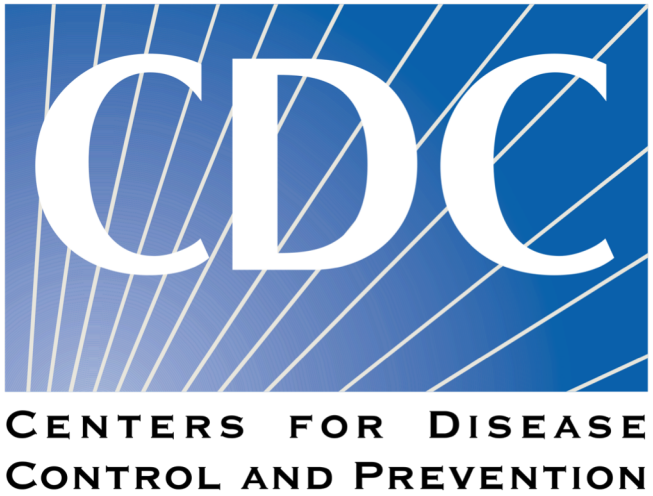
Using Cause-of-Death Literal Text from the Death Certificate for Classification

Anthony Lipphardt

Background

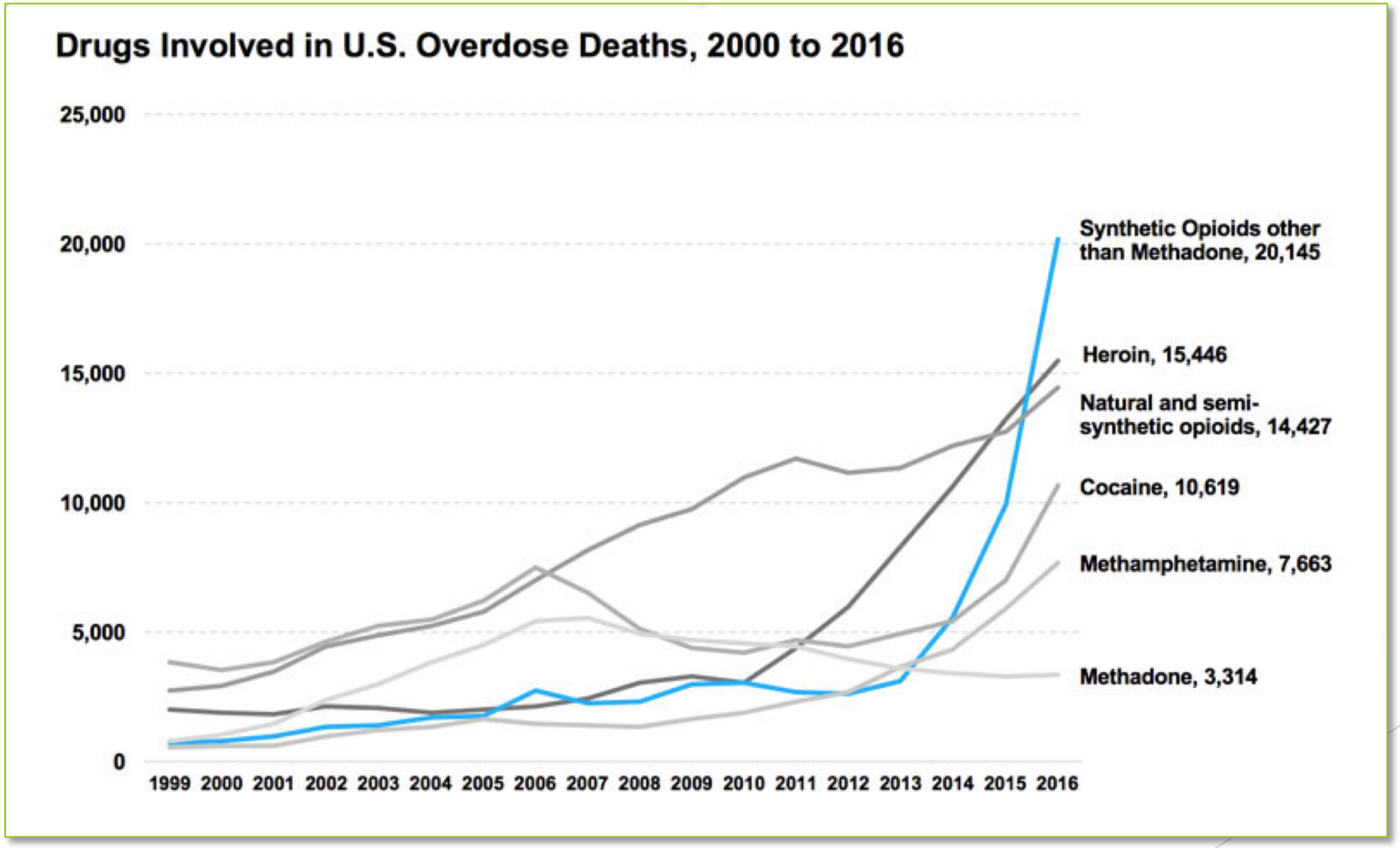
Mortality Surveillance at CDC

- CDC's National Center for Health Statistics
  - Oversees several major surveys and data collection systems to monitor health of the nation.
    - National Vital Statistics System
  - Partners with state and local governments to collect records on vital events (i.e. births, deaths)
  - Data is cleaned, coded, and released to the public annually for analysis.



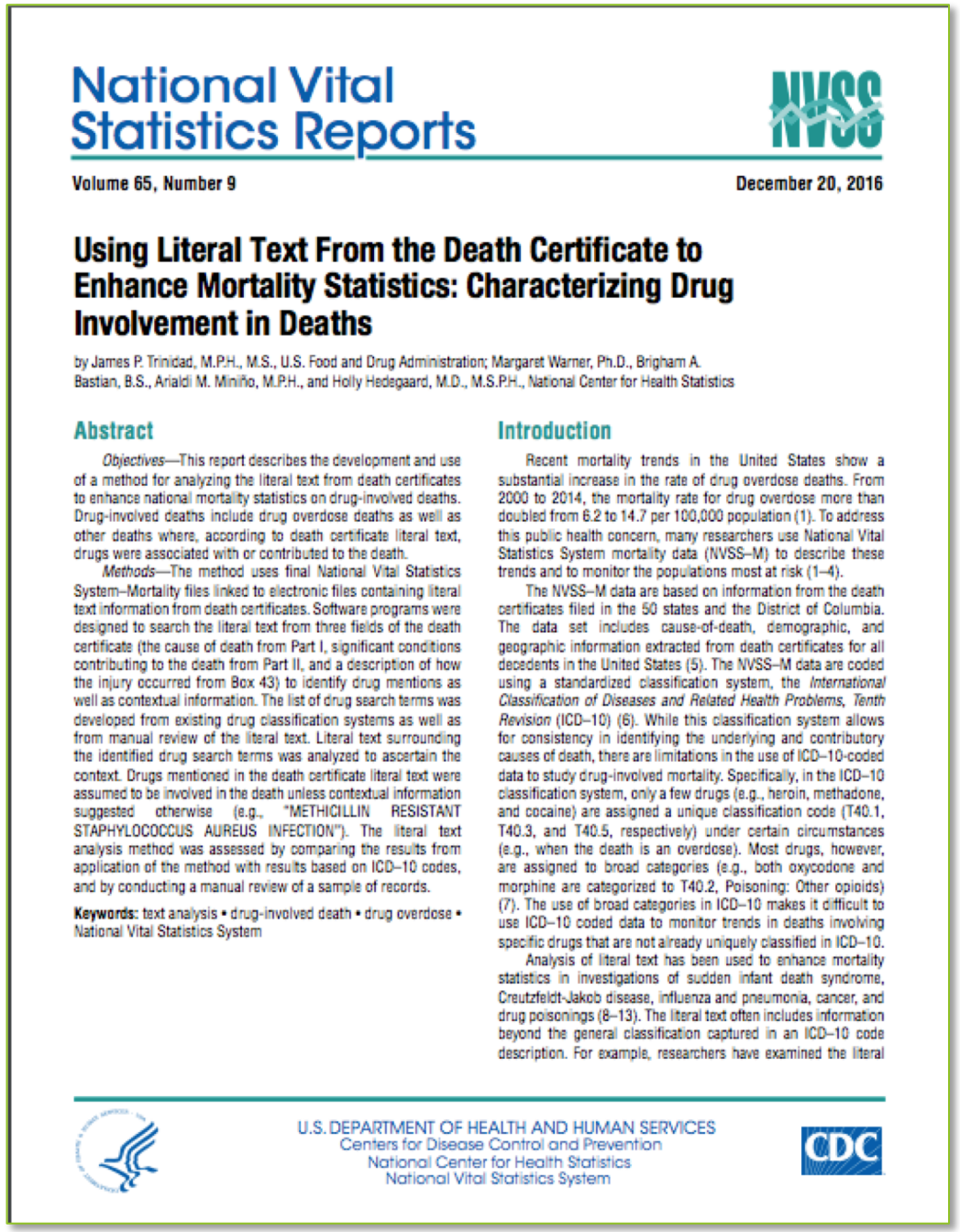
The Opioid Crisis

Deaths in 2016 over **5 times higher** than in 1999



Current Limitations with NCHS Data

- Limited ICD-10 codes for specific drugs
- Time involved in reviewing and manually coding records for drug-involved deaths



National Vital Statistics Reports

Volume 65, Number 9 December 20, 2016

Using Literal Text From the Death Certificate to Enhance Mortality Statistics: Characterizing Drug Involvement in Deaths

by James P. Trinidad, M.P.H., M.S., U.S. Food and Drug Administration; Margaret Warner, Ph.D., Brigham A. Bastian, B.S., Arialdi M. Miniño, M.P.H., and Holly Hedegaard, M.D., M.S.P.H., National Center for Health Statistics

Abstract

Objectives—This report describes the development and use of a method for analyzing the literal text from death certificates to enhance national mortality statistics on drug-involved deaths. Drug-involved deaths include drug overdose deaths as well as other deaths where, according to death certificate literal text, drugs were associated with or contributed to the death.

Methods—The method uses final National Vital Statistics System–Mortality files linked to electronic files containing literal text information from death certificates. Software programs were designed to search the literal text from three fields of the death certificate (the cause of death from Part I, significant conditions contributing to the death from Part II, and a description of how the injury occurred from Box 43) to identify drug mentions as well as contextual information. The list of drug search terms was developed from existing drug classification systems as well as from manual review of the literal text. Literal text surrounding the identified drug search terms was analyzed to ascertain the context. Drugs mentioned in the death certificate literal text were assumed to be involved in the death unless contextual information suggested otherwise (e.g., "METHICILLIN RESISTANT STAPHYLOCOCCUS AUREUS INFECTION"). The literal text analysis method was assessed by comparing the results from application of the method with results based on ICD–10 codes, and by conducting a manual review of a sample of records.

Keywords: text analysis • drug-involved death • drug overdose • National Vital Statistics System

Introduction

Recent mortality trends in the United States show a substantial increase in the rate of drug overdose deaths. From 2000 to 2014, the mortality rate for drug overdose more than doubled from 6.2 to 14.7 per 100,000 population (1). To address this public health concern, many researchers use National Vital Statistics System mortality data (NVSS–M) to describe these trends and to monitor the populations most at risk (1–4).

The NVSS–M data are based on information from the death certificates filed in the 50 states and the District of Columbia. The data set includes cause-of-death, demographic, and geographic information extracted from death certificates for all decedents in the United States (5). The NVSS–M data are coded using a standardized classification system, the *International Classification of Diseases and Related Health Problems, Tenth Revision* (ICD–10) (6). While this classification system allows for consistency in identifying the underlying and contributory causes of death, there are limitations in the use of ICD–10-coded data to study drug-involved mortality. Specifically, in the ICD–10 classification system, only a few drugs (e.g., heroin, methadone, and cocaine) are assigned a unique classification code (T40.1, T40.3, and T40.5, respectively) under certain circumstances (e.g., when the death is an overdose). Most drugs, however, are assigned to broad categories (e.g., both oxycodone and morphine are categorized to T40.2, Poisoning: Other opioids) (7). The use of broad categories in ICD–10 makes it difficult to use ICD–10 coded data to monitor trends in deaths involving specific drugs that are not already uniquely classified in ICD–10.

Analysis of literal text has been used to enhance mortality statistics in investigations of sudden infant death syndrome, Creutzfeldt-Jakob disease, influenza and pneumonia, cancer, and drug poisonings (8–13). The literal text often includes information beyond the general classification captured in an ICD–10 code description. For example, researchers have examined the literal

U.S. DEPARTMENT OF HEALTH AND HUMAN SERVICES
Centers for Disease Control and Prevention
National Center for Health Statistics
National Vital Statistics System

Recent Developments

- Details "Development and use of a method for analyzing the literal text from death certificates to enhance national mortality statistics on drug-involved deaths"
- Goal: Identify drug mentions with involvement (DMI death) and extract specific drugs involved in death.

Literal Text Fields from the Death Certificate

CAUSE OF DEATH (See instructions and examples)

32. **PART I.** Enter the chain of events--diseases, injuries, or complications--that directly caused the death. DO NOT enter terminal events such as cardiac arrest, respiratory arrest, or ventricular fibrillation without showing the etiology. DO NOT ABBREVIATE. Enter only one cause on a line. Add additional lines if necessary.

IMMEDIATE CAUSE (Final disease or condition ---------> resulting in death) a. _____

Due to (or as a consequence of):

Sequentially list conditions, if any, leading to the cause listed on line a. Enter the **UNDERLYING CAUSE** (disease or injury that initiated the events resulting in death) **LAST**

b. _____

Due to (or as a consequence of):

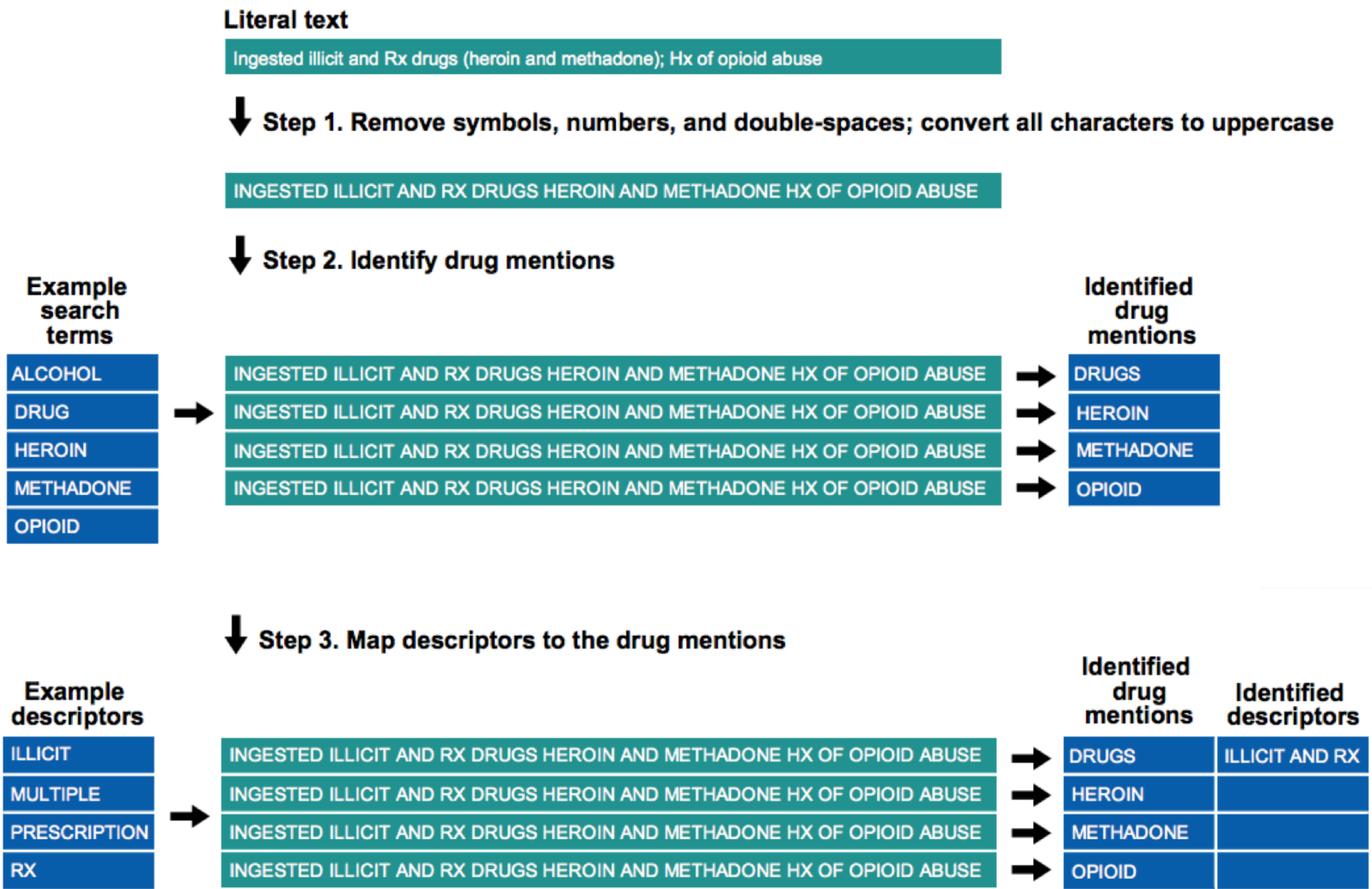
c. _____

Due to (or as a consequence of):

d. _____

PART II. Enter other significant conditions contributing to death but not resulting in the underlying cause given in PART I

43. DESCRIBE HOW INJURY OCCURRED:


Literal text
Ingested illicit and Rx drugs (heroin and methadone); Hx of opioid abuse
Step 1. Remove symbols, numbers, and double-spaces; convert all characters to uppercase
INGESTED ILLICIT AND RX DRUGS HEROIN AND METHADONE HX OF OPIOID ABUSE
Step 2. Identify drug mentions
Example search terms
ALCOHOL
DRUG
HEROIN
METHADONE
OPIOID
INGESTED ILLICIT AND RX DRUGS HEROIN AND METHADONE HX OF OPIOID ABUSE
INGESTED ILLICIT AND RX DRUGS HEROIN AND METHADONE HX OF OPIOID ABUSE
INGESTED ILLICIT AND RX DRUGS HEROIN AND METHADONE HX OF OPIOID ABUSE
INGESTED ILLICIT AND RX DRUGS HEROIN AND METHADONE HX OF OPIOID ABUSE
Identified drug mentions
DRUGS
HEROIN
METHADONE
OPIOID
Step 3. Map descriptors to the drug mentions
Example descriptors
ILLICIT
MULTIPLE
PRESCRIPTION
RX
INGESTED ILLICIT AND RX DRUGS HEROIN AND METHADONE HX OF OPIOID ABUSE
INGESTED ILLICIT AND RX DRUGS HEROIN AND METHADONE HX OF OPIOID ABUSE
INGESTED ILLICIT AND RX DRUGS HEROIN AND METHADONE HX OF OPIOID ABUSE
INGESTED ILLICIT AND RX DRUGS HEROIN AND METHADONE HX OF OPIOID ABUSE
Identified drug mentions
Identified descriptors
DRUGS
ILLICIT AND RX
HEROIN
METHADONE
OPIOID

Step 3 also identifies complex descriptions (e.g., "ILLICIT AND RX") by linking descriptors (e.g., "ILLICIT" and "RX") with joining phrases (e.g., "* AND *")

Step 4. Replace (consecutive) drug mentions and associated descriptors with a single asterisk ("*")

INGESTED ILLICIT AND RX DRUGS HEROIN AND METHADONE HX OF OPIOID ABUSE

consecutive drug mentions and associated descriptors — drug mention

	Identified drug mentions	Identified descriptors
INGESTED * HX OF * ABUSE	DRUGS	ILLICIT AND RX
INGESTED * HX OF * ABUSE	HEROIN	
INGESTED * HX OF * ABUSE	METHADONE	
INGESTED * HX OF * ABUSE	OPIOID	

Step 5. Identify and map contextual phrases to the appropriate drug mention(s)

Example contextual phrase		Identified drug mentions	Identified descriptors	Identified contextual phrase
* POISONING	INGESTED * HX OF * ABUSE	DRUGS	ILLICIT AND RX	INGESTED *
ABUSED *	INGESTED * HX OF * ABUSE	HEROIN		INGESTED *
HX OF * ABUSE	INGESTED * HX OF * ABUSE	METHADONE		INGESTED *
INGESTED *	INGESTED * HX OF * ABUSE	OPIOID		HX OF * ABUSE

Level of Precision with Current Approach

Table G. DMI programs' ability to identify DMI deaths among a random sample of 2,000 deaths having one or more ICD–10 entity axis codes or identified using the DMI programs: U.S. residents, 2013

Evaluation	DMI deaths identified from the manual review		
	Yes	No	Total
DMI deaths identified by the DMI programs	1,804	79	1,883
DMI deaths not identified by the DMI programs	17	100	117

NOTES: See Figure 4 for list of entity axis codes. Positive predictive value calculated as: 1,804 deaths/1,883 deaths = 95.8%. DMI is a drug mentioned with involvement in the death.

SOURCE: NCHS, National Vital Statistics System, Mortality files linked with death certificate literal text.

95.8% positive predictive rate (i.e. Precision)

Framing our Problem

The Question

Can machine learning techniques be used to train a classifier to predict with high precision and specificity whether a death record is a drug mention with involvement (DMI) death or not?

What Does Success Look Like?

- Exceed our baseline of 95.8 percent precision
- Benefits
  - Improve identification of DMI deaths
  - Generalized solution removes need for exhaustive lists of search terms

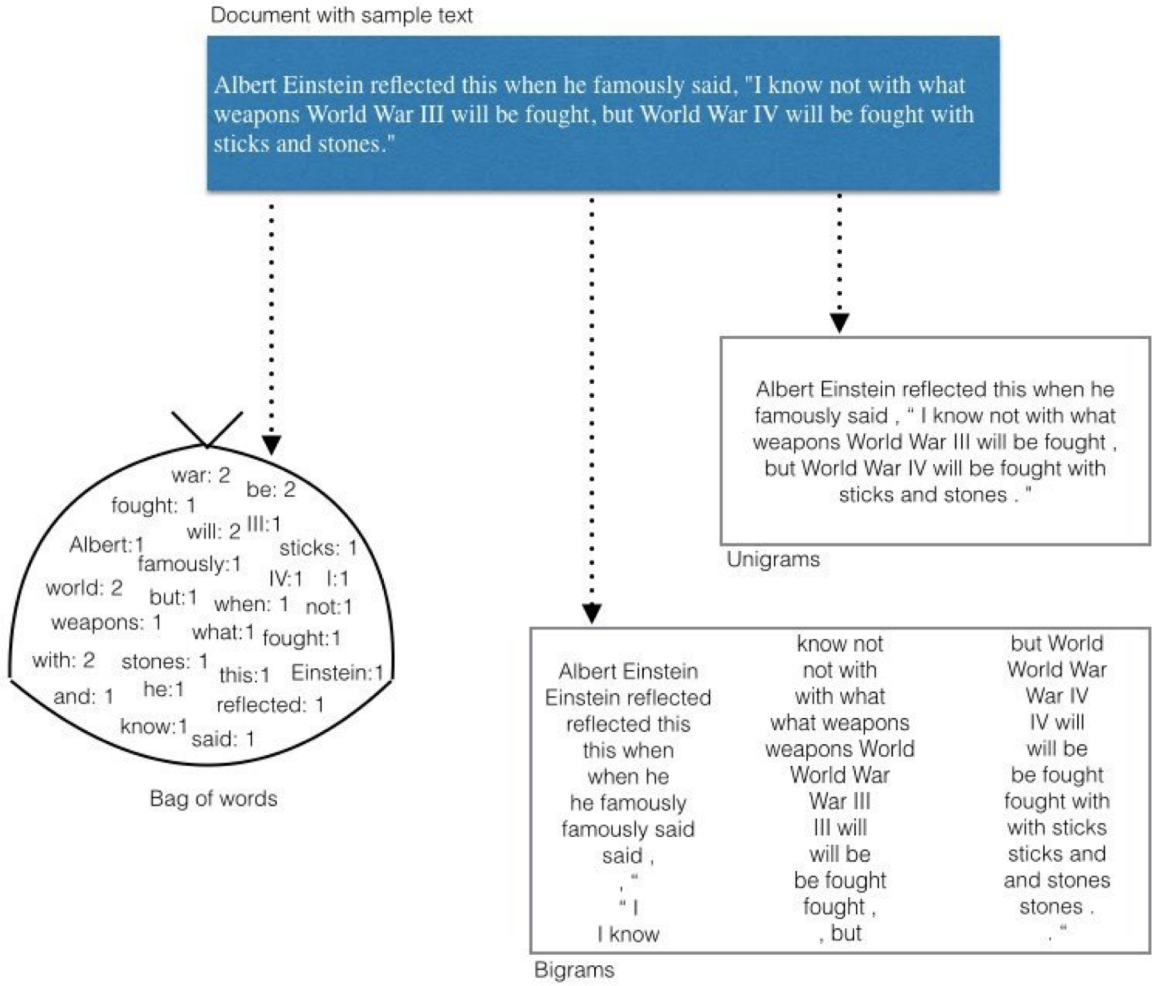
Data

- 2015-16 annual and cause of death literal text files from Washington State
- Annual files – More than 130 fields, including:
  - Certificate number
  - Underlying cause of death (ICD-10 Code)
  - Multiple Cause of Death (ICD-10 Code, up to 20)
- Literal text files
  - State File Number (i.e. certificate number)
  - Cause of Death – Line A/B/C/D
  - Interval Time – Line A/B/C/D
  - Other Significant Conditions
  - How Injury Occurred
  - Place of Injury

Applicable Machine Learning Techniques

- Bag-of-words Model/Term Frequency Matrix
- Support Vector Machine
  - Latent Semantic Analysis
- Naïve Bayes

Bag-of-Words Model

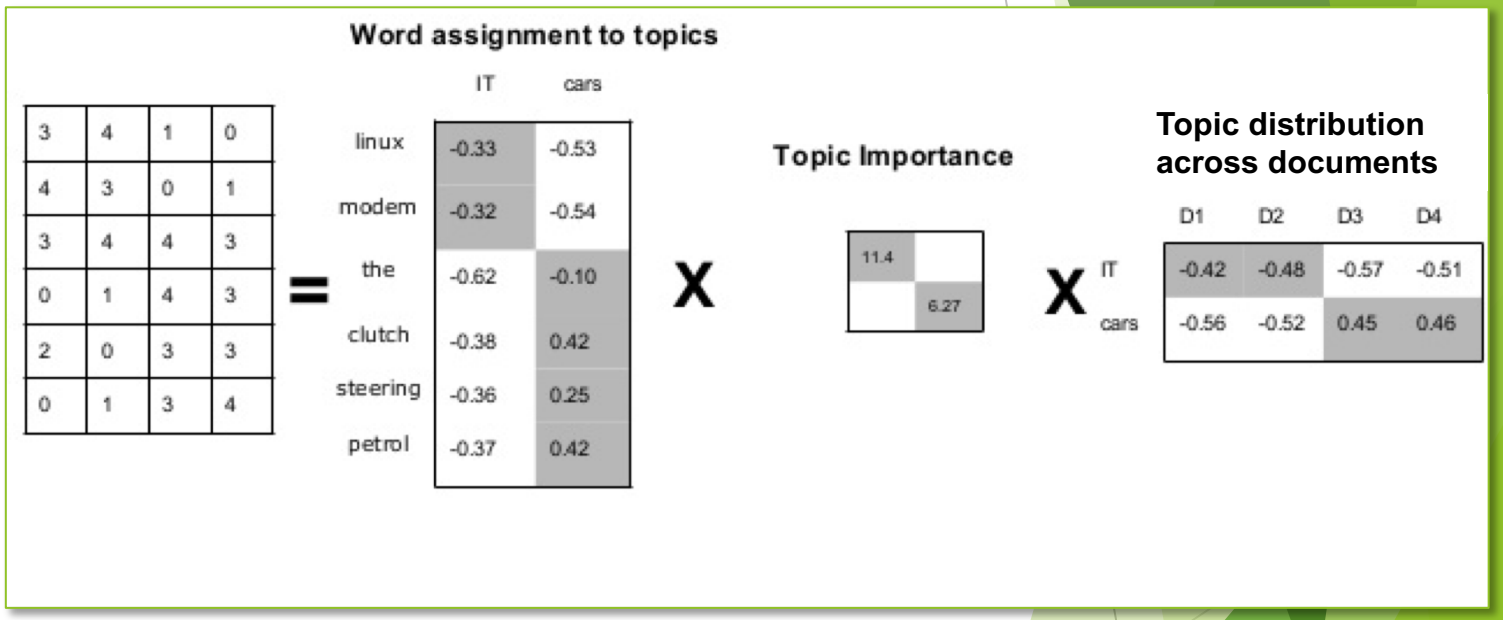


	DMI	literal text
0	0	GASTRIC CARCINOMA
1	0	SMALL CELL CARCINOMA PROSTATE METASTATIC PROST...
2	0	RESPIRATORY FAILURE METASTATIC ANAL SQUAMOUS C...
3	0	ASPIRATION PNEUMONIA ALZHEIMER DEMENTIA
4	0	PULMONARY EDEMA END STAGE RENAL FAILURE STOPPE...
5	0	CIRRHOSIS LIVER HEPATITIS C SEPSIS
6	0	CONGESTIVE HEART FAILURE CONGESTIVE CARDIOMYOP...
7	0	ACUTE RESPIRATORY FAILURE ASPIRATION PNEUMONIA...
8	0	METASTATIC LUNG CANCER
9	0	LUNG CANCER END STAGE CHRONIC OBSTRUCTIVE LUNG...
10	0	CEREBROVASCULAR ACCIDENT HYPERTENSION DIABETES

	AAA	ABCESS	ABDOMEN	ABDOMINAL	ABILITY	ABLATION
0	0	0	0	0	0	0
1	0	0	0	0	0	0
2	0	0	0	0	0	0
3	0	0	0	0	0	0
4	0	0	0	0	0	0

Latent Semantic Analysis (LSA)

- Similar to PCA:
  - Approximates original term-frequency matrix using singular value decomposition.
  - Choosing top k eigenvalues and eigenvectors retains a certain level of explained variance.
- Advantage:
  - Works well with sparse data such as term-frequency matrices.



Pre-processing: Data Cleaning

- Rename fields to conform to standard labels
- Drop irrelevant fields
- Link annual and literal text files
- Remove rows with missing literal text
- Remove rows with missing underlying and multiple causes of death
- Strip punctuation and numbers from literal text fields
- Balancing dataset through undersampling (100,000 records → 5,000 records)
- No standardization

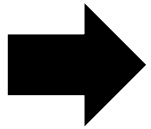
	State file number	Sex	Date of Death	Date of Death month	Date of Death day	Date of Death year	Race	Age Unit	Age	Hispanic NCHS bridge	...	ACME nature of injury flag 11	ACME nature of injury flag 12	ACME nature of injury flag 13	ACME nature of injury flag 14	ACME nature of injury flag 15
0	2016000001	M	1/1/2016	1	1	2016	1	1	87	0.0	...	NaN	NaN	NaN	NaN	NaN
1	2016000002	F	1/1/2016	1	1	2016	1	1	52	0.0	...	NaN	NaN	NaN	NaN	NaN
2	2016000003	M	1/3/2016	1	3	2016	1	1	52	0.0	...	NaN	NaN	NaN	NaN	NaN
3	2016000004	M	1/2/2016	1	2	2016	1	1	69	0.0	...	NaN	NaN	NaN	NaN	NaN
4	2016000005	M	1/2/2016	1	2	2016	1	1	82	0.0	...	NaN	NaN	NaN	NaN	NaN

	State file number	Cause of Death - Line A	Cause of Death - Line B	Cause of Death - Line C	Cause of Death - Line D	Interval Time - Line A	Interval Time - Line B	Interval Time - Line C	Interval Time - Line D	COD-DUE-TO-B	COD-DUE-TO-C	COD-DUE-TO-D
0	2015000001	GASTRIC CARCINOMA	NaN	NaN	NaN	YEARS	NaN	NaN	NaN	NaN	NaN	NaN
1	2015000002	SMALL CELL CARCINOMA OF THE PROSTATE	METASTATIC PROSTATE CANCER	NaN	NaN	15 MONTHS	7 YEARS	NaN	NaN	NaN	NaN	NaN
2	2015000003	RESPIRATORY FAILURE	METASTATIC ANAL SQUAMOUS CELL CARCINOMA	NaN	NaN	2 MINUTES	13 MONTHS	NaN	NaN	NaN	NaN	NaN

Pre-processing: Feature Engineering

- Add Year field to assist linkage (2015 or 2016)
- DMI field added as target label (1 for DMI, 0 for Non-DMI)
  - Check underlying cause of death and 20 multiple cause fields. If code is in the list of known drug overdose deaths (i.e., X40–X44, X60–X64, X85, or Y10–Y14), flag as a DMI Death
- Convert combined literal text field to term frequency matrix

	DMI	literal text
0	0	GASTRIC CARCINOMA
1	0	SMALL CELL CARCINOMA OF THE PROSTATE METASTATI...
2	0	RESPIRATORY FAILURE METASTATIC ANAL SQUAMOUS C...
3	0	ASPIRATION PNEUMONIA ALZHEIMER'S DEMENTIA
4	0	PULMONARY EDEMA END STAGE RENAL FAILURE- STOPP...

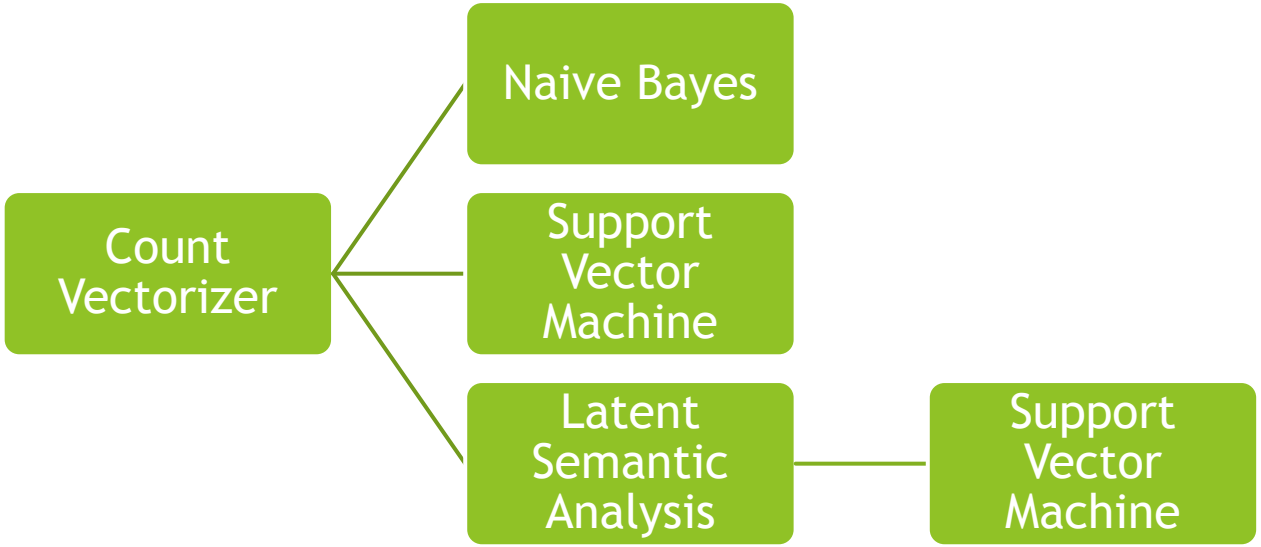


ABNORMALITY	...	YARD	YEAR	YEARS	YIELD	YORK	YRS	ZOLPIDEM	ZONE	ZOSTER	DMI
0	...	0	0	0	0	0	0	0	0	0	0.0
0	...	0	0	0	0	0	0	0	0	0	0.0
0	...	0	0	0	0	0	0	0	0	0	0.0
0	...	0	0	0	0	0	0	0	0	0	0.0
0	...	0	0	0	0	0	0	0	0	0	0.0

Experimental Plan

Basic Outline

- Split data into training and test data with 75:25 split.
- Three distinct workflows setup using Scikit-learn pipeline construct.
- Each workflow performs 10-fold cross validation using each combination of parameters from a supplied parameter grid.
- Optimal parameters are chosen using precision as our scoring metric.
- Conduct final test using optimal classifier and parameters.



CountVectorizer

- lowercase disabled as all records are in all caps
- Set term frequencies as 0 or 1
- Words must appear at least 3 or 5 times in the entire dataset in order to be considered.
- Unigrams (individual words) and bigrams (adjacent word pairs) will be included in term-frequency matrix.

Combination	lowercase	binary	min_df	ngram_range
1	False	True	3	(1,1)
2	False	True	5	(1,1)
3	False	True	3	(1,2)
4	False	True	5	(1,2)

Naïve Bayes

- Bernoulli Naïve Bayes classifier used with binary term-frequency matrix.
- binarize set to None as data is already converted to binary counts.
- Alpha parameter controls level of smoothing applied to account for features in the test set with probability of zero.

Combination	binarize	alpha
1	None	0
2	None	0.1
3	None	0.5
4	None	1

Support Vector Machine

- In text applications, data may often be linearly separable due to large size and dimensionality of dataset. Therefore, a linear kernel may be appropriate.
- Several test values for C, which controls the error penalty used when misclassifying an example.

Combination	Kernel	C
1	Linear	1
2	Linear	10
3	Linear	100
4	Linear	1000

Latent Semantic Analysis

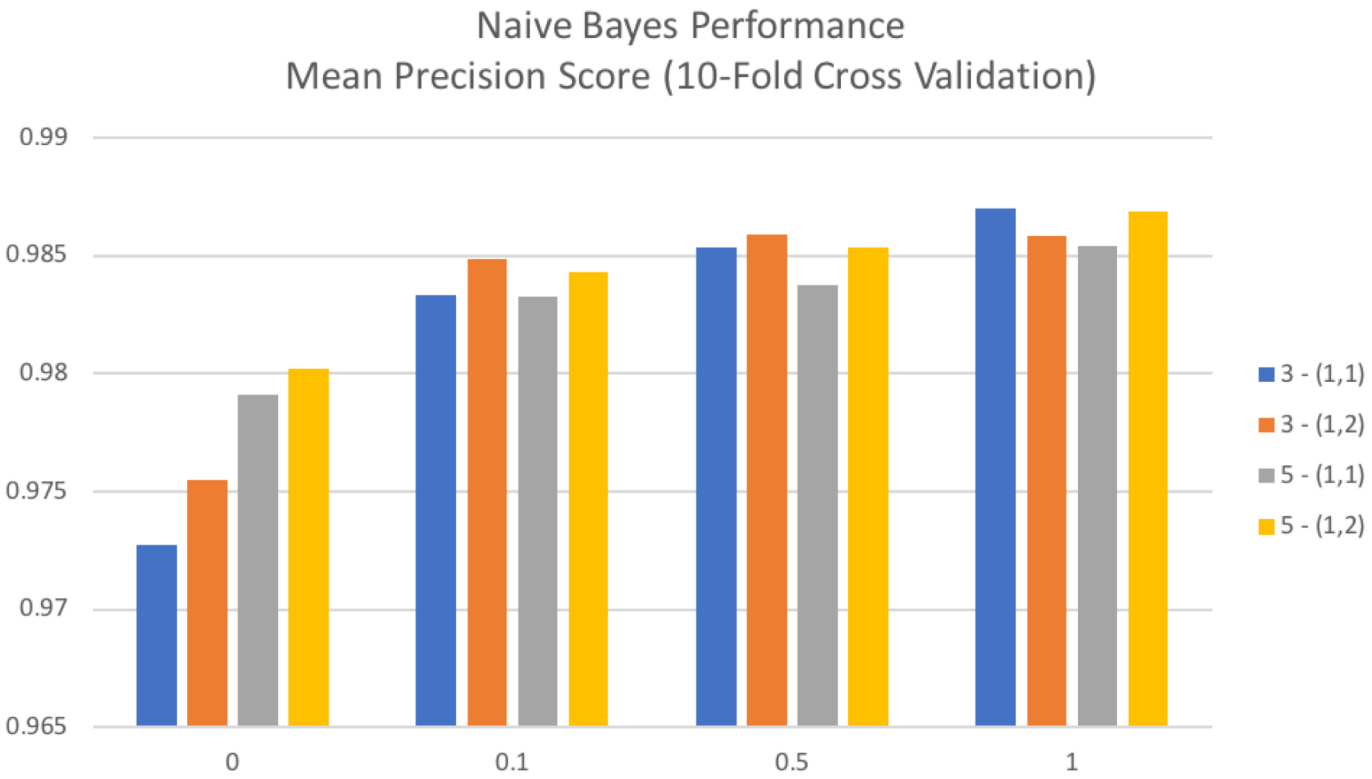
- With LSA, we must specify the number of components that will be selected.
- Number of components here are selected to retain approximately 85, 90, and 95 percent of explained variance in dataset.
- Numbers obtained by fitting training CountVectorizer with optimal parameters based on Naïve Bayes and SVM pipelines.

Combination	n_components
1	525
2	775
3	1180

Results of Grid Search

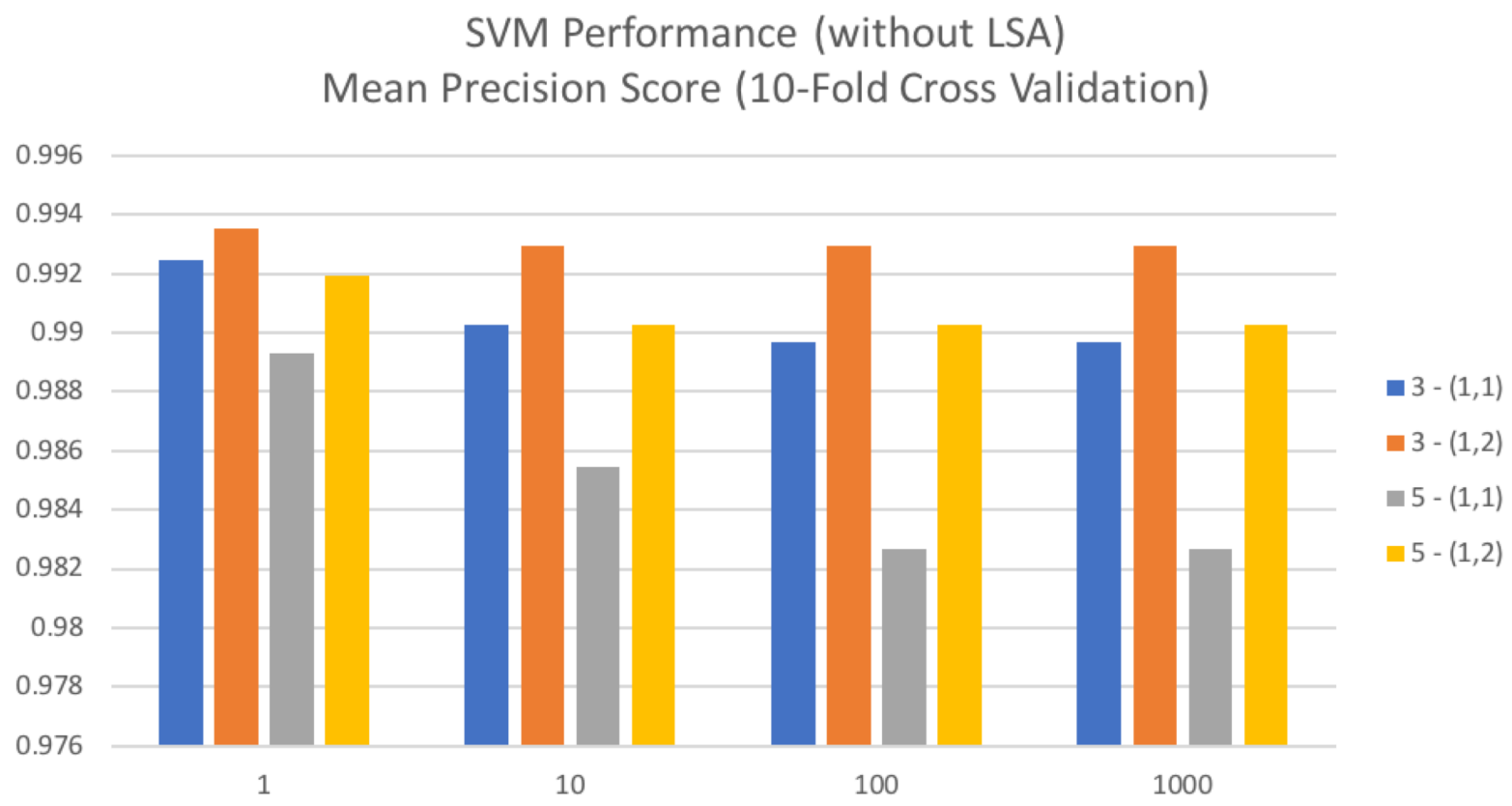
Naïve Bayes

Higher alpha parameter and bigrams → Higher precision



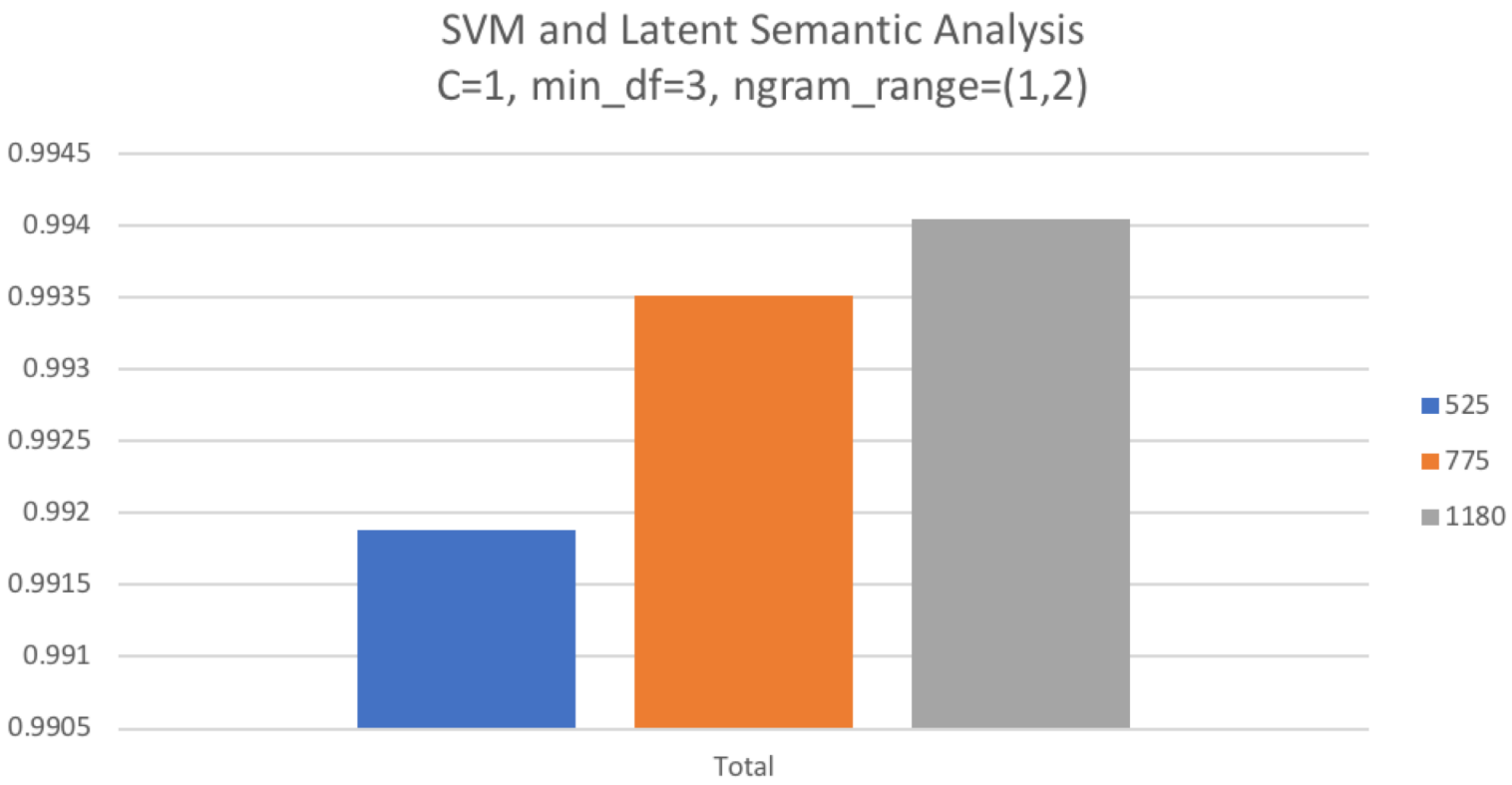
SVM without LSA

Lower C parameter, min_df=3, including bigrams → Higher precision

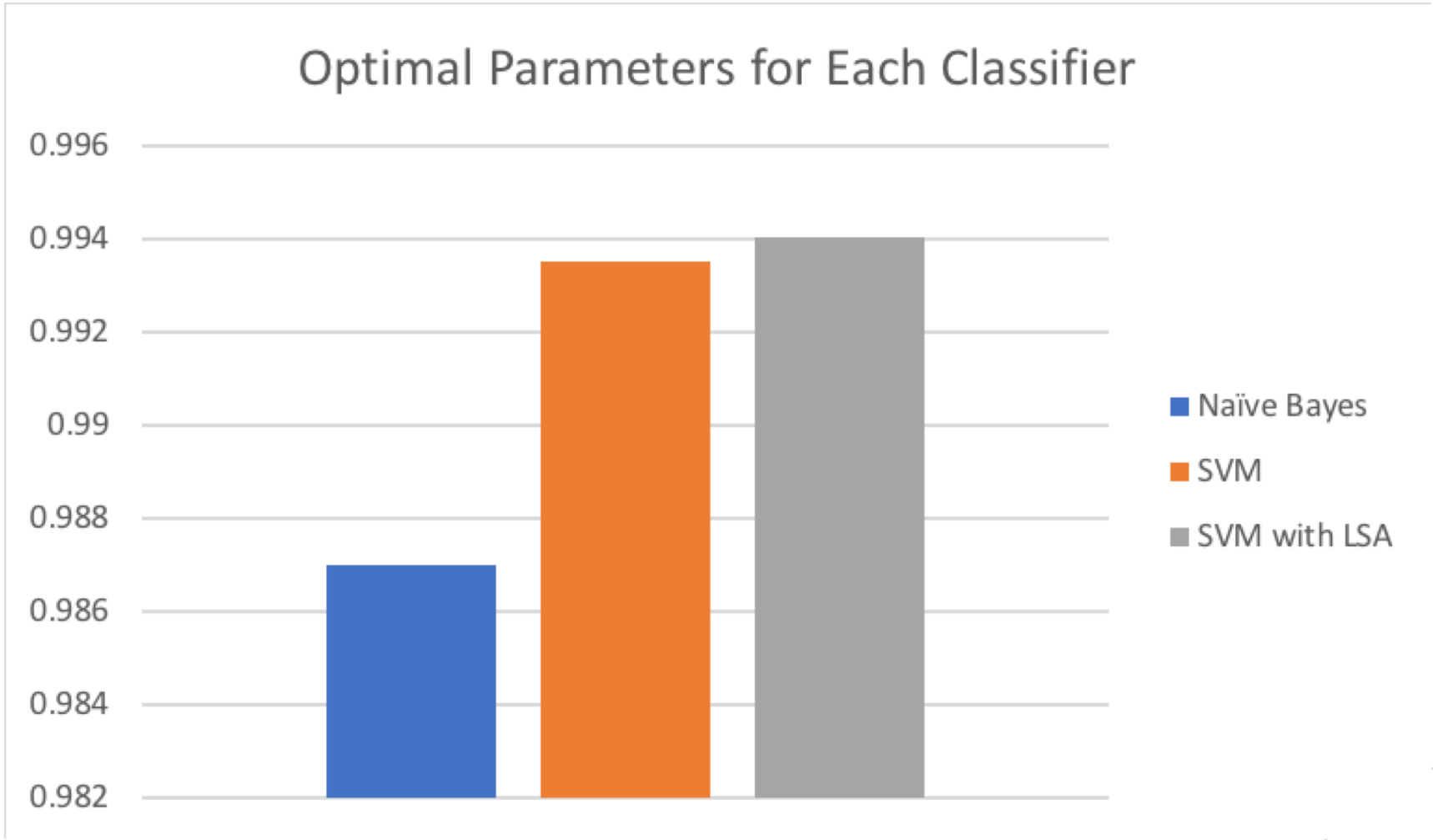


SVM with LSA

More explained variance → Higher precision



Comparing Three Workflows



Performance with Test Data

Predicting on the Test Data

- Fitting models on our training data with optimal parameters, we transform the test data and obtain a prediction for each record.
- Precision for the DMI class was approximately 99 percent.
- Specificity (i.e., Recall for the negative class in the binary case) was approximately 99 percent.

	Precision	Recall	F1 Score	Support
Non-DMI	0.98	0.99	0.98	621
DMI	0.99	0.98	0.98	627
Avg/Total	0.98	0.98	0.98	1248

Confusion Matrix

- For the Non-DMI class, only 8 records misclassified as DMI
- For the DMI class, only 15 records misclassified as Non-DMI

Actual\Predicted	Non-DMI	DMI
Non-DMI	613	8
DMI	15	612

RECORDS MISCLASSIFIED AS DMI
==========

APPARENT SUDDEN DEATH DUE DIFLUOROETHANE INHALATION HUFFING DECEASED FOUND DEAD BEDROOM FLOOR MULTIPLE CANS COMPRESSED AIR FOUND NEAR BODY

CERVICAL SPINE FRACTURE BLUNT IMPACT HEAD FALLEN HITTING HEAD BATHROOM WAL L ACUTE ETHANOL INTOXICATION

ASPHYXIA SUFFOCATION PLASTIC BAG INTENTIONALLY PLACED PLASTIC BAG HEAD SEC URED KNOT

ACUTE RESPIRATORY FAILURE ACUTE CHRONIC AORTIC ILIAC THROMBUS ISCHEMIC FOO T BILATERAL METASTASIS BONE LIVER GENRALIZED WEAKNESS DEHYDRATION SUSPECTE D PORTAL VEIN THROMBOSIS METABOLIC ACIDOSIS HYPONATREMIA SYSTEMIC INFLAMMA TORY RESPONSE SYNDROME TRANSAMINITIS MODERATE PROTEIN CALORIE MALNUTRITION

ANOXIC ENCEPHALOPATHY ACUTE ETHANOL INTOXICATION INGESTION LARGE QUANTITIE S ALCOHOL BULLOUS EMPHYSEMA CHRONIC ETHANOLISM

SELF INFLICTED HANDGUN GUNSHOT WOUND HEAD DECEDENT TOOK LIFE FIRING MM HAN DGUN BULLET HEAD SUICIDE NOTE RECOVERED SUICIDAL IDEATION PRIOR ATTEMPTS R EPORTED DEPRESSION ALCOHOLISM

ASPHYXIA DUE PLASTIC BAG HEAD FRESHWATER DROWNING FOUND CULVERT PLASTIC BA G HEAD

TOXIC ASPHYXIA INHALATION CARBON MONOXIDE ENGINE EXHAUST RAN VEHICLE ENCLO SED GARAGE CHRONIC DRUG ABUSE METHAMPHETAMINE

RECORDS MISCLASSIFIED AS NON-DMI

PULMONARY HEMORRHAGE ANTICOAGULATION COUMADIN THERAPY ATRIAL FIBRILLATION LEUKEMIA CONGESTIVE HEART FAILURE

DROWNING DECEDENT DROWN JACUZZI TUB ELEVATED GABAPENTIN BLOOD CONCENTRATION

RESPIRATORY ARREST SEPSIS DUE ASPIRATION PNEUMONITIS PNEUMONIA SEVERE CHRONIC OBSTRUCTIVE PULMONARY DISEASE ACUTE CHRONIC HYPOXEMIC HYPERCAPNIC RESPIRATORY FAILURE ENCEPHALOPATHY SECONDARY TOXIC MEDICATIONS HISTORY SEVERE PROTEIN CALORIE MALNUTRITION

ANATOMICAL CAUSE DEATH UNKNOWN EXTENT DIPHENHYDRAMINE CONTRIBUTED DEATH SUPRA THERAPEUTIC DIPHENHYDRAMINE LEVEL

ACUTE RENAL FAILURE PROBABLY MEDICATION NEPHROTOXICITY DEMENTIA H PYLORI INFECTION

ASPHYXIA FRESHWATER DROWING FOLLOWING INJECTED HEROIN DEMTHAMPHETAMINE USAGE DISCOVERED SUBMERGED BATHTUB WITHOUT INFLICTED INJURIES SUBSTANCE ABUSE MANY YEARS

MASSIVE HEMORRHAGIC SHOCK AORTOENTERIC FISTULA ESOPHAGEAL ULCER BOTOX INJECTION ESOPHAGUS MEDICAL PROCEDURE HYPERTENSION ESOPHAGEAL SPASM

HEMORRHAGIC STROKE ANTICOAGULATED ATRIAL FIBRILLATION MULTIPLE RECENT REMOTE GROUND LEVEL FALLS

ANOXIC BRAIN INJURY SMOKING HEROIN RESPIRATORY ARREST SMOKING HEROIN DOWNHILL COURSE COMPLICATIONS SUBSTANCE ABUSE MONTHS ADDICTION WEIGHT LOSS EMPHYSEMA

ASPHYXIA DUE EXCLUSION OXYGEN HELIUM FILLED BAG HEAD NECK SECURED PLASTIC BAG HEAD CONNECTED HELIUM TANKS

Transparency of Methods and Results

- Detailed data file layouts and documentation are made available for NCHS and Washington State datasets.
- Python scripts and annotated Jupyter notebooks are made available in the GitHub repository to reproduce the experiment.
- Perceived risks and limitations addressed:
  - Potential for classifier to misclassify Non-DMI deaths as DMI. Baseline set by current NCHS programs are 95.8% precision. Classifier exceeds baseline with precision of 99%.
  - Limitations discussed in the problem statement is the use of ICD-codes in underlying cause of death and multiple cause of death to set DMI targets. High-quality, manually reviewed labels would be preferable.

Conclusions

Recap

- Annual and cause-of-death literal text files from Washington State were linked. Records were flagged as DMI or Non-DMI.
- Three workflows were tested using Naïve Bayes, SVM, and SVM with Latent Semantic Analysis (LSA).
- SVM with LSA was used to achieve 99% precision and specificity for the DMI class on the test data, exceeding 95.8% precision of the DMI program.

Next Steps

- Train classifier on NCHS ‘gold standard’ data (i.e. our randomly sampled, manually reviewed death records)
- Examine whether this technique can be used to assign specific ICD codes with high precision.
- Research additional techniques in Natural Language Processing that may be used to reduce the number of terms in our dataset. UMLS thesaurus from NIH is one potential option for mapping synonyms or variants to biomedical concepts.