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

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 Atrain) and then applied to the remaining unseen documents (datasets Atest–D).

Results All three STM models obtained area under the receiver operating characteristic curve (AUC) scores above 0.950 on the four test datasets (Atest–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 included in 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.
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 integrators for integrating biology and the bedside (2b2 shared task)18 and determining sentiment found in suicide notes (fifth 2b2/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–E889) listed as the primary reason for the visit. The cohort was further reduced by eliminating records for poisonings (ICD-9-CM 960–979), 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).23

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 κ28 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 Atrain), while the remainder were held out as a test set (dataset Atest). 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 Atrain dataset using two machine learning algorithms: logistic regression (LR) using LogitBoost30 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: χ2, 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

<table>
<thead>
<tr>
<th>Algorithm</th>
<th>Parameters</th>
</tr>
</thead>
<tbody>
<tr>
<td>Logistic regression</td>
<td>GR, log OR, $\chi^2$</td>
</tr>
<tr>
<td>Top n terms</td>
<td>0, 50, 100, 250</td>
</tr>
<tr>
<td>SVD dimensions</td>
<td>0, 50, 100, 200</td>
</tr>
<tr>
<td>Support vector machines</td>
<td>GR, log OR, $\chi^2$</td>
</tr>
<tr>
<td>Top n terms</td>
<td>0, 50, 100, 250, 500</td>
</tr>
<tr>
<td>SVD dimensions</td>
<td>0, 50, 100, 200</td>
</tr>
<tr>
<td>C</td>
<td>$10^{-4}$, $10^{-3}$, ... , $10^0$, $10^1$, $10^4$</td>
</tr>
</tbody>
</table>

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. 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 RapidMiner (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, a cost-sensitive classification approach, was thus explored to minimize false negatives (FN) and hence increase sensitivity. MetaCost uses a form of bagging to re-label the training data with the estimated optimal class and then learns a single cost-sensitive model from the re-labeled data.

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 $\kappa$ 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 patients. The total number of terms is approximately 120 000. 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).

Table 2: Dataset descriptive statistics

| Dataset | Patients | Documents | | |
|---------|----------|-----------|----------|----------|----------|----------|
|         |          | Fall      | Not fall | Total    | Fall     | Not fall |
| Site A  | 841      | 1210 (18.2%) | 5431 (81.8%) | 6641 | 111 (43.5%) | 241 (94.5%) | 255 |
| Site B  | 841      | 1845 (19.7%) | 2327 (80.3%) | 2846 | 78 (40.6%) | 182 (94.5%) | 192 |
| Site C  | B        | 200       | 557 (14.8%) | 3214 (85.2%) | 3771 | 99 (45.2%) | 216 (98.5%) | 219 |
| Site D  | D        | 359       | 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: n=20%; site D: n=80%).
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 $\chi^2$ 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 1](image1.png) Ten document types with the most documents labeled ‘fall’. ED, emergency department; E&M

![Figure 2](image2.png) Cost-sensitive classification on $A_{train}$ (support vector machine cost sensitive; SVM-cost). FN, false negative; FP, false positive. Access the article online to view this figure in colour.

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 recommended 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

Table 3  Modeling results

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

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.

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. Access the article online to view this figure in colour.
in AUC between the LR, SVM, and SVM-cost models on any dataset was only 0.02 (dataset B). Compared to the Atrain dataset, all three models performed quite similarly on the Atest 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 Atrain 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 Atest 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.
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 classification 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 model or Morse fall scale, 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 ‘letting themselves down’ or ‘sliding 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

### Table 4 Misclassification categories

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

**Misclassification categories:**
- **UNSURE**
- **FALL RISK**
- **TEMPLATE**
- **JUDGMENT**
- **FALL-RELATED INFORMATION**
- **HISTORY**
- **INCORRECT ANNOTATIONS**
- **MOTOR VEHICLE ACCIDENT**
- **NEGATION**
- **SEMANTIC**
- **OTHER**
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. 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 Hahn, Philip R Fouls, 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 IRRS-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

Research and applications