DEVELOPMENT OF AN AUTOMATED SYSTEM FOR QUERYING RADIOLOGY REPORTS AND RECORDING DEEP VENOUS THROMBOSES AND PULMONARY EMBOLI

BY

Wazim R. Narain

A Dissertation Submitted to

Rutgers Biomedical and Health Sciences

School of Health Related Professions

Department of Health Informatics

[In partial fulfillment of the requirements for the Degree of Doctor of Philosophy]

April 2016

Final Dissertation Approval Form

DEVELOPMENT OF AN AUTOMATED SYSTEM FOR QUERYING RADIOLOGY REPORTS AND RECORDING DEEP VENOUS THROMBOSES AND PULMONARY EMBOLI

BY

Wazim R. Narain

Dissertation Committee:

Shankar Srinivasan, Ph.D., Advisor

Frederick Coffman, Ph.D.

Pete Stetson, MD, MA

Approved by the Dissertation Committee:

_____ Date_____

_____ Date_____

_____ Date_____

ABSTRACT

As the United States healthcare system transitions to a pay for performance model in response to increasing costs and utilization, assessing quality of care has come to the forefront. Venous thromboembolisms (VTE), which include deep vein thrombosis (DVT) and pulmonary embolism (PE), is a key measure of quality of hospital care and are associated with increased morbidity, mortality and cost in hospitalized patients. Traditional ways of measuring quality and identifying adverse events such as VTE using administrative data are convenient but lack accuracy. Manual review of clinical records is widely considered the gold standard but resource intensive. Consequently, this study sought to determine the accuracy of Natural Language Processing (NLP) and machine learning classifiers in identifying VTE from free text data.

This study used radiology reports performed within 30 days of surgery for hospital patients sampled from 2011 through 2014 as part of the American College of Surgeons-National Surgical Quality Improvement Program (ACS-NSQIP). Though records for this sample were previously reviewed and VTE cases identified, a total of 909 ultrasound reports and 1,837 computed tomography (CT) angiogram reports were again manually reviewed to identify DVT/PE within each report and served as the gold standard. The Naïve Bayes, k-Nearest Neighbors (kNN), C4.5 decision tree, and support vector machine (SVM) classifiers were trained on 70% of the total preprocessed reports and performance was assessed on the remaining 30%.

DVTs were identified in 16.8% of all ultrasound reports and PEs were identified in 5.0% of all CT angiogram reports. SVM yielded the best results in classifying both DVT and PE, with precision of 91.3%, recall of 95.5% and F-measure of 93.3% for DVT

classification and precision of 93.1%, recall of 87.1% and F-measure of 90.0% for PE classification. In conclusion, NLP along with statistical machine learning classifiers can accurately identify VTE from narrative radiology reports.

TABLE OF CONTENTS

ABSTRACT 3

LIST OF FIGURES 7

LIST OF TABLES 8

CHAPTER I: INTRODUCTION 9

1.1 Statement of the Problem 9

1.2 Background of the Problem 10

1.3 Research Purpose, Aims and Hypothesis 12

1.4 Significance of Study 13

CHAPTER II: LITERATURE REVIEW 14

2.1 General Overview of Previous Literature 14

2.2 Coding in Healthcare 14

2.3 Measuring Performance with Administrative Data 15

2.4 Reporting of Adverse Events/ Public Reporting of Quality of Care Indicators 16

2.5 Challenges in Capturing Significant Events 18

2.5.1 Mandatory Reporting and Administrative Data 18

2.5.2 Using Administrative Data to Assess Quality of Patient Care 20

2.6 Burden and Significance of DVT PE 25

2.7 Text Mining 27

2.7.1 Natural Language Processing, Text Mining and Machine Learning 27

2.7.2 Text Mining in Healthcare 28

2.7.3 Using Text Mining and NLP to Detect Adverse Events from Free Text 36

2.8 Natural Language Processing 40

2.9 Document Classification 41

2.10 Machine Learning 43

2.10.1 Naïve Bayes 44

2.10.2 Decision Trees 47
2.10.3 K- Nearest Neighbors 50
2.10.4 Support Vector Machines 51
CHAPTER III. RESEARCH METHODS 53
3.1 Overview 53
3.2 NSQIP 54
3.3 Analysis 54
CHAPTER IV: RESULTS OF DATA ANALYSIS 58
4.1 Introduction 58
4.2 Naïve Bayes 59
4.3 K-Nearest Neighbors 62
4.4 C4.5 Decision Tree 63
4.5 Support Vector Machine 67
CHAPTER V: DISCUSSION AND STUDY LIMITATIONS 68
CHAPTER VI: FUTURE RESEARCH 72
REFERENCES 74

LIST OF FIGURES

Figure 1: Positive predictive value and false positives analysis of DVT/PE, iPTX and APL cases 24
Figure 2: Content of Natural Language Query Elements 32
Figure 3: Most well known machine learning algorithms, or set of techniques that allow computers to learn from examples 34
Figure 4: Performance of machine learning algorithms for Study 1, assigning ICD-9 CM codes to radiology reports 35
Figure 5: Performance of machine learning algorithms for Study 2, identifying cases and non cases for liver disorders 35
Figure 6: An example of some optimizing tools changing a document from its natural form into a bag of words 42
Figure 7: Using information theory to choose attributes by calculating information gain 49
Figure 8: Final C4.5 decision tree based on the weather data 50
Figure 9: K- Nearest Neighbors classifier. 51
Figure 10: Support vectors defining the linear boundary in a SVM model. 52
Figure 11: Report classification workflow 56
Figure 12: Decision tree produced using the WEKA J48 classifier (C4.5) on DVT training documents. 64
Figure 13: Decision tree produced using the WEKA J48 classifier (C4.5) on PE training documents. 66
Figure 14: Classifier performance in detecting VTE from ultrasound and CT angiogram radiology reports 70

LIST OF TABLES

Table 1: Contents of a Uniform Hospital Discharge Dataset 15
Table 2: NYPORTS ICD-9 Codes with a >30 Percent Match Rate 20
Table 3: PPV and Percentage of Cases Present on Admission among Flagged Cases and False Positives..... 25
Table 4: AHRQ PSI Rate with and without POA Indicator 25
Table 5: Text Mining Accuracy Results in Identifying Follow-Up Appointment Elements 30
Table 6: Results of Natural Language Queries on the Validation Set (N=118) 32
Table 7: Accuracy of critical results algorithms 37
Table 8: Patient Symptoms and Outcomes 46
Table 9: Counts and Probabilities of Symptoms and Outcomes..... 46
Table 10: New Patient with Unknown Outcome 46
Table 11: Weather data (Witten, Hall and Frank)..... 48
Table 12: Confusion matrix using Naïve Bayes classifier for DVT training documents . 59
Table 13: Confusion matrix using Naïve Bayes classifier for DVT test documents 60
Table 14: Confusion matrix using Naïve Bayes classifier for PE training documents..... 61
Table 15: Confusion matrix using Naïve Bayes classifier for PE test documents..... 62
Table 16: Performance of KNN classifier on DVT training documents 62
Table 17: Performance of KNN classifier on DVT test documents 63
Table 18: Performance of KNN classifier on PE training documents 63
Table 19: Performance of KNN classifier on PE test documents 63
Table 20: Performance of C4.5 classifier on DVT training documents 64
Table 21: Performance of C4.5 classifier on DVT test documents 65
Table 22: Performance of C4.5 classifier on PE training documents 65
Table 23: Performance of C4.5 classifier on PE test documents 66
Table 24: Performance of SVM classifier on DVT training documents 67
Table 25: Performance of SVM classifier on DVT test documents 67
Table 26: Performance of SVM classifier on PE training documents 68
Table 27: Performance of SVM classifier on PE test documents 68

CHAPTER I: INTRODUCTION

1.1 Statement of the Problem

It is no surprise that the healthcare environment in the United States is going through dramatic changes. The economic burden and utilization of healthcare services in the U.S has become highly unsustainable and inefficient. As evidenced through such legislation as the Affordable Care Act (2010) and the Centers for Medicare and Medicaid's (CMS) Inpatient Quality Reporting (IQR) and Hospital Value-Based Purchasing (VBP) programs, healthcare providers are being held accountable for the quality of services provided.[1,2] Reimbursement in healthcare has now changed from paying for volume to paying for performance. The number of quality measures hospitals are asked to report and to be displayed for public consumption is increasing. These measures, many of which are identified through administrative data, can be viewed on sites such as Hospital Compare and are used by organizations such as U.S. News and World Report and Consumer Reports to rank hospitals and develop hospital "report cards".[3-5] For example, the Agency for Healthcare Research and Quality's (AHRQ) Patient Safety Indicators (PSI) and CMS's Hospital Acquired Conditions (HAC) Program use the International Classification of Diseases, Ninth Revision (ICD-9) codes to identify conditions related to quality of care during the hospital stay.[6,7] Measures such as these which rely heavily on quality of coding and integrity of data have come under scrutiny as they may not fully describe the complete patient clinical experience. Identifying these conditions through manual review of medical records can give a more accurate picture of these conditions and is considered by most as the "gold standard" but is extremely resource intensive.

Venous Thromboembolism, or VTE, refers to targeted conditions in the CMS HAC program and is also an AHRQ PSI. VTE refers to both Deep Vein Thrombosis (DVT) and Pulmonary Embolism (PE). DVT is a blood clot that forms in a vein deep in the body, mostly in the lower leg or thigh.[8] PE refers to when the blood clot in the deep vein breaks off and travels through the bloodstream, possibly to an artery in the lungs which can cause blockage of blood flow.[8] Incidence of hospital acquired DVT among patients who did not receive prophylaxis is between 10% and 40%, where 10% to 30% of all VTE patients suffer mortality within 30 days.[9] The outcome of a study assessing the accuracy of detecting adverse events such as VTEs using natural language processing and machine learning algorithms can provide insight into alternative methods of identifying conditions other than coding and manual chart abstraction. This study can also assist in identifying gaps in clinical documentation and coding of adverse events.

Previous studies have compared the accuracy of the ICD-9 diagnosis codes and AHRQ PSIs in identifying clinical events from administrative data to manual review of medical records by trained abstractors, revealing low positive predictive value for PSIs.[10-12]Some studies have also used NLP to identify adverse events, including VTE, and other aspects of the clinical experience from documentation in the medical record with high accuracies when compared to AHR PSIs.[13-17]However, few studies assess the performance of NLP methods in combination with multiple machine learning algorithms in detecting VTE from free text.

1.2 Background of the Problem

The attention to safety in healthcare intensified when the Institute of Medicine (IOM) released its report *To Err Is Human: Building a Safer Health System.*[18] In this

report IOM highlighted the cost and burden of medical errors in the U.S. healthcare system and suggested improving safety thorough a four tiered approach:

1. Establishing a national focus to create leadership, research, tools, and protocols to enhance the knowledge base about safety

2. Identifying and learning from errors by developing a nationwide public mandatory reporting system and by encouraging health care organizations and practitioners to develop and participate in voluntary reporting systems

3. Raising performance standards and expectations for improvements in safety through the actions of oversight organizations, professional groups, and group purchasers of health care

4. Implementing safety systems in health care organizations to en-sure safe practices at the delivery level

Since this report many hospitals now monitor and assess quality and safety through the collection and analysis of various measures. In the AHRQ published report *Making Health Care Safer: A Critical Analysis of Patient Safety Practices,* evidence of harm and burden due to VTEs was presented with the importance of surveillance and prophylaxis use was stressed.[19] DVT/PE are part of the CMS HAC program where preventable conditions deemed to be of high cost and high volume are tracked, publically reported and will impact reimbursement in the future.[7] In New York State, the Department of Health uses the New York Patient Occurrence Reporting and Tracking System (NYPORTS) to identify, correct and prevent patient safety issues.[20] NYPORTS is a mandatory reporting system that collects information from hospitals concerning adverse events. For these initiatives, events are identified using codes assigned for billing purposes, and cases may be manually reviewed for accuracy.

Many studies analyzing the accuracy of coded data have found that measuring quality performance with these methods can yield substandard results, with existing

methods needing considerable improvement.[21-23]These studies also acknowledge that though manual review of medical records is the gold standard, it also uses substantial resources. However, studies exploring text mining of various aspects of the medical record, such as discharge summaries to assess quality of care, have found positive results when compared to manual review.[16,17] Therefore applying text mining methods to clinical documentation with the goal of identifying quality indicators of care such as DVT/PE and comparing the results to those from coded data can provide insight into improving current methods as well as offer alternatives in monitoring quality of care.

1.3 Research Purpose, Aims and Hypothesis

1. RESEARCH PURPOSE

Based on the arguments above, the purpose of this research is to explore: 1) whether or not deep vein thromboses can be identified from radiology reports and 2) whether or not pulmonary emboli can be identified from radiology reports. In addition this research looks to explore natural language processing methods that can be used to preprocess text from radiology reports, apply machine learning algorithms to this preprocessed text based on what was done from previous research, and quantify the performance of each of these machine learning methods with the goal of assessing performance.

2. OBJECTIVES AND HYPOTHESES

1- To study the performance of the Naïve Bayes classifier in detecting deep vein thromboses and pulmonary emboli from radiology reports

Hypothesis: the Naïve Bayes classifier can accurately identify positive DVT and PE radiology reports.

2- To study the performance of the K-Nearest Neighbors (KNN) classifier in detecting deep vein thromboses and pulmonary emboli from radiology reports

Hypothesis: the K-Nearest Neighbors classifier can accurately identify positive DVT and PE radiology reports.

3- To study the performance of a decision tree learning method in detecting deep vein thromboses and pulmonary emboli from radiology reports

Hypothesis: decision trees can accurately identify positive DVT and PE radiology reports.

4- To study the performance of the support vector machines (SVM) learning method in detecting deep vein thromboses and pulmonary emboli from radiology reports

Hypothesis: Support Vector Machines accurately identify positive DVT and PE radiology reports.

1.4 Significance of Study

Studies such as this are essential in moving towards a healthcare system that delivers quality of care in an efficient manner while controlling costs. Hospitals are tasked with measuring quality, which includes reviewing aspects of the electronic medical record rich with information though limited in methods of extracting such information. Currently, manual review proves accurate but is time consuming and costly. Exploring methods to extract the same information in an automated fashion can have a meaningful impact in reducing healthcare costs.

CHAPTER II: LITERATURE REVIEW

2.1 General Overview of Previous Literature

Many studies have analyzed the accuracy of indicators assessing quality of care. Specifically, studies show that applying natural language processing techniques to free text can identify adverse events with higher accuracy when compared to identifying events from administrative data. Most of these studies use NLP in rule-based systems. However, studies assessing performance of NLP methods in combination with statistical machine learning classifiers in identifying adverse events such as VTE are lacking.

2.2 Coding in Healthcare

Medical coding involves translating narrative descriptions of diseases, injuries and procedures into numeric or alpha numeric codes.[24] Codes are commonly used for reimbursement purposes, administrative functions such as staffing and scheduling of services, and identifying patient symptoms or co morbidities. The vast majority of payments for healthcare providers are from filed insurance claims which require CPT-4 (Current Procedural Terminology, 4th Revision) and ICD9-CM codes.[25]These codes are derived from the medical records made during the course of the patient visit. Medical coders review the notes in the medical record and assign the codes.

Physician documentation is extremely important for coding. When reviewing the medical record, coders can only use documentation by physicians who are directly caring for the patient during the admission.[26] Coders can use documentation by resident physicians, physician assistants, or nurse practitioners, but with documented agreement from the attending. As a result, clinicians need to document using appropriate terminology for diagnosing conditions and symptoms.

2.3 Measuring Performance with Administrative Data

Administrative or claims data are readily accessible, fairly inexpensive to acquire and maintain, and contain diverse and large amounts of information. CMS describes administrative data as information that is collected, processed and stored in automated information systems.[27]These data contain enrollment or eligibility information as well as claims and encounter information. Hospital specific information may include claims, encounters and information on services pertaining to prescription drugs, laboratory tests and clinic visits. Some of these basic data elements are listed in Table 1.[27]

-Personal Identification
-Date of Birth
-Sex
-Race and ethnicity
-Residential zip code
-Hospital identification
-Admission date
-Discharge date
-Attending physician identification
-Operating physician identification
-Codes for principal diagnosis and other diagnoses
-Codes and dates for principal procedure and other procedures
-Disposition of the patient
-Expected principal source of payment

Table 1: Contents of a Uniform Hospital Discharge Dataset

Capturing information about the episode of care for billing and utilization is the main purpose of administrative data, though these datasets are increasingly being used for assessing quality of care, raising many concerns. Some argue that this information does not describe the full clinical experience and is subject to errors and omission of information. Despite these concerns, many state and federal organizations put a lot of

weight into measures derived from these data and are financially penalizing organizations based on performance on these measures.

2.4 Reporting of Adverse Events/ Public Reporting of Quality of Care Indicators

Stated earlier, one of the strategies mentioned in IOM's report *To Err Is Human: Building a Safer Health System* is the development of a mandatory nationwide public reporting system and encouraging healthcare organizations and practitioners to participate in voluntary reporting systems, with the goal of identifying and learning from errors.[18] The expectation at the time was that state governments would be required to collect standardized data about adverse events that result in serious patient harm or death. Currently, state and federal agencies require hospital and healthcare organizations to submit data on a regular basis. Some of these data are chart abstracted, where patient records are reviewed to look for adherence with process of care measures pertaining to, for example, heart failure or acute myocardial infarction. Administrative data are also submitted and used to determine events through codes such as in CMS's HAC program. For example, a code of 998.4 that is indicated to have occurred in the hospital setting is considered a preventable adverse event and can be publicly reported. Healthcare consumers and interested stakeholders can go to various websites such as Hospital Compare to see this information by hospitals.[3]

In an effort to promote patient safety, New York State established the New York Patient Occurrence Reporting System (NYPORTS).[28]One of the tools used by the Department of Health to identify, correct and prevent safety deficiencies; NYPORTS is a mandatory reporting system that collects information on adverse events from hospitals and diagnostic treatment centers. Serious NYPORTS occurrences, which are defined as

those with an impact on the patient, average about nine percent of all NYPORTS reports. These occurrences require a root cause analysis (RCA) of the human, equipment and/or system failures that led to the adverse event. A plan of correction that is approved by the Department of Health is also required to reduce the risk of future similar events. The Department of Health implemented several NYPORTS-related patient safety initiatives from 2005-2007 including Pulmonary Embolism (PE) Prevention. This project involved six hospitals in a study with the goal of improving physician compliance in using proper prophylaxis in the prevention of PE. In this study, records were selected that met the PE diagnostic criteria from the NYSDOH Statewide Planning and Research Cooperative System (SPARCS) inpatient discharge data. Potential PE cases were identified on a quarterly basis for each of the six study hospitals if they met any of three criteria:

1. PE and infarction (ICD-9-CM code of 415.1X) or obstetrical blood clot embolism (673.2X) reported in any of 14 secondary diagnosis fields as not present at the time of admission.
2. PE and infarction (415.1X) or obstetrical blood clot embolism (673.2X) as a principal diagnosis, along with a hospitalization less than 31 days prior to the admission date associated with the target discharge.
3. Any secondary diagnosis of PE and infarction (415.1X) or obstetrical blood clot embolism (673.2X) reported as present on admission, along with a hospitalization less than 31 days prior to the admission date associated with the target discharge.[28]

In addition hospitals were asked to submit detailed information on all adult PE patients including detailed descriptive data, type and timing of prophylaxis given, and the method of diagnosis. Intervention included implementation of a prophylaxis protocol and risk factor assessment at the six hospitals.

After reviewing each case, results showed a significant increase in the use of prophylaxis among PE patients, where post intervention prophylaxis rate was 88.9%, statistically significantly better than the baseline rate of 76.1%. Researchers state that

even with concerted efforts to increase prophylaxis use at the six hospitals, the rate of use did not pass 90%. They speculate that this could be due to clinicians perceiving patients to be at lower risk than they actually are, or lack of the hospital's reliability to respond and maintain protocols due to staff turnover.

This example demonstrates that mandatory reporting can facilitate change in delivery of care in improving quality and outcomes. However, these reported methods have weaknesses which may allow for underreporting and errors, as will be discussed later.

2.5 Challenges in Capturing Significant Events

2.5.1 Mandatory Reporting and Administrative Data

Though now mandated by government agencies and publicly reported, accuracy of reported adverse events is still a challenge for many organizations. A 2012 report by the Office of the Inspector General (OIG) concluded that hospital incidence reporting systems do not capture most patient harm.[29]In this study, the 189 hospitals sampled used incident reporting systems to capture adverse events. Of this sample, 34 hospitals that reported adverse events were interviewed. They indicated that they rely on incident reporting systems to capture the bulk of information on these events, which they use for activities surrounding improvement of patient safety. OIG found only 14% of incidents involving Medicare beneficiaries discharged in 2008 were captured by incident reporting systems. Because of staff misperception of what constitutes patient harm, 62% of events went unreported. An additional 25% went unreported because, in these instances, staff did not report these incidences that they would normally report. As this shows lack of reporting standards contributes to underreporting, OIG recommended that AHRQ and

CMS collaborate to create a list of potentially reportable events that can be used to educate hospital staff as well as medical and nursing students. Focusing on a subset of events rather than many events may lead to a higher rate of reporting by hospital staff.

Discussed earlier, NYPORTS has significantly improved the way hospitals identify and track adverse events while improving quality of patient care. Though as with many mandatory reporting systems it has weaknesses similar to those mentioned in OIG's report. Tuttle, Panzer and Baird analyzed NYPORTS data to describe how administrative data can improve identification and reporting of adverse events.[10] Stemming from a 2000 NYSDOH announcement that they would be using the Statewide Planning and Research Cooperative System (SPARCS) data to understand NYPORTS reporting rates, the authors sought to gauge the degree of underreporting in NYPORTS by looking at their own administrative data at Strong Memorial Hospital in Rochester, NY. The SPARCS dataset contains New York State billing discharge data, including ICD-9 codes. ICD-9 codes were identified for 24 NYPORTS categories. Using Strong Memorials Hospital's inpatient data, patient lists were developed for each code, excluding cases already identified in NYPORTS reporting. Pulling cases by ICD-9 defined events yielded a 30% or more match for 13 of the NYPORTS codes (Table 2), ranging from 35.7% to 100% with an average match rate of 56%. In total, 560 reviews identified 187 (33.4%) reportable events for the code the case was being screened for and 26 events for another NYPORTS code not being screened for.

Code	Description	Percentage
401	New pulmonary embolism	45.5
402	New DVT	64.3

601	New neurological deficit	40.0
603	Cardiac arrest with successful resuscitation	42.4
604	AMI	61.9
605	Death	47.0
701	2nd or 3rd degree burns	66.7
801	Injury Requiring repair, organ removal, or procedure	60.9
803	Hemorrhage or hemotoma requiring drainage	61.9
806	Breakage or shift of implant, device, or graft	35.7
808	Post-op wound infection	46.7
851	Hysterectomy in pregnant woman	100.0
854	Circumcision requiring repair	50.0

Table 2: NYPORTS ICD-9 Codes with a >30 Percent Match Rate[10]

The authors note that at a statewide level "underreporting revealed some continued ambiguity and lack of specificity in selection criteria, which resulted in further refinement of definitions." At their institution, there is a continued effort on a monthly basis to run a list of high yield ICD-9 codes not in their NYPORTS database. This is a great example of how administrative data being used to supplement a manual review process using less resources and time.

2.5.2 Using Administrative Data to Assess Quality of Patient Care

The inclination to use administrative data in quantifying quality of care is obvious as administrative data are readily available, inexpensive to acquire, and contains information on large populations.[27] As Iezonni states, "Administrative data cannot

elucidate the interpersonal quality of care, evaluate the technical quality of processes of care, determine most errors of omission or commission, or assess the appropriateness of care." This illustrates a major complaint from clinicians in that administrative data does not give the full picture of the patient-provider experience.

Quantifying quality of care is reliant on many factors including accuracy and completeness of coding, data quality standards across institutions, timing of events, and structure of administrative databases.[27] Measuring quality of care using administrative data is highly dependent on coding quality. The accuracy of coding has been scrutinized since codes impact reimbursement. For example, the concept of "DRG creep" is known as when diagnoses are coded to yield higher financially weighing Diagnosis Related Groups in an attempt to achieve optimal reimbursement, resulting in bias in coding practices. Coding standards across institutions also vary, where institutions that code more may capture more adverse events, therefore impacting quality and outcomes scores.

Kaafarani et al. sought to assess the validity of AHRQ's Patient Safety Indicators (PSI) as these ICD-9 code based indicators are being used to detect potential adverse events.[30]They examined the Positive Predictive Value (PPV) of three surgical PSIs; Postoperative Deep Vein Thrombosis and Pulmonary Embolism (DVT/PE), Iatrogenic Pneumothorax (iPTX), and Accidental Puncture and Laceration (APL). PPV was calculated by dividing the number of true positives by the number of records reviewed that were identified as positive. The AHRQ PSI software (version 3.1a) was applied to 2003-2007 Veterans Health Administration (VA) data from a sample of eight VA hospitals. Patients suspected of having one of the PSIs above were flagged and a

retrospective chart review was conducted on 336 charts (112 per PSI) by trained nursing staff to determine the number of true or false positives.

The VA data consisted of 2,343,088 admissions of which 6,080 were flagged for DVT/PE (0.28%), 1402 for iPTX (0.06%) and 7,203 for APL (0.31%). Results showed variance in the PPVs of the three PSIs with DVT/PE having the lowest PPV of 43%, iPTX with 73% and APL with the highest at 85% (Figure 1). The authors concluded that the PSIs studied have the potential to detect patient safety events though accuracy can be improved. Some of their suggestions included adjusting coding guidelines and increasing coders' clinical knowledge. They also highlight that ICD-9 CM codes are used for billing and using them for clinical and quality of care assessment will require changes in coding schemes. Ultimately they felt that using these quality measures for pay for performance and public reporting is premature. A follow up study using the same VA data broadened the focus to 12 PSIs where results showed great variability in PPV from 28% for Postoperative Hip Fracture to 87% for Postoperative Wound Dehiscence (Table 3). These results were comparable to those of other VA and non-VA studies.[11, 12]

The authors of this study also found limitations in using the AHRQ PSIs for detection of adverse events. Some of the reasons they list for the variation in results were differences in hospital coding practices, lack of POA codes, lack of precise or meaningful codes and poor documentation. As seen in Table 3, correct POA coding can increase PPV. For example, DVT/PE cases that are POA would be dropped therefore potentially increasing PPV by as much as 13% and decreasing the number of flagged cases. Using DVT/PE as an example again, the inability to "distinguish the type and nature of thrombosis" was a common reason for miscoding which was also stated in studies cited

earlier where cases flagged included arterial (not venous) thrombosis or had a history of thrombosis.[11,12] In conclusion the authors felt that PSIs are not ready to be used for public reporting or pay for performance metrics though they are a good tool for screening quality improvement and are a step in the right direction.

Bahl and others performed a study to assess AHRQ PSIs and their ability to flag conditions present on admission using University of Michigan Health System 2006 discharges.[31]They applied the AHRQ PSI software to 35,994 adult inpatient cases. PSIs of focus were those that use both principal and secondary diagnoses. Numerator and denominators were determined for 14 PSIs with and without POA. Table 4 shows the results of unadjusted rates of PSIs with and without POA and if they were significantly different. PSI rates were lower for all but one of the PSIs when considering POA. Though a more telling result of the study is that after nurse review of these cases, the agreement between the coders and the nurse reviewers for cases flagged for conditions that were POA was low (49%), compared with agreement for cases that were flagged for complications that happened during the hospital stay (89%). These results show the importance of the POA indicator in determining hospital acquired complications and also highlights the inconsistency of which POA is determined, as seen in the difference between coders and the nurse reviewers. As seen in the previous studies discussed, the validity of the AHRQ PSIs for use in public reporting and pay for performance is highly questioned. Administrative data in general and coded data meant to be used for billing can be used to identify problems though going as far as penalizing institutions and assessing performance is up for debate.

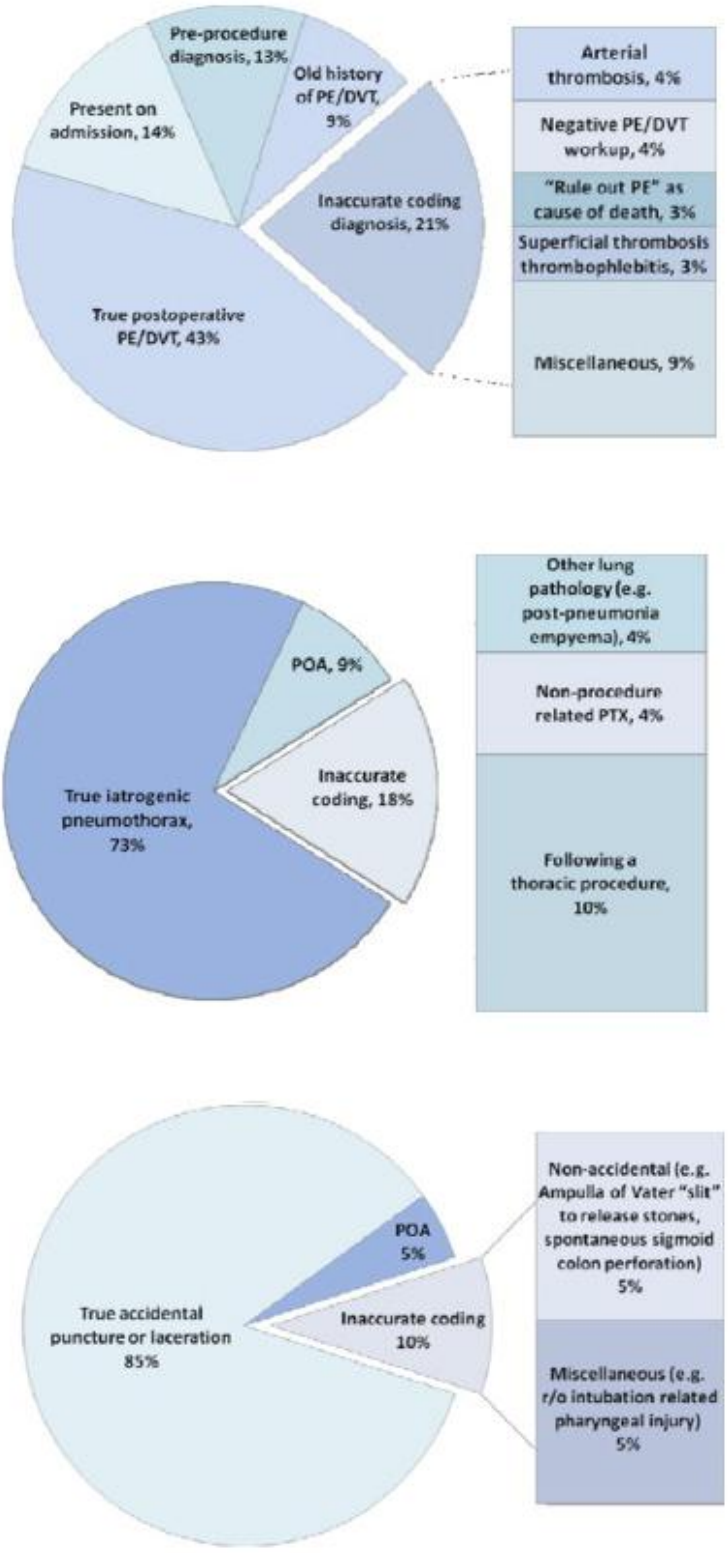


Figure 1: Positive predictive value and false positives analysis of DVT/PE, iPTX an APL cases.[30]

PSI No.	PSI Name	Sample (n)	PPV (%) (95% CI)	POA (%) in All Flagged Cases	POA (%) in False Positives	Sample w/o POA (n)	PPV w/o POA (%) (95% CI)
3	Decubitus Ulcer	112	30 (22–40)	59	83	46	74 (59–86)
5	Foreign Body Left in During Procedure	93	46 (36–55)	30	56	65	66 (53–77)
6	Iatrogenic Pneumothorax	112	73 (64–81)	8	33	103	80 (71–87)
7	Central Venous Catheter-related Bloodstream Infections	112	38 (29–47)	19	30	91	46 (36–57)
8	Postoperative Hip Fracture	46	28 (15–43)	52	73	21	62 (38–82)
9	Postoperative Hemorrhage or Hematoma	112	75 (66–83)	8	32	103	82 (73–89)
10	Postoperative Physiologic and Metabolic Derangements	119	63 (54–72)	18	47	98	77 (67–85)
11	Postoperative Respiratory Failure	112	67 (57–76)	0	0	112	67 (57–76)
12	Postoperative Pulmonary Embolism or Deep Vein Thrombosis	112	43 (34–53)	14	25	96	50 (40–60)
13	Postoperative Sepsis	112	53 (42–64)	14	30	96	61 (51–71)
14	Postoperative Wound Dehiscence	112	87 (79–92)	0	0	112	87 (79–92)
15	Accidental Puncture or Laceration	112	85 (77–91)	5	35	106	90 (82–95)

Table 3: PPV and Percentage of Cases Present on Admission among Flagged Cases and False Positives[30]

	Patient Safety Indicator	Without PoA Variable Rates (1)/1000	With PoA Variable Rates (2)/1000	Odds Ratio	95% CI†
1	Complications of anesthesia	0.95	0.68	1.40	0.63–3.25
3	Decubitus ulcer	33.24	8.77	3.89	3.080–4.954*
4	Failure to rescue	95.95	155.73	0.58	0.448–0.743*
5	Foreign body left during procedure	0.42	0.14	3.00	1.162–9.224*
6	Iatrogenic pneumothorax	1.18	0.90	1.31	0.787–2.198
7	Selected infections due to medical care	5.19	2.27	2.29	1.600–3.343*
8	Postoperative hip fracture	0.33	0.00	Undefined	—
9	Postoperative hemorrhage or hematoma	5.82	5.03	1.16	0.835–1.609
10	Postoperative physiologic and metabolic derangement	4.15	2.33	1.78	1.040–3.131*
11	Postoperative respiratory failure	14.59	14.45	1.01	0.766–1.331
12	Postoperative pulmonary embolism or deep vein thrombosis	22.13	11.49	1.95	1.601–2.378*
13	Postoperative sepsis	9.39	6.03	1.56	0.725–3.550
15	Accidental puncture or laceration	7.09	6.43	1.10	0.912–1.336
16	Transfusion reaction	0.00	0.00	Undefined	—

*Significant difference.

Table 4: AHRQ PSI Rate with and without POA Indicator[31]

2.6 Burden and Significance of DVT PE

Hospital acquired VTE is a major source of morbidity and mortality in the United States. Studies estimate that 10%-30% of all VTE patients suffer mortality within 30 days, with the majority of deaths occurring among those with PE; death occurs rapidly in

an estimated 20-25% of all PE cases.[9] Hospitalized patients have at least one risk factor for VTE and approximately 40% have three or more risk factors. Risk factors for VTE include, but are not limited to surgery, trauma, immobility, cancer, cancer therapy, previous VTE, age, obesity and central venous catheterization.[9,32-34] The incidence of hospital acquired DVT among patients who did not receive prophylaxis is 10 to 40% among medical or general surgical patients.[9] A study by the CDC using 2007-2009 National Hospital Discharge Survey (NHDS) data reports 547,596 hospitalizations with VTE occur each year among adults (>=18 years old) in the U.S., of which 348,558 of them are DVTs and 277,549 PEs, 78,511 have both DVT and PE. The average annual rates of DVT, PE or VTE among adults were 102,121 and 239 per 100,000 population respectively.[35]

A study using the AHRQ Healthcare Cost and Utilization Project Nationwide Inpatient Sample (HCUP NIS) assessed excess length of stay, charges and deaths attributable to medical injuries using AHRQ PSIs derived from 7.45 million hospital discharge abstracts from 994 acute-care hospitals across 28 states in the year 2000.[36] This study found that an excess length of stay of 5.36 days was attributable to postoperative pulmonary embolism or deep vein thrombosis, with excess charges of $21,709 and excess mortality of 6.56%.[36] In an effort to quantify the cost of preventable DVT/PEs, Mahan and others developed a cost model and calculated costs of DVT through literature searches.[37] Results, as reflected in 2010 U.S. dollars, show that the average annual cost of a DVT was $19,767 per patient. The authors estimated the annual cost to be in the range of $7.5 to $39.5 billion. The average annual cost of a hospital acquired DVT was $13,232 per patient leading to an estimated annual U.S. HAC DVT

cost of $5 to $26.5 billion. Estimated cost of HAC "preventable" DVTs ranges from $2.5 to $9.5 billion.[37]

Furthermore, DVT/PE rates are expected to rise in the U.S as many of the associated risks such as obesity, advanced age, chronic diseases and cancer are increasing.[37] Concerns over the cost and burden of community and hospital acquired DVT/PEs have led to policy changes such as public reporting of rates for most hospitals and a push for increased prophylaxis use.

2.7 Text Mining

2.7.1 Natural Language Processing, Text Mining and Machine Learning

Marti Hearst defines text mining as the discovery of new, previously unknown information by extracting information from different written resources, linking together the extracted information to form new facts or hypotheses.[38] Text mining can be accomplished in a few ways such as automatic text classification according to some fixed set of categories, text clustering, automatic summarization, extraction of topics from texts or groups of texts and analysis of trends in text streams.[39] Natural Language Processing or NLP differs from text mining in that NLP looks to break down text with the purpose of interpreting what the text is actually saying in an attempt to mimic language, where grammar, combination of words, and word relationships are taken into account.[39] Both text mining and NLP can be of great use in healthcare as providers are increasingly being asked to document more in the EHR. Machine learning is defined as knowledge for making predictions as obtained from processing training data through a computer.[40] Text mining applies machine learning techniques to accomplish the tasks mentioned above such as text classification.

2.7.2 Text Mining in Healthcare

There are growing examples of text mining and NLP being applied to the healthcare field. Providers spend a significant amount of time on documentation, much of which is never read or used though there may be valuable clinical information contained in these unstructured data. In one study researchers observed rates and times of authoring and viewing clinical documentation using audit logs from electronic health records from an urban academic medical center.[41] Their results showed that users spent 20-103 minutes per day authoring notes and 7-56 minutes per day viewing notes, with physicians spending 90 minutes per day total. Another study observing provider documentation at a 200 bed hospital in Austria showed that physicians spent 26.6% of their daily working time documenting, 27.5% on direct patient care, 36.2% for communication tasks and 9.7% on other tasks, showing that nearly as much time is spent on documentations as is on patient care.[42]

In one example researchers attempted to determine whether text mining can accurately detect follow-up appointment criteria in free text hospital discharge records, the relevance being that follow up appointments arranged at discharge can lower readmission rates.[16] In a retrospective cross sectional study, researchers at the Mayo Clinic in Rochester, MN manually reviewed textual hospital discharge summaries to determine whether records contained specific follow-up appointment elements such as date, time and physician or location. This was compared to data that was derived using text mining software (SAS) which has the capability to retrieve information from text using text parsing. Follow up appointment details are typed directly into an unstructured field of the EMR by a clinical assistant, attending physician or trained transcriptionist and

upon discharge a copy of the dismissal summary is given to the patient containing follow up arrangements.

The dataset consisted of 6,481 free text summaries from 2006 hospital discharges. To be considered complete for a follow up appointment, there had to be a specific date, time and physician name or location of appointment. There were 2 reviewers, one main reviewer who extracted the necessary data elements and another reviewer who reviewed a sample of the records to assess reviewer reliability. Researchers reported that the raw agreement between reviewers was high at 0.97. The data was then evaluated using the SAS text miner software that extracted words or phrases from large collections of unstructured documents. The analyst performing the electronic abstraction had to thoroughly review hundreds of terms from the documentation to select indicators for each appointment element be it date, time and location or physician.

In most text mining studies, outcome measures consist of agreement between text mining and manual review, true positives and negatives. The four main measures are:

- **Positive Predictive Value (PPV)-** percent of records flagged by text mining that actually contain information of focus, in this case percent of records flagged as containing follow-up appointment criteria by SAS Text Miner that actually have the elements identified by manual review
- **Negative Predictive Value (NPV)-** proportion of records not identified as containing follow-up appointment elements through text mining that were truly lacking appointment criteria
- **Sensitivity**- proportion of records containing follow-up appointment information that were identified via text mining

- **Specificity**- percent of records lacking follow up appointment elements not flagged through text mining

Using the text mining software, 96.6% of the records were in agreement with records identified through manual abstraction. Table 5 has the results expressed by the four measures described above for each appointment element. Researchers in this study also point out that manual abstraction (considered the gold standard) of the 6,481 electronic discharge summaries required 43 hours of the reviewer's time at a rate of about 150 records reviewed per hour. The analyst using the text miner software extracted the appointment information from the same records in a total of 14 hours.

Element	Sensitivity, % (95% CI)	Specificity, % (95% CI)	PPV, % (95% CI)	NPV, % (95% CI)
Date	99.6 (99.4–99.8)	84.2 (82.8–85.6)	90.5 (89.6–91.4)	99.3 (98.9–99.6)
Time	99.5 (99.2–99.7)	94.9 (94.0–95.6)	96.0 (95.4–96.6)	99.4 (99.0–99.6)
Physician	97.5 (97.0–97.9)	97.7 (94.3–96.1)	98.8 (98.4–99.1)	95.3 (96.1–94.3)
Location	77.8 (76.5–79.0)	94.2 (93.1–95.1)	96.7 (96.1–97.3)	65.7 (64.0–67.5)

Table 5: Text Mining Accuracy Results in Identifying Follow-Up Appointment Elements[16]

In a similar study, the text of inpatient and outpatient clinical reports was searched with natural language queries for evidence of neurological, vascular, and structural components of a foot exam, which is critical when caring for patients with diabetes.[17] Medical records for 401 eligible patients were randomly selected from a population of approximately 6,000 with a diabetes diagnosis in the Mayo Clinic diabetes registry (patients seen at a primary care clinic between July and September 2000 and July and September 2004). Compliance with the American Diabetes Association and National

Committee for Quality Assurance guidelines was assessed by examining medical records for these patients 12 months prior to index visit. The 401 patients were split into 3 sets; a development set used to determine terms for identifying foot examination, a validation set used to validate the methodology created based on the development set, and a reliability set to determine reliability of manual data abstraction from medical records.

Text queries were compiled from key words that show evidence of a foot exam. Natural language queries were constructed for the four aggregate query elements (Structural, Neurological, Vascular and Anatomy). Figure 2 shows the natural language content of these elements where each element consists of key words. Researchers used this to design three queries that 1) had an Anatomy element plus one other key foot examination element, 2) Anatomy plus 2 of the other elements and 3) Anatomy plus all 3 elements. Results of these queries are shown in Table 6. The query identifying 1 of the 3 components of a foot exam resulted in an overall accuracy (calculated as the proportion of true positives and true negatives to the total number of samples) of 89%, 88% overall accuracy for identifying 2 of the 3 components, and 75% for identifying all 3. The authors concluded that the methodology tested is a low-cost and scalable to monitoring large numbers of patients and can streamline quality of care reporting.

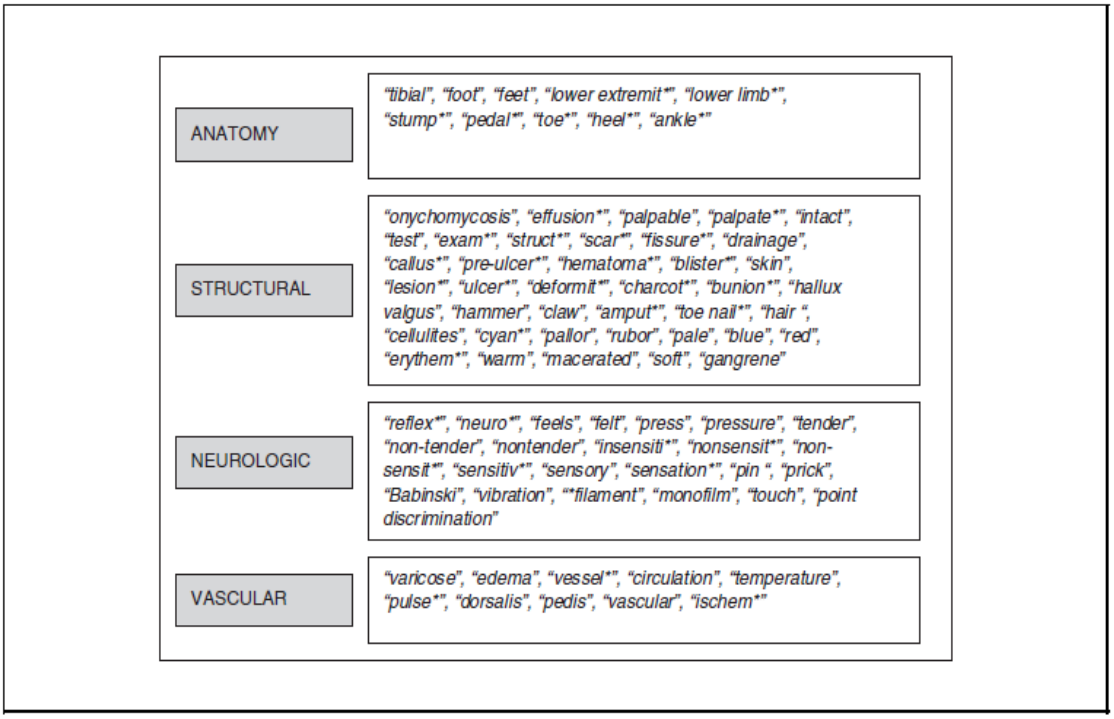


Figure 2: Content of Natural Language Query Elements[17]

Natural-Language Query Type	Positive				Negative				Sensitivity, % (95% CI)	Specificity, % (95% CI)	Accuracy, % (95% CI)
			Total				Total				
	TP	FP	*n*	%	TN	FN	*n*	%			
1 of 3 components	90	6	96	81	15	7	22	19	93 (88–98)	71 (52–90)	89 (83–95)
Any 2 components	88	5	93	79	16	9	25	21	91 (85–96)	76 (58–94)	88 (82–94)
All 3 components	73	5	78	66	16	24	30	34	75 (67–84)	76 (58–94)	75 (68–83)

Table 6: Results of Natural Language Queries on the Validation Set (N=118) [17]

A Dutch study published in 2012 investigated whether text mining can make unstructured narrative from EMRs suitable for epidemiological studies.[43] Researchers used machine learning algorithms to identify cases from non cases in 2 different datasets. The first study sought to automatically assign ICD-9 CM codes to radiology reports using a training set (n = 978) and a test set (n= 976). Each entry was annotated with ICD-9 CM

codes by the radiology department and two independent coding companies. The second study consisted of a collection of EHRs containing medical notes, prescriptions and indications for therapy, referrals, admissions and lab results of approximately 800,000 patients throughout the Netherlands. This study looked for signs of liver damage, which is one of the side effects of using drugs. Querying for specific terms denoting liver disorders (e.g. gall stones, cholecystitis, liver cirrhosis, etc.) returned 53,385 patient records of which 1,000 were randomly sampled for manual review. For both study sets, machine learning algorithms defined by the authors in Figure 3 were applied to the unstructured documentation. Cases and non cases identified by the computer were compared to manual identification of cases and results are represented graphically for each study in Figures 4 and 5. Performance results were varied by each algorithm, with the Ripper software performing the best when considering PPV and sensitivity. The authors conclude that machine learning algorithms are able to detect specific language used by physicians and can distinguish between cases and non cases.

• Naïve Bayes classifier is a simple probabilistic classifier basedonBayes' theorem. It assumes that the presence or absence of every feature independently contributes to the probability that the record belongs to a particular class (case or noncase).
• K-Nearest Neighbors (KNN) classifies records based on the most similar records in the training set. In our experiments, we found the best performance for k = 1, that is, a record is given the same classification as the most similar document in the training set. The similarity is calculated as the Euclidean distance between records.
• C4.5 is an algorithm that generates decision trees. The tree is constructed by first selecting the feature that best splits the set of examples in cases and noncases. Two branches, one for when the feature is found and one for when it is not found, are created, and the process is repeated for each branch. Later, the tree is pruned to remove noninformative branches.
• Random forest is an ensemble of many decision trees.
• MyC is a simple decision-tree learning algorithm that we developed, similar to C4.5. It is based solely on the chi-square test: iteratively, the feature with the highest chi-square score is used to split the data until the p-value becomes higher than a predetermined threshold (currently 0.00001).
• Support vector machine is a sophisticated mathematical approach that attempts to find a function that best separates cases from noncases.
• RIPPER is an algorithm that produces a set of decision rules similar to how C4.5 creates decision trees. An important aspect of RIPPER is that is retains part of the examples to test whether the learned rules are generalizable.

Figure 3: According to Schuemie et al., most well known machine learning algorithms, or set of techniques that allow computers to learn from examples.[43]

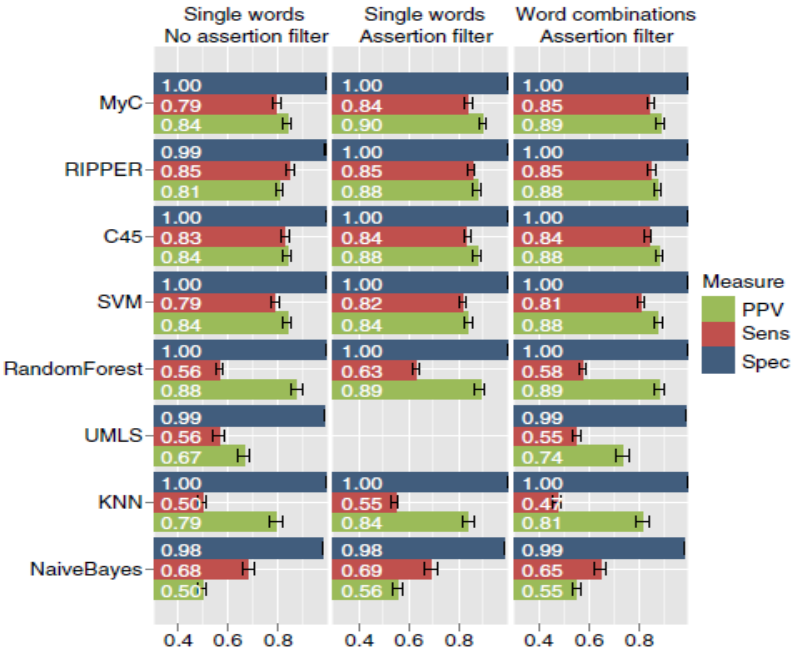


Figure 4: Performance of machine learning algorithms for Study 1, assigning ICD-9 CM codes to radiology reports[43]

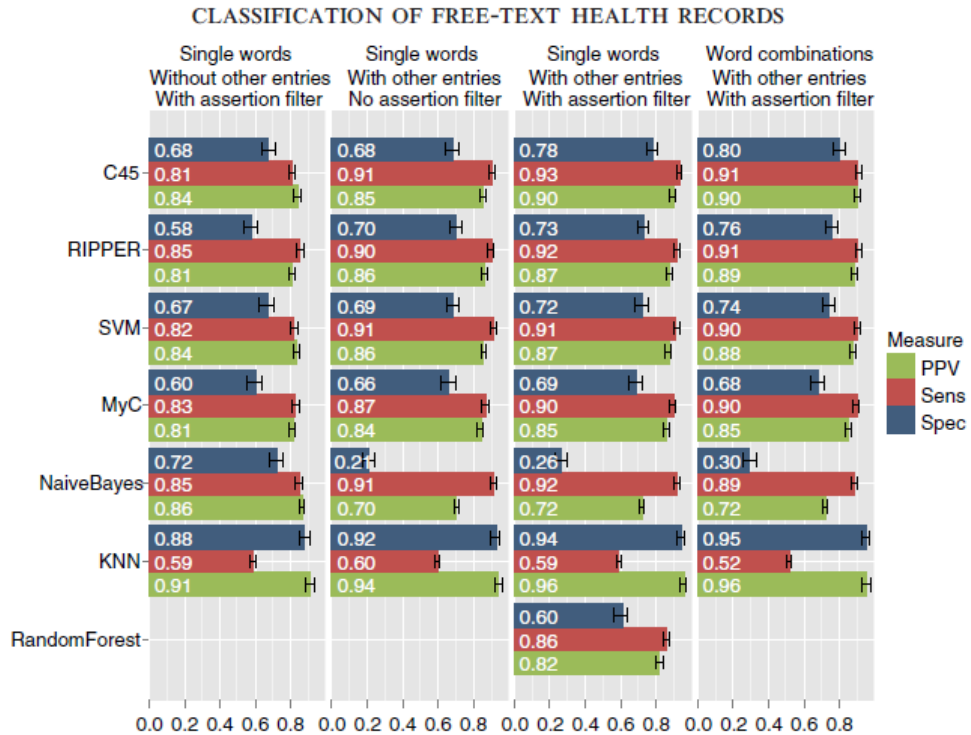


Figure 5: Performance of machine learning algorithms for Study 2, identifying cases and non cases for liver disorders[43]

2.7.3 Using Text Mining and NLP to Detect Adverse Events from Free Text

Accurately detecting adverse events remains a challenge for many institutions as observed in many of the studies cited earlier. Being able to accurately detect these events can help hospitals develop effective quality of care programs and improve coding practices. Text mining critical conditions from free text sources such as discharge summaries, radiology and pathology reports and lab values with accuracy is feasible. Lakhani, Kim and Langlotz developed text mining algorithms to detect critical results in radiology reports.[44] Conditions tested included pneumothorax, acute PE, acute cholecystitis, acute appendicitis, ectopic pregnancy, scrotal torsion, unexplained free intraperitoneal air, intracranial hemorrhage and malpositioned tubes and lines. Initial testing was performed on approximately 2.3 million radiology reports performed at The Hospital of The University of Pennsylvania from 1997-2005 and subsequent testing was done on approximately 10 million radiology reports from 1988-2011. Query algorithms were developed using SQL (Structured Query Language) and synonyms were used to expand the search. For example "ectopic pregnancy" and "extrauterine pregnancy" were considered the same. Other text mining concepts were applied to algorithms as well such as proximity searching (if the word "embolism" is a certain word distance from "pulmonary" then "pulmonary embolism") and wildcards ("embol%" searches for "embolism", "embolic", "emboli", and "embolus"). The "Impression" section was parsed out from the radiology report and most of the searches focused exclusively on the impression except for large pneumothorax. To refine algorithms, groups of 50-100 reports were selected by the algorithm then modified until precision and recall did not improve significantly. Precision refers to the percentage of radiology reports selected by

the algorithm that actually contained the critical result in question. Recall represents the percentage of reports selected by the algorithm of all possible reports in the database positive for that critical value. These two metrics are often combined to form harmonic mean, known as the *F*-measure.[45]After fine tuning the algorithms, precision, recall and the F-measure determined accuracy of these algorithms which were tested on 500-2000 new random reports for each algorithm. Results are displayed in Table 7.

Critical result	Precision	Recall
Malpositioned tubes*	98.0% (CI: 95.3–100.0%)	100% (CI: 92.2–100.0%)
Acute pulmonary embolism	99.0% (CI: 97.1–100%)	97.8% (CI: 81.4–100.0%)
Tension or new/increasing large pneumothorax	96.0% (CI: 92.2–100%)	84.2% (CI: 75.3–95.4%)
Acute cholecystitis	96.0% (CI: 92.2–100.0%)	88.6% (CI: 82.5–95.7%)
Ectopic pregnancy	98.0% (CI: 94.1–100.0%)	94.8% (CI: 86.1–100.0%)
Acute appendicitis	99.0% (CI: 97.1–100%)	89.0% (CI: 83.0–95.8%)
Unexplained free intraperitoneal air	87.0% (CI: 80.4–93.6%)	93.7% (CI: 80.3–100.0%)
New or increasing intracranial hemorrhage	94.0% (CI: 87.3–97.5%)	68.0% (CI: 59.7–79.0%)
Acute scrotal torsion	96.0% (CI: 88.3–100.0%)	100.0% (CI: 96.8–100.0%)

*(ET, NG, feeding tubes only)

Table 7: Accuracy of critical results algorithms[44]

All algorithms except for one had overall accuracies (*F*-measure) of greater than 90% with acute pulmonary embolism and malpositioned tubes yielding the highest accuracies (>98%). The lowest *F*-measure belonged to intracranial hemorrhage due to a recall of 68%, meaning the algorithm excluding a lot of cases that were actually positive for the result. This was due to the text mining concept of negation. The algorithm was built to exclude reports where the word "no" was near the critical value "hemorrhage" which was the case though it was meant to negate another term. The authors also noted that in the designing the algorithms there were occasional tradeoffs between precision and recall:

> That is, the more radiology reports recalled by the algorithms, the less precise they were by selecting some reports that did not contain critical findings. In such situations, the algorithms were preferentially tailored to have greater precision. That way, the reports selected by the

algorithm were more likely to contain critical findings and therefore to be relevant for quality improvement efforts in this area.[44]

The authors bring up an important concept here as institutions may want to be more inclusive of desired results and therefore accept false positives to achieve quality improvement goals. This study showed that it is possible to detect critical results from free text with reasonably high accuracy.

Additional studies show the effectiveness of detecting conditions from free text data. In one study at the University of Michigan, a locally developed electronic medical record search engine (EMERSE) was used to test the accuracy of automatically detected postoperative complications.[13] Cases that were reviewed as part of the American College of Surgeons National Surgical Quality Improvement Program (ACS-NSQIP) served as the gold standard as these cases have been manually reviewed. The NSQIP program provides reliable, risk-adjusted outcomes data using standardized definitions and end points.[47] Cases from 2001-2004 (5,894) were used to build the terminology while 4,898 cases from 2005-2006 were used for validation. Sensitivities of 100% to 93% for identifying postoperative myocardial infarction and pulmonary embolism were achieved respectively, with specificities of 93% and 95.9%, showing accurate identification of cases using the EMERSE tool compared with chart abstraction. The Veterans Affairs Surgical Quality Improvement Program, or VASQIP was used to test the accuracy of detecting post operative complications when using NLP and the AHRQ PSIs.[15] Again, with a random sample of inpatient admissions (1999-2006) from six Veterans Health Administration hospitals, the VASQIP data served as the gold standard. Source documents such as clinical notes, progress notes and discharge summaries were

processed through a Natural Language Processor which used an index schema based on medical concepts from the Systematized Nomenclature of Medicine- Clinical Terms (SNOMED-CT) terminology. The AHRQ PSI logic was applied to the same cohort of patients to evaluate the accuracy of both approaches (NLP vs. PSI). Results showed that the NLP processor detected adverse events with higher sensitivities. For example, VTE sensitivity was 0.59 with NLP versus 0.46 for the PSI. These studies are significant as they demonstrate increased accuracy of detecting post op complications using NLP when compared to the current and widely used PSIs.

Another study involving detection of VTE from narrative electronic health record data sampled 2,000 (from 2008-2012) narrative radiology reports from patients with suspected DVT/PE from the McGill University Health Centre (MUHC), a university health network located in Quebec, Canada.[47] DVT/PE events were manually identified within each report. A bag of words approach (to be discussed in more detail later) was used and 10 support vector machine (SVM) models were trained to detect DVT, and 10 SVM models trained to predict PE. Authors found that the best DVT model achieved an average sensitivity of 0.80, specificity of 0.98 and PPV of 0.89. The best PE model achieved a sensitivity of 0.79, specificity of 0.99 and PPV of 0.84, leading them to conclude that statistical NLP can accurately identify VTE from free text radiology reports. Similarly, one study used NLP to detect the presence, chronicity and location of pulmonary embolism from CT pulmonary angiography reports (CTPA).[48] Researchers used NILE which is an NLP library developed for information extraction from clinical narratives. As in the previous studies classifiers were trained on a set of reports and validated on a different set of reports. The Area Under the Curve (AUC) was reported

for each task. AUC is an effective and combined measure of sensitivity and specificity in assessing predictive validity of a classifier.[49]The classifiers achieved high accuracy for all four tasks with AUC being 0.998, 0.945, 0.987,0.986, for PE present, acute PE, central PE, and subsegmental PE, respectively.

The literature presented highlights some of the shortcomings with using administrative data, which can be unreliable, limited and biased. These data are the byproduct of the billing process and using them for assessment of hospital quality of care is understandable. Text mining, natural language processing and machine learning are growing fields and concepts but offer opportunities to get more out of free text and narrative data which are rich with information but expensive to mine.

2.8 Natural Language Processing

This section looks to expound further on common Natural Language Processing (NLP) tasks used when dealing with clinical text. NLP is a field of computer science concerned with the interactions between computers and human or natural languages. Some examples of NLP tasks include:

- Sentence boundary detection[50]
- Tokenization – given a character sequence, tokens are when the sequence is divided by each character[50-51]
- Part of Speech (POS)Tagging[50]
- Lemmatization and Stemming – goal is to reduce inflectional forms and derivationally related forms of a word into a common base form such as “am”, “are”, “is” converted to “be” (lemmatization) or “cars”, “car”, “car’s” converted to “car” (stemming)[51]
- Dropping stop words- very common words with little value in helping to classify documents[51]

- Named Entity Recognition- identifying specific words or phrases and categorizing them[50]

- Negation- inferring whether a named entity is absent or present, for example "No evidence of DVT" where DVT is the named entity[50,51]

- Temporal inferences and relationship extraction- for example, inferring something occurred in the past[50]

- Problem-specific segmentation- segmenting text into meaningful groups such as Chief complaint, Patient Medical History, etc.[50]

Regular expression (RegEx) in NLP concerns matching expressions within text by defining search patterns.[50]These expressions can be words, numbers or part of a sentence. With RegEx, rules are provided that can be applied to search for a match. For example, the RegEx expression "s(ei)?z" which would match the pattern "seiz" identifying related words like "seizure", "seizing", "seized", "seizes" and so on. The "ei" string in parenthesis followed by the question mark means search for expression with or without "ei", yielding "sz", for example, which can be a common abbreviation for seizure.[50] One study did just that where researchers developed a text search tool using RegEx to identify cases of children with first simple febrile seizure from the notes of 4,328 patient medical charts.[52]

These are just a few examples of the many NLP methods that are used to preprocess data. Once applied, the data can then be used to perform text mining tasks.

2.9 Document Classification

Document or text classification is a common task involving NLP and machine learning. As it relates to this dissertation, radiology reports of an unknown class will need to be placed into two classes where the conditions in question are either present or

absent. To achieve this goal, documents labeled with the correct classes will be used to train the algorithms (the learning phase), which will be our training dataset. A test dataset will then be applied to the algorithm and performance of the classifier determined. This is what is known as supervised learning. As Witten, Hall and Frank explain: "Classification learning is sometimes called *supervised*, because, in a sense, the scheme operates under supervision by being provided with the actual outcome for each of the training examples".[53] The outcome is the class and the success of classification learning can be judged by applying what is learned on an independent set of test data for which the true classifications are known but not made available to the machine.

To distinguish document classes, a classifier can use features within the document that set apart the two classes. These features can be each word in the document, an approach known as "bag of words". This "bag of words" approach takes into consideration the amount of times each words appears in the document, or word frequencies.[52] Stemming and other NLP concepts mentioned earlier can be applied to these words before documents are classified (Figure 6).

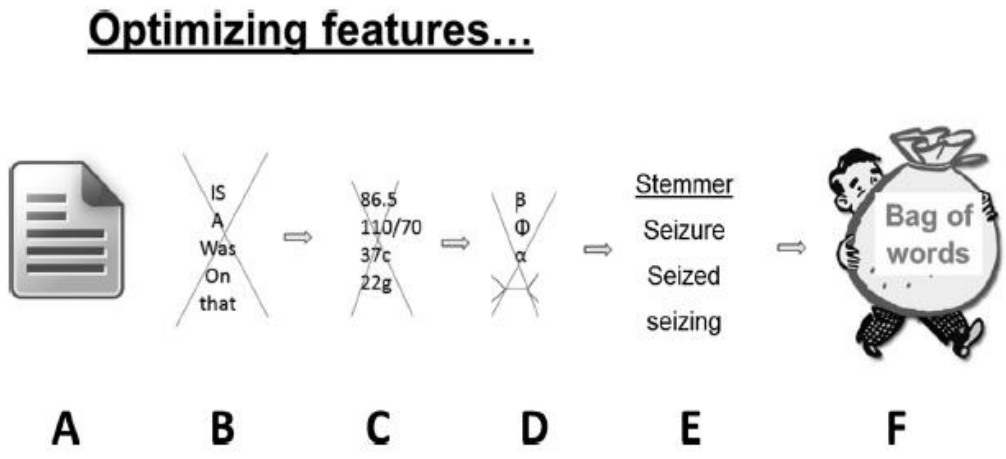


Figure 6: An example of some optimizing tools changing a document from its natural form into a bag of words. A, Initial documents. B, Eliminating the 200 most common

words in the English language (optional), also referred to as "stop words", which can significantly degrade document classification in some situations). C, Eliminating numeric characters or combined numeric/letters (optional). D, Eliminating non-α numeric features (optional). E, Running a stemmer (when counting the features [words], the word seizure and "seizing" will be considered the same adding to a count of 2 seizure words). F, Bag of words.[52]

Instead of using individual words, we can also use N-grams which are sequence of letters or words that appear together.[49,51] For example word "Thrombosis" can appear many times with the word "Vein" yielding the bigram "vein thrombosis". Once the textual data is preprocessed, using some of the methods described here, various classifiers can be applied which we will be discussed in the following sections.

2.10 Machine Learning

Machine learning is a rapidly growing field that integrates computer science and statistics where recent progress in the field has been driven by low cost computing and data availability.[54]Some examples of practical uses of machine learning can be seen in the field of Artificial Intelligence such as speech recognition, NLP and robot control. Most machine learning methods are supervised where the training data are a collection of input and output pairs (x, y) with the goal of predicting y given an unseen x.[50.54]The inputs can be vectors, documents, images or DNA sequences. As described by Jordan and Mitchell (2015), "Supervised learning systems generally form their predictions via a learned mapping f(x), which produces an output y for each input x (or a probability distribution over y given x). Many different forms of mapping f exist, including decision trees, decision forests, logistic regression, support vector machines, neural networks, kernel machines, and Bayesian classifiers.[54]

Machine learning methods can be generally classified as generative or discriminative. Generative classifiers learn what the data looks like in each category. Given that the input = X and the output/label = Y, generative models attempt to learn the model of joint probability p(X,Y) given p(Y)p(X|Y) and compute p(Y|X) (probability of Y given X) based on p(X|Y) and p(Y) (probability of y), which is Bayes Rule, p(X,Y) = p(Y)p(X|Y).[54] Discriminative classifiers model p(Y|X) directly by learning what features separate the categories.[54] Some examples of discriminative classifiers are Logistic Regression, Support Vector Machine (SVM) and k-Nearest Neighbors (kNN). In summary, a generative classifier is a likelihood function which indirectly measures training errors whereas a discriminative classifier learns exactly what features separate categories or labels and directly measures errors on training data.

In document classification, all machine learning methods rely on discriminative features to distinguish categories but differ in the way they measure errors on the training data.

2.10.1 Naïve Bayes

The Naïve Bayes classifier is described as a simple probabilistic classifier which assumes that the presence or absence of every feature contributes independently to the probability that a record belongs to a particular class.[43,56] As Witten, Hall and Frank explain (2011), Naïve Bayes is a simple and intuitive method is based on Bayes' rule of conditional probability.[56] Bayes' rule says that if you have a hypothesis H and evidence E that bears on that hypothesis, then

$$\Pr[H|E] = \frac{\Pr[E|H]\Pr[H]}{}$$

$\Pr[E]$

Table 8 shows fabricated data on patients who have or don't have food poisoning and "Yes" or "No" for the presence of three symptoms: *cramps*, *fever* and *vomiting*. Table 9 shows the fractions observed probabilities for each symptom and the outcome. For example, three patients have food poisoning = "Yes" and two of those three patients had *cramps* ("Yes") yielding a fraction of 2/3, or 0.67. Now if we wanted to predict the outcome of *food poisoning* for a new patient who exhibits the values for each symptoms (Table 10), the three symptoms, or features, and the overall likelihood that *food poisoning* = "Yes" or "No" are treated as equally important, independent pieces of evidence and multiply the fractions that correspond to each symptom and outcome. So for the outcome of yes from Table 9 gives

Likelihood of "Yes" = 0.67 x 0.67 x 1.00 x 0.60 = 0.27

Likelihood of "No" = 0.50 x 0.50 x 0.50 x 0.40 = 0.05

The resulting products of the observed probabilities show that it is more than five times more likely that Patient F has food poisoning than not. These numbers can be turned into probabilities by normalizing:

$$\text{Probability of "Yes"} = \frac{0.27}{0.27 + 0.05} = 84.2\%$$

$$\text{Probability of "No"} = \frac{0.05}{0.27 + 0.05} = 15.8\%$$

Referring back to Bayes' Rule, the hypothesis H is that Patient F has food poisoning and $\Pr[H|E]$ is 84.2%.

Patient	Cramps	Fever	Vomiting	Food Poisoning
A	Yes	Yes	Yes	Yes
B	Yes	No	Yes	Yes
C	No	Yes	Yes	Yes
D	Yes	No	No	No
E	No	Yes	Yes	No

Table 8: Patient Symptoms and Outcomes

	Cramps	Fever	Vomiting	Food Poisoning
Yes	0.67	0.67	1.00	0.60
No	0.50	0.50	0.50	0.40

Table 9: Counts and Probabilities of Symptoms and Outcomes

Patient	Cramps	Fever	Vomiting	Food Poisoning
F	No	No	Yes	?

Table 10: New Patient with Unknown Outcome

The Naïve Bayes Method is referred to as *Naïve* because it naively assumes independence, where each of the probabilities of the features is not related to one another. Though this assumption is pretty simplistic, the Naive Bayes method has been proven to work effectively on large datasets and can be easily applied to large datasets because of its simplistic nature.

2.10.2 Decision Trees

A decision tree refers to a graph or model that uses a tree-like structure to illustrate every possible outcome of a decision. The tree first starts out with a root node which consists of an attribute or feature. For each node there are branches for each value of that attribute and then another branch for another attribute and its values. This is repeated recursively for each branch. When all the instances at a node have the same classification then that part of that tree is fully developed. Witten, Hall and Frank provide an example using sample data in Table 11.[56] These fictitious data have weather condition attributes (*outlook, temperature, humidity and windy)* that determine what the outcome will be, or *play* = "yes" or "no". A popular decision tree classifier called C4.5 uses information theory to produce the purest nodes and smallest trees.[57] Information theory refers to quantifying information in bits based on entropy, providing the amount of information gained by knowing the value of each attribute, or difference between entropy of distribution before the split and entropy of distribution after the split.[58,59] Using the provided weather data, the with *outlook* having the highest amount of bits and therefore the attribute chosen for the root node. The tree is then further split by the other attributes depending on the attribute which yields the highest information gain. The resulting decision tree is shown in Figure 8. Calculating information gain, the tree starts with the outlook attribute. When the outlook is "overcast", play is always "yes". When outlook is "sunny", the next attribute selected is *humidity* and when humidity = "normal", there are two instances of play = "yes". This "divide and conquer" strategy is also performed with the *windy* attribute.

Decision trees are considered a simple classification method and easy to use. They are attractive because one can discern what went into making a decision by examining the tree structure. Furthermore, the C4.5 method which uses the information gain calculation assures the best feature/attributes to split on rather than splitting on all features which can increase efficiency in implementation.

Outlook	Temperature	Humidity	Windy	Play
Sunny	hot	high	false	no
Sunny	hot	high	true	no
Overcast	hot	high	false	yes
Rainy	mild	high	false	yes
Rainy	cool	normal	false	yes
Rainy	cool	normal	true	no
Overcast	cool	normal	true	yes
Sunny	mild	high	false	no
Sunny	cool	normal	false	yes
Rainy	mild	normal	false	yes
Sunny	mild	normal	true	yes
Overcast	mild	high	true	yes
Overcast	hot	normal	false	yes
Rainy	mild	high	true	no

Table 11: Weather data (Witten, Hall and Frank)[57]

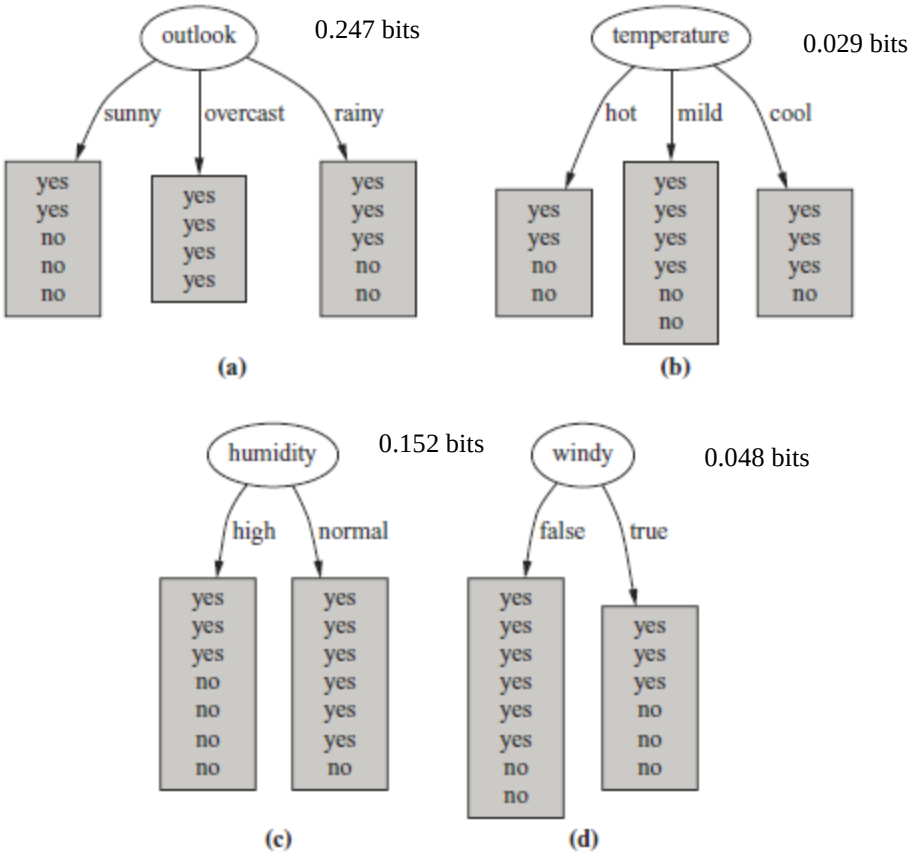


Figure 7: Using information theory to choose attributes by calculating information gain (expressed in bits).[57]

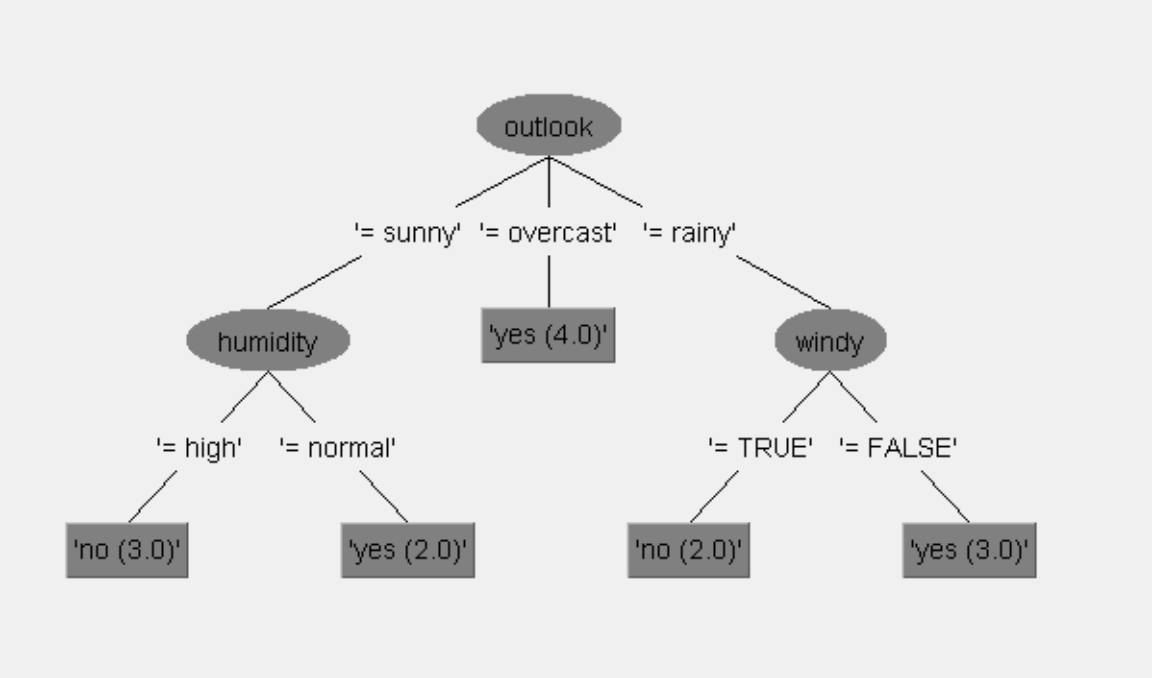


Figure 8: Final C4.5 decision tree based on the weather data.[57]

2.10.3 K- Nearest Neighbors

The k-Nearest Neighbors classifier, or kNN, classifies a new record based on the similarity between the new record and those in the training dataset.[55] This is also referred to as "instance- based learning" or "lazy learning". In document classification would mean the classifier does nothing until it gets a new document at which time it searches the training set for a document (or set of documents depending on *k*) for one that is most like the new document. Figure 9 illustrates the kNN method for classification where the shapes represent documents of different classes. When new document *X* is introduced, the nearest neighbors classifier looks for the document that is most like it, using distance. The smaller circle, k=1, determines the one neighbor closer to the new document (circle class) whereas the larger circle, k=4, looks for the nearest four neighbors to determine which document from the training set is the new document X most like (square class).

Most instance based learners use Euclidean distance to calculate the distance between two points on a plane.[57]

The kNN classifier is regarded as a discriminative classifier because it learns what features separate the categories. It is considered a very accurate but can have slow processing time since it scans the entire training data to make each prediction. It also assumes each feature or attribute is equally important so selecting important attributes or weighting more important attributes can improve classification.[56,60]

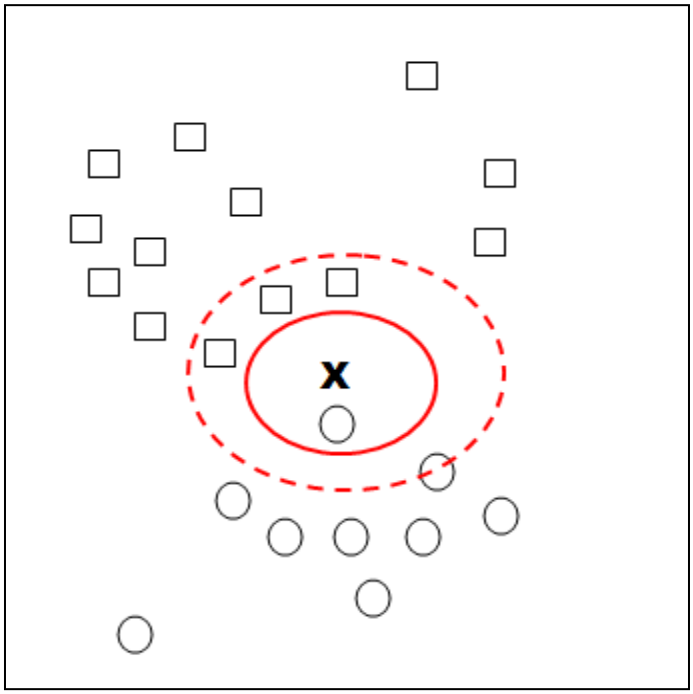


Figure 9: K- Nearest Neighbors classifier where k =1 (smaller solid red circle) and k=4 (larger, dashed red circle).

2.10.4 Support Vector Machines

A Support Vector Machine, or SVM, is another discriminative classifier that works well with two classes. The key concept here is that when given two classes in the training data, the classifier users a linear separator to divide the two classes, or a hyperplane in the case of multiple dimensions.[61,62] The margins between the linear

separator and the instances that separate each class are maximized, as shown in Figure 10. These instances, circled in red, are the support vectors and determine the class of a new instance, therefore taking the other instances in the model out of consideration.

Since SVM only relies on a few data points, the model is very resilient to overfitting. Overfitting refers to the concept of modifying a model based on a complex training dataset to yield to perform well in predicting outcomes on the training data but may not perform well on simpler test data because it is somewhat too "customized".

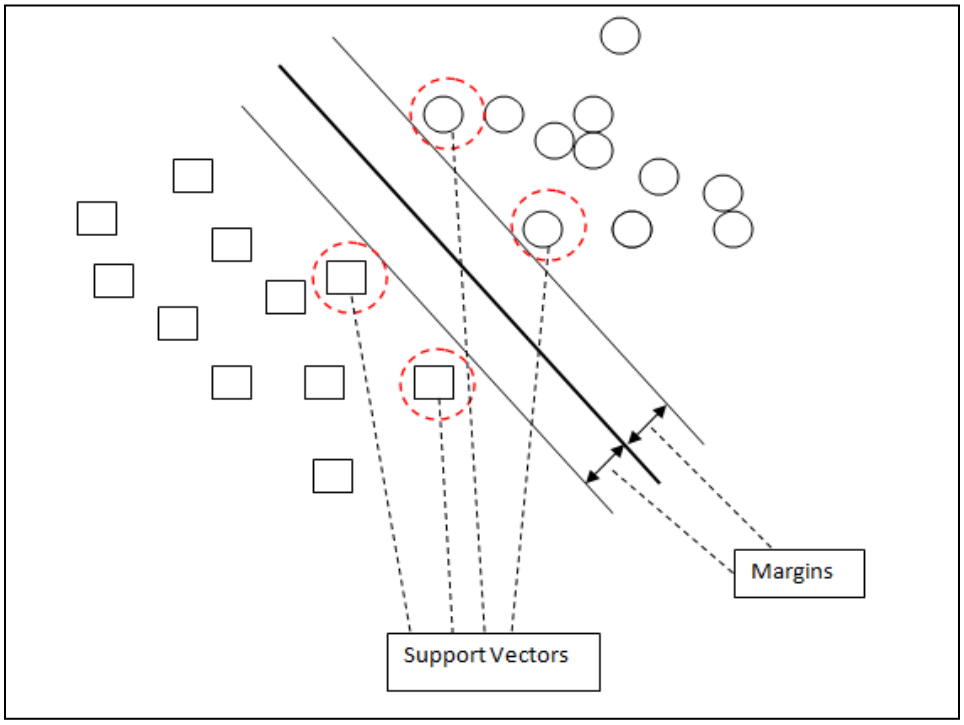


Figure 10: Support vectors (circled in red) defining the linear boundary in a SVM model.

In summary, prior work demonstrates the feasibility of using NLP and machine learning for adverse event detection, including for VTE. However, it is not clear what is the best performing solution for automated VTE detection from radiology notes.

Therefore, this study seeks to evaluate an NLP-enabled, head-to-head comparison of four leading machine learning approaches to this challenge to identify the best method for automated VTE detection.

CHAPTER III. RESEARCH METHODS

3.1 Overview

Studies related to document classification and identification of various clinical conditions were reviewed in detail in the previous chapter. A variety of rule based and machine learning classifiers were applied to different forms of clinical text. One of these studies detected critical conditions from radiology reports.[44]The impression section of the radiology report was parsed out and the algorithms were focused on this section only for all of the conditions except one. Query algorithms were developed using SQL and NLP methods such as proximity searching from key words and wildcard searches. Results were expressed in terms of precision, recall and the F-measure. Another study used various machine learning algorithms, including the four of interest for this dissertation, to automatically assign ICD-9 codes to radiology reports and identify liver disorder cases from the text of medical notes, prescriptions, referrals and lab results.[43] Similarly, one study uses cases reviewed as the gold standard to evaluate the performance of a locally developed medical record search engine on detecting complications. This dissertation looks to test the hypotheses stated earlier by using NLP methods along with four statistical machine learning to detect VTE from free text radiology reports.

3.2 NSQIP

This study uses cases reviewed and identified as a DVT or PE as part of the American College of Surgeons National Surgical Quality Improvement Program (ACS-NSQIP), of which Memorial Sloan Kettering Cancer center is a participating hospital. Since these cases have already been manually reviewed as part of the program, this served as the gold standard. NSQIP was started in the Department of Veterans Affairs (VA) in 1994 by the ACS and was expanded to private sector hospitals in 2004.[62]The program uses clinical data to assess outcomes at 30 days after the index surgery, including both inpatient and outpatient procedures. The data definitions are standardized and validated and data are collected by a trained and certified data collector. By participating, hospitals receive risk-adjusted comparisons of all ACS-NSQIP hospitals regarding morbidity, mortality and complications. Benefits of participating include identifying quality improvement targets, improving quality of patient care and reducing costs of care.

3.3 Analysis

This study was approved by the Institutional Review Board.

Database

Initial testing was done on radiology reports for surgery cases from Memorial Sloan Kettering Cancer (MSKCC) Center, a large cancer treatment facility in New York City. At this institution sampling is done on an eight day cycle assuring that cases from different surgery services have an equal chance of being selected. The report classification workflow is shown in Figure 11. The data set consisted of radiology reports performed within 30 days of surgery from 2011 to 2014, totaling 10,295 cases. The

radiology reports were transferred from the institution's radiology information system (RIS) to a Microsoft Structured Query Language (SQL) database. Since many critical results are contained within the "Impression" section of the radiology report, this section was parsed out for analyses. For DVT detection only the ultrasound reports were used and CT Angiogram reports for PE. There were 909 ultrasound reports for these 755 patients and 1,837 CT angiogram report for 1,451 patients performed within 30 days of surgery. These were divided into training (70%) and test sets (30%). The training set was used to train each of the classifiers while the test set evaluated performance of each classifier. Though cases were already identified through the NSQIP program, reports pulled were manually reviewed and flagged for absence or presence of a DVT or PE.

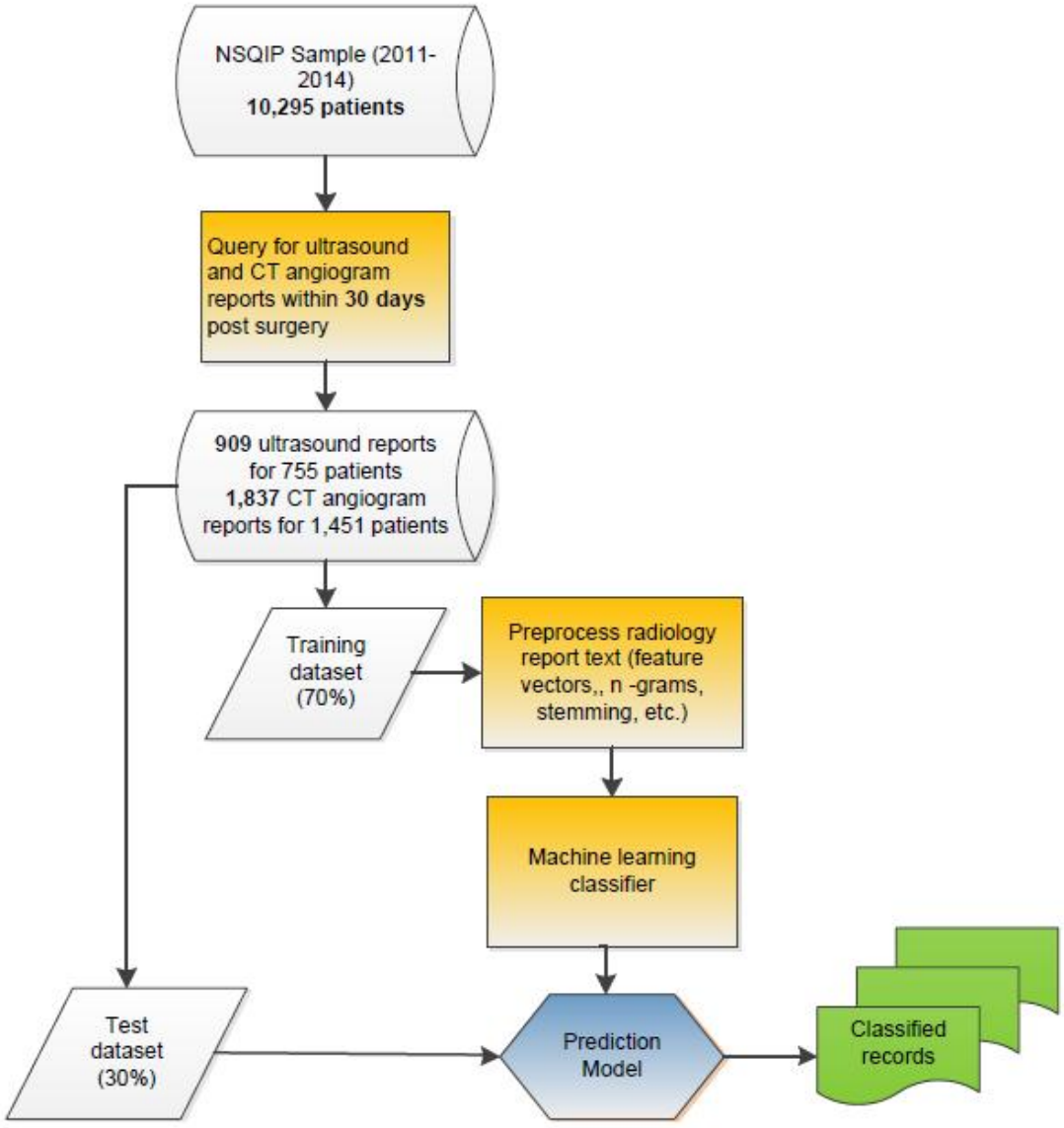


Figure 11: Report classification workflow

Data Preprocessing

Analyses were performed using the WEKA machine learning toolkit.[63] WEKA is an open source software consisting of a collection of machine learning algorithms with tools for data pre processing, classification, regression, clustering, association rules and

visualization. To prepare the dataset for input into the classifiers, the strings of text from radiology reports are converted to numeric data, or document vectors. Documents are represented by rows, sentences are split into tokens and each token is a column. For each document and word column, there is a number indicating the absence or presence of that word in the document (0,1) or frequency of that word in the document. Figure 12 shows a number NLP methods applied to the documents which include:

- Converting string to word vectors
- Converting all words to lower text
- Outputting word counts
- Stemming (using the LovinsStemmer)[64]
- Converting strings to n-grams
- Excluding stop words (no, the, of, etc)

Using the training data, each of the methods above were altered then performance was assessed on each iteration for each classifier. For example, when training the Naïve Bayes classifier, n grams were set to one, stop words were included and performance assessed. In another iteration n grams were set to two, stop words were included and performance assessed. Once it was determined which NLP settings performed the best when inputting the preprocessed text into each classifier, we applied the same methods to the test set, effectively using the model based on training data

Document	Impression Text
1	Deep venous thrombosis in bilateral calf veins, unchanged.
2	Deep vein thrombosis in the bilateral calf veins.
3	Deep venous thrombosis in left subclavian and axillary veins.
4	Deep venous thrombosis in right deep calf vein at the level of the right mid calf. Left mid calf probable hematoma.
5	No evidence of deep venous thrombosis in left upper extremity veins
6	No evidence of deep venous thrombosis in bilateral femoral popliteal system.



No.	@@class@@ Nominal	bilater Numeric	deep Numeric	evid Numeric	extrem Numeric	femor Numeric	in Numeric	left Numeric	no Numeric	of Numeric	poplite Numeric	system Numeric	thrombos Numeric	upper Numeric	vein Numeric
1	POS	1.0	1.0	0.0	0.0	0.0	1.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0	1.0
2	POS	1.0	1.0	0.0	0.0	0.0	1.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0	2.0
3	POS	0.0	1.0	0.0	0.0	0.0	1.0	1.0	0.0	0.0	0.0	0.0	1.0	0.0	1.0
4	POS	0.0	2.0	0.0	0.0	0.0	1.0	1.0	0.0	1.0	0.0	0.0	1.0	0.0	1.0
5	NEG	0.0	1.0	1.0	1.0	0.0	1.0	1.0	1.0	1.0	0.0	0.0	1.0	1.0	1.0
6	NEG	1.0	1.0	1.0	0.0	1.0	1.0	0.0	1.0	1.0	1.0	1.0	1.0	0.0	0.0

No.	@@class@@ Nominal	ven Numeric	and Numeric	at Numeric	axil Numeric	calf Numeric	hematom Numeric	level Numeric	mid Numeric	prob Numeric	right Numeric	subclav Numeric	th Numeric	**unchang** Numeric
1	POS	1.0	0.0	0.0	0.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0
2	POS	0.0	0.0	0.0	0.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0
3	POS	1.0	1.0	0.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0	0.0
4	POS	1.0	0.0	1.0	0.0	3.0	1.0	1.0	2.0	1.0	2.0	0.0	2.0	0.0
5	NEG	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
6	NEG	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0

Figure 12: Preprocessing of six documents from string to word vectors.

CHAPTER IV: RESULTS OF DATA ANALYSIS

4.1 Introduction

For identifying DVT, 909 ultrasound reports were randomly split into training and testing sets. For PE, 1,837 CT angiogram reports were randomly split into training and testing sets. Studies differ on exactly how to split the data or corpus with the most popular being holding out 70% of the data for training and 30% to test on, holding two thirds for training and one third for test, and holding out 80% for training and 20% for test.[50,65-66] The data can also be split into three groups; 1) a training set, 2) a development test set that serves as the set that "tuning" is performed on based on results from the training data, and 3) an unseen test set. A disadvantage to using the "holdout" method for training and testing is the possibility of uneven representation between the training

and test set. With cross validation, the training development, and test would switch roles repeatedly, or a fixed number of folds or partitions.[52] For example, in three fold cross validation, the data is split into three partitions, each in turn is used for testing development and training. This would be repeated three times to ensure a representative sample in each group. This study uses the 70/30 split for training and test. We used the same training and test datasets to train and test each of the four classifiers.

4.2 Naïve Bayes

DVT Analysis

When applying the Naïve Bayes classifier on the training data, the best performance, as evaluated with the F-Measure, occurs when all tokens are converted to lower case, word frequencies were counted per document, applying the stemmer and when n=3 for n-grams (trigrams). The confusion matrix for training is displayed in Table 12 with 95.9% of the documents classified correctly and yields a precision of 83.2%, recall of 95.4% and an F-measure of 88.9%.

		Classified as		
		NEG	**POS**	**TOTAL**
Actual	**NEG**	506	21	527
	POS	5	104	109
	TOTAL	**511**	**125**	**636**

Table 12: Confusion matrix using Naïve Bayes classifier for DVT training documents

Accuracy (Training)

$$\frac{506 \text{ NEG} + 104 \text{ POS}}{636} = 0.959$$

Precision (Training)

$$\frac{104 \text{ (Actual POS classified as POS)}}{125 \text{ (All classified as POS)}} = 0.832$$

Recall (Training)

$$\frac{104 \text{ (Actual POS)}}{109 \text{ (All Actual POS)}} = 0.954$$

F-Measure (Training)

$$\frac{2 \times (0.832)(0.954)}{(0.832)+(0.954)} = 0.889$$

Applying these settings to the test set yields results similar to those from the training set as computed from the confusion matrix in Table 13. With Naïve Bayes, 95.2% of the documents in the test set were classified correctly, with a precision of 80.4%, recall of 93.2% and an F-measure of 86.3%.

		Classified as		
		NEG	**POS**	**TOTAL**
Actual	**NEG**	219	10	229
	POS	3	41	44
	TOTAL	**222**	**51**	**273**

Table 13: Confusion matrix using Naïve Bayes classifier for DVT test documents

$$\textbf{Accuracy (Training)} = \frac{219\ \text{NEG} + 41\ \text{POS}}{273} = 0.952$$

$$\textbf{Precision} = \frac{41\ (\text{Actual POS classified as POS})}{51\ (\text{All classified as POS})} = 0.804$$

$$\textbf{Recall} = \frac{41\ (\text{Actual POS})}{44\ (\text{All Actual POS})} = 0.932$$

$$\textbf{F-Measure} = \frac{2 \times (0.804)(0.932)}{(0.804)+(0.932)} = 0.863$$

PE Analysis

For the evaluation of PE identification, the training set of the CT angiogram reports performed best when words were converted to lower case, stemming was applied and trigrams were used (Table 14), with an accuracy of 96.9%, precision of 61.0%, recall of 98.4% and F-measure of 75.3%.

		Classified as		
		NEG	**POS**	**TOTAL**
Actual	**NEG**	1184	39	1223
	POS	1	61	62
	TOTAL	**1185**	**100**	**1285**

Table 14: Confusion matrix using Naïve Bayes classifier for PE training documents

Using this model on the test set of documents gives an accuracy of 96.0%, and precision, recall and F-measure of 60.0%, 87.1% and 71.1% respectively, as calculated from the confusion matrix in Table 15.

		Classified as		
		NEG	**POS**	**TOTAL**
Actual	**NEG**	503	18	521
	POS	4	27	31
	TOTAL	**507**	**45**	**552**

Table 15: Confusion matrix using Naïve Bayes classifier for PE test documents

4.3 K-Nearest Neighbors

DVT Analysis

Applying the K- nearest neighbors classifier to the training set showed 100 % accuracy in classifying radiology reports as indicating DVT or not (Table 16). These results were without applying any changes to the text such as converting to lower case, and stemming with unigrams only and setting k=1 for neighbors.

		Classified as		
		NEG	**POS**	**TOTAL**
Actual	**NEG**	527	0	527
	POS	0	109	109
	TOTAL	**527**	**109**	**636**

Table 16: Performance of KNN classifier on DVT training documents

The confusion matrix in Table 17 shows the results of the KNN classifier with these settings applied to test documents. Classifier precision was 86.8%, recall was 75.0% and the F-measure was 80.5% with an overall accuracy of 94.1%.

		Classified as		
		NEG	**POS**	**TOTAL**
Actual	**NEG**	224	5	229
	POS	11	33	44
	TOTAL	**235**	**38**	**273**

Table 17: Performance of KNN classifier on DVT test documents

PE Analysis

With converting all words to lower case, stemming and using unigrams only, the KNN classifier achieved 100% accuracy as seen in Table 18. In applying KNN to the test set (Table 19), the accuracy was 96.2% with precision of 85.7%. recall of 38.7% an F-measure of 53.3%.

		Classified as		
		NEG	**POS**	**TOTAL**
Actual	**NEG**	1223	0	1223
	POS	0	62	62
	TOTAL	**1223**	**62**	**1285**

Table 18: Performance of KNN classifier on PE training documents

		Classified as		
		NEG	**POS**	**TOTAL**
Actual	**NEG**	519	2	521
	POS	19	12	31
	TOTAL	**538**	**14**	**552**

Table 19: Performance of KNN classifier on PE test documents

4.4 C4.5 Decision Tree

DVT Analysis

The C4.5 decision tree classifier, called J48 in WEKA as it is Java-based, performed best on training documents with stemming and unigrams only. This model yielded a precision of 98.1%, recall of 96.3% and the F-measure was 97.2% with an

overall accuracy of 99.1% (Table 20). Figure 13 illustrates the decision tree for this model which was applied to the test set of documents.

		Classified as		
		NEG	**POS**	**TOTAL**
Actual	**NEG**	525	2	527
	POS	4	105	109
	TOTAL	**529**	**107**	**636**

Table 20: Performance of C4.5 classifier on DVT training documents

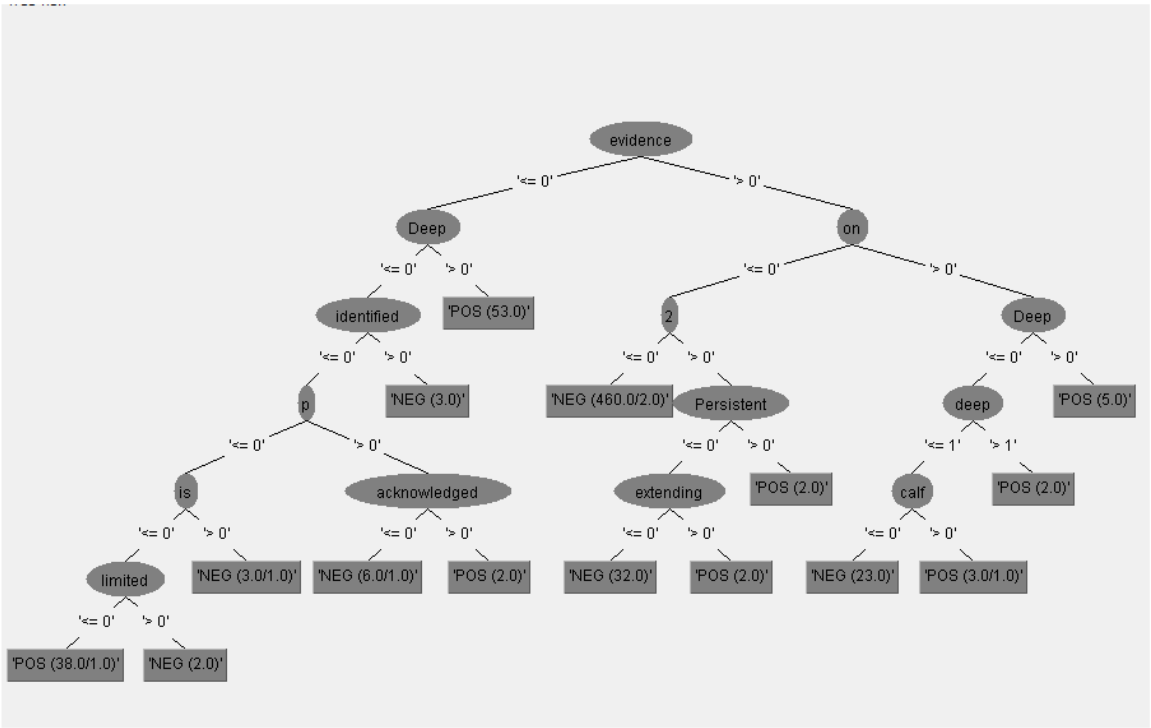


Figure 12: Decision tree produced using the WEKA J48 classifier (C4.5) on DVT training documents.

Applying this model to the test set yielded an overall accuracy of 96.7%. Precision, recall and the F-measure were 85.7%, 95.5% and 90.3% respectively (Table 21).

		Classified as		
		NEG	**POS**	**TOTAL**
Actual	**NEG**	222	7	229
	POS	2	42	44
	TOTAL	**224**	**49**	**273**

Table 21: Performance of C4.5 classifier on DVT test documents

PE Analysis

The C4.5 decision tree classifier, when applied to the training dataset of CT angiogram reports, produced best results of 98.2%, 90.3% and 94.1% for precision, recall and the f-measure respectively. As seen in Table 22, the overall accuracy was 99.5%. Figure 13 shows the tree decision tree produced from training. Table 23 show performance of the decision tree classifier on the test set of documents with an accuracy of 97.5%, precision of 75.8%, recall of 80.6% and F-measure of 78.1%.

		Classified as		
		NEG	**POS**	**TOTAL**
Actual	**NEG**	1222	1	1223
	POS	6	56	62
	TOTAL	**1228**	**57**	**1285**

Table 22: Performance of C4.5 classifier on PE training documents

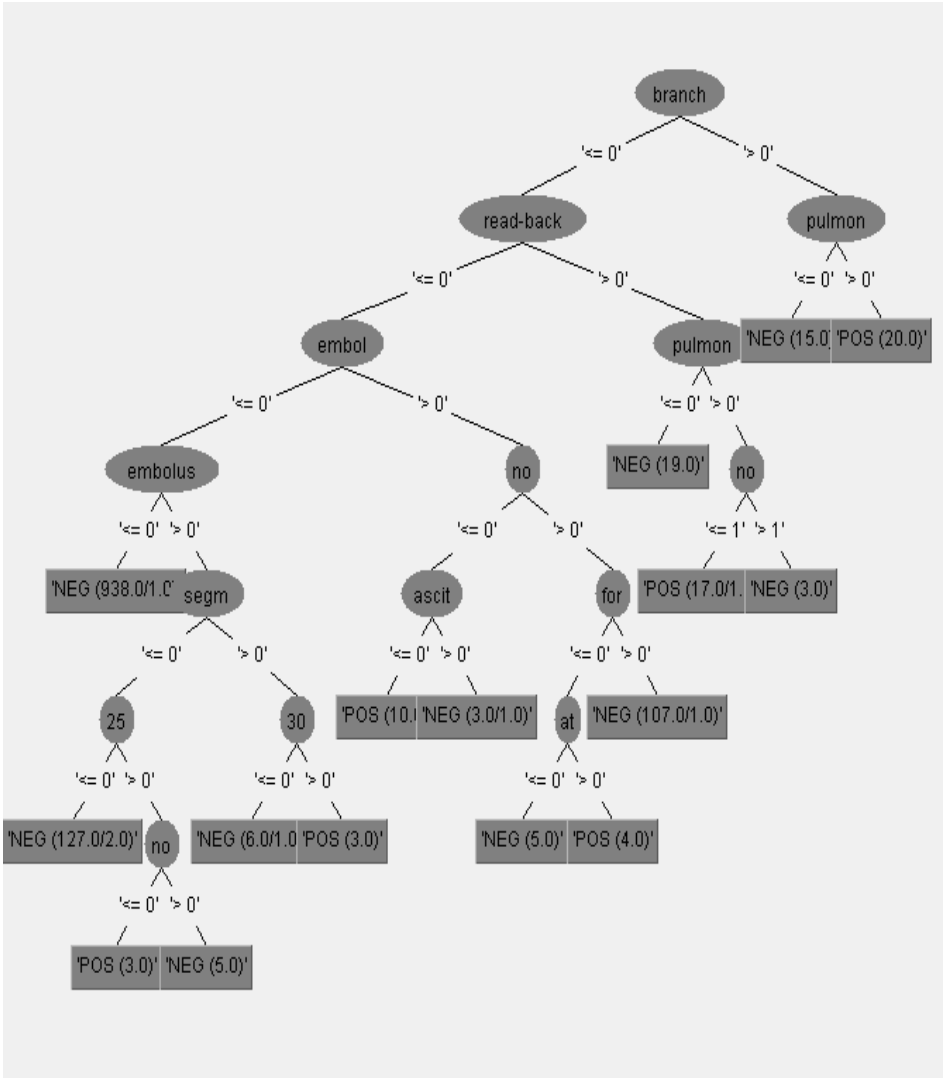


Figure 13: Decision tree produced using the WEKA J48 classifier (C4.5) on PE training documents.

		Classified as		
		NEG	**POS**	**TOTAL**
Actual	**NEG**	513	8	521
	POS	6	25	31
	TOTAL	**519**	**33**	**552**

Table 23: Performance of C4.5 classifier on PE test documents

4.5 Support Vector Machine

DVT Analysis

The best performance for classifying documents using SVM on the training set when a simple unigram bag of words approach was used. This model resulted in 100% correct classification as seen in the confusion matrix (Table 24).

		Classified as		
		NEG	**POS**	**TOTAL**
Actual	**NEG**	527	0	527
	POS	0	109	109
	TOTAL	**527**	**109**	**636**

Table 24: Performance of SVM classifier on DVT training documents

Application of this model to the test set of ultrasound reports resulted in 97.8% accuracy. Precision was 91.3%%, recall was 95.5% and the F-measure was 93.3% (Table 25).

		Classified as		
		NEG	**POS**	**TOTAL**
Actual	**NEG**	225	4	229
	POS	2	42	44
	TOTAL	**227**	**46**	**273**

Table 25: Performance of SVM classifier on DVT test documents

PE Analysis

With the CT angiogram reports, the SVM model used unigrams only to achieve 100% accuracy with the training set of documents (Table 26).

		Classified as		
		NEG	**POS**	**TOTAL**
Actual	**NEG**	1223	0	1223
	POS	0	62	62
	TOTAL	**1223**	**62**	**1285**

Table 26: Performance of SVM classifier on PE training documents

Using the SVM classifier on the test set of documents yielded an accuracy of 98.9%, precision of 93.1%, recall of 87.1% and F-measure of 90.0% (Table 27).

		Classified as		
		NEG	**POS**	**TOTAL**
Actual	**NEG**	519	2	521
	POS	4	27	31
	TOTAL	**523**	**29**	**552**

Table 27: Performance of SVM classifier on PE test documents

CHAPTER V: DISCUSSION AND STUDY LIMITATIONS

The aim of this research was to assess the performance of four different classifiers on detecting critical results from radiology reports. These classifiers were different in nature as there was one probabilistic classifier in Naïve Bayes, discriminative classifiers such as the Support Vector Machine and K-Nearest Neighbors, and decision trees based on the C4.5 induction, or top down classifier. Data consisted of the impression sections of 909 ultrasound and 1,837 CT angiogram radiology reports from cases that were sampled as part of the NSQIP program from 2011 through 2014. Ultrasounds are usually performed when there is suspicion of DVT, and CT angiograms where there may be a PE present. The NSQIP sample was used as these cases have already been flagged for critical conditions, DVT and PE included, by trained reviewers. Radiology reports within

30 days of surgery for these cases were queried and reviewed again for indications of DVT and PE. As done in most document classification studies, a part of the sample was held out for training and one part for testing. In this study 70% of the sample was designated for training and 30% for testing. When training each classifier, NLP methods such as lower case conversion, excluding stop words, stemming and n-grams were used, then this preprocessed data was input into each of the four classifiers. Settings of each classifier that performed best on training data were then applied to the test set.

The F-measure was used to evaluate overall performance of each classifier as it is the harmonic mean of both the recall and precision. As can be seen on Figure 14, the SVM classifier predicted class (positive or negative for DVT/PE) for the test set of radiology reports with the highest accuracy, with an F-measure of 93.3% for DVT and 90.0% for PE. In order of decreasing performance for detecting both DVT and PE were the decision tree classier, Naïve Bayes and k-Nearest Neighbors. There are a few reasons why the SVM classifier performed the best. SVM is an instance based classifier therefore only uses document features that are closest to the linear separator or hyperplane dividing the two classes. In this study a linear separator was used and proved sufficient in separating classes. Since SVM is an instance based learner and does not use all features in the dataset, it is resilient to overfitting. It also uses a so called kernel function for calculating the best space between two classes.[61] Although Naïve Bayes did not perform as well as SVM, recall is very similar for DVT detection, with 95.5% for SVM and 93.2% for Naïve Bayes. Similarly, recall for PE detection is 87.1% for both the SVM and Naïve Bayes classifiers. As a result of low precision, the F-measure for Naïve Bayes is decreased. In this study, both precision and recall are weighted the same,

where β = 1. Though the Naïve Bayes classifier predicted a high number of false positives (low precision), it can still reduce the number of cases for review and is proven to be accurate in detecting positive reports. In this case β can be adjusted to give precision a higher weight in the resulting F-measure. One important preprocessing aspect in using the Naïve Bayes classifier is that using trigrams increased performance. This is not surprising since Naïve Bayes calculates probability of an outcome without considering how features relate to one another. For example when an ultrasound is negative for a DVT, many of the reports have the same template text, "No evidence of DVT in…". Since negation was not applied in preprocessing, using trigrams would capture "no" and "DVT" together rather than treat them as separate features.

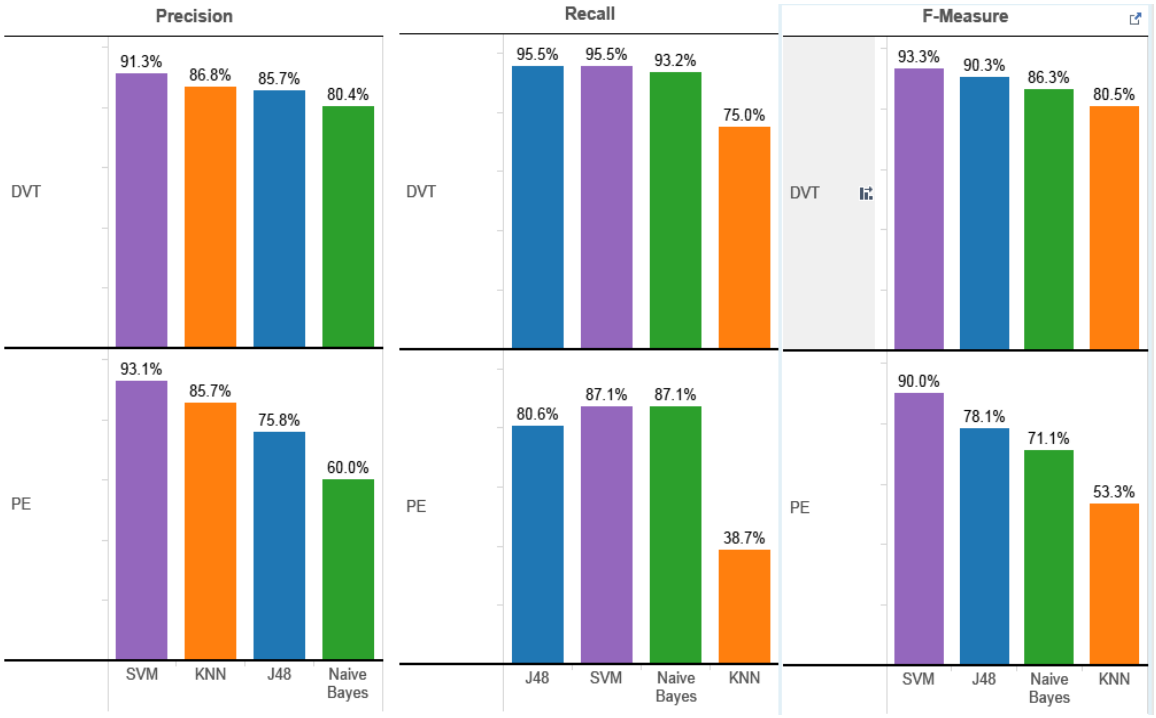


Figure 14: Classifier performance in detecting VTE from ultrasound and CT angiogram radiology reports

The kNN and C4.5 (J48 in Weka) classifiers varied in accuracy of detecting DVT and PE. The kNN classifier performed well for DVT detection. This is possible since the text in these ultrasound reports are similar in the phrasing of a non-finding, resulting in a uniform dataset. For PE, the CT angiogram reports are not as uniform as many other conditions can be detected and documented from this diagnostic test, resulting in poor performance of this instance based classifier. The decision tree classifier performed well in detecting DVT and followed only SVM in performance of detecting PE. The C4.5 classifier is an induction based classifier and only uses the "best" features to split the tree on, based on the information theory discussed earlier. Again, as in kNN, since the CT angiogram reports contain more heterogeneous text compared to the ultrasounds, overall performance was not as high as in the SVM classifier.

A limitation of this study is that the radiology reports used for testing and training are from one cancer center and these methods should be applied to a more general patient base as models may be skewed by the high volume of cancer/surgery patients. In addition, classifiers were trained on a single source of data, the impression sections of ultrasounds and CT angiogram radiology reports. Accuracy may differ if other clinical narratives are included such as other radiology report types, physician notes or discharge summaries. There was some possible overfitting as seen in the PE analysis when using kNN, where the training set was classified with 100% accuracy but the test set yielded an F-measure of 53.3%. Using cross validation or a larger sample may improve performance.

In the event that a report indicated both the absence and presence of a condition such as "1.Deep vein thrombosis in the left superior calf. 2. No evidence of deep venous

thrombosis in bilateral femoral popliteal system.", the classifiers categorized this as a negative report. Accuracy could potentially be increased by including logic that accounts for a positive and negative indication in the same report.

Although we used the NSQIP sample of cases as the gold standard, classifiers were able to detect positive reports for cases that were not identified in the NSQIP database. To demonstrate, there were a total of 44 positive DVT reports of which the SVM classifier identified all but two. In addition, the SVM classifier also found six reports for patients that were not flagged as having a DVT in the NSQIP database. The SVM classifier also found four additional CT angiogram reports positive for PE not flagged in the NSQIP database.

In this study performance of four different machine learning classifiers was assessed. The results show that VTEs can be identified from narrative radiology reports with high accuracy. Accurate identification of VTEs could assist hospitals in quality improvement, adverse event surveillance, evaluation of interventions and improving documentation and coding.

CHAPTER VI: FUTURE RESEARCH

The research performed in this study contributes to the growing literature of NLP and machine learning in healthcare. There are research opportunities that can follow up on this study. For example, a larger dataset and different sampling methods such as cross validation could improve classifier performance and prevent overfitting.

In this study there were limited NLP methods used and applying negation, named entity recognition, keyword searches, semantic analysis, etc. to narrative text can

potentially increase classifier performance. It would be valuable to assess performance using other existing NLP tools such as the Apache clinical Text Analysis and Knowledge Extraction System (cTAKES) and the Medial Language Extraction and Encoding System (MedLEE).[67,68] Both of these tools are NLP systems for extracting information from electronic medical record free text.

Natural language processing and machine learning are being used to extract information from unstructured text in all disciplines. Healthcare is extremely rich and reliant on free text data and it is essential that researchers, clinicians, data scientists and all involved continue improve and contribute to the field of text mining and computational linguistics to maximize potential applications and benefits, with the ultimate goal of providing optimal patient care.

REFERENCES

1. Centers for Medicare & Medicaid Services. *Medicare*. Accessed on November 3, 2012 from http://www.cms.gov/Medicare/Medicare.html

2. The Joint Commission. *Accountability Measures*. Accessed on November 20, 2012 from http://www.jointcommission.org/accountability_measures.aspx

3. Medicare.gov. *Hospital Compare*. Accessed November 21, 2012 from http://www.medicare.gov/hospitalcompare/search.html

4. US News and World Report. HEALTH: Hospitals. Accessed November 21, 2012 from http://health.usnews.com/best-hospitals/rankings

5. The Joint Commission. *Accountability Measures.* Accessed November 21, 2012 from http://www.jointcommission.org/accountability_measures.aspx

6. Agency for Healthcare Research and Quality. *Patient Safety Indicators Overview*. Accessed December 9, 2012 from http://www.qualityindicators.ahrq.gov/Modules/psi_overview.aspx

7. Center for Medicare & Medicaid. *Hospital-Acquired Conditions*. Accessed November 29 from http://www.cms.gov/Medicare/Medicare-Fee-for-Service-Payment/HospitalAcqCond/Hospital-Acquired_Conditions.html

8. Accessed February 27, 2013 from http://www.nhlbi.nih.gov/health/health-topics/topics/dvt/

9. Beckman MG, Hooper WC, Critchley SE, Ortel TL. Venous thromboembolism: a public health concern. Am J Prev Med 2010;38(4 Suppl):S495-501.

10. Tuttle, D., et al. (2002). "Using administrative data to improve compliance with mandatory state event reporting." *Jt Comm J Qual Improv* **28**(6): 349-358.

11. Henderson KE, Recktenwald Aj, Reichley RM, Bailey TC. Clinical validation of the AHRQ postoperative venous thromboembolism patient safety indicator. Joint Commission journal on quality and patient safety 2009;**35**(7):370-6.

12. White RH, Sadeghi B, Tancredi DJ, Zrelak P. How Valid is the ICD-9-CM Based AHRQ Patient Safety Indicator for Postoperative Venous Thromboembolism? Medical care 2009;**47**(12):1237-43.

13. Hanauer DA, Englesbe MJ, Cowan JA, Jr., Campbell DA. Informatics and the American College of Surgeons National Surgical Quality Improvement Program: automated processes could replace manual record review. J Am Coll Surg. United States, 2009:37 41.

14. Melton GB, Hripcsak G. Automated detection of adverse events using natural language processing of discharge summaries. J Am Med Inform Assoc 2005;**12**(4).

15. Murff HJ, FitzHenry F, Matheny ME, et al. Automated identification of postoperative complications within an electronic medical record using natural language processing. JAMA 2011;**306**(8).

16. Ruud KL, Johnson MG, Liesinger JT, Grafft CA, Naessens JM. Automated detection of follow-up appointments using text mining of discharge records. Int J Qual Health Care 2010;**22**(3):229-35.

17. Pakhomov S, Bjornsen S, Hanson P, Smith S. Quality performance measurement using the text of electronic medical records. Med Decis Making 2008;**28**(4):462-70.

18. Kohn, L.T., J. Corrigan, and M.S. Donaldson, To err is human : building a safer health system. 1999, Washington, D.C.: National Academy Press.

19. Agency for Health Care Research and Quality. *Making Health Care Safer II: An Updated Critical Analysis of the Evidence for Patient Safety Practices*. Accessed March 27, 2013 from http://www.ahrq.gov/research/findings/evidence-based-reports/ptsafetyuptp.html

20. State of New York Department of Health. *NYPORTS*. Accessed March 3 from http://www.health.ny.gov/facilities/hospital/nyports/annual_report/2002-2004/docs/2002-2004_nyports_annual_report.pdf

21. Rosen AK, Itani KM, Cevasco M, et al. Validating the patient safety indicators in the Veterans Health Administration: do they accurately identify true safety events? Med Care 2012;**50**(1):74-85.
22. Campbell SE, Campbell MK, Grimshaw JM, Walker AE. A systematic review of discharge coding accuracy. J Public Health Med 2001;**23**(3):205-11

23. American Health Information Management Association. *Medical Coding*. Retrieved on December 28, 2012 from http://www.ahima.org/coding/

24. Gartee R. *Electronic health records : understanding and using computerized medical records*. 2nd ed. Boston: Pearson, 2011.

25. Ballentine NH. Coding and documentation: Medicare severity diagnosis-related groups and present-on-admission documentation. J Hosp Med 2009;**4**(2):124-30.

26. Center for Medicare & Medicaid Services. *Glossary*. Accessed January 20, 2013 from http://www.cms.gov/apps/glossary/default.asp?Letter=ALL&Language=English

27. Iezzoni LI. Assessing quality using administrative data. Ann Intern Med 1997;**127**(8 Pt 2):666-74

28. New York State Department of Health. *The New York Patient Occurrence and Tracking System (NYPORTS).* Accessed January 20, 2013 from https://www.health.ny.gov/facilities/hospital/nyports/

29. U.S. Department of Health and Human Services: Office of Inspector General. Accessed January 20, 2013 from http://oig.hhs.gov/oei/reports/oei-06-09-00091.asp

30. Kaafarani HM, Borzecki AM, Itani KM, et al. Validity of selected Patient Safety Indicators: opportunities and concerns. J Am Coll Surg 2011;**212**(6):924-34.

31. Bahl, V., et al. (2008). "Do the AHRQ patient safety indicators flag conditions that are present at the time of hospital admission?" *Medical Care* **46**(5): 516-522.

32. Geerts WH, Bergqvist D, Pineo GF, et al. Prevention of venous thromboembolism: American College of Chest Physicians Evidence-Based Clinical Practice Guidelines (8th Edition). Chest 2008;**133**(6 Suppl):381S-453S.

33. Goldhaber SZ. Venous thromboembolism: epidemiology and magnitude of the problem. Best Pract Res Clin Haematol 2012;**25**(3):235-42 doi: 10.1016/j.beha.2012.06.007[published Online First: Epub Date]|.

34. Raskob GE, Silverstein R, Bratzler DW, Heit JA, White RH. Surveillance for deep vein thrombosis and pulmonary embolism: recommendations from a national workshop. Am J Prev Med 2010;**38**(4 Suppl):S502-9 doi: 10.1016/j.amepre.2010.01.010[published Online First: Epub Date]|.

35. The Centers for Disease Control and Prevention. *Morbidity and Mortality Weekly Report (MMWR): Venous Thromboembolism in Adult Hospitalizations-United States, 2007–2009.* Retrieved on August 5, 2013 from http://www.cdc.gov/mmwr/preview/mmwrhtml/mm6122a1.htm?s_cid=mm6122a1_w

36. Zhan C, Miller MR. Excess length of stay, charges, and mortality attributable to medical injuries during hospitalization. JAMA 2003;**290**(14):1868-74 doi: 10.1001/jama.290.14.1868[published Online First: Epub Date]|.

37. Mahan CE, Holdsworth MT, Welch SM, Borrego M, Spyropoulos AC. Deep-vein thrombosis: a United States cost model for a preventable and costly adverse event. Thromb Haemost 2011;**106**(3):405-15 doi: 10.1160/TH11-02-0132[published Online First: Epub Date]|.The Centers for Disease Control and Prevention. *Deep Vein Thrombosis and Pulmonary Embolism.* Retrieved on August 6, 2013 from http://www.cdc.gov/ncbddd/blooddisorders/documents/bbv_pnv_c0_1159_thro

mbosis_r1mtr.pdf

38. University of Berkeley, California. *What is Text Mining?* Retrieved on August 20 2013 from http://people.ischool.berkeley.edu/~hearst/text-mining.html

39. Kao A, Poteet SR. *Natural language processing and text mining*. London: Springer, 2007.

40. Cleophas TJM, Zwinderman AH, SpringerLink (Online service). Machine Learning in Medicine:1 online resource.

41. Hripcsak G, Vawdrey DK, Fred MR, Bostwick SB. Use of electronic clinical documentation: time spent and team interactions. J Am Med Inform Assoc 2011;**18**(2):112-7.

42. Ammenwerth E, Spötl HP. The time needed for clinical documentation versus direct patient care. A work-sampling analysis of physicians' activities. Methods Inf Med 2009;**48**(1):84-91

43. Schuemie, M. J., et al. (2012). "Automating classification of free-text electronic health records for epidemiological studies." *Pharmacoepidemiol Drug Saf* **21**(6): 651-658.

44. Lakhani P, Kim W, Langlotz CP. Automated Detection of Critical Results in Radiology Reports. Journal of digital imaging 2012;**25**(1):30

45. Hripcsak G, Rothschild AS. Agreement, the f-measure, and reliability in information retrieval. J Am Med Inform Assoc 2005;**12**(3):296-8.

46. Khuri SF, Daley J, Henderson W, et al. The Department of Veterans Affairs' NSQIP: the first national, validated, outcome based, risk-adjusted, and peer-controlled program for the measurement and enhancement of the quality of surgical care. National VA Surgical Quality Improvement Program. Ann Surg 1998;228:491–507.

47. Rochefort CM, Verma AD, Eguale T, Lee TC, Buckeridge DL. A novel method of adverse event detection can accurately identify venous thromboembolisms (VTEs) from narrative electronic health record data. J Am Med Inform Assoc 2015;**22**(1):155-65 doi: 10.1136/amiajnl-2014-002768[published Online First: Epub Date]|.

48. Yu S, Kumamaru KK, George E, et al. Classification of CT pulmonary angiography reports by presence, chronicity, and location of pulmonary embolism with natural language processing. J Biomed Inform 2014;**52**:386-93.

49. Kumar R, Indrayan A. Receiver operating characteristic (ROC) curve for medical researchers. Indian Pediatr 2011;**48**(4):277-87

50. Nadkarni PM, Ohno-Machado L, Chapman WW. Natural language processing: an introduction. J Am Med Inform Assoc 2011;**18**(5):544-51.

51. Introduction to Information Retrieval. Accessed on October 25, 2013 from http://www-nlp.stanford.edu/IR-book/

52. Kimia AA, Savova G, Landschaft A, Harper MB. An Introduction to Natural Language Processing: How You Can Get More From Those Electronic Notes You Are Generating. Pediatr Emerg Care 2015;**31**(7):536-41 .

53. Witten IH, Frank E, Hall MA, Books24x7 Inc. Data mining practical machine learning tools and techniques, third edition. Morgan Kaufmann series in data management systems. 3rd ed. Burlington, Mass.: Morgan Kaufmann Publishers, 2011: 39-60.

54. Jordan MI, Mitchell TM. Machine learning: Trends, perspectives, and prospects. Science 2015;**349**(6245):255-60

55. Ng, Andrew Y., and Michael I. Jordan. 2001. On discriminative vs. generative classifiers: A comparison of logistic regression and naive Bayes. In Proc. NIPS, pp. 841–848. URL: www-2.cs.cmu.edu/Groups/NIPS/NIPS2001/papers/psgz/AA28.ps.gz

56. Witten IH, Frank E, Hall MA, Books24x7 Inc. Data mining- practical machine learning tools and techniques, third edition. Morgan Kaufmann series in data management systems. 3rd ed. Burlington, Mass.: Morgan Kaufmann Publishers, 2011: 85-145.

57. Witten IH, Frank E, Hall MA, Books24x7 Inc. Data mining- practical machine learning tools and techniques, third edition. Morgan Kaufmann series in data management systems. 3rd ed. Burlington, Mass.: Morgan Kaufmann Publishers, 2011: 3-36.

58. Quinlan, J.R. (1986). Induction of Decision Trees. Machine Learning 1:1, 81-106.

59. Shannon CE. The mathematical theory of communication. 1963. MD Comput 1997;**14**(4):306-17

60. Han ES, Karypis G, Kumar V. Text categorization using weight adjusted k-nearest neighbor classification. Adv Knowl Discov Data Min, 2001. 53–65.

61. Witten IH, Frank E, Hall MA, Books24x7 Inc. Data mining- practical machine learning tools and techniques, third edition. Morgan Kaufmann series in data management systems. 3rd ed. Burlington, Mass.: Morgan Kaufmann Publishers, 2011: 191-303.

62. Drucker H, Wu D, Vapnik VN. Support vector machines for spam categorization. IEEE Trans Neural Netw 1999;**10**(5):1048-54

63. Hall M, Frank E, Holmes G, Pfahringer B, Reutemann P, Witten IH. The WEKA data mining software: An update. ACM SIGKDD Explorations Newsletter. 2009; 11(1): 10–18.

64. Lovins, J. B. Development of a Stemming Algorithm, Mechanical Translation and Computational Linguistics, 1968; **11**: 22-31.

65. Chen H, Fuller S, Friedman C. Medical informatics: knowledge management and data mining in biomedicine. 1st edn. New York: Springer-Verlag, 2005: 6-11.

66. Evaluating Classifiers. Accessed on October 30, 2015 from http://cseweb.ucsd.edu/~elkan/250Bwinter2012/classifiereval.pdf

67. Apache cTAKES. Accessed on March 10, 2016 from http://ctakes.apache.org/index.html

68. MedLingMap. Accessed on March 10, 2016 from http://www.medlingmap.org/taxonomy/term/80