer program]. Vienna, Austria: R Foundation for Statistical Computing; 2020.

123. Ratinov L, Roth D. Design Challenges and Misconceptions in Named Entity Recognition. Conference on Computational Natural Language Learning; 2009; Boulder, Colorado.

124. Akbik A, Blythe D, Vollgraf R. Contextual string embeddings for sequence labeling. Paper presented at: International Conference on Computational Linguistics2018; Santa Fe, New Mexico.

125. Pennington J, Socher R, Manning CD. Glove: Global vectors for word representation. Paper presented at: Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)2014.

126. Akbik A, Bergmann T, Vollgraf R. Pooled contextualized embeddings for named entity recognition. Paper presented at: 2019 Conference of the North American Chapter of the Association for Computational Linguistics2019; Minneapolis, Minnesota.

127. Paszke A, Gross S, Massa F, et al. Pytorch: An imperative style, high-performance deep learning library. Paper presented at: Advances in neural information processing systems2019.

128. Hedegaard H, Bastian BA, Trinidad JP, Spencer M, Warner M. Regional differences in the drugs most frequently involved in drug overdose deaths: United States, 2017. *National Vital Statistics Reports.* 2019;68(12).

129. Landis JR, Koch GG. The measurement of observer agreement for categorical data. *biometrics.* 1977:159-174.

130. Fidacaro GA, Jr., Patel P, Carroll G, Bartimus H, Hunter K, Hong R. Do Patients Require Emergency Department Interventions After Prehospital Naloxone? *J Addict Med.* 2019.

131. Jeffery RM, Dickinson L, Ng ND, DeGeorge LM, Nable JV. Naloxone administration for suspected opioid overdose: An expanded scope of practice by a basic life support collegiate-based emergency medical services agency. *J Am Coll Health.* 2017;65(3):212-216.

132. Ising A, Proescholdbell S, Harmon KJ, Sachdeva N, Marshall SW, Waller AE. Use of syndromic surveillance data to monitor poisonings and drug overdoses in state and local public health agencies. *Injury Prevention.* 2016;22(Suppl 1):i43-i49.

133. Koh PW, Liang P. Understanding black-box predictions via influence functions. Paper presented at: International Conference on Machine Learning 2017.

VITA

EDUCATION

University of Kentucky
College of Public Health
- PhD, Epidemiology and Biostatistics Expected Spring 2021
- Dissertation Title: "Enhancing Drug Overdose Mortality Surveillance Through Natural Language Processing And Machine Learning"

Saint Louis University
College for Public Health and Social Justice
- Master of Public Health, Epidemiology May 2016
- Capstone Title: "Racial Disparities and the Effect of Maternal Sexually Transmitted Infections on Adverse Birth Outcomes"

University of Missouri—St. Louis
College of Arts and Sciences
- Bachelor of Science, Biology December 2013
- Pierre Laclede Honors College Certificate

PROFESSIONAL EXPERIENCE

Senior Biostatistician August 2020-Present
Collaborative Studies Coordinating Center
Department of Biostatistics
University of North Carolina at Chapel Hill (Chapel Hill, NC)

Part Time Instructor January 2020-Present
Department of Biostatistics
University of Kentucky College of Public Health (Lexington, KY)
Co-taught and co-developed "BST-535: Databases and SAS Programming"

Statistician April 2019-July 2020
Kentucky Injury Prevention and Research Center
University of Kentucky College of Public Health (Lexington, KY)

Epidemiologist January 2017- March 2019

Kentucky Injury Prevention and Research Center
University of Kentucky College of Public Health (Lexington, KY)

Research Assistant August 2016-December 2016
Department of Epidemiology
University of Kentucky College of Public Health (Lexington, KY)
TA for "CPH-310: Disease Detectives: Epidemiology in Action"

Epidemiology Intern May 2015-December 2015
Missouri Department of Health and Senior Services (St. Louis, MO)

Teaching Assistant August 2015-December 2015
Department of Epidemiology
Saint Louis University College for Public Health and Social Justice (St. Louis, MO)
TA for "EPI-5000: Introduction to Epidemiology"

Graduate Research Assistant August 2014-July 2015
Department of Epidemiology
Saint Louis University College for Public Health and Social Justice (St. Louis, MO)

Associate Competitor Information Analyst December 2014-July 2016
Sigma Aldrich Corporation (St. Louis, MO)

PUBLICATIONS

Published

1. **Ward, P.**, Rock, P., Slavova, S., Young, A., Bunn, T., & Kavuluru, R. (2019). Enhancing Timeliness of Drug Overdose Mortality Surveillance: A Machine Learning Approach. *PLoS ONE*, 14(10): e0223318. https://doi.org/10.1371/journal.pone.0223318.
2. **Ward, P.**, Anderson, J., Gordon, R., Salas, J., & Xaverius, P. (2019). Racial Disparities and the Effect of Maternal Sexually Transmitted Infections on Adverse Birth Outcomes. *Women's Health Science Journal*, 3(1), 1-8.
3. Algarin, B.*, **Ward, P**., Christian, W., Rudolph, A., Holloway, I., & Young, A. (2018). Spatial Distribution of Partner-Seeking Men Who Have Sex With Men Using Geosocial Networking Apps: Epidemiologic Study. *Journal of Medical Internet Research*, 20(5), e173.
4. Qian, Z., Liang, S., Yang, S., Trevathan, E., Huang, Z., Yang, R., Wang, J., Hu, K., Zhang, Y., Vaughn, M., Shen, L., Liu, W., Li, P., **Ward, P**., Yang, L., Zhang, W., Chen, W., Dong, G., Zheng, T., Xu, S., & Zhang, B. (2016). Ambient air pollution and preterm birth: A prospective birth cohort study in Wuhan, China. *International journal of hygiene and environmental health*, 219(2), 195-203.
5. Qin, X. D., Qian, Z., Vaughn, M. G., Huang, J., **Ward, P**., Zeng, X., Zhou, Y., Zhu, Y., Yuan, P., Li, M., Bai, Z., Paul, G., Hao, Y., Chen, W., Chen, P., Dong, G., & Lee,

Y. (2016). Positive associations of serum perfluoroalkyl substances with uric acid and hyperuricemia in children from Taiwan. *Environmental Pollution*, 212, 519-524.

6. Yuan, P., Qian, Z., Vaughn, M., Huang, J., **Ward, P**., Zhu, Y., Qin, X., Zhou, Y., Li, M., Xu, S., Zhang, Y., Bao, W., Hao, Y., Zeng, X., & Dong, G. (2016). Comparison of body mass index with abdominal obesity for identifying elevated blood pressure in children and adolescents: The SNEC study. *Obesity Research & Clinical Practice.*
7. Li, M., Qian, Z., Vaughn, M., Boutwell, B., **Ward, P**., Lu, T., Lin, S., Zhao, Y., Zeng, X., Liu, R., Qin, X., Zhu, Y., Chen, W., & Dong, G. (2015). Sex-specific difference of the association between ambient air pollution and the prevalence of obesity in Chinese adults from a high pollution range area: 33 communities Chinese health study. *Atmospheric Environment*, 117, 227-233.

*Indicates student mentee

PRESENTATIONS

1. **Ward, P**. Enhancing Drug Overdose Surveillance with Machine Learning. Presentation to National Center for Health Statistics, February 2020, Hyattsville, Maryland, online.
2. **Ward, P**. Enhancing Drug Overdose Surveillance with Machine Learning. *Rural Opioid Initiative, Epidemiology, Law, and Policy Workgroup*, November 2019, online.
3. **Ward, P**. Enhancing Drug Overdose Surveillance with Machine Learning. Presentation to National Center for Health Statistics leadership, November 2019, University of Kentucky, Lexington, Kentucky.
4. **Ward, P.**, Rock, P., Slavova, S., Young, A., Bunn, T., & Kavuluru, R. Improving Drug Overdose Surveillance with Machine Learning. *Council of State and Territorial Epidemiologists Annual Meeting*, June 2019, Raleigh, NC.
5. **Ward, P.**, Rock, P., Slavova, S., Young, A., Bunn, T., & Kavuluru, R. Enhancing Timeliness of Drug Overdose Mortality Surveillance: A Machine Learning Approach. *College of Public Health Research Day*, April 2019, University of Kentucky, Lexington, KY.
6. **Ward, P.**, Hargrove, S., Akers, A., Singleton, M., & Bunn, T. Re-emergence of Methamphetamine: Evidence from Kentucky Drug Overdose Mortality Surveillance. *Substance Use Research Day*, March 2019, University of Kentucky, Lexington, KY.
7. Slavova, S., Rock, P., & **Ward, P**. Drug Overdose Surveillance in Kentucky. *BERD Pizza and Pilots*, January 2019, University of Kentucky, Lexington, KY.
8. **Ward, P**., Hargrove, S., Akers, D., & Bunn, T. Trends in amphetamine-class surveillance from Kentucky Drug Overdose Deaths, 2013-2016. *Council of State and Territorial Epidemiologists Annual Meeting*, June 2018, West Palm Beach, FL.
9. **Ward, P**. & Bunn, T. Developing an individual, comprehensive, county-level index to assess the burden of the opioid epidemic. *College of Public Health Research Day*, March 2017, Lexington, KY.
10. **Ward, P**., Anderson, J., Gordon, R., Salas, J., & Xaverius, P. Racial disparities and the effect of maternal sexually transmitted infections on Adverse Birth Outcomes. *Pediatric Science Days*, May 2016, St. Louis, MO.

11. **Ward, P**., Blank, C., Schornheuser, O., & Koch, Z. Driven to distraction: Correlations between teen texting while driving and risk behaviors. *American Public Health Association Annual Meeting*, November 2015, Chicago, IL.
12. **Ward, P**., Blank, C., Schornheuser, O., & Koch, Z. Driven to distraction: Correlations between teen texting while driving and risk behaviors. *SLU Graduate Research Symposium*, April 2015, St. Louis, MO.

FUNDING

1. **1NU17CE924971-01-00** **Bunn (PI)** **9/01/2019-8/30/2022**
   CDC
   Overdose Data to Action
   Role: Drug Overdose Surveillance Statistician
2. **1UM1DA049406-01** **Walsh (PI)** **4/17/2019-3/31/2023**
   NIH
   Kentucky Can Heal (Communities and Networks Helping End Addiction Long-Term)
   Role: Data and Informatics Core
3. **5NU17CE924880-03** **Bunn (PI)** **9/01/2016-8/30/2019**
   CDC
   Enhance State Opioid Overdose Surveillance
   Role: Drug Overdose Epidemiologist

REPORTS

1. **Ward, P**. (2019). Kentucky Drug Overdose Mortality Dashboards. *Kentucky Injury Prevention and Research Center*.
2. **Ward, P.** & Rock, P. (2019). Kentucky Resident Drug Overdose Deaths by County of Residence, 2014-2018. *Kentucky Injury Prevention and Research Center*.
3. **Ward, P**. (2019). Kentucky Drug Overdose Fatality Surveillance System Quarterly Update, Q3/2018. *Kentucky Injury Prevention and Research Center*.
4. **Ward, P**., Hargrove, S., Mitchell, L., & Bunn, T. (2018). Kentucky Drug Overdose Fatality Surveillance System (DOFSS) Highlights, 2017. *Kentucky Injury Prevention and Research Center*.
5. Akers, D. & **Ward, P**. (2018). Kentucky Resident Opioid-Involved Drug Overdose Deaths by County of Residence, 2011-2017. *Kentucky Injury Prevention and Research Center*.
6. Akers, D. & **Ward, P**. (2018). Kentucky Resident Drug Overdose Deaths by County of Residence, 2008-2017. *Kentucky Injury Prevention and Research Center*.
7. Bush, A., Bunn, T., Quesinberry, D., & **Ward, P**. (2018). Kentucky's Drug Overdose Burden, 2016-2017. *Kentucky Injury Prevention and Research Center*.
8. Hargrove, S., **Ward, P**., Mitchell, L., & Bunn, T. (2018). Kentucky Drug Overdose Fatality Surveillance System 2016 Annual Report. *Kentucky Injury Prevention and Research Center*.

9. **Ward, P**. (2018). Kentucky Drug Overdose Fatality Surveillance System Quarterly Update, Q1/2018. *Kentucky Injury Prevention and Research Center*.
10. Akers, D. & **Ward, P**. (2018). Kentucky Resident Drug Overdose Deaths by County of Residence, 2013-2017. *Kentucky Injury Prevention and Research Center*.
11. **Ward, P**. (2018). Update on Kentucky Resident Drug Overdose Deaths. *Kentucky Injury Prevention and Research Center*.
12. Hargrove, S., **Ward P**., & Bunn, T. (2017). Kentucky Drug Overdose Fatality Surveillance System 2015 Annual Report. *Kentucky Injury Prevention and Research Center*.
13. **Ward, P**. (2017). Fatal Drug Overdoses in the Kentucky Military and Veteran Population, 2010-2015. *Kentucky Injury Prevention and Research Center.*
14. Akers, D. & **Ward, P**. (2017). Kentucky Resident Drug Overdose Deaths by County of Residence, 2012-2016. *Kentucky Injury Prevention and Research Center*.

AWARDS AND HONORS

- College of Public Health Research Day Award, Best Doctoral Poster Presentation, April 2019
- Book Award for High Score on PhD Comprehensive Exam, January 2018
- Inducted into Delta Omega, the Honor Society in Public Health, April 2015
- University of Missouri—St. Louis Curator's Scholarship (2010-2013)
- Pierre Laclede Honors College Scholarship (2010-2013)