Selection of Clinical Text Features for Classifying Suicide Attempts

By
Ryan Buckland
B.A., The College of William and Mary, 2013

Thesis
Submitted in partial fulfillment of the requirements for the
Degree of Master of Science in the Department of Biostatistics at Brown University

PROVIDENCE, RHODE ISLAND
MAY 2020

This thesis by Ryan Buckland is accepted in its present form
by the Department of Biostatistics as satisfying the
thesis requirements for the degree of Master of Science

Date_____ _____
Professor Elizabeth S. Chen, PhD, Advisor

Date_____ _____
Professor Joseph W. Hogan, ScD, Advisor

Date_____ _____
Professor Christopher H. Schmid, PhD, Director

Approved by the Graduate Council

Date_____ _____
Andrew G. Campbell, Dean of the Graduate School

Acknowledgments

To my parents, who forever inspire and encourage me.

To Professor Elizabeth Chen, who welcomed me to her lab and introduced me to clinical informatics research. I am thankful for the guidance she has offered during our research, this thesis, and my development as a scholar, and I deeply respect the seemingly infinite support and empathy she has for her students.

To Professor Joseph Hogan, who graciously helped me contextualize and refine the statistical methods I used. I am grateful for his reassurance and expertise, and I could not have asked for a better Biostatistics Department advisor.

To Professor Adam Sullivan, who is the reason I applied and came to the Biostatistics Department here at Brown University.

To Liz Lirakis, Beth Clark, and Denise Arver, who add an undeniable warmth to the Biostatistics Department and have constantly gone the extra mile to make my experience here a great one.

To Professor Christopher Schmid, who immediately and continually engaged Biostatistics students to ensure our well-being during a global pandemic, demonstrating true leadership as Department Chair.

This work was funded in part by National Institutes of Health grant U54GM115677. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

Abstract of Selection of Clinical Text Features for Classifying Suicide Attempts

Ryan Buckland, ScM, Brown University, May 2020

Research has demonstrated cohort misclassification when studies of suicidal thoughts and behaviors (STBs) rely on ICD-9/10-CM diagnosis codes. Electronic health record (EHR) data are being explored to better identify patients, a process called EHR phenotyping. Most STB phenotyping studies have used structured EHR data, but some are beginning to incorporate unstructured clinical text. In this study, we used a publicly accessible natural language processing (NLP) program for biomedical text (MetaMap) and iterative elastic net regression to extract and select predictive text features from the discharge summaries of 810 inpatient admissions of interest. Initial sets of 5,866 and 2,709 text features were reduced to 18 and 11, respectively. The two models fit with these features obtained an area under the receiver operating characteristic curve of 0.866-0.895 and an area under the precision-recall curve of 0.800-0.838, demonstrating the approach's potential to identify textual features to incorporate in phenotyping models.

Table of Contents

List of Tables

List of Figures

Introduction

From 1999 to 2017, the age-adjusted suicide rate in the United States (U.S.) increased 33%, making suicide the 10th leading cause of death since 2008.[1,2] Along with opioid overdoses, suicide is part of the "deaths of despair" that reversed progress in U.S. life expectancy down for the third year in a row in 2017.[3,4] A complicated web of risk factors has slowed progress toward preventing suicide, though work so far has found that some risk factors such as a suicide attempt in the past year may be more important than others.[5] However, underdiagnosis of suicide attempts and the broader category of suicidal thoughts and behaviors (STBs) continues to undermine a complete etiology.

Diagnoses during healthcare encounters are indicated in electronic health records (EHR) and insurance claims data using the International Classification of Diseases (ICD). The U.S. used the 9th Revision, Clinical Modification (ICD-9-CM), which contains over 14,000 diagnosis codes, until October 2015 when it switched to the 10th Revision, Clinical Modification (ICD-10-CM) that contains over 69,000 codes. Each code is for a different diagnosis.[6] Researchers, clinical decision support tools, and many others use these ICD codes to identify cohorts of patients with diagnosis(es) of interest.

While using these codes is an efficient way to create a cohort to study and may be sensitive enough for other conditions, research suggests they may not capture all or even most patients with STBs.[7] A recent study in the emergency department setting performed manual chart review and found that only 30% of charts that had suicide attempt documented had the ICD code associated with suicide or intentional self-inflicted injury.[8] A similar study of a network of primary care clinics found only 19% of patients whose physician had

documented suicide attempt had the corresponding ICD code.[9] In response, other EHR data are being explored to better identify patient cohorts, a process called case detection or "EHR phenotyping." While many of these efforts have focused on structured EHR data, a growing body of literature has incorporated unstructured data (e.g., clinical notes) by applying natural language processing (NLP) to extract, represent, and analyze information captured within narrative text.

In what follows, we explore and characterize existing literature on EHR phenotyping of STBs with unstructured text. Then, we describe a novel approach that builds on these previous studies and focuses on reproducibility by using an accessible feature extraction and selection method and clinical NLP tool. Finally, we discuss the performance of the approach and opportunities for future work.

Literature Review

Through a targeted literature review, 12 articles published between 2012–2019 were identified that use NLP with clinical notes from the EHR for STB phenotyping.[10-21] Two common reasons for exclusion emerged. Either (1) the study used unstructured text from social media or suicide notes or (2) researchers focused on a clinical condition that was not suicide or STBs. There were no criteria for publication date.

Article Characteristics

There were six core characteristics analyzed for each article: (1) first author's primary discipline, (2) care setting or clinical sub-population, (3) classification approach, (4) NLP tools used, (5) section of EHR notes, and (6) performance. Some articles provided well-detailed explanations or diagrams of their methods. Where possible, the information was

translated into Unified Modeling Language (UML) diagrams to provide insight into the steps each study took in their approach. These UML diagrams can be found in the Appendix.

The types of journals the articles were published in could be categorized as informatics journals, clinical medicine journals, epidemiology journals, or others (e.g., PLOS One). Similarly, across studies the first author was either an epidemiologist or a physician—often a psychiatrist—sometimes with specialized informatics training. The list of articles and first-author discipline can be seen in Table 6 (Appendix).

Care Setting or Clinical Sub-population

Studies typically focused on a specific care setting (e.g., inpatient psychiatric unit), clinical condition (e.g., autism spectrum disorder), or both. Most studies looked at inpatient care, emergency department, or patients receiving mental health care. Only one study was done that concentrated on patients seen in primary care. The breakout of clinical areas of focus for each article can be seen in Table 6 (Appendix).

Data Collection

Studies often focused on inpatient EHRs because these data offer the largest amount of unstructured text due to multiple notes being documented throughout a patient's stay in a hospital. The more unstructured notes ingested and processed, the more likely there is something relevant to identifying a patient who ideated about or attempted suicide. Additionally, studies often selected a population of interest based on a clinical condition theoretically at higher risk for STBs. For example, one study focused on adolescents with autism spectrum disorder because "over 1 in 6 young people with autism spectrum

disorders will contemplate or attempt suicide during childhood, making them 30 times more at risk than typically developing children."[11] The empirical strategy behind the selection of a population at higher risk is to create a data set with a theoretically higher prevalence of STBs that increases the chances text mining will uncover missed diagnoses. Additionally, low prevalence limits the positive predictive value of the EHR phenotyping algorithm, so a sensitive data set that contains data suggestive of the phenotype, such as a co-occurring diagnosis that is an STB risk factor, can increase prevalence and address this limitation.[22]

Section of EHR

Most studies alluded to using all clinical notes that were available and machine-readable, which would include several different types of progress and consult notes from physicians, nurses, and social workers, among others. Other studies used chief complaints, a list of symptoms the patient communicates when they first speak to a clinician, and the discharge diagnoses, a list of the final medical diagnoses that were the reason for the patient's visit. The type of clinical text studies used can be found in Table 6 (Appendix).

NLP System

To process clinical text for purposes beyond keyword searches, such as part-of-speech, temporality, or negation, requires an NLP program. In several cases, studies used pre-existing, well-known NLP tools to extract clinically meaningful concepts from text. In other cases, studies used a custom system developed specifically for the study. Seven of the 12 articles used an established program, with most of those seven using clinical Text Analysis Knowledge Extraction System (cTAKES). Table 7 (Appendix) shows which studies used which systems.

Classification Approach

Almost half of the articles used rule-based algorithms for classification, including articles published as recently as 2018. In this context, rule-based algorithms are heuristic approaches based on criteria, often the presence of keywords (e.g., suicide), decided by the study team. For example, one article used a rule-based system that searched clinician notes for positive mention or negation using a curated list of keywords.[9]

The complexity of the rule-based algorithms varied widely. Most straightforward was a study that performed a keyword search in chief complaints and discharge diagnoses list. On the other hand, one article employed a multi-step process that first screened patients based on the number of documents in the EHR, performed a search for mentions of suicidality in documents, and then aggregated document-level assessment to perform patient classification.

The rest of the articles used machine learning approaches for classification, including naïve Bayes, logistic regression, random forests, and support vector machines. In one case, a study used multiple approaches. Which approach each study used can be found in Table 8 (Appendix).

Performance

The typical performance metrics for EHR phenotyping studies are precision (i.e., positive predictive value), recall (i.e., sensitivity), and F1 score. In some cases, area under the curve was also reported. For studies that reported performance statistics, precision ranged from 0.30 to 0.97, recall ranged from 0.58 to 0.98, and F1 scores ranged from 0.70 to 0.95. Since a few studies tried more than one model, they had multiple corresponding performance statistics. What studies reported can be found in Table 9 (Appendix).

Study Objectives

Given the sophistication of NLP, the complex nature of STBs, and the relative novelty of these techniques to mental health, no consensus exists for a standard approach. The optimal level of NLP standardization is unclear in part due to how features of suicidality, including text features, may differ across geographic, clinical, and sociodemographic factors, among others. The approaches of several studies have involved collecting an initial sample of the clinical notes of potential STB-positive cases using keyword searches (e.g., 'suic') and heuristic rules (e.g., a potential STB-positive case from clinical text must contain two or more occurrences of the word stem 'suic'). Other studies have developed custom text processing and machine learning programs (e.g., support vector machines, random forests).

Such pioneering approaches have had promising results that are key in empirically demonstrating the value of unstructured data. However, these approaches likely take significant time and cognitive effort for technical and clinical teams to pre-process the text and develop exhaustive lists of potential keywords and rules. Such pre-processing and lists may not transfer well to other settings with different expertise and documentation trends. Therefore, it is worthwhile to explore other text feature selection methods that may be more easily reproduced and modifiable. The present study's contribution toward this goal is an approach that combines a publicly available NLP tool and EHR data source along with an accessible statistical feature selection method for STB prediction studies using text features. Patient admissions with intentional or unintentional drug overdose were selected as the initial application because intent is particularly difficult to determine and the potential to identify discriminative text features was of interest.[23]

Methods

Overview

The methodological approach consisted of four core steps: (1) development of a sensitive data set, (2) text feature extraction, (3) text feature selection, and (4) evaluation. Discharge summaries from relevant admissions in the Medical Information Mart for Intensive Care III (MIMIC-III) database were extracted and run through the MetaMap NLP system. A set of salient concepts identified by MetaMap were used as features to represent each document. The resulting high-dimensional feature matrix was further reduced via feature selection with penalized regression. Logistic regression models were fit based on the features selected by the penalized regression process, and then these models were used to predict classification in the validation dataset. The details of the data and each step in the approach are described in the sections below.

Data source

The study used discharge summaries from the MIMIC-III database containing EHR data for more than 60,000 intensive care admissions between 2001 and 2012 at Beth Israel Deaconess Medical Center in Boston, Massachusetts.[24] As the name suggests, the sickest patients in a hospital receive intensive care, which is broadly defined as the comprehensive management of patients having, or at risk of developing, acute, life-threatening organ dysfunction.[25]

As in other STB phenotyping studies, we identify a subpopulation of patients at higher risk for STBs, patients treated for an overdose. Bohnert et al.[23] discuss the relationship between opioid use and overdose with suicidal intent—more than 40% of suicide and overdose

deaths in 2017 were known to involve opioids. By selecting a patient population at higher risk for STBs, we follow an established practice that addresses the low prevalence of STBs in the broader EHR population while also acknowledging that different clinical subpopulations may have STBs documented differently. While a previous study has done similar work before, the researchers used a custom application they developed.[20] We also narrowed the scope of our study to nonfatal overdoses to exclude completed suicides because many do not end up in the hospital.

There is a variety of different types of clinical text in the EHR, including chief complaints, admission notes, progress notes, consult notes, discharge summary, and others. The discharge summary is perhaps the most complete because it is designed to communicate to the patient and post-hospital care team a detailed overview of the care the patient received while in the hospital, including reason for hospitalization, significant findings, and discharge condition.[26] Therefore, discharge summaries were selected over other EHR document types because of their comprehensiveness and level of detail as well as their availability in MIMIC-III.

Given the observations of interest were admissions of patients with a non-fatal overdose or suicide attempt by poisoning, the inclusion criteria for an admission were: (1) at least one discharge summary document, (2) patient discharged alive, and (3) at least one diagnosis code matching the ICD-9-CM criteria in Table 1. The diagnosis criteria used ICD-9-CM codes because the admissions are from 2001 to 2012, before the United States changed to ICD-10-CM in 2015. These diagnosis code criteria were adapted from a guide published by the Centers for Disease Control and Prevention.[27]

Table 1. Diagnosis Criteria and Outcome Labels

Code	Description	Outcome Label
96X.X-97X.X	Poisoning by drugs, medicinal substances, and biologicals	*No*
E85X.X	Accidental poisoning by drugs, medicinal substances, and biologicals	
E980.0-E980.5, E980.9	Poisoning by solid or liquid substance, undetermined whether accidentally or purposely inflicted	
E950.0-E950.5, E950.9	Suicide and self-inflicted poisoning by solid or liquid substances	*Yes*

Text feature extraction

The discharge summaries for these admissions were extracted from MIMIC-III and each discharge summary was processed using MetaMap. MetaMap is a publicly available program developed and maintained by the National Library of Medicine (NLM) to map biomedical text to the Unified Medical Language System (UMLS) Metathesaurus. The UMLS Metathesaurus is a thesaurus organized by concept, or meaning, that links similar names for the same concept from nearly 200 different vocabularies.[28]

MetaMap is robust and flexible in how it processes input text, evaluates potential UMLS Metathesaurus mappings, and generates output.[29] The present study used MetaMap Indexing (MMI) fielded output with default MetaMap settings.[30] The process resulted in one output document for each input discharge summary document, where each output document contained a list of all the UMLS concepts assigned by MetaMap to the discharge summary. The text feature extraction process is shown in Figure 1.

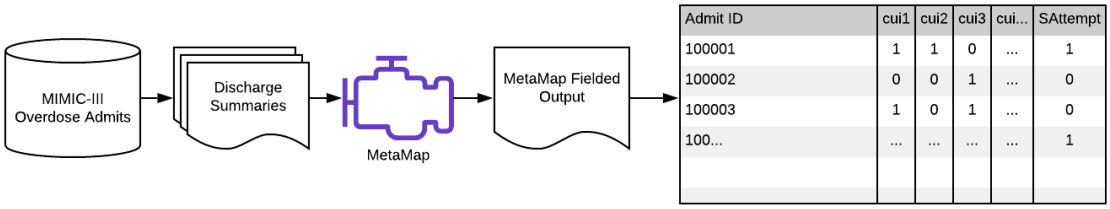

Figure 1. Methods Pipeline: Text Feature Extraction

A fabricated example of how MetaMap processes text and returns MMI output is in Figure 2. Each row in the output of Figure 2 represents a mapping between a UMLS concept and a trigger word or phrase in the input document.

Input:

CHIEF COMPLAINT: The patient does say, "I screwed up my whole life and wrecked my car." The patient claims he is med compliant, although his mother, and stepfather saying he is off his meds. He had a two-day stay at XYZ Hospital for medical clearance after his car accident, and ***no injuries*** *were found other than a sore back, which was negative by x-ray and CT scan.*

Output:

```
5.18|Today (temporal qualifier)|C0750526|[tmco]|["Today"-tx-6-"Today"-noun-0]|TX|565/5|
5.18|Traumatic injury|C3263723|[inpo]|["Injuries"-tx-3-"injuries"-noun-1]|TX|282/8|
5.18|subscriber - self|C1551994|[inpr]|["self"-tx-4-"self"-noun-0]|TX|443/4|
```

Concept Unique Identifier
Semantic Type
Negation

Figure 2. Example of MetaMap Output

Within each row are pipe-delimited fields, each containing information about attributes of the concept mapped. For the present study, the UMLS concept unique identifier (CUI), UMLS preferred name of the concept, negation, and the concept's UMLS semantic type

were most relevant. In the UMLS Metathesaurus, each concept is assigned a unique ID called a CUI. Similarly, each concept is assigned a biomedically meaningful category called a semantic type.[31]

For example, the mapping for “no injuries” from the input text can be seen in the second line of MMI output in Figure 2. The preferred name of the concept is “Traumatic injury”, the UMLS CUI is in the black box, the semantic type is in the green box, and the flag for negation is in the red box. The “Traumatic injury” concept has the semantic type “inpo” (i.e., “Injury or Poisoning”), which is also a common semantic type of the concepts identified in the discharge summaries in this study. Finally, the negation flag in the red box is 1, indicating the “Traumatic injury” is negated. The negation flag would be 0 if the concept were not negated.

Each MetaMap output file was ingested and combined into a single dataset of one-hot encoded UMLS CUI features, where each CUI was a binary feature indicated by 1 if the CUI was present in the discharge summary for a given hospital admission or 0 if it was not present. Similar to a bag-of-words representation that describes the occurrence of words in a document, this is a bag-of-CUIs approach that does the same by describing the occurrence of CUIs in a discharge summary document. Before using any machine learning methods, semantic type and negation criteria for CUI feature inclusion were applied to reduce the number of features, p. A CUI had to be (1) non-negated and (2) belong to one of the semantic types listed in Table 2, which were selected based on their prima facie relevance to suicide attempts.

This step reduced the number of features to 5,866. The breakdown by semantic type can be seen in Table 2. Initially excluded, the “Finding” semantic type was retained because

five of the seven suicide concepts extracted belonged to it. Even after pragmatic steps to reduce the number of features, the two datasets had 5,866 or 2,709 features (5,866 − 3,157 "Finding"), making suicide attempt classification among these admissions a p >> n problem. The outcome labels were also binary and based on the ICD-9-CM codes in Table 1, where an admission with a suicide attempt code (i.e., E950.0-E950.5 or E950.9) was labeled with 1 and 0 if not.

Table 2. Semantic Type, by CUI Predictor Frequency

Semantic Type	**Frequency (%)**
Finding (fndg)	3157 (54)
Disease or Syndrome (dsyn)	1675 (29)
Injury or Poisoning (inpo)	391 (7)
Mental or Behavioral Dysfunction (mobd)	320 (6)
Mental Process (menp)	142 (2)
Individual Behavior (inbe)	93 (2)
Social Behavior (socb)	88 (2)

Text feature selection

Figure 3 provides an overview of the text feature selection and evaluation processes. Elastic net logistic regression was used for further feature selection.

The elastic net is a method that combines L1- and L2-regularized regression, also known as least absolute shrinkage and selection operator (LASSO) regression and ridge regression, respectively.[32] Thus, the elastic net takes advantage of the strengths of LASSO and ridge regression.

The elastic net equation to be minimized is in Figure 4. The balance between the LASSO and ridge penalties is controlled by the hyperparameter α and λ determines the magnitude of the combined penalty.

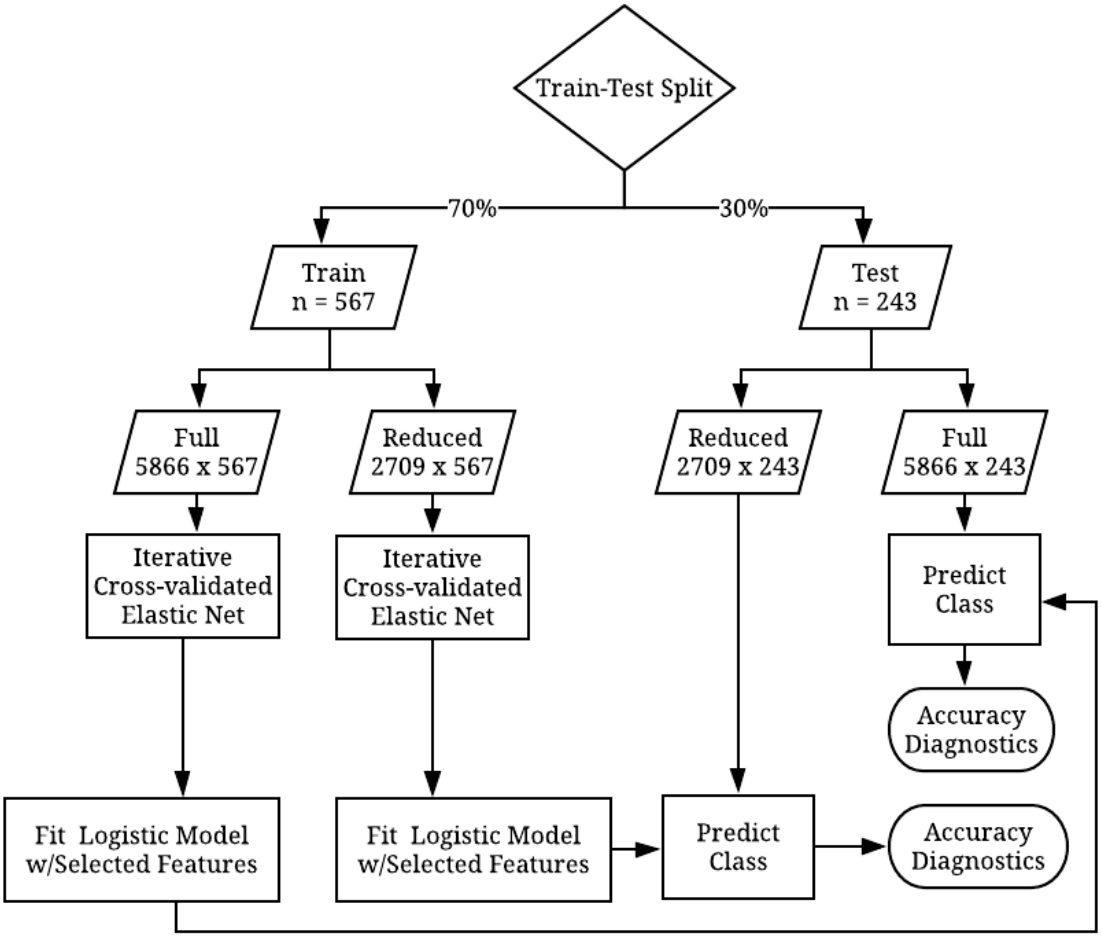

Figure 3. Methods Pipeline: Text Feature Selection and Evaluation

In practical terms, elastic net uses the LASSO's L1 penalty ($\|\beta\|$) to be able to shrink features all the way to 0, which enables model sparsity. The ridge L2 penalty ($\|\beta\|^2$) cannot achieve this because it squares the coefficient. However, this characteristic that enables LASSO to encourage model sparsity by shrinking coefficients to zero does not account well for groups of collinear features. Consider an example case with two collinear features, x and z, that are centered and scaled. Because they are collinear, these two features would be able to substitute each

$$\min_{\beta_0,\beta} \frac{1}{N} \sum_{i=1}^{N} w_i l(y_i, \beta_0 + \beta^T x_i) + \lambda \left[(1-\alpha)||\beta||_2^2/2 + \alpha||\beta||_1 \right]$$

Figure 4. Equation to be Minimized

other in predicting Y. Under the L1 norm that LASSO uses, $0.2x + 0.8z$ and $0.5x + 0.5z$ would yield the same penalty. For a given ƛ, one feature could be shrunk to zero while the other is kept in the model, ignoring they are collinear and should be selected in or out of the model as a pair. Fortunately, the ridge penalty accounts for groups of collinear features because the L2 norm would yield different penalties for $0.2x + 0.8z$ and $0.5x + 0.5z$ and tend to choose equal weights. Therefore, the elastic net can be helpful in selecting features and groups of features for a sparse model when p >> n.

In addition to the ƛ that minimizes the equation in Figure 4, a second option is ƛ1se, the value that most regularizes the model within one standard error of the minimum. Given a core objective of the present study was to provide a parsimonious set of CUI features that are most relevant to classification, ƛ1se is used.[33] Two other important method specifications were setting α to 0.5 and performing elastic net over many iterations to identify the most relevant CUI features. The R package *glmnet* uses k-fold cross-validation to tune ƛ. The default is ten folds and that is maintained in this study. It is important to consider that each cross-validated fit splits the data into a different set of k folds, which means the optimal ƛ and the CUI features kept in the model will change from one fit to another. To address this instability, 100 iterations of the 10-fold cross-validated fit were run. A CUI feature needed to be non-zero in all 100 of the iterations in order to be used in the final logistic regression model.

Since more than 54% of CUI features belonged to "Finding" and these were included because of a hesitancy to exclude the five suicide concepts among them, an important objective of the study was to explore the benefit of the "Finding" CUIs. To that end, the iterative fit process was performed separately to two datasets, a "Full" dataset containing

all 5,866 features, including the "Finding" CUI features, and a "Reduced" dataset without the "Finding" CUI features. The two processes each yielded a set of CUI features that appeared in all 100 iterations. Two final logistic regression models were fit, one for each set, and their performance compared.

Evaluation

The performance of the two final logistic regression models was evaluated using precision, recall (i.e., sensitivity), the precision-recall (PR) curve, area under the PR curve (AUC-PR), specificity, the receiver operating characteristic (ROC) curve, and area under the ROC curve (AUC-ROC). Precision, also known as positive predictive value, is the ratio of the correct positive predictions to the total predicted positives. In other words, precision measures the probability that a positive prediction from the classifier is correct. Recall, also known as sensitivity or the true positive rate, is the ratio of the correct positive predictions to the overall number of truly positive cases. Recall expresses the probability that the prediction will be positive, given that the true status of the case is positive. While broadly useful, precision and recall can be particularly helpful when there are a small number of cases among many controls because they focus on the performance of a classifier to identify positive cases correctly. Such a situation is common for STB classification and a naive model that predicts all observations in a test set as negatives could be perceived to have high accuracy because of how well it identifies true negatives. Since precision and recall do not measure true negatives, they avoid this pitfall.

It is often desirable to consider precision and recall together, which is done with the F1 score, PR curve, and AUC-PR. The F1 score is the harmonic mean of precision and recall and provides a measure of the tradeoff between them at a set probability threshold. While

the F1 score measures performance at a set probability threshold, the PR curve plots precision and recall at varying probability thresholds. Additionally, AUC-PR is a summary of how well the model identifies cases and is essentially the precision averaged over recall values from 0 to 1. The baseline for AUC-PR is the prevalence of cases.

The prevalence of suicide attempts among admissions with an overdose, the population of focus in the present study, was high and meant the data were more balanced than what might be observed among a broader population. Therefore, ROC curves and AUC-ROC were also used for evaluation. A ROC curve plots the recall (i.e., sensitivity) against (1 − specificity) of a classification model. Specificity, also known as negative predictive value, is the ratio of correct negative predictions to the correct negative predictions and incorrect positive predictions (i.e., how many actual negative cases are predicted correctly). Subtracting specificity from 1 gives the false positive rate. The ROC curve shows the tradeoff between sensitivity and (1 − specificity) at different outcome probability thresholds predicted by a model. Additionally, the AUC-ROC is a value between 0 and 1 that is used to summarize the accuracy of the classifier in discriminating between cases and controls. The closer to 1 the AUC is, the better, with an AUC of 0.5 being random chance. Still, precision and recall are most useful when class imbalance exists and identification of positive cases is particularly important, as is the case for STB classification.

Results

There were 810 admissions that met the inclusion criteria. Of the 810 admissions, there were 383 coded suicide attempts (i.e., prevalence = 47.3%) and 427 overdoses not explicitly coded as suicide attempts. A 70/30 train-test split yielded a training set of 567

admissions and a test set of 243 admissions. The test set contained 109 coded suicide attempts (i.e., prevalence = 44.9%) and 134 admissions not coded as suicide attempts.

In the elastic net fit process for the full model, 18 CUI features were kept in all 100 iterations. The average λ1se was 0.086. In the fit process for the reduced model, 11 CUI features were kept in all 100 iterations. The average λ1se was 0.092. The sets of 18 and 11 CUI features were selected out of 5,866 total CUI features that met the initial semantic type inclusion criteria.

The logistic regression model fits with each set of CUI features can be seen in Table 3. The table also includes the frequency counts of the presence of the CUI by outcome as well as the semantic type of the CUI. Seven of the 18 features had the "Finding" semantic type, and those seven are the difference between the set of features selected from the full data and the set of features selected from the reduced data. There are no CUI features in the set from the reduced data that are not included in the set from the full data. In the full model, 11 features have statistically significant odds ratios (OR) that range from 0.11 to 11.18. In the reduced model, 9 features have statistically significant odds ratio (OR) coefficients that range from 0.14 to 11.49. For example, in the reduced model an admission which has the CUI "C0038663 Suicide Attempt" in the discharge summary is 11.49 times as likely to have been coded as a suicide attempt as one that does not have the CUI in the discharge summary. Another example is the CUI feature "C1306597 Psychiatric Problem," which has an OR of 1.85 in the full model. This can be interpreted as follows: an admission with the CUI in the discharge summary is 1.85 times as likely to have been coded as a suicide attempt as one that does not have the CUI in the discharge summary.

Table 3. Logistic Regression Full and Reduced Model Summaries

Feature	*Odds Ratios [95% CI]*		*Frequency of CUI = 1*		*Semantic Type*
	Full	**Reduced**	**Suicide Attempt (n = 109)**	**Not Attempt (n = 134)**	
Intercept	0.26 [0.12, 0.58]	0.15 [0.08, 0.26]	NA	NA	NA
C0011570 Mental Depression	1.14 [0.39, 3.37]	1.37 [0.57, 3.34]	72	61	Mental or Behavioral Dysfunction
C0020538 Hypertensive Disease	0.49 [0.26, 0.90]	0.42 [0.24, 0.73]	26	61	Disease or Syndrome
C0023890 Liver Cirrhosis	0.11 [0.02, 0.46]	0.14 [0.03, 0.50]	1	12	Disease or Syndrome
C0029944 Drug Overdose	1.59 [0.22, 9.33]	2.58 [1.52, 4.42]	85	69	Injury or Poisoning
C0038661 Suicide	7.15 [2.59, 23.97]	NA	23	3	Finding
C0038663 Suicide Attempt	1.83 [0.74, 4.46]	11.49 [6.83, 19.85]	82	25	Injury or Poisoning
C0085281 Addictive Behavior	0.30 [0.08, 0.87]	0.19 [0.06, 0.55]	1	17	Mental or Behavioral Dysfunction
C0344315 Depressed Mood	1.69 [0.56, 4.90]	1.66 [0.67, 4.08]	75	65	Mental or Behavioral Dysfunction
C0438696 Suicidal	3.74 [1.12, 14.83]	5.78 [1.88, 20.74]	11	2	Mental or Behavioral Dysfunction
C0455503 History of (H.O.) Depression	1.42 [0.72, 2.81]	NA	24	17	Finding
C0748061 Psychiatric Hospitalization	9.82 [2.95, 39.31]	8.46 [2.85, 30.90]	17	4	Mental or Behavioral Dysfunction
C1306597 Psychiatric Problem	1.85 [1.02, 3.37]	2.16 [1.29, 3.65]	65	35	Mental or Behavioral Dysfunction
C1535939 Pneumocystis Jiroveci Pneumonia	0.58 [0.32, 1.03]	0.52 [0.31, 0.89]	45	59	Disease or Syndrome
C3838679 4+ Answer to Question	0.43 [0.20, 0.92]	NA	83	110	Finding
C4018909 Overdose	1.73 [0.30, 11.94]	NA	82	63	Finding
C4084795 PSA Level Less than Five	0.63 [0.30, 1.29]	NA	76	106	Finding
C4554104 Suicidal Ideation, Common Terminology Criteria for Adverse Events (CTCAE)	3.36 [1.40, 8.41]	NA	27	8	Finding
C4554105 Suicide Attempt, CTCAE	11.18 [4.35, 29.76]	NA	78	18	Finding

A threshold predicted probability of 0.5 or greater was used, and the resulting confusion matrix, F1 score, precision, recall, AUC-PR, and AUC-ROC for each model can be seen in Tables 4 and 5. The model fit from the full model has slightly worse precision, meaning the suicide attempts it predicts are less often actual suicide attempts than the reduced model. Conversely, the full model has better recall than the reduced model, meaning its predicted suicide attempts identify more of the actual suicide attempts. The difference in the two models' recall scores is larger than their difference in precision, which is why the F1 score of the full data model is larger.

The PR curves for each model are in Figures 5 and 6. The full model had an AUC-PR of 0.838 and the reduced model had an AUC-PR of 0.800. The color scale on the right side of each plot indicates the classification threshold at a given point on the curve. Both are a marked improvement over the baseline of 0.449 (i.e., prevalence in test data). The ROC curves for each model are in Figure 7. The full model had an AUC-ROC of 0.895 and the reduced model had an AUC-ROC of 0.866.

Table 4. Confusion Matrix and Performance, Full Model

	Predicted Positive	**Predicted Negative**
Positive	91	18
Negative	26	108

F1:	0.805	**Precision:**	0.778
AUC:	0.838 (PR), 0.895 (ROC)	**Recall:**	0.835

Table 5. Confusion Matrix and Performance, Reduced Model

	Predicted Positive	Predicted Negative
Positive	83	26
Negative	23	111

F1:	0.772	**Precision:**	0.783
AUC:	0.800 (PR), 0.866 (ROC)	**Recall:**	0.761

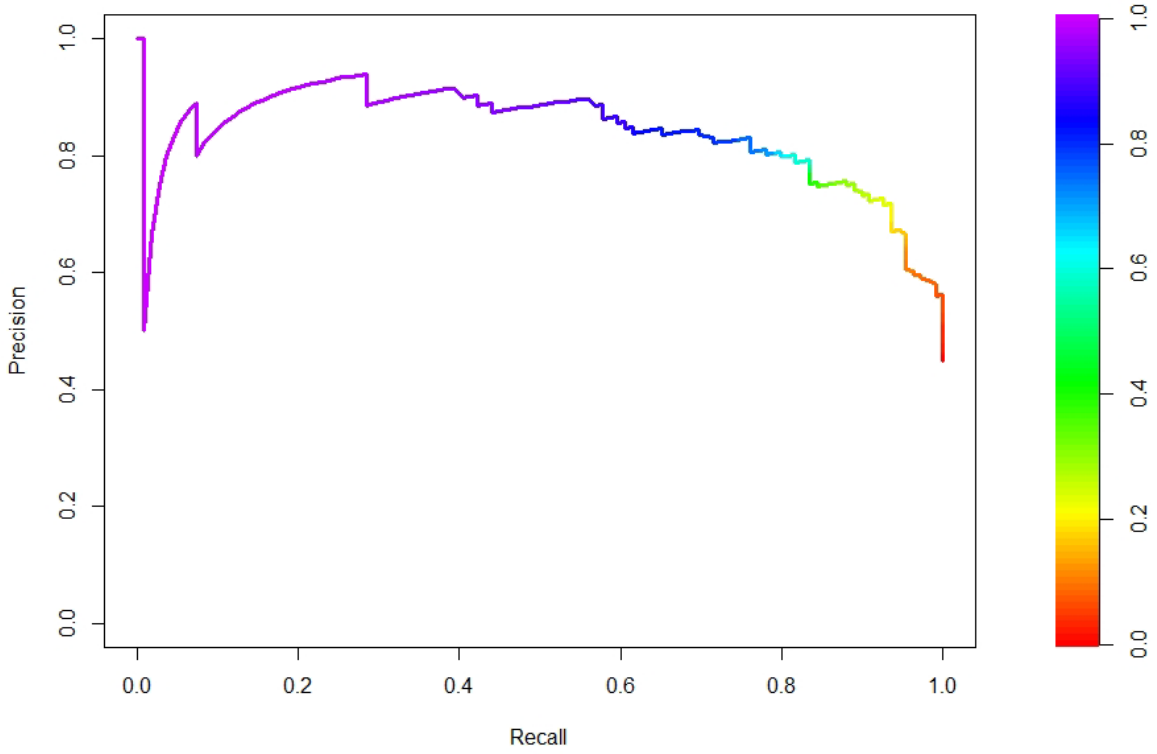

Figure 5. Precision-Recall Curve for Full Model

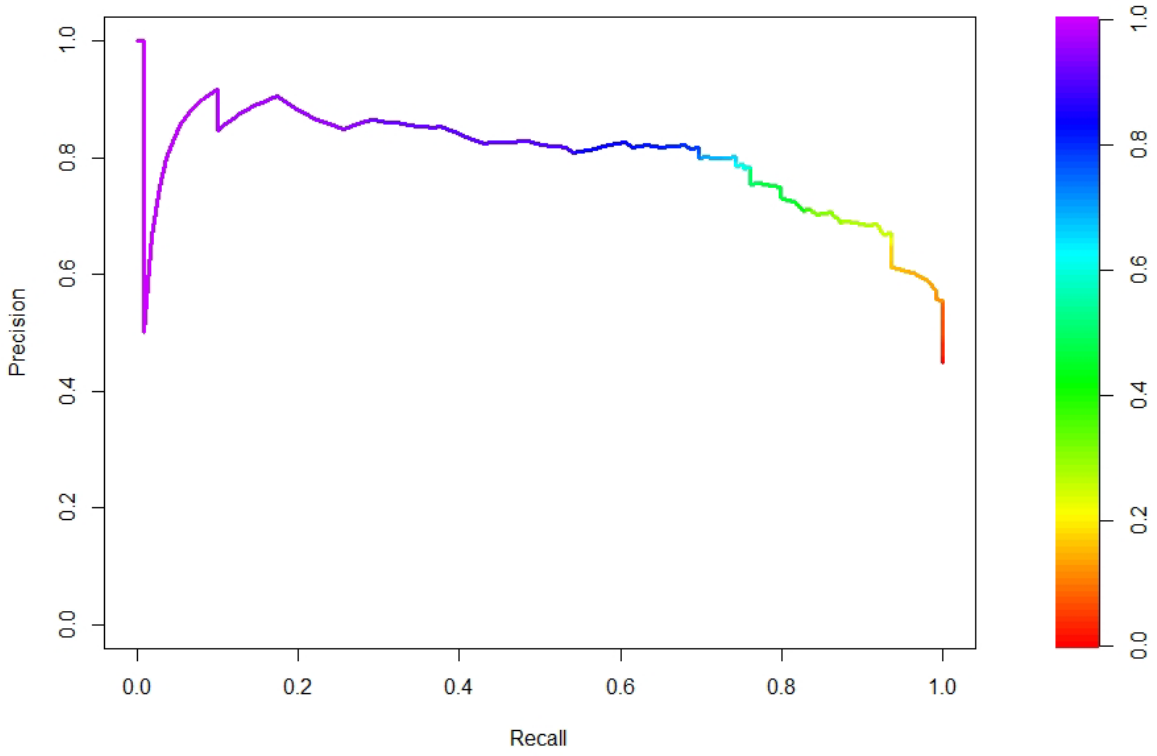

Figure 6. Precision-Recall Curve for Reduced Model

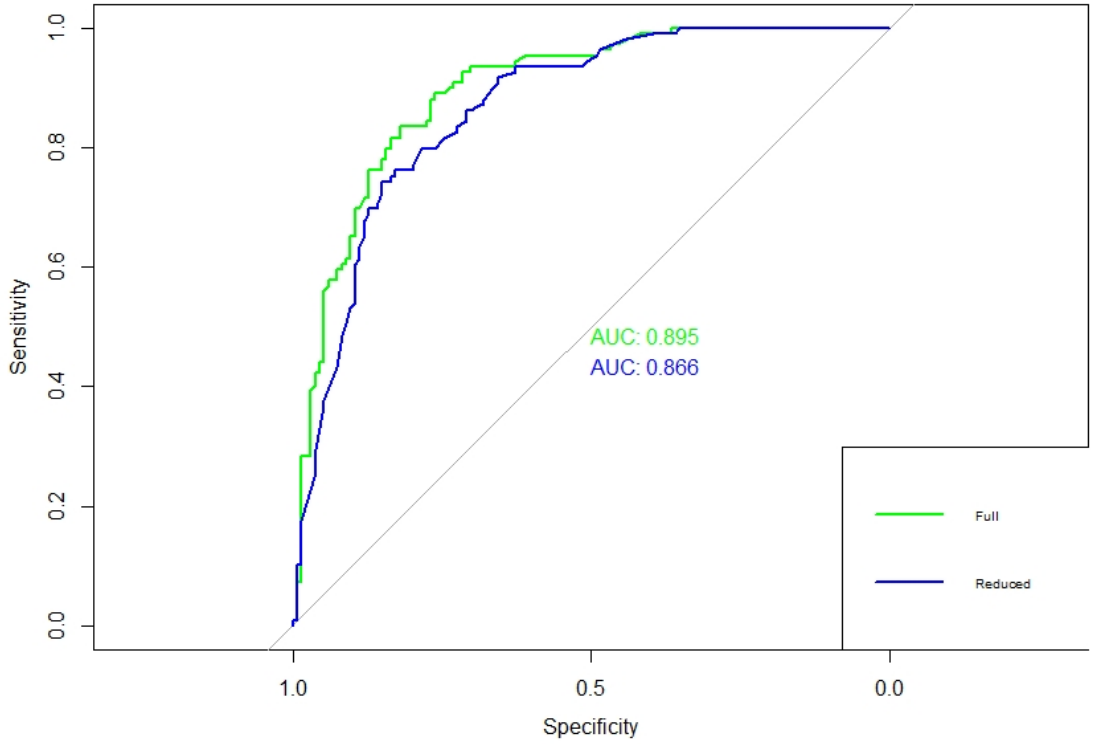

Figure 7. Receiver Operating Curve

Discussion

The present study evaluated an approach to text feature extraction and selection that blends two strategies. First, it takes advantage of MetaMap, a publicly available NLP tool and MetaMap's output as decision rules for initial CUI feature selection. Second, the approach uses elastic net penalized regression to reduce the number of features to as few as 11. In addition to being conceptually accessible and easily modified, the hybrid strategy allows for the exploration of downstream effects, including performance, of a decision like the CUI inclusion criteria based on semantic type. For example, the present study explored what performance tradeoff, if any, exists for including those suicide-related "Finding" CUIs that intuitively seem important but whose inclusion means sifting through many other "Finding" CUIs that are likely unimportant for the context. Interestingly, the average magnitude of $\lambda 1se$ was quite small at 0.086 and 0.092 for the full and reduced data, respectively, but resulted in more than a 99.6% reduction in features selected for the models. For models that include only CUIs, AUC-PRs of more than 0.80 and an AUC-ROC of almost 0.90 were better than expected. The PR curve for the reduced model drops after recall of around 0.20 and steadily decreases until 0.70 when it drops again and sharply decreases. In comparison, the PR curve for the full model maintains a higher precision with occasional stepped drops and a sharp decrease after recall of 0.80. This could suggest the full model's ability to classify as positive only those cases actually positive while continuing to increase its coverage of the actual positive cases.

The two sets of CUI features produced by the elastic net processes included intuitive features: CUIs that explicitly mention suicide and suicide attempts as well as CUIs that relate to psychiatric conditions and hospitalization. However, they also include features

that do not immediately make sense and warrant further exploration. For example, C3838679 4+ Answer to Question and C4084795 PSA Level Less than Five do not appear related to suicide attempt classifications. A review of the MetaMap output for these two CUIs showed that the trigger words were exclusively "4" for C3838679 4+ Answer to Question and "5" for C4084795 PSA Level Less than Five. The triggers did not include any other words. Discharge summaries, including those in the present study, often have numbers in the form of numbered lists or quantities, and that is likely what caused these two errant CUIs. It is likely these occurrences could be managed using MetaMap's processing options or pre-processing (e.g., removal of numbering or selecting specific sections in the discharge summary). However, it is important to reiterate the intent of the study was to avoid barriers to reproducibility by relying on statistical feature selection techniques in lieu of significant pre-processing expertise.

The primary goal was to identify CUIs for inclusion in more comprehensive models that contain structured sociodemographic and clinical data. The intent is not to suggest classification models which only contain CUIs or text features. Therefore, while the models' ORs are not at their core crucial to the overall objective, they are still worth briefly discussing. Such wide confidence intervals for many of the ORs is due to the sparsity of the data and how well the presence of certain features separated cases and controls for suicide attempt.

Within the context of the present study, performance and diagnostic metrics serve as a proof of concept for a pipeline that is able to process unstructured text and perform feature selection while maintaining replicability, customization, and explainability to stakeholders who may not be experts in statistics or informatics.

Future work will focus on five areas that should improve the study: (1) obtaining gold standard outcome labels, (2) using a different type of clinical document or different NLP system, (3) incorporating structured data, (4) exploring effects of adjustments to pragmatic inclusion criteria around negation and semantic type, and (5) additional tuning of the elastic net and final model fitting process.

It is important to acknowledge the source of the outcome labels. The labels were ICD-9-CM codes that are assigned by hospital coders based on their review of the patient chart and documentation. They are not labels designated by trained clinicians with a structured process of retrospective chart review and inter-rater agreement evaluation via Cohen's kappa. This means the study itself did not have the objective of identifying missed cases of suicide attempt, though the pipeline and methods are in service of that effort. It is likely performance may change if gold standard labelling by trained clinicians was performed with the training and test data.

There are several other sources of clinical text besides the discharge summary, including chief complaints, progress notes, case management notes, and consult notes. Discharge summaries were selected because of their comprehensive nature. Each of these other types of notes are documented for different purposes. Consult notes could be particularly useful for patients with STBs as a consultation by a psychiatrist may be requested to determine the nature of an injury and suicidality of a patient. Processing and using the specialized text of the consult note could reveal nuanced insights.

MetaMap was selected out of pragmatism and availability. It has several options that impact how it maps text to UMLS concepts. Exploring these options may be beneficial, especially if they can serve as further upstream feature selection. There are also other NLP

systems besides MetaMap, including cTAKES and Clinical Language Annotation, Modeling, and Processing (CLAMP). Both cTAKES and CLAMP are accessible, intended for clinical text, and merit further consideration and comparison.[34, 35]

The present study focused on text feature extraction and selection techniques. The resulting feature matrices were sparse and incorporating structured EHR data would likely improve the performance of the classifier. There are myriad structured EHR features, ranging from administrative and demographic information such as length of stay or insurance type to clinical features such as lab values and survey tool responses. Several studies have explored which of those features are most promising in predicting STB risk, and a useful next step may be adding text features selected using a process similar to the one demonstrated here.

As mentioned, CUI features were only included if they were not negated in order to maintain clear interpretation of ORs that resulted from final model fits. Especially if there were situations where a CUI was negated for one admission and not for another. Interpreting these would be challenging. Additionally, the choice of which semantic types are most relevant is subjective and exploring further the effect of using them as inclusion criteria is worthwhile.

Areas of computational future work are tuning α to find the optimal value rather than selecting 0.5. This can be done via k-fold cross-validation in the same way $\lambda 1se$ was tuned. Additionally, the selection of 100 iterations versus 1000 iterations or some other number was in part due to the time to run the program. Finally, moving from one-hot encoded features to term frequency inverse document frequency or a more complex representation like word2vec embedding would add useful information to the process and final model fits.

Conclusion

Predicting STBs and identifying those most at risk continues to be difficult and there is evidence that STBs are missed when case definitions rely solely on ICD codes. Researchers have increasingly used machine learning approaches with structured EHR and claims data for STB phenotyping. Several are adding unstructured clinical notes data to their efforts, but there have been persistent barriers to efficiently processing these data into something usable for models that might end up deployed in support of real-time clinical decision making. The present study presents a flexible approach to doing so by using both pragmatic choices and statistical feature selection methods, demonstrating promising results. As NLP becomes more mainstream in STB research, processing pipelines that are reproducible and modifiable will be key.

References

1. Products - Data Briefs - Number 330 - September 2018 [Internet]. Centers for Disease Control and Prevention. Centers for Disease Control and Prevention; 2018 [cited 2020Mar14]. Available from: https://www.cdc.gov/nchs/products/databriefs/db330.htm

2. Products - Data Briefs - Number 309 - June 2018 [Internet]. Centers for Disease Control and Prevention. Centers for Disease Control and Prevention; 2018 [cited 2020Mar14]. Available from: https://www.cdc.gov/nchs/products/databriefs/db309.htm

3. Products - Data Briefs - Number 328 - November 2018 [Internet]. Centers for Disease Control and Prevention. Centers for Disease Control and Prevention; 2018 [cited 2020Mar14]. Available from: https://www.cdc.gov/nchs/products/databriefs/db328.htm

4. CDC Director's Media Statement on U.S. Life Expectancy [Internet]. Centers for Disease Control and Prevention. Centers for Disease Control and Prevention; 2018 [cited 2020Mar14]. Available from: https://www.cdc.gov/media/releases/2018/s1129-US-life-expectancy.html

5. Simon GE, Johnson E, Lawrence JM, Rossom RC, Ahmedani B, Lynch FL, et al. Predicting Suicide Attempts and Suicide Deaths Following Outpatient Visits Using Electronic Health Records. American Journal of Psychiatry. 2018;175(10):951–60.

6. International Classification of Diseases, 11th Revision (ICD-11) [Internet]. World Health Organization. World Health Organization; 2019 [cited 2020Mar14]. Available from: https://www.who.int/classifications/icd/en/

7. Lu, C. Y., Stewart, C., Ahmed, A. T., Ahmedani, B. K., Coleman, K., Copeland, L. A., . . . Soumerai, S. B. (2014). How complete are E-codes in commercial plan claims databases? Pharmacoepidemiology and Drug Safety, 23(2), 218-220. doi:10.1002/pds.3551

8. Arias, S. A., Boudreaux, E. D., Chen, E., Miller, I., Camargo, C. A., Jr., Jones, R. N., & Uebelacker, L. (2018). Which Chart Elements Accurately Identify Emergency Department Visits for Suicidal Ideation or Behavior? Arch Suicide Res, 1-14. doi:10.1080/13811118.2018.1472691

9. Anderson HD, Pace WD, Brandt E, Nielsen RD, Allen RR, Libby AM, West DR, Valuck RJ. Monitoring suicidal patients in primary care using electronic health records. J Am Board Fam Med. 2015 Jan 1;28(1):65-71.

10. Carson NJ, Mullin B, Sanchez MJ, Lu F, Yang K, Menezes M, et al. Identification of suicidal behavior among psychiatrically hospitalized adolescents using natural

language processing and machine learning of electronic health records. Plos One. 2019;14(2).

11. Downs J, Velupillai S, George G, Holden R, Kikoler M, Dean H, Fernandes A, Dutta R. Detection of suicidality in adolescents with autism spectrum disorders: developing a natural language processing approach for use in electronic health records. InAMIA annual symposium proceedings 2017 (Vol. 2017, p. 641). American Medical Informatics Association.

12. Fernandes AC, Dutta R, Velupillai S, Sanyal J, Stewart R, Chandran D. Identifying suicide ideation and suicidal attempts in a psychiatric clinical research database using natural language processing. Scientific reports. 2018 May 9;8(1):1-0.

13. 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. Journal of the American Medical Informatics Association. 2016 Sep 1;23(5):1007-15.

14. Haerian K, Salmasian H, Friedman C. Methods for identifying suicide or suicidal ideation in EHRs. InAMIA annual symposium proceedings 2012 (Vol. 2012, p. 1244). American Medical Informatics Association.

15. Hammond KW, Laundry RJ, OLeary TM, Jones WP. Use of text search to effectively identify lifetime prevalence of suicide attempts among veterans. In 2013 46th Hawaii International Conference on System Sciences 2013 Jan 7 (pp. 2676-2683). IEEE.

16. Hammond KW, Laundry RJ. Application of a hybrid text mining approach to the study of suicidal behavior in a large population. In2014 47th Hawaii International Conference on System Sciences 2014 Jan 6 (pp. 2555-2561). IEEE.

17. 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. International journal of methods in psychiatric research. 2017 Jun;26(2):e1522.

18. Zhong QY, Karlson EW, Gelaye B, Finan S, Avillach P, Smoller JW, Cai T, Williams MA. Screening pregnant women for suicidal behavior in electronic medical records: diagnostic codes vs. clinical notes processed by natural language processing. BMC medical informatics and decision making. 2018 Dec;18(1):30.

19. Zhong QY, Mittal LP, Nathan MD, Brown KM, González DK, Cai T, Finan S, Gelaye B, Avillach P, Smoller JW, Karlson EW. Use of natural language processing in electronic medical records to identify pregnant women with suicidal behavior: towards a solution to the complex classification problem. European journal of epidemiology. 2019 Feb 15;34(2):153-62.

20. Hazlehurst B, Green CA, Perrin NA, Brandes J, Carrell DS, Baer A, DeVeaugh-Geiss A, Coplan PM. Using natural language processing of clinical text to enhance identification of opioid-related overdoses in electronic health records data. Pharmacoepidemiology and drug safety. 2019 Aug;28(8):1143-51.

21. Kuramoto-Crawford SJ, Spies EL, Davies-Cole J. Detecting suicide-related emergency department visits among adults using the District of Columbia syndromic surveillance system. Public Health Reports. 2017 Jul;132(1_suppl):88S-94S.

22. Liao KP, Cai T, Savova GK, Murphy SN, Karlson EW, Ananthakrishnan AN, Gainer VS, Shaw SY, Xia Z, Szolovits P, Churchill S. Development of phenotype algorithms using electronic medical records and incorporating natural language processing. bmj. 2015 Apr 24;350:h1885.

23. Bohnert AS, Ilgen MA. Understanding links among opioid use, overdose, and suicide. New England journal of medicine. 2019 Jan 3;380(1):71-9.

24. Johnson AE, Pollard TJ, Shen L, Li-wei HL, Feng M, Ghassemi M, Moody B, Szolovits P, Celi LA, Mark RG. MIMIC-III, a freely accessible critical care database. Scientific data. 2016 May 24;3:160035.

25. Marshall JC, Bosco L, Adhikari NK, Connolly B, Diaz JV, Dorman T, Fowler RA, Meyfroidt G, Nakagawa S, Pelosi P, Vincent JL. What is an intensive care unit? A report of the task force of the World Federation of Societies of Intensive and Critical Care Medicine. Journal of critical care. 2017 Feb 1;37:270-6.

26. Kind AJ, Smith MA. Documentation of mandated discharge summary components in transitions from acute to subacute care. In Advances in patient safety: New directions and alternative approaches (vol. 2: Culture and redesign) 2008 Aug. Agency for Healthcare Research and Quality (US).

27. Prescription Drug Overdose Data & Statistics: Guide to ICD-9-CM and ICD-10 Codes Related to Poisoning and Pain [Internet]. The Centers for Disease Control and Prevention; Available from: https://www.cdc.gov/drugoverdose/pdf/pdo_guide_to_icd-9-cm_and_icd-10_codes-a.pdf

28. UMLS - Metathesaurus [Internet]. U.S. National Library of Medicine. National Institutes of Health; [cited 2020Apr26]. Available from: https://www.nlm.nih.gov/research/umls/knowledge_sources/metathesaurus/index.html

29. Aronson AR, Lang FM. An overview of MetaMap: historical perspective and recent advances. Journal of the American Medical Informatics Association. 2010 May 1;17(3):229-36.

30. Fielded MetaMap Indexing (MMI) Output Explained (Updated for MetaMap 2016 Output) [Internet]. Available from: https://metamap.nlm.nih.gov/Docs/MMI_Output_2016.pdf

31. Semantic Types and Groups [Internet]. U.S. National Library of Medicine. National Institutes of Health; [cited 2020Mar14]. Available from: https://metamap.nlm.nih.gov/SemanticTypesAndGroups.shtml

32. Zou H, Hastie T. Regularization and variable selection via the elastic net. Journal of the royal statistical society: series B (statistical methodology). 2005 Apr 1;67(2):301-20.

33. Hastie T, Qian J. Glmnet vignette. Retrieve from http://www. web. stanford. edu/~ hastie/Papers/Glmnet_Vignette. pdf. Accessed September. 2014 Jun;20:2016.

34. Savova GK, Masanz JJ, Ogren PV, Zheng J, Sohn S, Kipper-Schuler KC, Chute CG. Mayo clinical Text Analysis and Knowledge Extraction System (cTAKES): architecture, component evaluation and applications. Journal of the American Medical Informatics Association. 2010 Sep 1;17(5):507-13.

35. Soysal E, Wang J, Jiang M, Wu Y, Pakhomov S, Liu H, Xu H. CLAMP–a toolkit for efficiently building customized clinical natural language processing pipelines. Journal of the American Medical Informatics Association. 2018 Mar 1;25(3):331-6.

Appendix

Table 6. Clinical Discipline, Care Setting, and EHR Section

Clinical Discipline	Care Setting	EHR Section	Article
Epidemiologist	Pregnant Women (Inpatient)	All clinical notes	Zhong (2018)
	Pregnant Women (Inpatient)	All clinical notes	Zhong (2018)
	Patients with depression (Primary Care Clinics)	History of present illness	Anderson (2015)
Physician	Adolescents (Inpatient Mental Health)	All clinical notes	Carson (2019)
	All (Emergency Department)	Chief complaint, clinical observation, laboratory tests, diagnostic/therapeutic intervention, specialists' notes	Metzger (2017)
	Adolescents (Inpatient and Outpatient Mental Health)	Events and Correspondence	Downs (2017)
	All (Inpatient)	Discharge summaries	Haerian (2012)
	All (Veterans Affairs)	All clinical notes	Hammond (2013)
	Persian Gulf Veterans (Veterans Affairs)	Clinical documents	Hammond (2014)
Undetermined/ Miscellaneous	All (Inpatient and Outpatient Mental Health)	Events and Correspondence*	Fernandes (2018)
	All (Emergency Departments)	Chief complaint and discharge diagnoses	Kuramoto-Crawford (2017)
	All (Inpatient and Outpatient)	All machine-readable notes	Hazlehurst (2019)

*Events and Correspondence are the two different types of notes in the EHR of South London and Maudsley National Health Service Trust. Events are the same as progress and consulting notes, while correspondence is all patient-provider and provider-provider communication, including things like letters that have been scanned in.

Table 7. NLP Systems Used

NLP Tool	Article
cTAKES	Zhong (2018) Zhong (2018) Carson (2019) Hammond (2014)
urgindex and a French-language medical multi-terminology indexer called ECMT	Metzger (2017)
General architecture for text engineering (GATE) and TextHunter	Fernandes (2018)
Homegrown and manual	Downs (2017)
MedLEE	Haerian (2012)
Keyword search	Hammond (2013) Kuramoto-Crawford (2017) Anderson (2015)
Mediclass	Hazlehurst (2019)

Table 8. Phenotyping Approach

Approach	Article
Rule-based	Zhong (2018) Downs (2017) Haerian (2012) Kuramoto-Crawford (2017) Hammond (2013) Hazlehurst (2019)
K-statistic agreement between ICD-9 code and data source (e.g., HPI, PHQ-9 item)	Anderson (2015)
Rule-based (ideation) and hybrid (attempt) rule-based with SVMs with classic bag of words	Fernandes (2018)
Predictive association rules, decision trees, neural networks, logistic regression, random forest, support vector machine, naïve Bayes	Metzger (2017)
Conditional random fields classifier	Hammond (2014)
Adaptive elastic net	Zhong (2018)
Random forest	Carson (2019)

Table 9. Reported Performance

Article	PPV (precision)	F1 Score	Sensitivity (recall)	Area Under the Curve
Zhong (2018)	NLP-identified 0.30, diagnostic codes 0.76	NA	NA	NA
Zhong (2018)	0.63	NA	0.58	0.83
Carson (2019)	0.42	NA	0.83	0.68
Kuramoto-Crawford (2017)	NA	NA	NA	NA
Metzger (2017)	Random forest 93.1, naïve Bayes 96.9, support vector machine 92.6, predictive association rules 91.7, decision trees 91.4, neural networks 89.5, logistic regression 93.6	Random forest 95.3, naïve Bayes 95.3, support vector machine 90.4, predictive association rules 90.2, decision trees 87.6, neural networks 80.2, logistic regression 70.4	Random forest 95.9, naïve Bayes 94.9, support vector machine 89.8, predictive association rules 89.8, decision trees 86.7, neural networks 78.6, logistic regression 89.8	NA
Anderson (2015)	NA	NA	NA	NA
Fernandes (2018)	Suicide ideation (heuristic) 91.7, suicide attempt (machine learning) 82.8	NA	Suicide ideation (heuristic) 87.8, suicide attempt (machine learning) 98.2	NA
Downs (2017)	0.81	0.88	0.95	NA
Haerian (2012)	0.97	NA	NA	NA
Hammond (2013)	NA	NA	NA	NA
Hammond (2014)	0.797	0.831	0.872	NA
Hazlehurst (2019)	0.82	NA	0.74	NA

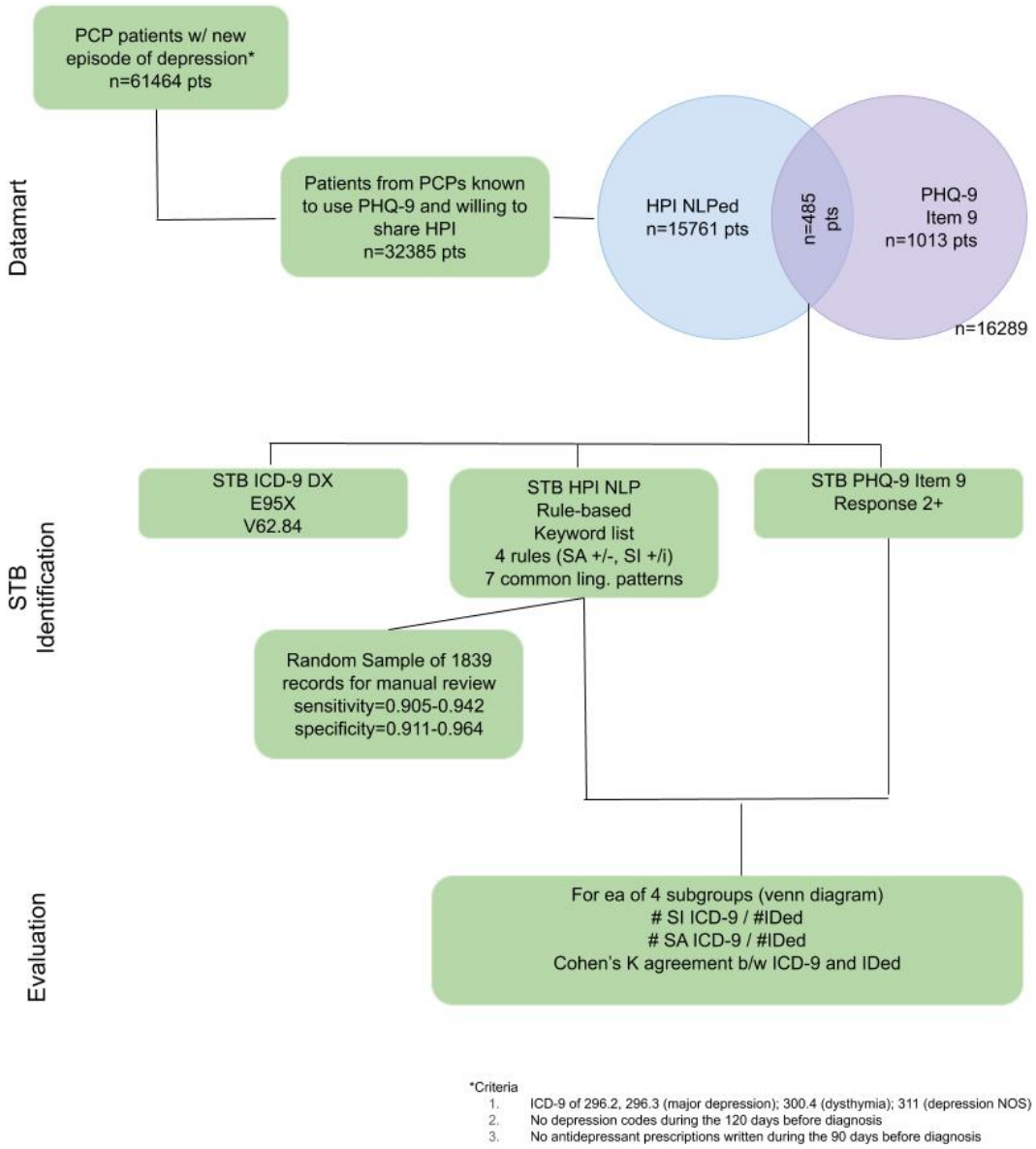

Figure 8. UML Diagram for Anderson (2015)

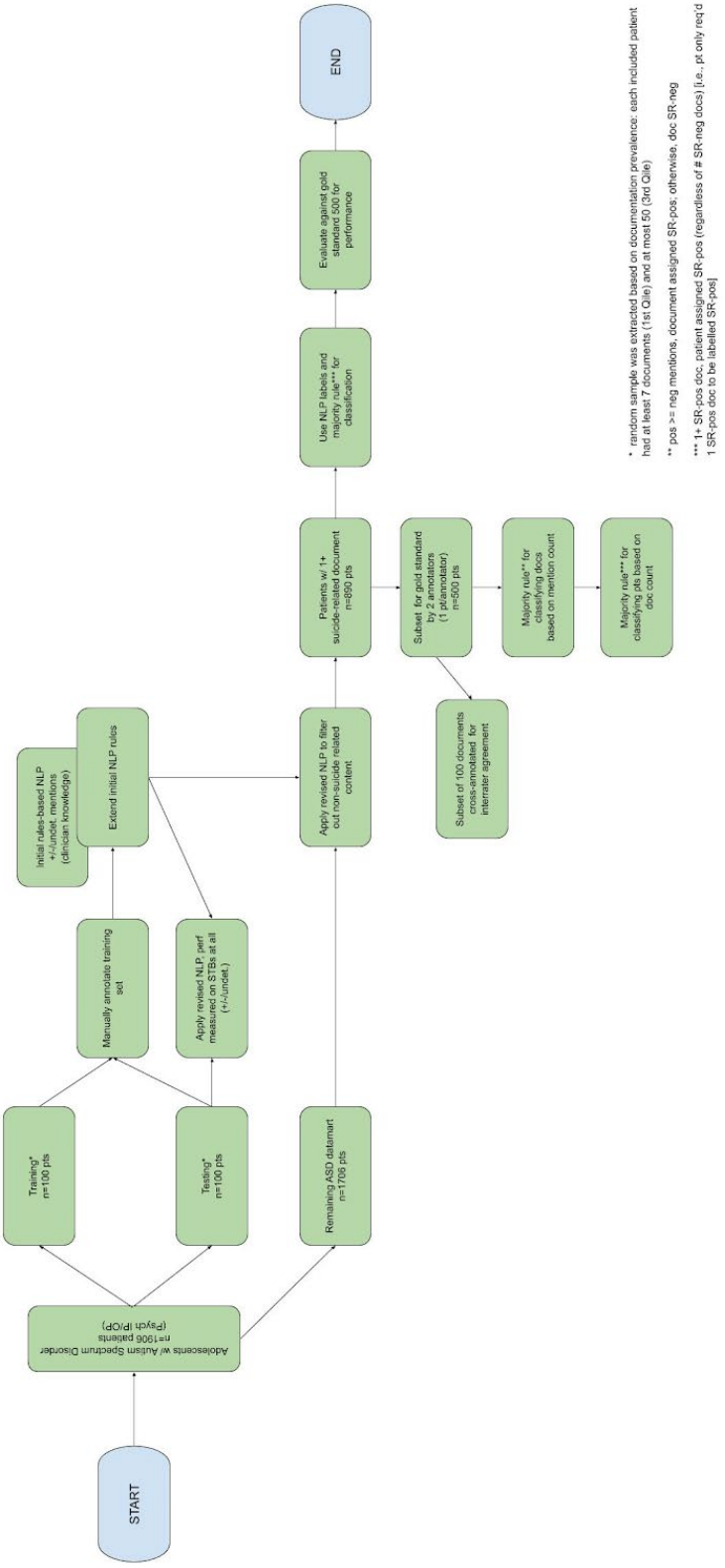

Figure 9. UML Diagram for Downs (2017)

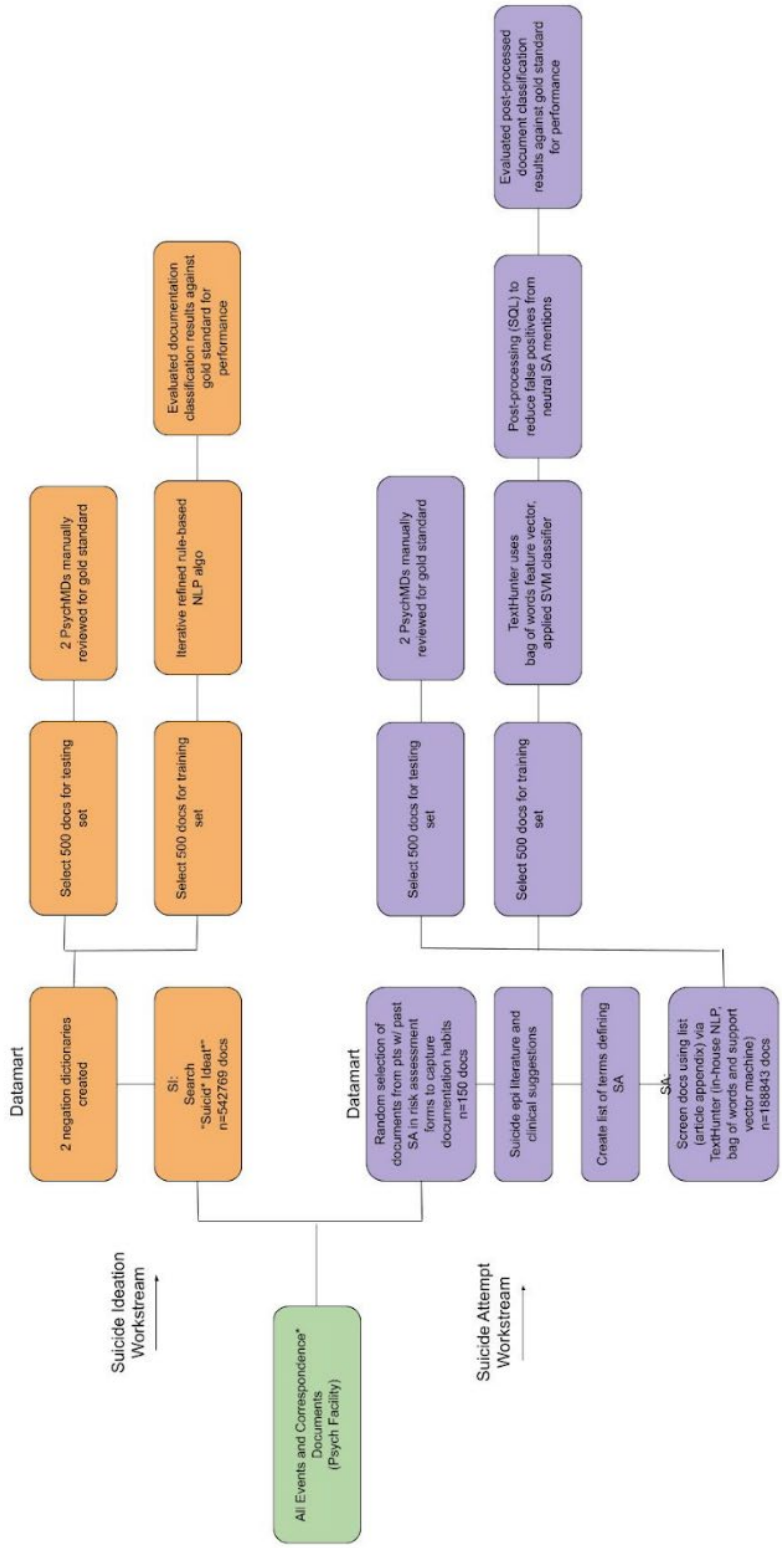

Figure 10. UML Diagram for Fernandex (2018)

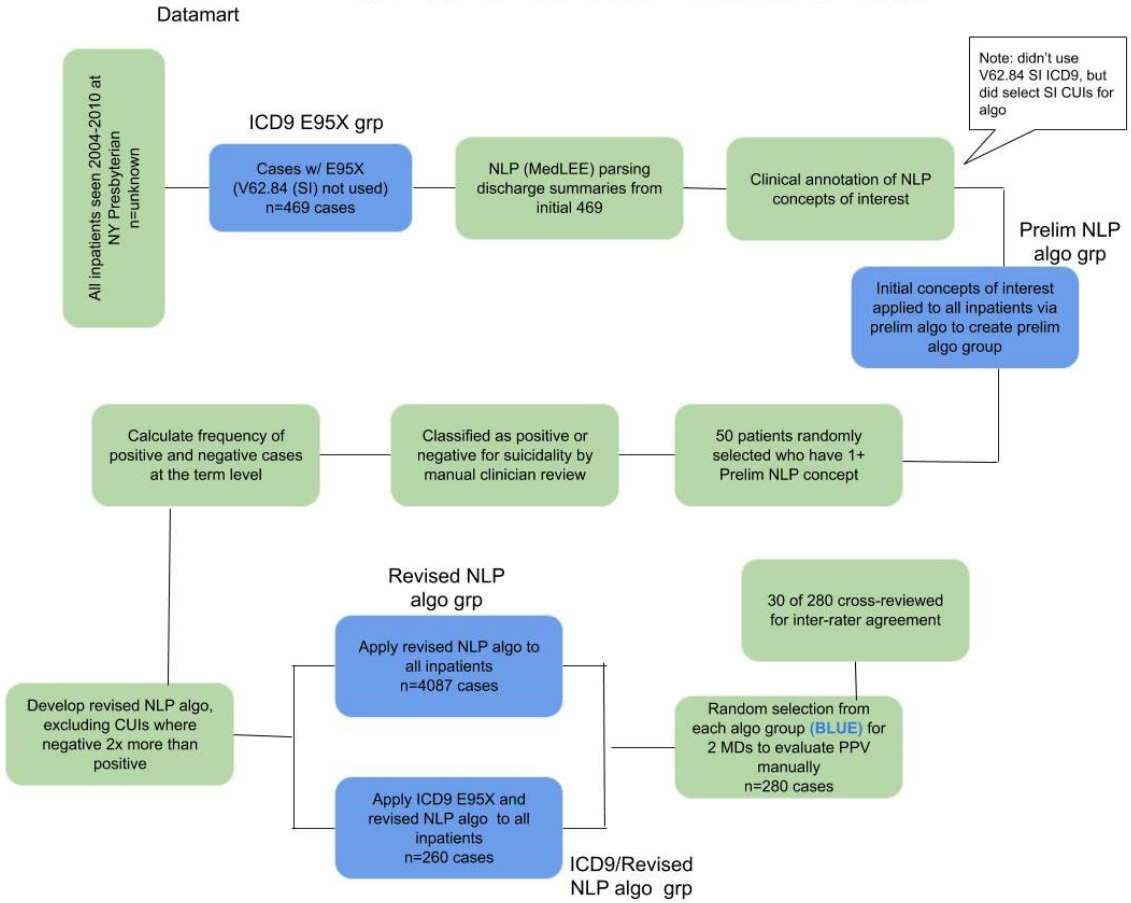

Figure 11. UML Diagram for Hearian (2012)

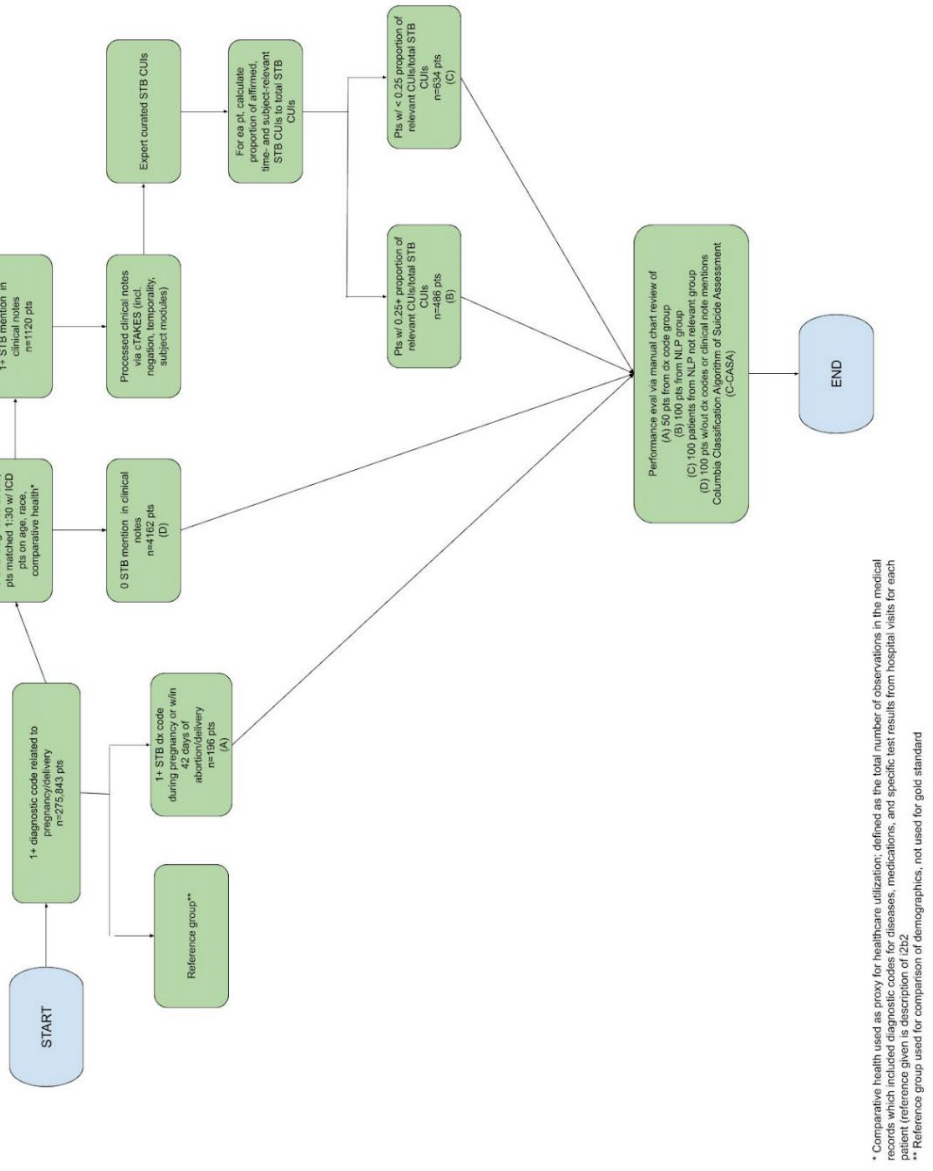

Figure 12. UML Diagram for Zhong (2018)