Automated Extraction of VTE Events From Narrative Radiology Reports in Electronic Health Records

A Validation Study

Zhe Tian, BSc,[*][†] Simon Sun, MD,[‡] Tewodros Eguale, MD, PhD,[†][§] and Christian M. Rochefort, RN, PhD*[*][†][‖]

Background: Surveillance of venous thromboembolisms (VTEs) is necessary for improving patient safety in acute care hospitals, but current detection methods are inaccurate and inefficient. With the growing availability of clinical narratives in an electronic format, automated surveillance using natural language processing (NLP) techniques may represent a better method.

Objective: We assessed the accuracy of using symbolic NLP for identifying the 2 clinical manifestations of VTE, deep vein thrombosis (DVT) and pulmonary embolism (PE), from narrative radiology reports.

Methods: A random sample of 4000 narrative reports was selected among imaging studies that could diagnose DVT or PE, and that were performed between 2008 and 2012 in a university health network of 5 adult-care hospitals in Montreal (Canada). The reports were coded by clinical experts to identify positive and negative cases of DVT and PE, which served as the reference standard. Using data from the largest hospital (n = 2788), 2 symbolic NLP classifiers were trained; one for DVT, the other for PE. The accuracy of these classifiers was tested on data from the other 4 hospitals (n = 1212).

Results: On manual review, 663 DVT-positive and 272 PE-positive reports were identified. In the testing dataset, the DVT classifier achieved 94% sensitivity (95% CI, 88%-97%), 96% specificity (95% CI, 94%-97%), and 73% positive predictive value (95% CI, 65%-80%), whereas the PE classifier achieved 94% sensitivity (95% CI, 89%-97%), 96% specificity (95% CI, 95%-97%), and 80% positive predictive value (95% CI, 73%-85%).

Conclusions: Symbolic NLP can accurately identify VTEs from narrative radiology reports. This method could facilitate VTE surveillance and the evaluation of preventive measures.

Key Words: natural language processing, NLP, VTE, pulmonary embolism, deep vein thrombosis, electronic health record, EHR, narrative radiology report

(*Med Care* 2017;55: e73–e80)

Venous thromboembolism (VTE), which encompasses both deep vein thrombosis (DVT) and pulmonary embolism (PE),[1,2] is the leading cause of hospital death in the United States,[3] and is associated with significant morbidity and cost.[4–6] Although effective preventive measures are available, their underuse and inappropriate use have led patient safety experts to identify VTE prevention as the most important priority for improving patient safety in hospitals.[7] However, at the present time, there is no effective surveillance system for monitoring the occurrence of VTE and for evaluating the success of preventive measures.

Indeed, VTE rates are currently monitored by screening administrative data for International Classification of Disease, Ninth Revision, Clinical Modification (ICD-9-CM) discharge diagnostic codes.[8,9] However, recent studies have questioned the validity and the accuracy of using these codes for monitoring VTEs.[10,11] In addition, in many jurisdictions, discharge diagnostic codes are not dated with precision. As a consequence, it is often difficult to determine whether a given code (eg, PE) represents an event that occurred before the patient was hospitalized (ie, a comorbid condition) or during the actual hospitalization (ie, an acute event).[12] Moreover, discharge diagnostic codes are typically available only several months after discharge, which hinders both timely surveillance of VTE and prompt interventions.

According to evidence-based practice guidelines,[13,14] VTEs can be objectively identified through imaging studies from diagnostic radiology (eg, venous ultrasound, CT scan of the chest). The results of these imaging studies are summarized in an unstructured narrative report written by the radiologist and stored in the electronic health record (EHR).

From the *Department of Epidemiology, Biostatics and Occupational Health, Faculty of Medicine; †McGill Clinical and Health Informatics Research Group, McGill University; ‡Department of Radiology, McGill University Health Centre, Montreal, QC, Canada; §Brigham & Women's Hospital, Boston, MA; and ‖Ingram School of Nursing, Faculty of Medicine, McGill University, Montreal, QC, Canada.

Supported by the Canadian Institutes of Health Research (CIHR). Z.T. received a Master's training award from CIHR and C.M.R. holds a career award from the Fonds de recherche du Québec en santé (FRQ-S). The research ethics committee of the MUHC approved this study, and the Director of Professional Services authorized access to the EHR data.

The authors declare no conflict of interest.

Reprints: Zhe Tian, BSc, Department of Epidemiology, Biostatistics and Occupational Health, 1020 Pine Avenue, West Montreal, QC, Canada H3A 1A2. E-mail: zhe.tian@mail.mcgill.ca.

ISSN: 0025-7079/17/5510-e073

Although clinical experts can accurately identify VTE by reading these reports, it is a time-consuming and costly process. Automating this process through natural language processing (NLP) techniques could greatly reduce the time and cost required for monitoring VTE rates.

NLP refers to a set of automated techniques that convert free-text data into a computer-processable format. Among the various NLP techniques, symbolic NLP uses the structure and the semantic meaning of the written language to construct logical rules for classifying narrative documents.[15,16] Although recent studies have provided evidence that symbolic NLP can accurately identify some medical conditions from narrative EHR documents,[17–19] comparatively little attention has been given to the detection of VTEs. The objective of this study was to determine the accuracy of using symbolic NLP classifiers to identify DVT and PE from narrative radiology reports.

METHODS

Setting and Study Population

The study was conducted at the McGill University Health Centre (MUHC), a university health network located in the Canadian province of Quebec. The MUHC is composed of 5 adult-care hospitals and has more than 800 beds. It serves a population of 1.7 million people (22% of the provincial population), with an annual volume of approximately 735,000 ambulatory visits, 33,300 surgeries, and 40,000 hospitalizations.[20] The research ethics committee of the MUHC approved this study, and the Director of Professional Services authorized access to EHR data.

Data Source

Data for this study were extracted from 3 electronic databases at the MUHC and were linked by unit, patient, and hospital admission date. The *Discharge Abstract Database* provided patient age and sex, dates of hospital admission and discharge, and diagnostic codes. The *Radiology Report Database* provided data on all radiologic examinations that were performed over the study period, including dates when these examinations were performed, a text description of the radiologic findings, and the radiologist's interpretation. At the time of this study, no other clinical narratives were available in an electronic format at the MUHC.

Study Design

To assess the accuracy of using symbolic NLP for detecting VTE from narrative radiology reports, a validation study was conducted. First, a random sample of 4000 narrative reports was selected among all reports of radiologic examinations that were performed at the MUHC, between 2008 and 2012, in patients undergoing an imaging study that could diagnose DVT or PE (eg, venous Doppler, CT scan of the chest). The 4000 narrative radiology reports were then manually coded by clinical experts, which served as the reference standard. These clinical experts included a general practitioner (T.E.), a radiology fellow (S.S.), and a nurse epidemiologist (C.M.R.). Then, using data from the largest MUHC hospital site as a training set (n = 2788), 2 symbolic NLP classifiers were iteratively developed and tested; one for detecting DVT, and another for detecting PE. The accuracy of the best performing NLP classifiers developed on the training set was then measured on the testing set, which included data from the other 4 MUHC hospital sites (n = 1212).

Reference Standard Development

The 4000 reports were initially coded by a clinical expert (C.M.R.) and assigned 2 codes: (1) positive or negative for DVT of the lower or upper extremities and (2) positive or negative for PE. Positive radiology reports for a DVT were those where a thrombus was identified in the proximal deep veins of the lower extremities (eg, external iliac, common femoral, deep femoral, or popliteal veins), in the deep distal veins of the lower extremities (eg, peroneal and posterior tibial veins), or in the deep veins of the upper extremities (eg, brachial, radial, ulnar, axillary, subclavian). Negative cases included those where no thrombus was identified or where a thrombus was identified in a superficial vein of the lower extremity (eg, saphenous), in a superficial vein of the upper extremity (eg, cephalic), or in a perforating vein of the lower extremity but not extending into a deep vein.[21] Radiologic examinations finding evidence of chronic thrombosis were coded as negative.

Similarly, positive radiology reports for a PE included those where a filling defect was identified in the central, segmental, or subsegmental pulmonary arteries. Radiologic reports describing evidence of chronic PE were coded as negative, as were those reporting no evidence of the disease.[20] To assess the reliability of the reference standard, 2 clinical experts (T.E. and S.S.) blindly recoded a random 20% of the radiology reports. Intercoder reliability was assessed using the κ statistic; yielding near perfect agreement ($\kappa = 0.98$).

DATASET PREPARATION

In preparation for symbolic NLP classifier development and validation, the 4000 narrative reports were broken down into sentences using the period as the breakpoint. Words within these sentences were converted to lower case and their radical form, and punctuation marks were removed. This process generated sequences of unigrams, which included word radicals (eg, thromb, embol) and acronyms (eg, DVT, PE).

FEATURE SELECTION: DISEASE REFERENCES AND NEGATION WORD MODIFIERS

Then, with the input of clinical experts, unigrams that referred to DVT and PE were selected from the training set and listed in 2 different disease reference sets; one that referred to DVT concepts, and the other to PE concepts. Throughout the initial stages of the NLP classifiers development it was observed that bigrams increased the accuracy of DVT and PE detection. Bigrams, which are 2 unigrams that appear in the same sentence and that are separated by none (eg, pulmonary embolism) to many other unigrams (eg, the words *thrombus* and *brachial* in the sentence: "a thrombus is observed in the brachial vein"), were thus included in the

TABLE 1. Disease Reference Sets for Defining NLP Models Identifying DVT and PE in Narrative Radiology Reports*

DVT			PE		
	Bigram[‡]			Bigram	
Unigram[†]	Pathologic Manifestation	Anatomic Reference	Unigram	Pathologic Manifestation	Anatomic Reference
DVT	Clot	Axillary	PE	Clot	Artery
	Defect	Brachial	Embol	Defect	LLL
	Occlusion	Brachiocephalic		Thromb	Lobe
	Thromb	Deep			LUL
		Femoral			Lung
		Illiac			Pulmon
		Profunda			RLL
		Radial			Segmental
		Subclavian			
		SVC			
		Ulnar			

*Words underlined are stems (ie, thromb is a stem for words such as thrombus, thrombosis, thrombosed).

[†] Unigrams are unique words and also include abbreviations and acronyms.

[‡] Bigrams are 2 unigrams that appear in the same sentence and that are separated by none (eg, pulmonary embolism) to many other unigrams (eg, the words thrombus and brachial in the sentence: "a thrombus is observed in the brachial vein.").

DVT indicates deep vein thrombosis; LLL, left lower lobe; LUL, left upper lobe; NLP, natural language processing; PE, pulmonary embolism; RLL, right lower lobe.

disease reference sets (Table 1). These bigrams combined unigrams that referred to pathologic manifestations of DVT or PE (eg, thromb, clot, embol) and unigrams that referred to relevant anatomic body parts [eg, femoral (vein), segmental (artery)] (Table 1).

Then a list of negation modifier words was created with the input of the clinical experts. These negation modifiers aimed at identifying narrative radiology reports that described findings that were: (1) historical (eg, past medical history of lower extremity DVT); (2) chronic; (3) possible but not confirmed; or (4) negative (eg, no evidence of DVT or PE could be identified) (Table 2).

SYMBOLIC NLP CLASSIFIERS DEVELOPMENT AND TESTING

Lastly, a set of classification rules was defined to determine if a given sentence contained a positive reference to DVT or PE by automatically searching for entries in the disease reference sets. If a reference to DVT or PE was found and no negation modifier could be identified in the same sentence, the sentence was classified as positive for DVT or PE. When a negation modifier word was present, the sentence was coded as negative. In the event that 2 distinct negation modifier words followed each other (eg, not rule-out), they were deleted and the sentence was coded as positive for DVT or PE. Lastly, any narrative report with at least 1 positive sentence was classified as a positive report. Figure 1 shows a flow chart of how the classification rules function.

In developing these classification rules, we iteratively reviewed false-positive and false-negative reports in the training data to further perfect the disease reference sets and the set of negation modifier words. A new disease reference or negation modifier word was included in their respective sets only if it increased the net number of correctly classified reports by 5 or more in the training data. We continued this process until our manual review of false positives and false negatives could no longer produce such modifications. As a sensitivity analysis, we explored if using alternative cutoff values (eg, 3, 6, or 7) for including a new disease reference or negation modifier word in their respective set improved the accuracy of DVT and PE detection. Throughout the development and initial testing of the NLP classifiers, the testing set was not used to avoid overfitting the NLP classifiers to the data.

VALIDATION OF THE SYMBOLIC NLP CLASSIFIERS

The accuracy of the best performing NLP classification rules developed on the training set was assessed on the testing set. To further assess the robustness of these rules, a second sensitivity analysis was performed, stratifying the testing set by years when the radiologic examination was performed. Sensitivity, specificity, positive predictive value (PPV), and negative predictive value (NPV) along with their 95% confidence intervals (CIs), were calculated and reported for DVT and PE separately. Sensitivity is the proportion of correctly classified positive reports among all positive reports identified by clinical experts. Specificity is the proportion of correctly classified negative reports among all negative reports. PPV is the proportion of correctly classified positive reports among all reports classified to be positive. NPV is the proportion of correctly classified negative reports among all reports classified to be negative. Confidence intervals were constructed using the binomial distribution. To assess whether there were any differences in writing styles across institutions, we compared the percentages of positive reports identified by each unigram and bigram in the training and the testing datasets. The unit of analysis was the narrative radiology report, and multiple reports from the same patient were included in the analyses. R 3.0.1 and Java 1.7.1 were used for the analyses.

RESULTS

The 4000 radiology reports came from 2819 patients over 3140 distinct hospitalizations. The mean age of these

TABLE 2. Negation Word Modifiers

Unigram*		Bigram†	
Occurring Before a Disease Reference	**Occurring After a Disease Reference**	**Occurring Before a Disease Reference**	**Occurring After a Disease Reference**
Absence	Artifact	Rule out	Ruled out
Assess	Excluded	Ruling out	
Chronic	Negative		
Exclude	Protocol		
History	Study		
No			
Non			
Not			
Possible			
Previous			
R/o			
Suspected			
Suspicious			
Without			

*Unigrams are unique words and also include abbreviations and acronyms.
†Bigrams are sets of 2 words where the second word must immediately follow the first.
R/o indicates rule out.

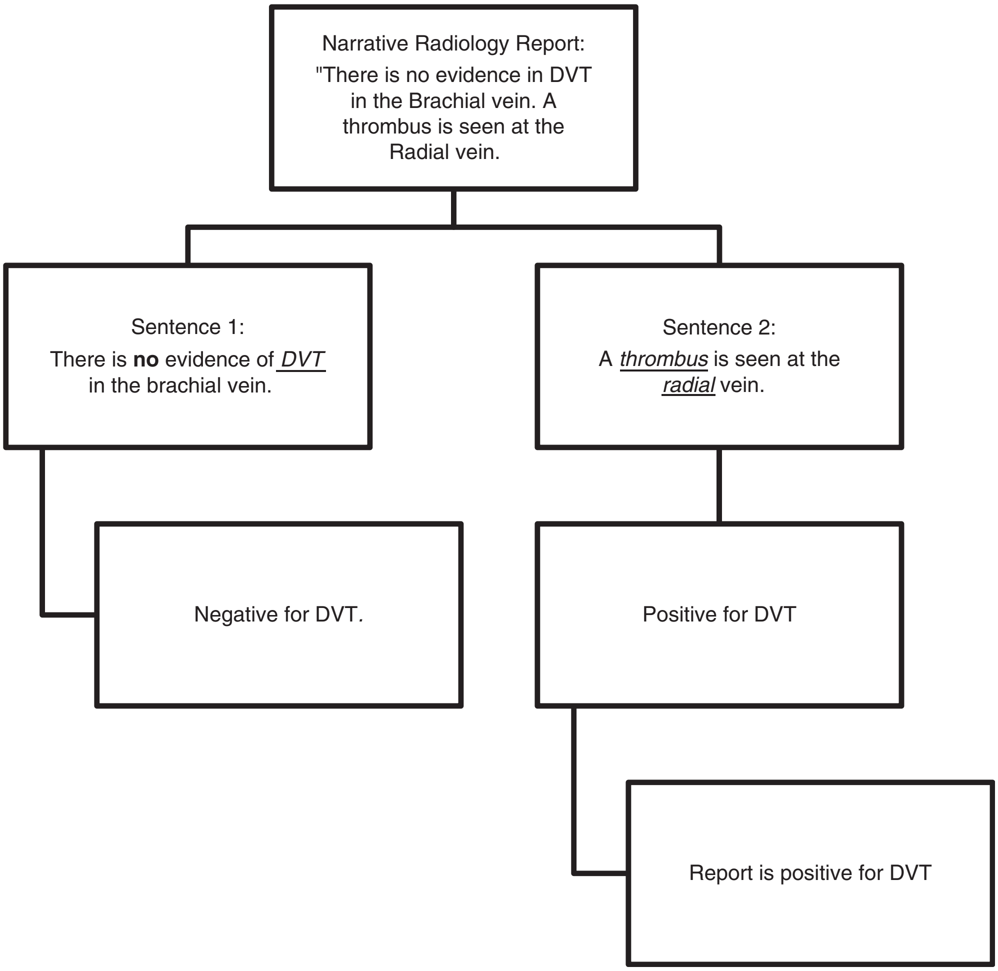
Narrative Radiology Report:
"There is no evidence in DVT in the Brachial vein. A thrombus is seen at the Radial vein.
Sentence 1:
There is **no** evidence of DVT in the brachial vein.
Sentence 2:
A thrombus is seen at the radial vein.
Negative for DVT.
Positive for DVT
Report is positive for DVT

FIGURE 1. Conceptualization of the algorithm in a simple example. Word in bold mark a negation modifier word. Words underlined mark a reference to venous thromboembolism.

TABLE 3. Descriptive Statistics of Radiologic Reports Between Training Data and Testing Data, Stratified by Year

Characteristics	Overall (4000)	Training (2788)	Testing (1212)	2008 (533)	2009 (469)	2010 (526)	2011 (1966)	2012 (506)
Age, [mean (SD)]	66.8 (16.1)	67.1 (16.1)	66.2 (16.4)	68.7 (14.4)	70.2 (13.7)	70.5 (12.2)	63.5 (17.9)	71.1 (12.6)
Sex [n (%)]								
Female	1980 (49.5)	1302 (46.7)	677 (55.8)	265 (49.7)	251 (53.5)	250 (47.5)	913 (46.4)	301 (59.4)
Male	2020 (50.5)	1486 (53.2)	530 (43.7)	268 (50.2)	218 (46.5)	276 (52.4)	1053 (53.5)	205 (60.6)
Comorbidity [mean (SD)]*	2.84 (2.31)	2.97 (2.35)	2.76 (2.22)	3.27 (2.44)	3.23 (2.52)	2.95 (2.33)	2.51 (2.09)	3.15 (2.60)
VTE status [n (%)]								
DVT positive	662 (16.5)	538 (19.3)	124 (10.2)	98 (18.3)	91 (19.4)	111 (21.1)	290 (14.7)	65 (12.8)
PE positive	272 (6.8)	114 (4.0)	148 (12.2)	19 (3.5)	28 (5.9)	36 (6.8)	152 (7.7)	37 (7.3)

*Comorbidities were measured using the Charlson's comorbidity index.

DVT indicates deep vein thrombosis; PE, pulmonary embolism; VTE, venous thromboembolism.

patients was 66.8 years, and 49.5% were female. On manual review, 662 (16.5%) reports were DVT positive and 272 (6.8%) were PE positive (Table 3).

The accuracy of the NLP classifiers for identifying DVT and PE in the training and testing sets are presented in Table 4. In the testing set, the NLP rules achieved 94% sensitivity (95% CI, 88%-97%), 96% specificity (95% CI, 94%-97%), and 73% PPV (95% CI, 65%-79%) in classifying DVT, and 94% sensitivity (95% CI, 89%-97%), 96% specificity (95% CI, 95%-97%), and 80% PPV (95% CI, 73%-85%) in classifying PE. Interestingly, some striking differences were noted, between the training and the testing sets, in the percentages of true positive radiology reports for DVT and PE that were correctly identified by some unigrams and bigrams (Appendix 1). For instance the word "defect" identified 27% of the true positive reports for DVT in the testing set and only 1% of such reports in the training set. Conversely, "clot" identified 26% of the true positive reports for PE in the training set as opposed to only 5% in the testing set (Appendix 1).

With the exception of PPV, results of the sensitivity analysis, stratifying the testing set by years of radiologic examination, generated very little change in the accuracy of the NLP classifiers within each stratum (Table 4). Of note, the year 2011 contained about 3 times more radiology reports than the other years; a pattern that reflects the observations in the complete radiology database. Although accuracy of the NLP classifiers did not change with the year of the radiologic examination, it was influenced by the cutoff value used for retaining a disease reference or a negation word modifier. Using a cutoff of 3 resulted in a slight increase in the accuracy in the training data as several more bigrams were included in the negation word modifiers. However, this increase was not observed in the testing data. Using 6 as the cutoff resulted in no change in accuracy, whereas using 7 resulted in the loss of several negation word modifiers and a subsequent decrease in accuracy in both the training and testing data.

Lastly, we examined the falsely classified cases in the test set. Among the false positive for PE and DVT, 48% were due to statements that mention only a possibility of VTE (eg, "filling defect seen in the pulmonary arteries may be an artifact"), 20% referred to thrombus in nonrelevant body parts (eg, "clot in the arteries above the elbow"), 18% were due to references to previous or chronic disease (eg, "observed DVT may be of chronic nature"), and 13% were due to complex sentences and grammar (eg, "probable compression of a segmental pulmonary artery from a tiny node rather than a pulmonary embolus"). As for the false negatives, they were all due to complex sentencing.

DISCUSSION

In this study, we measured the accuracy of using symbolic NLP classifiers for identifying DVT and PE from narrative radiology reports. We found that a classifier using grammatical and semantic rules to encode domain expert knowledge can identify DVT and PE as accurately as a manual review of the narrative reports.

Recently, Hanauer et al[22] achieved 93% sensitivity, 96% specificity, and 20% PPV in a single-center study that identified postoperative PE from dictated clinical notes (excluding radiology reports). In Hanauer's analysis, only 6 terms were used to identify PE (ie, *PE*, *pulmonary embolism*, *pulmonary embolus*, *pulmonary emboli*, *pulmonic emboli*, *pulmonary thromboembolism*). Although the sensitivity and the specificity observed in Hanauer's study compare to those measured in this study, their lower PPV could potentially be explained by their usage of a limited number of disease references and negation word modifiers, which may have inflated the number of false positives. Indeed, while our disease references for PE included all of Hanauer's terms, additional pathologic manifestations (eg, defect, clot), the use of anatomic references (eg, artery, segmental), and additional negation word modifiers were also required to maximize accuracy. Alternatively, it is also possible that both the number of distinct words used to refer to a given condition (eg, pulmonary embolism) and the prevalence of that condition may vary, within a given institution, across the numerous sources of narrative documents available in the EHR (eg, progress notes vs. narrative radiology reports).

These technical issues are further illustrated in recent work by Murff et al.[17] These researchers used an all-purpose symbolic NLP classifier, *Multi-threaded Clinical Vocabulary Server NLP system*, to detect a variety of adverse events, including VTE, from an integrated EHR. Using pooled data from 5 Veteran Health Affairs hospitals, and a variety of narrative reports (eg, radiology, surgery, outpatient visits), they reported 59% sensitivity, 91% specificity, and 11% PPV for the detection of VTEs. In a replication study based on a larger sample of Veteran Health Affairs hospitals and

TABLE 4. Accuracy of Using Symbolic Natural Language Processing Models to Identify DVT and PE From Narrative Radiology Reports; Overall All Years in the Training and Testing Sets, and by Years

	DVT				PE			
	Sensitivity	Specificity	PPV	NPV	Sensitivity	Specificity	PPV	NPV
Data (n)	Overall all years in the training and testing sets							
Training (2788)	96% (94%-97%)	96% (95%-97%)	86% (83%-88%)	99% (99%-100%)	96% (90%-98%)	99% (98%-99%)	74% (66%-81%)	100% (99%-100%)
Testing (1212)	94% (88%-97%)	96% (95%-97%)	73% (66%-80%)	96% (98%-100%)	94% (89%-97%)	96% (95%-97%)	80% (73%-85%)	99% (98%-100%)
Exam year (n)	By years when the radiologic examination was conducted within the test data							
2008 (165)	93% (61%-98%)	97% (92%-98%)	74% (48%-90%)	99% (95%-100%)	95% (73%-99%)	97% (92%-98%)	78% (56%-92%)	99% (96%-100%)
2009 (172)	95% (77%-99%)	93% (88%-96%)	68% (48%-83%)	99% (96%-99%)	88% (69%-97%)	97% (93%-99%)	85% (66%-95%)	98% (94%-100%)
2010 (176)	96% (79%-99%)	95% (89%-97%)	75% (56%-88%)	99% (96%-100%)	96% (80%-99%)	97% (93%-99%)	86% (68%-96%)	99% (96%-100%)
2011 (540)	93% (82%-98%)	98% (96%-99%)	79% (67%-89%)	99% (98%-100%)	94% (85%-98%)	96% (93%-97%)	75% (64%-84%)	99% (97%-100%)
2012 (154)	92% (66%-99%)	94% (89%-97%)	60% (36%-60%)	99% (95%-99%)	100% (82%-100%)	97% (92%-99%)	83% (61%-95%)	100% (97%-100%)

DVT indicates deep vein thrombosis; NPV, negative predictive value; PE, pulmonary embolism; PPV, positive predictive value.

patients, Fitzhenry et al[23] reported 56% sensitivity, 94% specificity, and 15% PPV for DVT (disease prevalence: 2%), and 80% sensitivity, 97% specificity, and 23% PPV for PE (disease prevalence: 1%). Although the PPVs observed in our study are higher than those noted in Murff's and Fitzhenry's studies, it is likely attributable to the higher disease prevalence in radiologic reports targeting VTE (prevalence: 16.5% for DVT and 6.8% for PE). Nonetheless, it is also plausible that the lower overall accuracy observed in Murff's and Fitzhenry's studies could be attributed to their multinarrative multicenter approach. Indeed, numerous and very complex symbolic NLP rules may be required for capturing the idiosyncrasies of different sources of clinical narratives, as well as the potential variations in the structure and content of these sources in each of the institutions involved.[15,16] As the number and the complexity of the rules increase, it is possible that interactions and contradictions among these rules could reduce the overall accuracy of a symbolic NLP classifier; a hypothesis that warrants further investigation.

Moreover, the hospitals included in Murff's and Fitzhenry's studies came from distinct geographical regions of the United States, whereas the 5 hospitals included in our study were all located in the same Canadian city. Although we observed some heterogeneity in writing styles across these sites, it is also possible that this heterogeneity increases as the distance between sites grows. Greater heterogeneity in writing styles could then negatively influence the performances of an NLP classifier. To our knowledge, this hypothesis has never been explored empirically.

An important strength of the method used in this study was the definition of a criterion to avoid overfitting the NLP classifiers to the data, an aspect that has received scant attention in the literature. Indeed, we observed during our iterative training process that the majority of falsely classified reports came from uncommon and complex ways of referencing VTEs and negations, which were observed only once or twice in the training set. Defining rules to code such local idiosyncrasies would have contributed toward overfitting the symbolic NLP classifiers. To guard against this, a classification rule was accepted only if it increased the net number of correctly classified reports by 5 or more in the training data. Moreover, we provided empirical evidence that this criterion maximized the accuracy of VTE detection. Future studies should examine if this criterion also applies to the detection of other conditions (eg, pneumonia), to other types of narrative documents, and other institutions.

Compared with the discharge diagnostic codes, the current reference standard for detecting VTEs and other adverse events, our proposed method, based on symbolic NLP, has a number of other important strengths. First, it uses data that are date and time stamped. Such temporal information provides a mean for determining the timing of adverse event occurrence, a critical requirement for differentiating past medical events from acute complications of hospitalization. Second, narrative radiology reports are available in near real time allowing for the timely monitoring of adverse events. Indeed, the classifiers developed in this study could potentially be integrated into existing EHR systems and assist with real-time surveillance of VTE events. This, would allow hospital administrators to rapidly investigate the likely causes of the events (eg, underuse or inappropriate use of thromboprophylactic measures), and institute the appropriate interventions to minimize their incidence. These represent major advantages over discharge diagnostic codes that often take several months before being available.

Our study is not without limitations. Our proposed approach was only applied to one type of adverse event, namely VTE. Although we have provided evidence that symbolic NLP can accurately identify the 2 clinical manifestations of VTE (ie, DVT and PE), it has yet to be demonstrated if this approach can be generalized to other types of events. Moreover, the symbolic NLP classifiers were developed and tested using a single source of narrative documents. Although this approach was appropriate for the detection of VTEs, it may not be generalizable to other adverse events, such as catheter-associated bloodstream infection or hospital-acquired pneumonia. Indeed, several sources of EHR data may need to be combined to accurately identify these events. In addition, while we have provided evidence of some variability in writing styles, it is nonetheless possible that the accuracy of the NLP classifiers would have been different if they were validated on data from a geographically distant hospital. Lastly, because the NLP classifiers are dependent on the language used in the clinical narratives, recalibration would be advisable before using them in a new clinical environment. Periodic recalibrations are also recommended as technology and clinical guidelines change over time.

CONCLUSIONS

In conclusion, our findings suggest that symbolic NLP classifiers are accurate for identifying DVT and PE from narrative radiology reports. Symbolic NLP classifiers could potentially be used by hospitals to leverage existing EHR data and conduct surveillance of VTE rates, and assess the effectiveness of preventive measures. Additional studies are required to determine if the approach developed and validated in this study can be successfully applied to other types of adverse events, and if the accuracy of symbolic NLP classifiers, in general, can be improved by including additional information from the EHR, such as laboratory and microbiology data, or additional clinical narratives.

APPENDIX

Table A1. Percentage of True Positive Narrative Radiology Reports Identified by Selected Unigrams and Bigrams From the Disease Reference Sets

	True Positive Identification Rate*	
	Training (%)	**Testing (%)**
Deep vein thrombosis (DVT)		
Unigram		
DVT	23	19
Bigram		
Thromb	87	89
Clot	3	5
Occlusion	4	7
Defect	1	27
Deep	29	38
Axillary	7	5
Radial	1	0
Brachiocephalic	1	1
Subclavian	5	7
Brachial	3	0
Ulnar	1	0
SVC	1	4
Femoral	1	32
Pulmonary embolism (PE)		
Unigram		
PE	6	5
Embol	53	65
Bigram		
Defect	44	44
Thromb	11	9
Clot	26	5
Pulmon	53	30
Lung	1	3
Lobe	17	19
RUL	2	0
LLL	3	1
LUL	2	0
RLL	2	1

*The total percentage will exceed 100% as many reports contain >1 sentence that states the occurrence of a venous thromboembolism.

Words in italics are stems (ie, thromb is a stem for words such as thrombus, thrombosis, thrombosed).

LLL indicates left lower lobe; LUL, left upper lobe; RLL, right lower lobe; RUL, right upper lobe; SVC, superior vena cava.

REFERENCES

1. Raskob GE, Silverstein R, Bratzler DW, et al. Surveillance for deep vein thrombosis and pulmonary embolism: recommendations from a national workshop. *Am J Prev Med*. 2010;38(suppl):S502–S509.
2. Centers for Disease Control and Prevention (CDC). Venous thromboembolism in adult hospitalizations—United States, 2007-2009. *MMWR Morb Mortal Wkly Rep*. 2012;61:401–404.
3. G. Maynard JS. *Preventing Hospital-acquired Venous Thromboembolism: A Guide for Effective Quality Improvement (AHRQ Publication No 08-0075)*. RockVille, MD: Agency for Healthcare Research and Quality; 2008.
4. Spyropoulos AC, Lin J. Direct medical costs of venous thromboembolism and subsequent hospital readmission rates: an administrative claims analysis from 30 managed care organizations. *J Manag Care Pharm*. 2007;13:475–486.
5. LaMori JC, Shoheiber O, Mody SH, et al. Inpatient resource use and cost burden of deep vein thrombosis and pulmonary embolism in the United States. *Clin Ther*. 2015;37:62–70.
6. Pendergraft T, Atwood M, Liu X, et al. Cost of venous thromboembolism in hospitalized medically ill patients. *Am J Health Syst Pharm*. 2013;70:1681–1687.
7. Streiff MB, Brady JP, Grant AM, et al. CDC grand rounds: preventing hospital-associated venous thromboembolism. *MMWR Morb Mortal Wkly Rep*. 2014;63:190–193.
8. Zhan C, Miller MR. Administrative data based patient safety research: a critical review. *Qual Saf Health Care*. 2003;12(suppl 2):ii58–ii63.
9. Kaafarani HM, Rosen AK. Using administrative data to identify surgical adverse events: an introduction to the Patient Safety Indicators. *Am J Surg*. 2009;198(suppl):S63–S68.
10. White RH, Sadeghi B, Tancredi DJ, et al. How valid is the ICD-9-CM based AHRQ patient safety indicator for postoperative venous thromboembolism? *Med Care*. 2009;47:1237–1243.
11. Romano PS, Mull HJ, Rivard PE, et al. Validity of selected AHRQ patient safety indicators based on VA National Surgical Quality Improvement Program data. *Health Serv Res*. 2009;44:182–204.
12. H.A.H. SERVICES. *Hospital-Acquired Conditions and Present on Admission Indicator Reporting Provision*. Baltimore, MD: Medicare Learning Network (MLN); 2014.
13. Bates SM, Jaeschke R, Stevens SM, et al. *Diagnosis of DVT: Antithrombotic Therapy and Prevention of Thrombosis, 9th ed: American College of Chest Physicians Evidence-Based Clinical Practice Guidelines*. *Chest*. 2012;141(suppl):e351S–e418S.
14. Kearon C, Akl EA, Comerota AJ, et al. Antithrombotic therapy for VTE disease: Antithrombotic Therapy and Prevention of Thrombosis, 9th ed: American College of Chest Physicians Evidence-Based Clinical Practice Guidelines. *Chest*. 2012;141(suppl):e419S–e494S.
15. Doan S, Conway M, Phuong TM, et al. Natural language processing in biomedicine: a unified system architecture overview. *Methods Mol Biol*. 2014;1168:275–294.
16. Nadkarni PM, Ohno-Machado L, Chapman WW. Natural language processing: an introduction. *J Am Med Inform Assoc*. 2011;18:544–551.
17. Murff HJ, FitzHenry F, Matheny ME, et al. Automated identification of postoperative complications within an electronic medical record using natural language processing. *JAMA*. 2011;306:848–855.
18. Wu CY, Chang CK, Robson D, et al. Evaluation of smoking status identification using electronic health records and open-text information in a large mental health case register. *PLoS One*. 2013;8:e74262.
19. Carrell DS, Halgrim S, Tran DT, et al. Using natural language processing to improve efficiency of manual chart abstraction in research: the case of breast cancer recurrence. *Am J Epidemiol*. 2014;179: 749–758.
20. Rochefort CM, Verma AD, Eguale T, et al. 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;22:155–165.
21. Henderson KE, Recktenwald A, Reichley RM, et al. Clinical validation of the AHRQ postoperative venous thromboembolism patient safety indicator. *Jt Comm J Qual Patient Saf*. 2009;35:370–376.
22. Hanauer DA, Englesbe MJ, Cowan JA Jr, et al. Informatics and the American College of Surgeons National Surgical Quality Improvement Program: automated processes could replace manual record review. *J Am Coll Surg*. 2009;208:37–41.
23. FitzHenry F, Murff HJ, Matheny ME, et al. Exploring the frontier of electronic health record surveillance: the case of postoperative complications. *Med Care*. 2013;51:509–516.