1-1-2005

Utilizing Text Mining Techniques to Identify Fall Related Injuries

Monica Chiarini-Tremblay
University of South Florida, mtrembla@coba.usf.edu

Donald J. Berndt
University of South Florida, dberndt@coba.usf.edu

Philip Foulis
University of South Florida, philip.foulis@med.va.gov

Stephen Luther
University of South Florida, sluther@hsc.usf.edu

Follow this and additional works at: http://aisel.aisnet.org/amcis2005

Recommended Citation
Chiarini-Tremblay, Monica; Berndt, Donald J.; Foulis, Philip; and Luther, Stephen, "Utilizing Text Mining Techniques to Identify Fall Related Injuries" (2005). AMCIS 2005 Proceedings. Paper 109.
http://aisel.aisnet.org/amcis2005/109

This material is brought to you by the Americas Conference on Information Systems (AMCIS) at AIS Electronic Library (AISeL). It has been accepted for inclusion in AMCIS 2005 Proceedings by an authorized administrator of AIS Electronic Library (AISeL). For more information, please contact elibrary@aisnet.org.
Utilizing Text Mining Techniques to Identify Fall Related Injuries

Monica Chiarini-Tremblay  
Information Systems and Decision Sciences  
University of South Florida  
m trembla@coba.usf.edu

Donald J Berndt  
Information Systems and Decision Sciences  
University of South Florida  
dberndt@coba.usf.edu

Philip Foulis  
James A Haley VA Hospital/College of Medicine  
University of South Florida  
philip.foulis@med.va.gov

Stephen Luther  
College of Public Health  
University of South Florida  
sluther@hsc.usf.edu

ABSTRACT

This study focuses on investigating the computerized medical record, including textual progress notes, using data and text mining techniques to examine patient fall-related injuries (FRIs) in the Veterans Administration (VA) ambulatory care setting. FRIs are high cost, high volume adverse events in the VA that are difficult to identify from VA administrative databases. Recognizing patterns in progress notes can aid in identifying those records that should have been coded as FRIs in the administrative data. This facilitates understanding the frequency and nature of fall related injuries for implementation of prevention programs at the VA. Latent semantic indexing is used to create structured data from large textual fields found in the medical records. Unsupervised learning (cluster analysis) is used to assess the potential predictive power of the textual descriptions. Two predictive data mining approaches are then used, in combination with supervised text mining, to classify patient records as fall-related injuries.

Keywords

data mining, text mining, neural networks, decision trees, cluster analysis, knowledge discovery, decision support systems, latent semantic indexing, healthcare informatics

INTRODUCTION

Unintentional injury due to falls is a serious health problem among elderly people in the United States, with fall-related injuries incurring an estimated 20.2 billion dollars in healthcare expenses alone in 1994 (Nevitt et al. 1991; Nevitt et al. 1989; Rubenstein et al. 2002; Tinetti et al. 1998). In 1999, approximately 38% of male veterans were age 65 and over (compared to 13% for males in the United States) (www.cdc.gov/ncipc/factsheets/fallcost.htm 2003). The number of “oldest old” veterans (age 85 and over) who are at highest risk of suffering injurious falls, is projected to increase dramatically from 154,000 in 1990 to 1.3 million in 2010 (Klein et al. 2000). Due to the high incidence of falls among those aged 65 and over, it is likely that the treatment of fall related injuries represents a large volume of service in the Veterans Administration (VA) healthcare system.

Given the aging population served by the VA healthcare system, and the high rate of unintentional injurious falls among older adults, accurate information about the incidence, prevalence and epidemiology of injurious falls is essential to clinicians, researchers and policy makers (Luther et al. 2005). E-codes, or “event codes” should provide an efficient way to identify fall-related injuries (FRIs) from the administrative data, however, E-codes are known to be severely under utilized by providers and coders. The written medical record presumably contains references to falls, however, these data are not easily searched for clinical and research purposes. For the VA databases to be fully utilized to study fall-related injuries, new techniques need to be validated to more completely identify these injuries.

Text mining is an emerging technology characterized by a set of techniques and tools which allow for the extraction of structured information from text (Feldman et al. 1995; Loh et al. 2003). This research focuses on using text mining techniques to extract knowledge that can then be used to better classify those records that were not properly assigned E-
codes. The results could be embedded in decision support systems, prompting clinicians to assign the correct E-code based on the progress notes just entered, or used to improve the coding in existing electronic medical records. Correctly coded data can then aid the VA in identifying the frequency and nature of fall related injuries in order to organize and implement prevention strategies. The technique we utilize for text mining is latent semantic indexing (LSI) (Deerwester 1990).

The paper is organized as follows. The Knowledge Discovery Approach section outlines the general data mining process and provides a description of how data was staged into a database that integrated the relevant data from administrative sources and the electronic medical records for all veterans treated for injuries. We then describe an initial strategy for identifying information from the record related specifically to the treatment of the indexed injury “episode of care”; with a goal of approximating which records were to be combined for a patient in order to accurately illustrate a treatment cycle for the injury. The Text Mining section describes the initial results of mining the progress notes, with some examples of the discovered clusters, stop lists and concepts formed. Two approaches are taken: unsupervised learning and supervised learning. Unsupervised learning investigates clusters created from the terms (using an entropy weighting scheme) extracted from the progress notes to assess the potential predictive power of these textual descriptions in identifying FRIs. The supervised learning approach attempts to form improved clusters by using a dataset that includes class labels for FRI and non-FRI cases, based on an information gain weighting scheme. The Predictive Model section describes the preliminary results obtained from combining general administrative data and text mining clusters (derived from the progress notes) to the classification of FRI and non-FRI cases. Two predictive data mining approaches based on an artificial neural network and a decision tree induction algorithm are described. We conclude with closing remarks and a description of the future steps for this ongoing study.

KNOWLEDGE DISCOVERY APPROACH

Fayyad and his co-authors outline the process of knowledge discovery: learning the application domain, creating a target dataset, data cleaning and preprocessing, data reduction and projection, choosing the function of data, choosing the data mining algorithm, model interpretation and using discovered knowledge (Fayyad et al. 1996). The amount of time and effort required for each of these steps is not evenly distributed; the initial steps that focus on data cleansing and preprocessing almost always account for a large share of the effort. As Fayyad points out “In practice, a large portion of the applications effort can go into properly formulating the problem (asking the right question) rather than optimizing the algorithmic details of a particular data mining method.”(Fayyad et al. 1996). In this study there was a large amount of available data, which is normally an ideal condition, but one which can introduce many challenges. The processes of understanding which data were important, what techniques to use to combine data from two very different sources, and understanding how to manipulate and process very large text fields (such as progress notes) were difficult challenges that are still being addressed. In addition, the availability of many rich textual fields in the VA data made it crucial to fully understand how to best extract keywords and concepts with the purpose of building predictive models.

Data Staging/Preprocessing/Transformation

Two databases were utilized, the VA Ambulatory Events Database and the VA electronic medical record. The VA Ambulatory Events Database captures ambulatory encounters within hospital outpatient clinics as well as smaller satellite facilities and contains information on diagnosis, procedure, type of clinic visited, and demographic characteristics of patients. Each encounter (i.e., an interaction with a clinical department/provider) produces a separate record in the event database, which is linked to an individual patient by a scrambled identifier. All records for patients with injuries identified as the primary diagnosis code from the James A. Haley Veterans Hospital Corporate facilities for FY 2001-2003 were extracted. In addition a dataset containing text-based clinical information from the VA electronic medical record system (CPRS) was obtained for all unique patients identified as being treated for injuries.

The first step in preprocessing involved extracting this data and loading it into a relational database. In order to link the Ambulatory Events data to the text based medical records a patient identifier and date were used. Multiple encounters may appear for each patient in a selected time period. Thus, a patient could have a series of progress notes for events that were in fact related, describing the normal process of a patient proceeding through the outpatient facility. Deciding how to group these records, creating an “episode of care” is crucial. A brute force approach is to read each record and to manually combine these records to provide accurate representations with which to train a model. The obvious problem with this approach is that it is extremely time consuming and would not be feasible in regional or national data. Another approach is to create a “sliding window”, allowing for some overlap and deciding on a time frame in order to automate this process. Other approaches include clustering on other data, essentially creating “types” of progress notes, and mining them as separate inputs. An example of this is to cluster based on a field called “progress note title” which essentially outlines the source of
the progress note (for example “Ambulatory Care” or “Nursing”). While we intend to pursue more promising strategies, our initial model used a simple approach of defining the episode of care as all the notes and data collected during a day.

For the current study “unique falls” were determined by examining the encounter dates for each patient. The International Classification of Disease, 9th Revision, Clinical Modification, (ICD-9-CM) “external causes of injury” codes E800-E999 (E-codes) were used to identify fall-related encounters. E-codes are designed to permit the classifications of environmental events, circumstances, and conditions as the cause of injury, poisoning and other adverse event. Specifically E-codes (E880-E888) signify “fall-related injuries” due to slips, trips, or falls unrelated to transportation. Records were included in the analysis if an E-code appeared in any of the 10 diagnosis fields included in the database (one primary diagnosis and nine secondary diagnosis fields) (Luther et al. 2005). The challenge encountered in this step was that though those records that contained fall related ICD-9 codes were almost always correct, some of those that did not contain fall-related ICD-9 codes were in-fact miscoded (or rather un-coded) and should have been identified as falls. As described above, E-codes are often under utilized in the absence of clear directions and incentives for their use. The unidentified FRIs introduce some concerns when training a predictive model, which are further discussed in the results section of this paper.

TEXT MINING

Most text mining approaches “count” occurrences of words in documents. To simplify this, most algorithms generally remove specified words (stop list) or keep specified words (start list). Words that have a common root are stemmed, and common words are removed since they have little power in discriminating documents (Woodfield 2003). A term-by-document frequency matrix is built and serves as the foundation for analysis of the document collection. To improve performance entries can be adjusted by utilizing weighting functions for certain words (e.g., infrequent words may be weighed more heavily, or words that are highly correlated to a target variable). This matrix can grow quite large. Additionally, most of these terms are not used in all the records (thus many columns will have a value of 0), requiring a large amount of computing resources to analyze. Thus, it becomes necessary to reduce the dimensionality of this matrix. Latent semantic indexing (LSI) reduces dimensionality by using singular value decomposition (SVD). LSI is a technique that transforms the large matrix into a much lower dimensional form (Berry et al. 1999; Deerwester 1990). SVD allows the arrangement of the space to reflect the major associative patterns in the data, and ignore the smaller, less important influences. Singular value decomposition is closely related to eigenvector decomposition. Similar to factor analysis, the frequency matrix is decomposed into eigenvalues and eigenvectors that create linearly independent components of the data. The smaller components can be ignored and the similarity between two documents can be determined by the values of the remaining factors. The result can be represented geometrically by a spatial configuration in which the dot product or cosine between vectors representing two documents corresponds to their estimated similarity (Deerwester 1990). For more detailed coverage of these issues, see (Berry et al. 1999; Deerwester 1990).

These algorithms were applied to the progress notes found in the medical record using SAS Enterprise Miner. This tool automates stemming of terms (for example BIG; BIGGER, BIGGEST), as well as providing initial synonym lists. It also has an initial stop list. These synonym and stop lists can then modified by applying domain knowledge. The lists are examined, and terms that have the same meaning are combined, removing words from the stop list (e.g., the word outpatient appears often because these are outpatient records and is not a necessary stop word). A term by document frequency matrix is created, with the row dimension of the matrix limited to the 100 most frequent terms.

An important decision is selecting a weighting scheme that can help emphasize discrimination between document groups (Woodfield 2003). The weight of each entry is determined by the frequency weight and term weight as follows: \( a_{ij} = L_{ij} G_{ij} \), where \( L_{ij} \) is the frequency weight and \( G_{ij} \) is the term weight. For this first experiment, log frequency weighting was used where \( L_{ij} = \log_2(a_{ij} + 1) \), and \( a_{ij} \) is the frequency with which term \( i \) appears in document \( j \). Several term importance weightings exist. Research has shown that the best results are obtained using entropy or inverse document frequency. The weightings can also be target-based. Target-based formulas require the specification of a target (in our case whether a given record had a fall-related injury E-Code or not). For the unsupervised task of cluster analysis, entropy was used. This is described with more detail in the next section. Since the final goal was to predict a fall related injury, several term weightings were tried, with the best results obtained by using information gain. Information gain indicates how well the term, or the absence of that term, predicts the category (calculating reduction in entropy) (Woodfield 2003). Entries in the term-document frequency matrix are weighted and then used as input for the singular value decomposition. Table 1 lists some of the selected terms sorted by weight. As expected, the term “fall” has a large weight of 1.0, since all progress notes were in-fact miscoded (or rather un-coded) and should have been identified as falls. As described above, E-codes are often under utilized in the absence of clear directions and incentives for their use. The unidentified FRIs introduce some concerns when training a predictive model, which are further discussed in the results section of this paper.
for fall related injuries contained the verb “fall”. Also the noun form of the word “fall” received a heavy weight. The resulting term by document weighted matrix is added to the dataset and is utilized for predicting fall related injuries from the progress notes.

<table>
<thead>
<tr>
<th>Term</th>
<th>Freq</th>
<th># Documents</th>
<th>Keep</th>
<th>Weight</th>
<th>Role</th>
<th>Attribute</th>
</tr>
</thead>
<tbody>
<tr>
<td>+ fall</td>
<td>880</td>
<td>682</td>
<td>Y</td>
<td>1.000</td>
<td>Verb</td>
<td>Alpha</td>
</tr>
<tr>
<td>+ fall</td>
<td>550</td>
<td>394</td>
<td>Y</td>
<td>0.460</td>
<td>Noun</td>
<td>Alpha</td>
</tr>
<tr>
<td>xray</td>
<td>202</td>
<td>152</td>
<td>Y</td>
<td>0.251</td>
<td>Prop</td>
<td>Alpha</td>
</tr>
<tr>
<td>+ abrasion</td>
<td>176</td>
<td>113</td>
<td>Y</td>
<td>0.183</td>
<td>Noun</td>
<td>Alpha</td>
</tr>
<tr>
<td>+ month</td>
<td>354</td>
<td>305</td>
<td>Y</td>
<td>0.170</td>
<td>Noun</td>
<td>Alpha</td>
</tr>
<tr>
<td>+ he</td>
<td>4503</td>
<td>1132</td>
<td>Y</td>
<td>0.161</td>
<td>Pron</td>
<td>Alpha</td>
</tr>
<tr>
<td>+ injury</td>
<td>328</td>
<td>268</td>
<td>Y</td>
<td>0.161</td>
<td>Noun</td>
<td>Alpha</td>
</tr>
<tr>
<td>+ continue</td>
<td>467</td>
<td>373</td>
<td>Y</td>
<td>0.141</td>
<td>Verb</td>
<td>Alpha</td>
</tr>
</tbody>
</table>

Table 1 – Partial Example of Frequency Matrix with Weights using Information Gain

Unsupervised Learning – Cluster Analysis

Clustering is a form of unsupervised learning. In unsupervised learning one does not rely on predefined events, with assigned class labels, as a training example. Thus, learning is achieved by observation rather then by example. Clustering is the process of grouping the data into classes or clusters so that objects within a cluster have high similarity in comparison with one another, but are dissimilar to objects in other clusters (Han et al. 2001). Clustering classifies the data into groups based on measures of distance or similarity. This technique is utilized because it does not take into account a target variable, such as the classification of FRI or non-FRI case. The goal of this exploratory step is to see if it is possible to identify clusters of terms that are indicative of a FRI, providing evidence that the progress note text will have some predictive power. The clusters are based on the SVD dimensions calculated with using the entropy-based weighting scheme. Entropy is a concept from communication theory (Shannon 1948) and is a measure of information content (i.e., disorder). Entropy gives high weights to terms that are infrequent in all the data, but frequent in a few documents (Woodfield 2003). Table 2 shows how using entropy for unsupervised clustering calculates different weights then those shown in Table 1, where information gain was used.

<table>
<thead>
<tr>
<th>Term</th>
<th>Freq</th>
<th># Documents</th>
<th>Keep</th>
<th>Weight</th>
<th>Role</th>
<th>Attribute</th>
</tr>
</thead>
<tbody>
<tr>
<td>+ fall</td>
<td>646</td>
<td>480</td>
<td>Y</td>
<td>1.000</td>
<td>Verb</td>
<td>Alpha</td>
</tr>
<tr>
<td>+ fall</td>
<td>402</td>
<td>272</td>
<td>Y</td>
<td>0.479</td>
<td>Noun</td>
<td>Alpha</td>
</tr>
<tr>
<td>+ treatment</td>
<td>115</td>
<td>90</td>
<td>Y</td>
<td>0.282</td>
<td>Noun</td>
<td>Alpha</td>
</tr>
<tr>
<td>+ injury</td>
<td>182</td>
<td>151</td>
<td>Y</td>
<td>0.253</td>
<td>Noun</td>
<td>Alpha</td>
</tr>
<tr>
<td>+ abrasion</td>
<td>149</td>
<td>84</td>
<td>Y</td>
<td>0.214</td>
<td>Noun</td>
<td>Alpha</td>
</tr>
<tr>
<td>xray</td>
<td>129</td>
<td>90</td>
<td>Y</td>
<td>0.183</td>
<td>Prop</td>
<td>Alpha</td>
</tr>
<tr>
<td>reason</td>
<td>42</td>
<td>42</td>
<td>Y</td>
<td>0.159</td>
<td>Prop</td>
<td>Alpha</td>
</tr>
<tr>
<td>reactions</td>
<td>123</td>
<td>123</td>
<td>Y</td>
<td>0.154</td>
<td>Prop</td>
<td>Alpha</td>
</tr>
<tr>
<td>+ continue</td>
<td>214</td>
<td>168</td>
<td>Y</td>
<td>0.142</td>
<td>Verb</td>
<td>Alpha</td>
</tr>
</tbody>
</table>

Table 2 – Partial Example of Frequency Table with Weights using Entropy
Table 3 shows the clusters identified using k-means clustering. The columns Flag=Y and Flag=N categorize what percentage of the records in this clusters were or were not FRIs. Clusters 2, 3, 4, and 7 are of interest because the terms seem indicative of a FRI and contain a large proportion of FRI records. Another interesting cluster is number 6, which contains 57% FRIs and also has terms indicative of a FRI. This gives an indication that progress notes can be utilized to predict FRIs.

<table>
<thead>
<tr>
<th>CLUS</th>
<th>Flag=N</th>
<th>Flag=Y</th>
<th>Descriptive Terms</th>
<th>Freq</th>
<th>Percent</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>70%</td>
<td>30%</td>
<td>+ year, alcohol, + he, past, clinical, screen, during, some, + will, + referral, + make, mental, regular, + complete, + call, + symptom, + comment</td>
<td>319</td>
<td>25%</td>
</tr>
<tr>
<td>2</td>
<td>19%</td>
<td>81%</td>
<td>+ knee, + abrasion, + pain, + fall, + swell, + he, + leg, + walker, + fall, + injure, some, examination, regular, + injury</td>
<td>113</td>
<td>9%</td>
</tr>
<tr>
<td>3</td>
<td>16%</td>
<td>84%</td>
<td>+ wrist, + hand, + swell, + fall, floor, + fall, + pain, + call, some, + continue, + he, + injure, xray, + abrasion, examination, + loss, + orient, + injury, + note, meds</td>
<td>64</td>
<td>5%</td>
</tr>
<tr>
<td>4</td>
<td>25%</td>
<td>75%</td>
<td>male, past, reactions, xray, + vital, water, hydrocortisone, + list, active, adverse, supplies, vital, meds, medications, + tablet, pain, back, + injury, + abrasion, + swell</td>
<td>286</td>
<td>22%</td>
</tr>
<tr>
<td>5</td>
<td>88%</td>
<td>13%</td>
<td>+ referral, education, + make, + comment, education, clinical, + instruct, + walker, + month, during, understanding, + note</td>
<td>40</td>
<td>3%</td>
</tr>
<tr>
<td>6</td>
<td>43%</td>
<td>57%</td>
<td>+ artery, coronary, coronary artery disease, + movement, regular, examination, + month, floor, past, alcohol, + loss, + orient, some, + walker, + abrasion, + symptom, + injure, + tablet, + he, + hand</td>
<td>23</td>
<td>2%</td>
</tr>
<tr>
<td>7</td>
<td>25%</td>
<td>75%</td>
<td>+ loss, + orient, + fall, + leg, + walker, + movement, regular, floor, + tablet, + injury, + pain, + fall, + injure, back, + complete, + he, mental, + symptom, + referral, + instruct</td>
<td>153</td>
<td>12%</td>
</tr>
<tr>
<td>8</td>
<td>71%</td>
<td>29%</td>
<td>understanding, + instruct, + call, + symptom, + will, + he, + make, education, during, clinical, + injure, regular, + swell, + complete</td>
<td>107</td>
<td>8%</td>
</tr>
</tbody>
</table>

Table 3 - Clusters Identified from Terms in the Progress Notes

**PREDICTIVE MODELS**

Preliminary predictive models were created using a combination of general administrative data and text mining clusters, as described above. In these models, approximately 50 attributes are available: age, gender, ethnicity, race, as well as the SVD terms created using supervised text clustering based on information gain. The data available from the VA has many other fields that can be used for predictive modeling. Future work will include more complete attribute selection from the abundant structured and unstructured data available. Also, more precise “episode of care” groupings of the progress notes would undoubtedly enhance the performance of the predictive models. Thus, the results of these current models are preliminary and should improve as our data preprocessing techniques evolve. Several predictive models were created; two are outlined in this paper: a neural network model and a decision tree model. The data utilized contained 648 non-FRI records and 648 FRI records. This is a higher percentage of FRI than would occur naturally in the database. This over sampling was done to “boost” the percentage of FRI records for model construction, thereby more adequately representing the somewhat rare FRI event. The data was divided into training (40%), validation (30%) and test (30%) datasets.

**Neural Networks**

An artificial neural network (NN) is an information-processing paradigm that is inspired by the way biological nervous systems, such as the brain, process information. Artificial neural networks, like people, learn by example. They are trained for a specific application, such as pattern recognition or data classification. Neural networks are used to extract patterns and detect complex trends. While these tools owe their origin to research in artificial intelligence (AI), they are now looked upon as nonlinear statistical tools that act as “universal function approximators” (UFA) and tend to converge faster than other
commonly used UFAs, like linear sums of nonlinear functions (Bishop 1995; Han et al. 2001; Hertz et al. 1991). The most commonly used NN architecture is the multi-layer perceptron (MLP), which is a special type of feed-forward network. A MLP is composed of an input layer, a hidden layer composed of hidden units, and an output layer (Walsh 2002). An MLP was used as an initial predictive model, feeding the structured data and the clusters identified with the text mining algorithm. The results for predictions made on the test dataset are shown in Table 4. The model did well in predicting a FRI with an 88% hit rate, but of some concern is the high false positive rate (23%).

<table>
<thead>
<tr>
<th>Actual/Predicted</th>
<th>N</th>
<th>Y</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>N</td>
<td>149</td>
<td>45</td>
<td>194</td>
</tr>
<tr>
<td>Y</td>
<td>23</td>
<td>171</td>
<td>194</td>
</tr>
<tr>
<td>Total</td>
<td>172</td>
<td>216</td>
<td>388</td>
</tr>
</tbody>
</table>

Sensitivity (Hit Rate) 88% - Specificity 77% - False Positive Rate 23% - False Negative Rate 12%

Table 4 - Confusion Matrix for Neural Network

**Decision Trees**

Decision trees are another widely used data mining technique, with many different algorithms available for tree construction. Among the most commonly used approaches are CHAID (chi-squared automatic interaction detection), C4.5/5.0, and CART (classification and regression trees) (Han et al. 2001). There are two main variations: regression trees which estimate the numeric value of a target variable, and classification trees that assign class labels to cases. In this experiment, cases are coded as FRI or non-FRI, a binary classification problem. As with artificial neural networks, decision tree induction is a supervised learning technique that requires a training data set annotated with correct class labels. As noted above, the data set being used in this experiment was created by over sampling FRI cases to create a fairly balanced training set with both FRI and non-FRI cases. In addition, the data set certainly contains some coding errors, especially since E-codes are known to be under utilized, which would lead to un-coded falls in the non-FRI cases.

Decision trees have many advantages, including the interpretability of the results that may be easily represented as a set of explicit rules. Decision trees also handle many different types of inputs, missing values, and are fairly robust in the face of outliers. Ultimately, the fitted model partitions the space into bins or box-shaped regions (using discontinuous planes), limiting the ability to easily model continuous functions or surfaces. However, classification problems, such as FRI/non-FRI labeling, may sometimes be modeled very effectively by decision trees. Therefore, a decision tree induction approach based on the CART algorithm was employed as an alternative prediction strategy. The data contained structured data as well as the SVDs derived in the text mining exercise. The results were promising, improving on the initial neural network results with a substantial reduction in the number of false positives. This provides some indication that this simple binary classification task may be more appropriate for a decision tree model, while a neural network approach may prove useful for other planned tasks.

<table>
<thead>
<tr>
<th>Actual/Predicted</th>
<th>N</th>
<th>Y</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>N</td>
<td>183</td>
<td>11</td>
<td>194</td>
</tr>
<tr>
<td>Y</td>
<td>28</td>
<td>166</td>
<td>194</td>
</tr>
<tr>
<td>Total</td>
<td>211</td>
<td>177</td>
<td>388</td>
</tr>
</tbody>
</table>

Sensitivity (Hit Rate) 86% - Specificity 94% - False Positive Rate 6% - False Negative Rate 14%

Table 5 - Confusion Matrix for Decision Tree

**Exploring the False Positives**

The false positives were explored, since as previously described, some of the FRI records are actually miscoded. Table 6 shows two examples of the text associated with a false positive. The first record contains words such as hip, pain, knee that could also be found in a FRI, but it is clearly not a FRI record. The second record, however, is clearly a miscoded FRI. It is these incorrectly coded records that may cause some trouble during the model training process. Currently, we are exploring methods for creating a cleaner training set to reduce this threat.
This is a pleasant male, he has a history of some left hip pain, which was evaluated and was thought to be left trochanteric bursitis based on x-ray findings. He has no history of diabetes or hypertension, or any other chronic medical problem such as the above. The patient relates a 3 month history of some lower back pain and left hip pain, which is intermittent. It is at a point where it does go down below the level of the knee, and follows an S1 pattern. 

Patient stated she was calling to let us know she had fallen again. She denies any injury, denied needing medical evaluation. Spoke to her about assisted living. About one hour later I received a phone call from the patient’s son, XXXXNAMEXXXX. The nurse called him about the patient’s frequent falls. She has fallen twice this week. Advised son that patient has been referred to Falls Clinic for consultation 

**Table 6- False Positives**

**FUTURE DIRECTIONS**

The initial results are encouraging and demonstrate that conducting text mining on the progress notes obtained from electronic medical records will be helpful in correctly identifying FRIs. Several steps need to be taken to build more accurate training and evaluation data sets. A method has to be devised to create more realistic “episodes of care”. Several approaches are currently being investigated and discussed with domain experts at the VA. This is important because it would allow E-code assignment to the whole patient care episode, rather than to the individual progress note or an artificially restricted grouping. Typically it is the clinician, physician or nurse, at the first clinical department (e.g. Primary Care Clinic) that would describe the circumstances of how the injury occurred and assign an E-code to the record. Progress notes generated at subsequent clinics encountered throughout the patient’s visit to the facility for the injury (e.g. Radiology) likely would not describe the circumstances that lead to the injury. This fact clearly contributes to the low values for specificity described above. Once a method to identify and automate the selection of “episodes of care” is found, a clean data set needs to be created to train better predictive models. For instance, we plan to have nurses review the medical charts used for training and verify the E-codes assigned. Another area to be explored is identifying why both the neural net and decision tree models placed heavier weights on variables identified from the text mining process than those in the administrative data. Better care needs to be taken in selecting the attributes from the administrative data. However, it appears that text mining alone may be useful for identifying fall-related injuries, at least as a binary classification problem. Future research will also focus on predicting more detailed E-codes that describe the type of fall-related injury, the location of the event, and other information useful in better understanding these health challenges.

Once more research is conducted and more robust models are constructed, the resulting models can be imbedded in decision support tools, with the possibility of prompting nurses or clinicians entering data to assign a suggested E-Code based on the progress notes being written. Another possibility is to use these data mining techniques to post-process medical records and add structured codes in a fully or semi-automated manner. Correctly coded data can then aid the VA in identifying the frequency and nature of fall related injuries in order to organize and implement prevention programs. 

**REFERENCES**

17. Woodfield, T. "Text Mining Using SAS Software Course Notes."