

Finding falls in ambulatory care clinical documents using statistical text mining

James A McCart,¹ Donald J Berndt,² Jay Jarman,³ Dezon K Finch,¹ Stephen L Luther¹

¹Consortium for Healthcare Informatics Research (CHIR) and the HSR&D/RR&D Center of Excellence: Maximizing Rehabilitation Outcomes, James A Haley Veterans' Hospital, Tampa, Florida, USA

²Consortium for Healthcare Informatics Research (CHIR) and the University of South Florida College of Business, Tampa, Florida, USA

³Consortium for Healthcare Informatics Research (CHIR) and East Tennessee State University College of Business and Technology, Johnson City, Tennessee, USA

Correspondence to

Dr James A McCart, Consortium for Healthcare Informatics Research (CHIR) and the HSR&D/RR&D Center of Excellence: Maximizing Rehabilitation Outcomes, James A Haley Veterans' Hospital, 8900 Grand Oak Circle, 118M, Tampa, FL 33637-1022, USA; james.mccart@va.gov

Received 13 September 2012

Revised 1 November 2012

Accepted 21 November 2012

ABSTRACT

Objective To determine how well statistical text mining (STM) models can identify falls within clinical text associated with an ambulatory encounter.

Materials and Methods 2241 patients were selected with a fall-related ICD-9-CM E-code or matched injury diagnosis code while being treated as an outpatient at one of four sites within the Veterans Health Administration. All clinical documents within a 48-h window of the recorded E-code or injury diagnosis code for each patient were obtained (n=26 010; 611 distinct document titles) and annotated for falls. Logistic regression, support vector machine, and cost-sensitive support vector machine (SVM-cost) models were trained on a stratified sample of 70% of documents from one location (dataset A_{train}) and then applied to the remaining unseen documents (datasets A_{test}-D).

Results All three STM models obtained area under the receiver operating characteristic curve (AUC) scores above 0.950 on the four test datasets (A_{test}-D). The SVM-cost model obtained the highest AUC scores, ranging from 0.953 to 0.978. The SVM-cost model also achieved F-measure values ranging from 0.745 to 0.853, sensitivity from 0.890 to 0.931, and specificity from 0.877 to 0.944.

Discussion The STM models performed well across a large heterogeneous collection of document titles. In addition, the models also generalized across other sites, including a traditionally bilingual site that had distinctly different grammatical patterns.

Conclusions The results of this study suggest STM-based models have the potential to improve surveillance of falls. Furthermore, the encouraging evidence shown here that STM is a robust technique for mining clinical documents bodes well for other surveillance-related topics.

INTRODUCTION

Fall-related injuries are an important healthcare issue, especially among aging populations. A recent study found approximately 40% of people age 70 years and older reported falling during a 1-year period.¹ Injuries due to falls are a leading cause of death and disability among older adults,² with direct costs totaling almost US\$20 billion annually for people 65 years and older in the USA.³ A national study of emergency department (ED) visits between 2001 and 2008 found that fall-related hospitalizations in older adults increased 50%, from 373 128 to 559 355 cases with the age-adjusted incidence rate, expressed per 100 000 population, increased from 1046 to 1368.⁴ While most estimates of treatment for fall-related injuries come from hospital ED data, a recent national survey estimated that treatment for more than 50%

of 76 million non-fatal acute injuries (most of which were fall injuries) occurred in ambulatory care settings outside of hospital ED.⁵

A history of a previous fall is one of the most important clinical indicators that identify an elderly patient as being at high risk of future falls.⁶ However, falls have been found to be under-coded in administrative databases,⁷ making it difficult to identify at-risk patients and thus take steps to help prevent falls. An alternative source of fall-related information may be found in the clinical text associated with a patient's electronic health record (EHR). This study explores how well falls, associated with an ambulatory encounter for the treatment of an injury, can be identified in clinical text.

The Veterans Health Administration's (VHA) EHR contains almost two billion clinical documents (Scott DuVall, personal communication, 2012) and provides a rich repository to assess the effectiveness of automated text-based surveillance systems. Given this resource, we explored how well statistical text mining (STM), a machine learning approach that represents documents as a 'bag of words', could classify individual clinical documents (progress notes, reports, etc.) as being fall related or not. Within the VHA's EHR, clinical documents are assigned a title that reflects either the place of service or clinical author. Example titles include 'Emergency Department' progress notes, 'Nursing Triage' progress notes, or 'Orthopedic Surgery Consult' progress notes. As patients may receive care for a fall from multiple sources (eg, ED, outpatient clinic), fall-related information is likely to be found in a variety of document titles. We therefore selected a heterogeneous collection of documents, representing a wide variety of document titles, each with varying clinical sublanguages,⁸ to help maximize the discovery of fall-related documents and assess model performance across a variety of document types. Finally, we also explored how well our models generalized by building models using a single site and then applying those models to three other sites. This method mimics situations in which a system is built and used at a single facility and then later rolled out to other facilities.

BACKGROUND

STM is an inductive approach based on machine learning that uncovers patterns from textual documents such as clinical progress notes. STM induces structure on unstructured text by representing each document within a term-by-document matrix, in which each term from the document collection is represented in the matrix. The cells of the term-by-document matrix consist of counts or weightings of

To cite: McCart JA, Berndt DJ, Jarman J, et al. *J Am Med Inform Assoc*. Published Online First: 14 December 2012
doi:10.1136/amiainjnl-2012-001334

Research and applications

how many times a particular term occurs within a document. Patterns based on these counts can then be derived and used for both unsupervised (eg, clustering) and supervised (eg, classification) tasks. An overview of text classification using machine learning can be found in Sebastiani.⁹

STM has been commonly used to prioritize and/or categorize biomedical literature to help reduce the effort associated with curating articles.^{10–12} In this area, numerous article-based classification challenges and tasks have been introduced to help spur the development and comparison of tools, as well as to provide readily available datasets. For instance, the 2004 and 2005 Text REtrieval Conference (TREC) Genomics track contained tasks to determine whether articles should be examined for curation in the Mouse Genome Informatics (MGI) system.^{13–15} The BioCreative II and III challenges also had tasks to classify and rank articles based on their likelihood of having protein–protein interactions of interest.^{16 17}

In other medically related areas, challenges have also been organized that are well suited for STM. For instance, classifying patient smoking status (first informatics for integrating biology and the bedside (i2b2) shared task)¹⁸ and determining sentiment found in suicide notes (fifth i2b2/VA/Cincinnati shared task).¹⁹ Outside of community organized challenges, STM has also been used to augment terms for ontology development,²⁰ classify foot examination findings,²¹ detect infectious disease outbreaks from data on the web,^{22 23} and highlight adverse events of anaphylaxis in US vaccine adverse event reporting system reports about H1N1.²⁴ Related specifically to falls, our previous work used both supervised and unsupervised STM to identify falls from clinical text.⁷ In this current work, we focus strictly on supervised STM, and conduct a more comprehensive analysis across a larger dataset spanning multiple facilities.

METHODS

Sample

We followed a four-step process to obtain our corpus. In step 1, we identified our initial cohort based on International Classification of Disease, version 9, clinical modification (ICD-9-CM) codes from the ambulatory care encounter records of four VHA hospital outpatient clinics and community-based outpatient centers in the southeastern USA and Puerto Rico during fiscal year 2007. The patients selected had an ICD-9-CM code for treatment of injury (ICD-9-CM 800-999) or a ‘fall-related’ E-code (ICD-9-CM E880-889) listed as the primary reason for the visit. The cohort was further reduced by eliminating records for poisonings (ICD-9-CM 960-979), toxic effects of non-medical substances (ICD-9-CM 980-989), other unspecified effects of external causes (ICD-9-CM 990-995), complications of surgical and medical care, not elsewhere classified (ICD-9-CM 996-999), and records in which the primary reason for the encounter was due to spinal cord injuries (ICD-9-CM 952) or dealing with occasions other than disease or injury-classifiable categories (ICD-9-CM V01-V86).²⁵ In step 2, patients with a primary injury code and a supplementary ‘fall-related’ E-code were matched with up to two controls drawn from patients with only a primary injury code based on facility, gender, type of injury (ICD-9-CM code), and age (within 10 years). Patients with only a primary ‘fall-related’ E-code were unable to be matched due to a missing injury code, but were still included in the cohort. In step 3, all outpatient administrative encounters for each patient in the cohort over the study period with primary injury codes and/or ‘fall-related’ E-codes were selected and grouped by day to form ambulatory visits. In step 4, all documents authored within a 48-h window of each ambulatory visit were extracted and placed in our corpus to be

annotated. The 48-h window was necessary because the VHA EHR does not link ambulatory administrative data and clinical documents (ie, a clinical document is not associated with a particular encounter), thus the window helped ensure documents relevant to the ambulatory visit were selected.

Annotation

All documents in the corpus were annotated to provide a reference standard for STM with a document-level classification of ‘fall’ or ‘not fall’. To create the reference standard, written guidelines defining what constituted a fall were first developed and reviewed by a clinical expert. A fall was defined according to the World Health Organization as ‘inadvertently coming to rest on the ground, floor or other lower level, excluding intentional change in position to rest in furniture, wall or other objects’.²⁶ The guidelines also contained examples of text indicating a fall occurred, such as ‘pt fell and broke his hip’ and ‘pt slipped and fell in shower’. An annotation schema based on the guidelines was then created using Knowtator.²⁷ Three clinicians, already experienced with annotation, were then trained using the guidelines and a set of 50 documents randomly selected from the first available site. The annotators were instructed to identify spans of text indicating the patient had suffered a fall. Based on feedback from the initial training, the guidelines were revised and 100 additional documents were randomly selected for further training. The 150 documents were used as a training reference standard (based on the clinical expert’s annotations) to assess how well the clinicians were performing according to the guidelines. The annotation of the final dataset began once all three clinicians achieved a Cohen’s κ ²⁸ of 0.80 or above compared to the training reference standard. To monitor adherence to the written guidelines after training, spot checks using the clinical expert were performed on 10 documents for every 1000 documents annotated. A final κ was then calculated based on all spot checks.

Modeling

A stratified random sample of 70% of documents from the site with the most documents was used for training STM models (dataset A_{train}), while the remainder were held out as a test set (dataset A_{test}). All documents from the other sites were used as additional test sets (datasets B–D). The first step of the STM process consisted of transforming documents into a term-by-document matrix by converting all text to lower case; tokenizing; removing tokens with fewer than three characters or no alphabetical characters; normalizing terms using the National Library of Medicine lexical tool Norm;²⁹ removing stop words; and removing terms that only occur once in the matrix. A model selection process was then performed on the A_{train} dataset using two machine learning algorithms: logistic regression (LR) using LogitBoost^{30 31} and linear support vector machines (SVM).^{32 33} Features for the models were developed by weighting the term-by-document matrix and then performing dimension reduction. Three factors were used in weighting the matrix: (1) term frequency for local weight; (2) collection frequency for global weight; and (3) a normalization factor.³⁴ A log transformation on term frequency and cosine normalization was used for the first and third weighting factors, respectively. For the second weighting factor three commonly used methods were examined: χ^2 , gain ratio, and log OR.

Features were selected from the top n weighted terms and/or generated using latent semantic analysis.³⁵ Latent semantic analysis uses singular value decomposition (SVD) to decompose the term-by-document matrix so documents with different, but related, term usage map to the same dimension. The top m

Table 1 Modeling parameters

Algorithm	Parameters
Logistic regression	
Global weighting:	GR, log OR, χ^2
Top n terms:	0, 50, 100, 250
SVD dimensions:	0, 50, 100, 200
Support vector machines	
Global weighting:	GR, log OR, χ^2
Top n terms:	0, 50, 100, 250, 500
SVD dimensions:	0, 50, 100, 200
C:	10^{-4} , 10^{-3} , ..., 10^0 , ..., 10^3 , 10^4

GR, gain ratio; SVD, singular value decomposition.

SVD dimensions, which represent the best approximation of the matrix in m dimensions, are then selected as features for a model.

Table 1 summarizes the parameter values used to build models for each algorithm. A model was built for each combination of parameters (excluding when the top n terms and m SVD dimensions were both zero). The performance of each model was evaluated using 10-fold stratified cross-validation on A_{train} , with the 'best' model determined based on the area under the receiver operating characteristic curve (AUC). After finding the best SVM model based on a global weighting option, with the top n terms and m SVD dimensions, an additional round of model refinement was performed by varying the C parameter, which adjusts between minimizing error and maximizing the margin between classes.³⁶ The performance of each refined SVM model was again calculated using 10-fold stratified cross-validation on A_{train} . The parameter combinations that obtained the highest AUC for LR and SVM were then used to train models on the entire A_{train} dataset, and were subsequently applied to the test datasets (A_{test} –D). All analyses were performed using RapidMiner³⁷ (V5.2) with the Text Processing (V5.2.1) and Weka (V5.1.1) extensions, as well as some custom-built components.

Cost-sensitive learning

In this study we were concerned with achieving high levels of sensitivity, while still maintaining relatively high levels of specificity. MetaCost,³⁸ a cost-sensitive classification approach, was thus explored to minimize false negatives (FN) and hence increase sensitivity. MetaCost uses a form of bagging³⁹ to re-label the training data with the estimated optimal class and then learns a single cost-sensitive model from the re-labeled data.³⁸

The same parameter combination that resulted in the best SVM model on A_{train} was used for the model in MetaCost (SVM-cost). The MetaCost procedure, using an ensemble of 10 models, was performed within each fold of a 10-fold stratified cross-validation process on A_{train} . This process was repeated five times for cost ratios of false positives (FP) to FN ranging from 1:1 to 1:2 in 0.25 increments. The cost ratio that demonstrated the greatest balance between sensitivity and specificity was used to build a model on the entire A_{train} dataset and then applied to the test datasets (A_{test} –D).

Error analysis

An error analysis of misclassified documents was undertaken to understand better the limitations of the trained models. Incorrect classifications for each site were categorized into seven groups based on whether the error was unique to a particular model or combination of models (eg, {LR}, {SVM}, ..., {LR, SVM-cost}, ..., {LR, SVM, SVM-cost}). A maximum of 10 documents (five FP and five FN) were randomly sampled from each group, and patterns of errors were noted.

RESULTS

Table 2 lists the datasets, number of patients, how many documents were annotated as 'fall' and 'not fall', and the number of distinct document titles. The final dataset contained 26 010 documents with 611 distinct titles from 2241 patients, with 5034 (19.4%) of those documents labeled as 'fall' (representing approximately half of all distinct document titles). Documents from all available patients were used from sites A and B. Approximately 20% of available patients were used at site C because of annotation resource constraints. During spot checks of site D it was determined one annotator was having difficulty because the documents were extracted without any formatting (eg, line breaks and paragraphs were not preserved). Therefore, any patients with documents annotated by that annotator at site D were removed, resulting in approximately 80% of the patients being retained. Only formatted documents were used for the other sites, and no further annotation difficulties were noted. The κ across all four sites and three annotators was 0.90 ($n=275$). Figure 1 illustrates the number of documents labeled as 'fall' versus 'not fall' for the 10 document types with the most fall documents, excluding those with the title 'addendum'. Not surprisingly, the majority of the labels reflect services provided in the emergency setting, although documents in both primary care and internal medicine were also common.

Processing the documents from the A_{train} dataset resulted in a sparse term-by-document matrix of 14 762 terms from 6641

Table 2 Dataset descriptive statistics

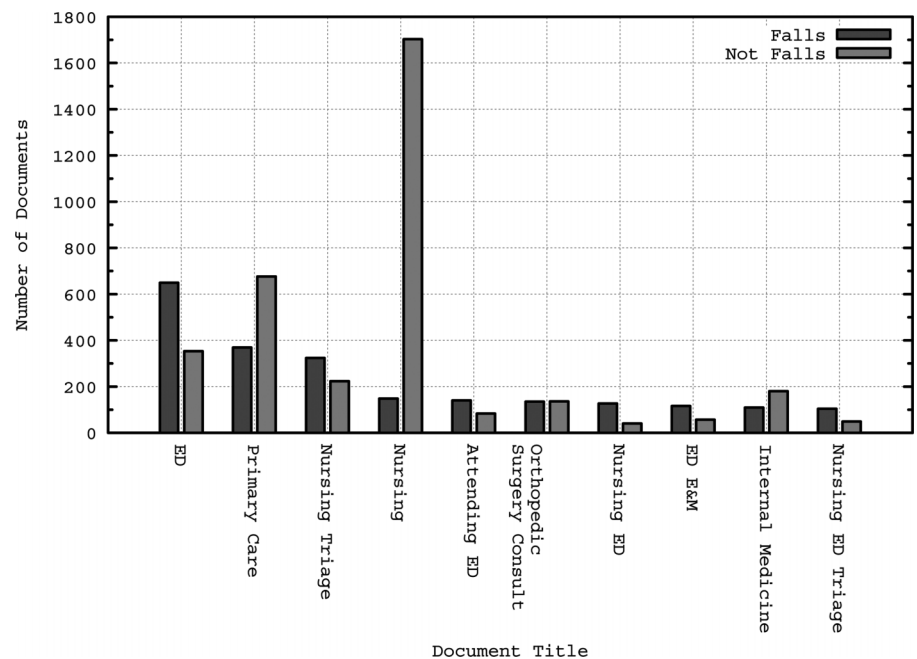
Dataset	Patients	Documents			Document titles*		
		Fall	Not fall	Total	Fall	Not fall	All
Site A	841						
Train	A_{train}	1210 (18.2%)	5431 (81.8%)	6641	111 (43.5%)	241 (94.5%)	255
Test	A_{test}	519 (18.2%)	2327 (81.8%)	2846	78 (40.6%)	182 (94.8%)	192
Site B	B	1845 (19.7%)	7506 (80.3%)	9351	127 (47.2%)	257 (95.5%)	269
Site C†	C	557 (14.8%)	3214 (85.2%)	3771	99 (45.2%)	216 (98.6%)	219
Site D†	D	903 (26.6%)	2498 (73.4%)	3401	96 (44.0%)	203 (93.1%)	218
Total	2241	5034 (19.4%)	20976 (80.6%)	26010	302 (49.4%)	586 (95.9%)	611

*Number of distinct document titles. The total row represents the distinct number of document titles across all sites.

†Sample of patients (site C $\approx$ 20%; site D $\approx$ 80%).

Research and applications

Figure 1 Ten document types with the most documents labeled 'fall'. ED, emergency department; E&M



documents, with over 99% of the cells having a value of zero. The LR and SVM models with the highest AUC values used features weighted via gain ratio and χ^2 in the term-by-document matrix, respectively. The top 100 terms and 200 SVD dimensions were supplied to the LR model, of which 15 terms and 52 SVD dimensions were selected by the LogitBoost algorithm for inclusion in the final model. The SVM model used the top 50 terms and 200 dimensions in its final model. No performance increases for the SVM model were seen with the additional

round of model refinement. Therefore, the C parameter was kept at the default value of 1. The SVM-cost model used the same parameter settings as the SVM model. In addition, a cost ratio of 1:1.5 FP to FN was used for the SVM-cost model because it resulted in the most balanced performance between sensitivity and specificity (see figure 2).

Table 3 details the results obtained for each model type for both the training and testing datasets. The results from the best models from the model selection and refinement process on the

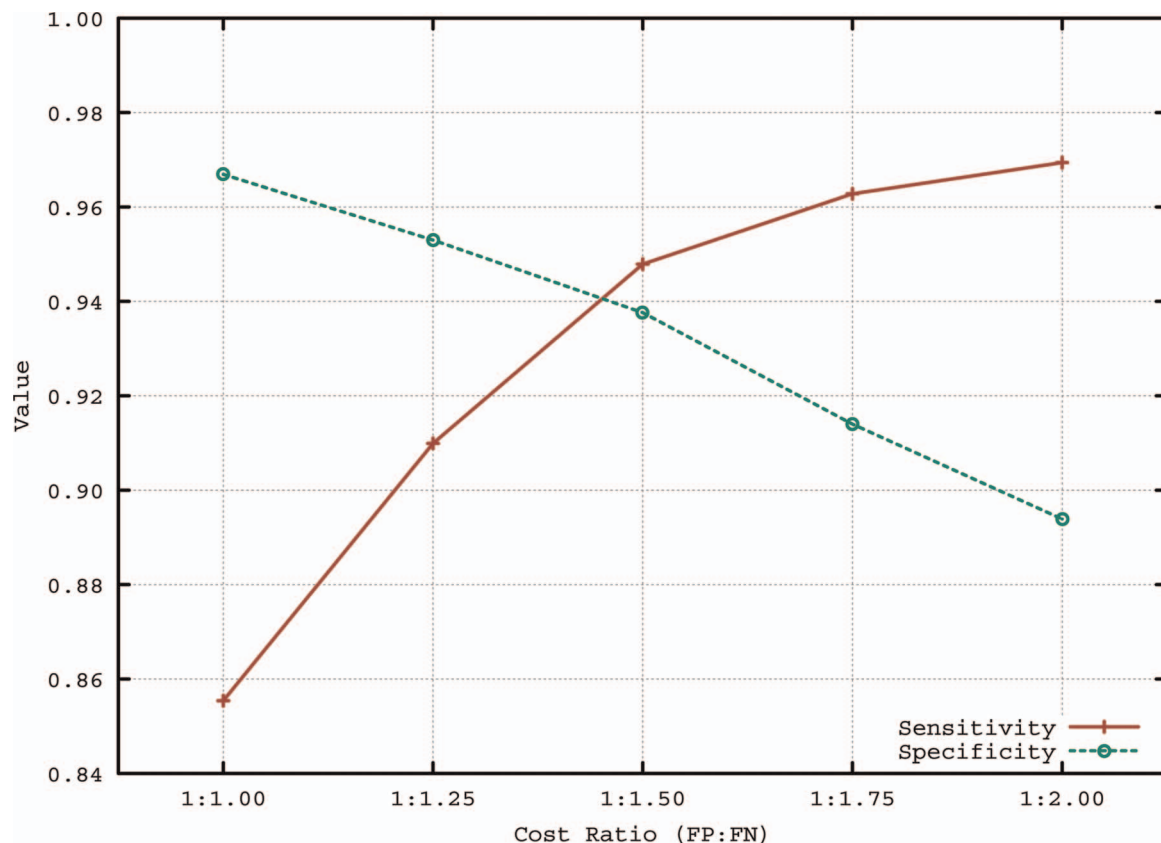


Figure 2 Cost-sensitive classification on A_{train} (support vector machine cost sensitive; SVM-cost). FN, false negative; FP, false positive.

Table 3 Modeling results

	AUC	F-measure	Accuracy	Sensitivity	Specificity	PPV	NPV
Site A _{train}							
LR	0.977	0.850	0.945	0.853	0.966	0.847	0.967
SVM	0.976	0.858	0.948	0.860	0.968	0.856	0.969
SVM-cost	0.977	0.851	0.939	0.948	0.938	0.772	0.988
Site A _{test}							
LR	0.978	0.849	0.945	0.846	0.967	0.852	0.966
SVM	0.978	0.850	0.946	0.840	0.969	0.860	0.965
SVM-cost	0.978	0.853	0.941	0.931	0.944	0.787	0.984
Site B							
LR	0.951	0.767	0.909	0.760	0.946	0.775	0.941
SVM	0.951	0.778	0.916	0.743	0.959	0.817	0.938
SVM-cost	0.953	0.794	0.909	0.890	0.914	0.717	0.971
Site C							
LR	0.957	0.725	0.917	0.741	0.947	0.708	0.955
SVM	0.957	0.749	0.921	0.797	0.942	0.706	0.964
SVM-cost	0.957	0.745	0.907	0.925	0.904	0.624	0.986
Site D							
LR	0.952	0.823	0.906	0.828	0.934	0.818	0.938
SVM	0.953	0.832	0.912	0.818	0.946	0.847	0.935
SVM-cost	0.953	0.808	0.885	0.909	0.877	0.727	0.964

AUC, area under the receiver operating characteristic curve; LR, logistic regression; NPV, negative predictive value; PPV, positive predictive value; SVM-cost, support vector machine cost sensitive.

A_{train} dataset are shown at the top followed by the results of applying these models to the various test datasets. (AUC values for A_{train} were averaged over all folds as recommend by Forman and Scholz.⁴⁰ All other measures for A_{train} were calculated based on the sum of true positives, true negatives, FP, and FN from all

folds.) Figure 3 displays the AUC, F-measure, positive predictive value (PPV), and negative predictive value (NPV) values obtained from each model across all the datasets. The SVM-cost model had the highest AUC across all five datasets, with AUC values ranging from 0.953 to 0.978. However, the maximum difference

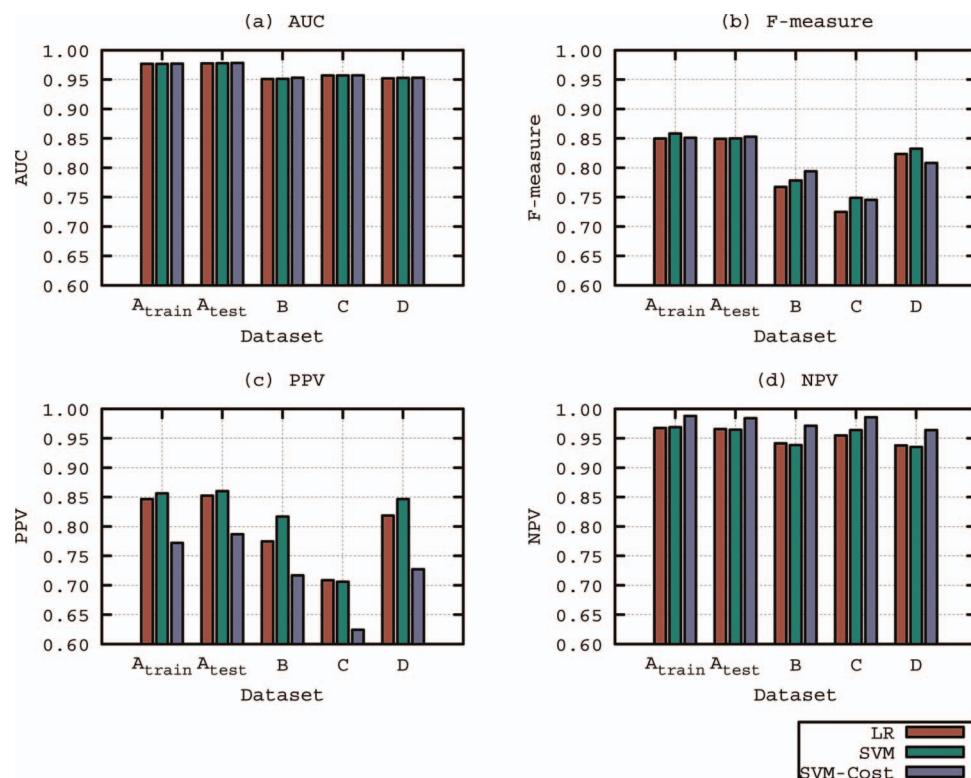
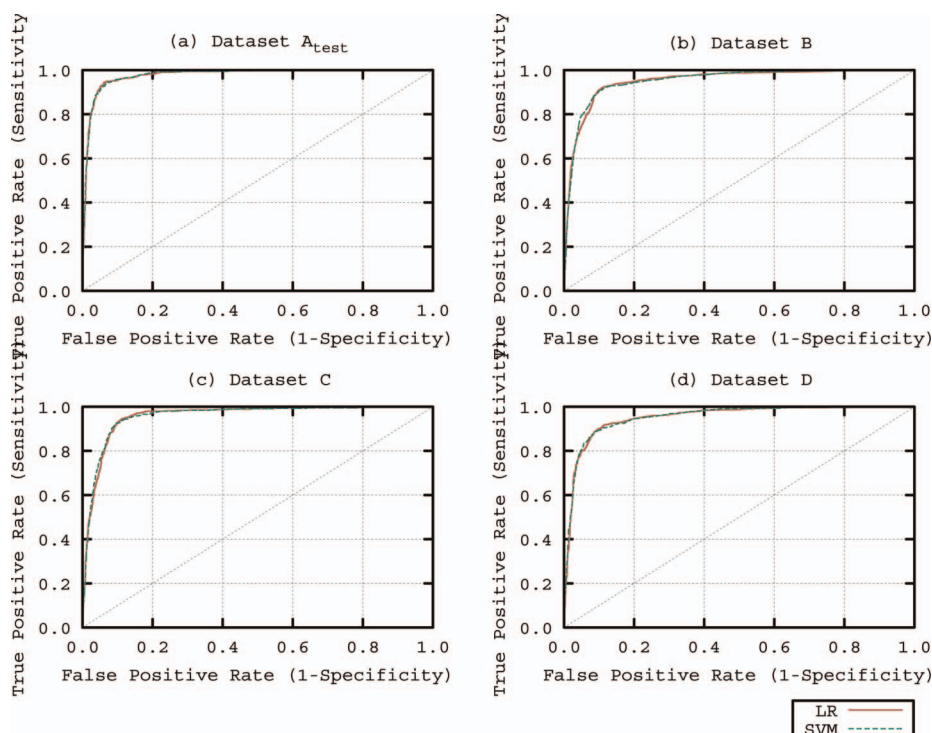


Figure 3 Performance measure charts by model and dataset. AUC, area under the receiver operating characteristic curve; LR, logistic regression; NPV, negative predictive value; PPV, positive predictive value; SVM, support vector machine; SVM-cost, support vector machine cost sensitive.

Research and applications

Figure 4 Receiver operating characteristic curves by model and dataset. LR, logistic regression; SVM, support vector machine.



in AUC between the LR, SVM, and SVM-cost models on any dataset was only 0.02 (dataset B). Compared to the A_{train} dataset, all three models performed quite similarly on the A_{test} dataset, with the largest absolute change being a decrease in sensitivity by 0.02 points for the SVM model. The models performed less well on the other test datasets, with the largest decreases from A_{train} being in sensitivity for all three models for dataset B (0.058–0.117), PPV for all three models (0.139–0.150) for dataset C, and accuracy (0.039), sensitivity (0.042), and specificity (0.061), for the LR, SVM, and SVM-cost models for dataset D. As expected, assigning a greater cost to FN for the SVM-cost model caused measures heavily influenced by FP (eg, specificity and PPV) to be consistently lower than the other two models (average difference of 0.046 and 0.094 for specificity and PPV vs the best performing models from LR and SVM). However, the SVM-cost model was able to maintain an average increase of 0.106 in sensitivity over the next highest model's sensitivity across the test datasets.

Figure 4 shows the receiver operating characteristic curves for the LR and SVM models for all four test datasets. (The

receiver operating characteristic curves for the SVM-cost models are very similar to the SVM models and are not shown to ease readability.) Very little difference is seen between the LR and SVM models for almost all sensitivity values in the A_{test} dataset. For the other test datasets, the LR and SVM models also perform similarly to one another in the very high sensitivity ranges. The overall performance levels are remarkably similar, even though the models use different weighting schemes and overlapping, rather than identical, feature sets. There are some slight variations, with the LR model doing a bit better than the SVM model when the sensitivity ranges from approximately 0.90 to 0.95, 0.94 to 0.97, and 0.90 to 0.93 in the B, C, and D datasets, respectively. Conversely, the SVM model performs slightly better than the LR model at lower sensitivity levels; in particular, when the sensitivity ranges from approximately 0.70 to 0.87, 0.65 to 0.80, and 0.80 to 0.85 in the B, C, and D datasets, respectively. However, these differences are minor and both models represent appropriate choices for this classification task.

Figure 5 displays the unique number of misclassified documents for each individual model and all combinations of models

Figure 5 Misclassification counts by model across all test datasets (A_{test} –D). LR, logistic regression; SVM, support vector machine; SVM-cost, support vector machine cost sensitive.

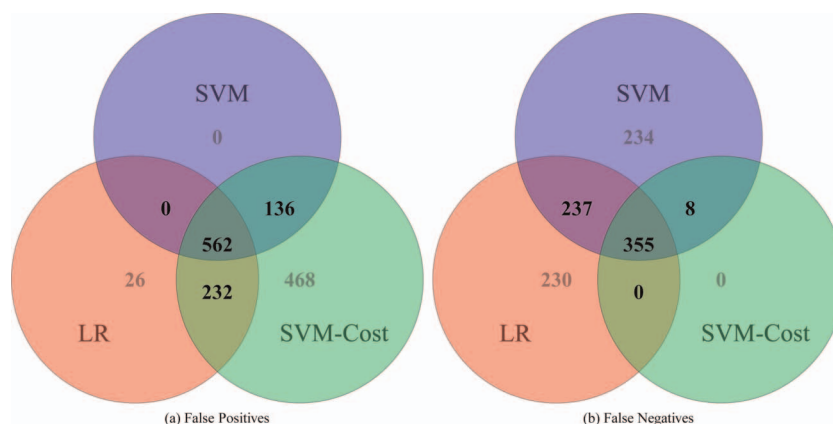


Table 4 Misclassification categories

Misclassification category	Occurrences within error sample		Example document text
UNSURE	56	31.1%	87-year-old woman s/p ORIF Right hip due to fall and fracture
FALL RISK	32	17.8%	Fall-risk
TEMPLATE	30	16.7%	Hendrich II fall risk model Morse fall scale score
JUDGMENT	29	16.1%	Accidentally slide himself from the wall to the floor
FALL-RELATED INFORMATION	18	10.0%	PT GOALS: BE ABLE TO PERFORMED SAFELY WITHOUT FALLING DOWN
HISTORY	15	8.3%	Pt describe 3 falls in the last few years
INCORRECT ANNOTATIONS	8	4.4%	hx of LBP since fall off of fire truck in 1986
MOTOR VEHICLE ACCIDENT	5	2.8%	PT REFERS CAME TO ER DUE TO AN FALL OFF HIS MOTORCYCLE THIS MORNING
NEGATION	4	2.2%	Very painful, didn't fall. Heard a pop Denies falls
SEMANTIC	4	2.2%	Felll (sic) hit head and left hip and leg
OTHER	4	2.2%	Can't find his mouth with the fork and green beans fall off the fork

by error type (FP and FN). For example, figure 5A shows 562 documents (2.9%) were misclassified as FP by all three models, while only 26 documents (0.1%) were solely misclassified by the LR model. An error analysis on a sample of 180 documents (92 FP and 88 FN) from all the groups shown in figure 5 revealed 11 categories of errors. Table 4 lists each error category and an example. Except for the UNSURE category, a document could be classified into more than one category. For instance, several documents contained templates that were used to identify patients at risk of a fall. Therefore, these documents were classified in both the TEMPLATE and FALL RISK categories. Overall, the majority of errors were described in six categories: UNSURE, FALL RISK, TEMPLATE, JUDGMENT, FALL-RELATED INFORMATION, and HISTORY.

The UNSURE category included documents in which the reason for a misclassification was not apparent. Many documents within the UNSURE category contained clear documentation of a fall or fall-related injury occurring (see example in table 4). Documents were classified into the FALL RISK category when information identified a patient as at risk of a fall, but the reason for the patient's visit was not for treatment of an injury due to a fall. How patients were flagged as being at risk of a fall differed by site. One site commonly included the text 'Fall-risk' in the patient's document, whereas other sites included fall risk models, such as the Hendrich II fall risk model⁴¹ or Morse fall scale,⁴² as templates. In addition to fall risk models, the TEMPLATE category also included templates associated with fall prevention. These templates often contained terms predictive of a fall, thus leading to FP.

The JUDGMENT category included documents in which the classification of 'fall' or 'not fall' was not obvious and required further clinical judgment. The example shown in table 4 highlights a scenario in which an elderly patient may not want to be identified as having fallen. The patient may thus report a fall by saying they 'let themselves down' or 'slide himself from the wall to the floor'. The FALL-RELATED INFORMATION category contained documents with information related to, but not specifically about, a fall occurring. For instance, the example shown in table 4 is from a physical therapy document describing a patient's goal to complete an exercise without falling down. The final major category, HISTORY, contained documents in which a patient had fallen in the past; however, the current visit was not related to a fall.

The less frequently occurring categories included errors in annotations (INCORRECT ANNOTATIONS), presence of negation (NEGATION), misspellings (eg, 'fall'), incorrect word usage (eg, 'felt', 'feel'), and wrong word sense (eg, fall from one plane to another vs the season) (SEMANTIC), and other errors such as asserting a fall on someone or something other than the patient (eg, 'green beans fell to the floor') (OTHER). In addition, this project did not consider an accident from a motor vehicle as a fall (MOTOR VEHICLE ACCIDENT). Therefore, documents describing a patient falling off a motorcycle or from a golf cart were not considered to be falls.

DISCUSSION

This study demonstrated the effectiveness of STM in identifying falls in clinical text by maintaining AUC scores above 0.95 across multiple sites with diverse document types. Of the 611 distinct document types in the corpus, almost half had at least one document annotated as a fall, highlighting the fact that fall-related information may be found in numerous places within the EHR. If the analysis had been restricted to the 10 most frequent document types (by number of falls, see figure 1), only half the available fall documents would have been found. If instead document types were selected based on prevalence, the top 10 document types (with at least 10 fall documents) would have resulted in less than 30% of available fall documents being found. Therefore, surveillance systems seeking to uncover evidence of fall-related information should not be constrained to a small subset of document types, considering the potential difficulty in selecting document types a priori. In addition, such systems must be capable of handling a large heterogeneous collection of document types, while still retaining acceptable performance.

In addition to the STM models being able to account for content variation within document types at a particular site, we also assessed the generalizability of the STM models across sites. The goal was to provide an indication of how well a STM-based fall surveillance system would perform outside of the facility in which it was developed. Overall, our results were promising, with AUC scores in the 0.95 range for the three additional sites. However, because our three test sites were all located within the same region, additional work is needed to determine how well our STM models would perform in a nationwide system. An interesting aspect of our dataset is that one of the sites was

Research and applications

located in an area that is traditionally bilingual (Spanish/English). The STM models at this site still performed on par with the other sites, even though the annotators found distinctly different grammatical patterns used by clinicians, highlighting an advantage of STM and the 'bag of words' approach to classification. However, a disadvantage of the 'bag of words' approach is the loss of context surrounding a word or phrase. Therefore, the use of natural language processing (NLP) in conjunction with STM may be useful in reducing errors found within the context-dependent error categories uncovered during our error analysis (eg, HISTORY, NEGATION, SEMANTIC). Given the presence of templates and a general lack of grammatically correct sentences in the documents, hybrid NLP-STM systems should be evaluated carefully because the use of NLP may also introduce additional errors into the process.

A limitation of STM, and all techniques relying on documentation in the EHR, is that only reported and documented falls can be identified. Therefore, STM is not able to find falls when the patient does not mention or recall a fall occurring or the clinician does not document the fall. Several authors have suggested lack of clinician inquiry or patient recall impacts the identification of a previous fall as a risk factor.^{43 44} The design of this study probably limits this concern in our data because inclusion was based on seeking care for an injury. Typically, a fall was documented as part of the 'chief complaint' or 'history of present illness' section within a document. This represents a relatively straightforward observation, relevant to the primary diagnoses related to the ambulatory care visit. Therefore, it is likely that clinicians would remember to include this information in the text even if they are unaware of or forget to include specialized ICD-9-CM codes designed to document the occurrence of a fall (E-880-889). This study supports this contention, as over 25% of the documents classified as positive for a fall in the test datasets were from patients with no recorded fall E-code. We believe STM represents an opportunity to supplement information available from the EHR to clinicians and policy makers on this important healthcare issue.

CONCLUSION

This study demonstrated that STM could reliably identify falls in clinical text related to ambulatory events. Using a dataset of over 26 000 documents, the STM models were able to obtain AUC scores above 0.95 across four sites and over 600 distinct document titles. The results of this study suggest that STM-based fall models have the potential to improve surveillance of falls. For instance, patients could be flagged in real time as being at a higher risk of a fall given a history of falls documented in their EHR. In addition, results can be rolled up to the patient level of analysis to obtain regional or national statistics of fall prevalence for safety reporting measures. Finally, STM may also be useful in identifying other fall-related issues that may be buried within text, such as the place the fall occurred (for reimbursement purposes) or the type of injury sustained. Finally, the encouraging evidence shown here that STM is a robust technique for mining clinical documents bodes well for other surveillance-related topics.

Acknowledgements The authors would like to thank Bridget Hahm, Philip R Foulis, Gail Powell-Cope, Ronald I Shorr, Keryl Motta Valencia, and Blesila R Vasquez for their assistance with the completion of this study.

Funding Funding for this work was provided by the Veterans Healthcare Administration Health Services Research & Development grants IIR05-120-3 and SDR HIR 09-002. The views expressed in this paper are those of the authors and do not necessarily reflect the position or policy of the Department of Veterans Affairs or the USA government.

Competing interests None.

Ethics approval Ethics approval for this study was granted by the institutional review boards at all participating VHA hospitals.

Provenance and peer review Not commissioned; externally peer reviewed.

REFERENCES

- Hausdorff JM, Rios DA, Edelberg HK. Gait variability and fall risk in community-living older adults: a 1-year prospective study. *Arch Phys Med Rehabil* 2001;82:1050-6.
- Alamgir H, Muazzam S, Nasrullah M. Unintentional falls mortality among elderly in the United States: time for action. *Injury* Published Online First: 20 January 2012. doi:10.1016/j.injury.2011.12.001
- Stevens JA, Corso PS, Finkelstein EA, et al. The costs of fatal and non-fatal falls among older adults. *Inj Prev* 2006;12:290-5.
- Hartholt KA, Stevens JA, Polinder S, et al. Increase in fall-related hospitalizations in the United States, 2001-2008. *J Trauma* 2011;71:255-8.
- Betz ME, Li G. Epidemiologic patterns of injuries treated in ambulatory care settings. *Ann Emerg Med* 2005;46:544-51.
- Ganz DA, Bao Y, Shekelle PG, et al. Will my patient fall? *J Am Med Inform Assoc* 2007;297:77-86.
- Tremblay M, Berndt D, Luther S, et al. Identifying fall-related injuries: text mining the electronic medical record. *Info Technol Manag* 2009;10:253-65.
- Zeng QT, Redd D, Divita G, et al. Characterizing clinical text and sublanguages: a case study of the VA clinical notes. *J Health Med Informat* 2011;53.
- Sebastiani F. Machine learning in automated text categorization. *ACM Comput Surv* 2002;34:1-47.
- Donaldson I, Martin J, de Bruijn B, et al. PreBIND and Textomy—mining the biomedical literature for protein-protein interactions using a support vector machine. *BMC Bioinformatics* 2003;4:11.
- Miotto O, Tan TW, Brusci V. Supporting the curation of biological databases with reusable text mining. *Genome Inform* 2005;16:32-44.
- Wang P, Morgan AA, Zhang A, et al. Automating document classification for the Immune Epitope Database. *BMC Bioinformatics* 2007;8:269.
- Hersh WR, Bhupitiraju RT, Ross AM, et al. TREC 2004 genomics track overview. Proceedings of the 13th Text Retrieval Conference 2004; Gaithersburg, MD.
- Hersh WR, Cohen A, Yang J, et al. TREC 2005 genomics track overview. Proceedings of the 14th Text Retrieval Conference 2005; Gaithersburg, MD.
- Hersh WR, Voorhees EM. TREC genomics special issue overview. *Inf Retr Boston* 2009;12:1-15.
- Krallinger M, Morgan A, Smith L, et al. Evaluation of text-mining systems for biology: overview of the second BioCreative community challenge. *Genome Biol* 2008;9:S1.
- Krallinger M, Vazquez M, Leitner F, et al. The protein-protein interaction tasks of BioCreative III: classification/ranking of articles and linking bio-ontology concepts to full text. *BMC Bioinformatics* 2011;12:S3.
- Uzuner O, Goldstein I, Luo Y, et al. Identifying patient smoking status from medical discharge records. *J Am Med Inform Assoc* 2008;15:14-24.
- Pestian J, Matykiewicz P, Linn-Gust M, et al. Sentiment analysis of suicide notes: a shared task. *Biomed Inform Insights* 2012;5:3-16.
- Luther S, Berndt D, Finch D, et al. Using statistical text mining to supplement the development of an ontology. *J Biomed Inform* 2011;44:S86-93.
- Pakhomov SV, Hanson PL, Bjornsen SS, et al. Automatic classification of foot examination findings using clinical notes and machine learning. *J Am Med Inform Assoc* 2008;15:198-202.
- Collier N, Doan S, Kawazoe A, et al. BioCaster: detecting public health rumors with a web-based text mining system. *Bioinformatics* 2008;24:2940-1.
- Conway M, Doan S, Kawazoe A, et al. Classifying disease outbreak reports using n-grams and semantic features. *Int J Med Inform* 2009;78:e47-58.
- Botsis T, Nguyen MD, Woo EJ, et al. Text mining for the vaccine adverse event reporting system: medical text classification using informative feature selection. *J Am Med Inform Assoc* 2011;18:631-8.
- Centers for Medicare and Medicaid Services and National Center for Health Statistics. ICD-9-CM official guidelines for coding and reporting, 2011. http://www.cdc.gov/nchs/data/icd9/icd9cm_guidelines_2011.pdf (accessed 4 Sep 2012).
- World Health Organization. WHO global report on falls prevention in older age. http://www.who.int/ageing/publications/Falls_prevention7March.pdf (accessed 31 Oct 2012).
- Ogren PV. Knowtator: a protege plug-in for annotated corpus construction. Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology 2006; 273-5.
- Cohen J. A coefficient of agreement for nominal scales. *Educ Psychol Meas* 1960;20:37-46.
- McCray AT, Srinivasan S, Browne AC. Lexical methods for managing variation in biomedical terminologies. Proceedings of the 18th Annu Symp Comput Appl Med Care 1994; Washington, DC: 235-9.
- Friedman J, Hastie T, Tibshirani R. Additive logistic regression: a statistical view of boosting. *Ann Stat* 2000;28:337-407.

- 31 Witten IH, Frank E. *Data mining: practical machine learning tools and techniques*, 2nd edn. San Francisco, CA: Morgan Kaufmann Publishers Inc, 2005.
- 32 Cortes C, Vapnik V. Support-vector networks. *Mach Learn* 1995;20:273–97.
- 33 Fan RE, Chang KW, Hsieh CJ, *et al*. LIBLINEAR: a library for large linear classification. *J Mach Learn Res* 2008;9:1871–4.
- 34 Salton G, Buckley C. Term-weighting approaches in automatic text retrieval. *Inf Process Manag* 1988;24:513–23.
- 35 Deerwester S, Dumais ST, Furnas GW, *et al*. Indexing by latent semantic analysis. *J Am Soc Inf Sci* 1990;41:391–407.
- 36 Bennett KP, Campbell C. Support vector machines: hype or hallelujah? *SIGKDD Explor* 2000;2:1–13.
- 37 Mierswa I, Wurst M, Klinkenberg R, *et al*. YALE: rapid prototyping for complex data mining tasks. Proceedings of the 12th Conference on KDD 2006; Philadelphia, PA: 935–40.
- 38 Domingos P. MetaCost: a general method for making classifiers cost-sensitive. Proceedings of the 5th Conference on KDD 1999; San Diego, CA: 155–64.
- 39 Breiman L. Bagging predictors. *Mach Learn* 1996;24:123–40.
- 40 Forman G, Scholz M. Apples-to-apples in cross-validation studies. *SIGKDD Explor* 2010;12:49–57.
- 41 Hendrich AL, Bender PS, Nyhuis A. Validation of the Hendrich II Fall Risk Model: a large concurrent case/control study of hospitalized patients. *Appl Nurs Res* 2003;16:9–21.
- 42 Morse JM. *Preventing patient falls: establishing a fall intervention program*, 2nd edn. New York, NY: Springer Publishing Company, 2008.
- 43 Masud T, Morris RO. Epidemiology of falls. *Age Ageing* 2001;30-S4:3–7.
- 44 Chang JT, Ganz DA. Quality indicators for falls and mobility problems in vulnerable elders. *J Am Geriatr Soc* 2007;55:S327–34.



Finding falls in ambulatory care clinical documents using statistical text mining

James A McCart, Donald J Berndt, Jay Jarman, et al.

J Am Med Inform Assoc published online December 15, 2012

doi: 10.1136/amiajnl-2012-001334

Updated information and services can be found at:

<http://jamia.bmj.com/content/early/2012/12/14/amiajnl-2012-001334.full.html>

These include:

References

This article cites 32 articles, 5 of which can be accessed free at:

<http://jamia.bmj.com/content/early/2012/12/14/amiajnl-2012-001334.full.html#ref-list-1>

P<P

Published online December 15, 2012 in advance of the print journal.

Email alerting service

Receive free email alerts when new articles cite this article. Sign up in the box at the top right corner of the online article.

Notes

Advance online articles have been peer reviewed, accepted for publication, edited and typeset, but have not yet appeared in the paper journal. Advance online articles are citable and establish publication priority; they are indexed by PubMed from initial publication. Citations to Advance online articles must include the digital object identifier (DOIs) and date of initial publication.

To request permissions go to:

<http://group.bmj.com/group/rights-licensing/permissions>

To order reprints go to:

<http://journals.bmj.com/cgi/reprintform>

To subscribe to BMJ go to:

<http://group.bmj.com/subscribe/>