Classification of radiology reports for falls in an HIV study cohort

Jonathan Bates¹,³, Samah J Fodeh¹, Cynthia A Brandt¹,³, Julie A Womack²,³

ABSTRACT

Objective To identify patients in a human immunodeficiency virus (HIV) study cohort who have fallen by applying supervised machine learning methods to radiology reports of the cohort.

Methods We used the Veterans Aging Cohort Study Virtual Cohort (VACS-VC), an electronic health record-based cohort of 146 530 veterans for whom radiology reports were available (N = 2,977,739). We created a reference standard of radiology reports, represented each report by a feature set of words and Unified Medical Language System concepts, and then developed several support vector machine (SVM) classifiers for falls. We compared mutual information (MI) ranking and embedded feature selection approaches. The SVM classifier with MI feature selection was chosen to classify all radiology reports in VACS-VC.

Results Our SVM classifier with MI feature selection achieved an area under the curve score of 97.04 on the test set. When applied to all the radiology reports in VACS-VC, 80,416 of these reports were classified as positive for a fall. Of these, 11,484 were associated with a fall-related external cause of injury code (E-code) and 68,932 were not, corresponding to 29,280 patients with potential fall-related injuries who could not have been found using E-codes.

Discussion Feature selection was crucial to improving the classifier’s performance. Feature selection with MI allowed us to select the number of discriminative features to use for classification, in contrast to the embedded feature selection method, in which the number of features is chosen automatically.

Conclusion Machine learning is an effective method of identifying patients who have suffered a fall. The development of this classifier supplements the clinician’s toolkit and reduces dependence on un-coded structured electronic health record data.

Keywords: information retrieval, text mining, falls, aging, HIV

INTRODUCTION

Our motivation for identifying patients with falling injuries began with a study of human immunodeficiency virus (HIV) and aging, which arose from the following considerations. There are more than 1 million individuals living with HIV in the United States, more than half of whom are over the age of 50.¹ Unadjusted fragility fracture (hip, spine, upper arm fractures) rates are higher among HIV-infected individuals, particularly as they age, than among uninfected individuals.²,³ Fragility fractures are an important public health problem because of their cost, both to the healthcare system and to individuals. It has been established that HIV-infected individuals are far more likely than their uninfected counterparts to experience decreased bone mineral density. However, in the elderly population, falls, not decreased bone mineral density, are the primary cause of unadjusted fragility fractures. Unfortunately, little is known about falls among the HIV-infected population.

Much of the work on falls has been done in cohorts recruited specifically for research purposes. The Veterans Aging Cohort Study Virtual Cohort (VACS-VC) is an electronic health record (EHR)-based cohort that presents novel challenges to exploring falls in HIV-infected veterans (Veterans Administration Institutional Review Board AJ001, Yale HC 0309025943). There are 146 530 patients in VACS-VC, for whom a total of 2,977,739 radiology reports were available. We have mined these records using machine learning methods to identify patients who have suffered a fall.

In administrative datasets, falls are typically identified by International Statistical Classification of Diseases, Ninth Edition, Clinical Modification external cause of injury codes (E-codes). E-codes are supplementary codes and, thus, can only be used in conjunction with an injury code and are not required for billing purposes. This makes E-codes an unreliable source of information in the EHR for identifying falls (ie, they have a low sensitivity). We propose to identify non-coded falls in VACS-VC by training a classifier to find evidence of a patient having fallen in their radiology reports. Although falls may be documented elsewhere in a patient’s EHR, it is natural to look in their radiology reports. Falls often result in injuries – from bumps and bruises to more serious injuries like fractures and traumatic brain injury. Radiologic studies are often ordered to rule out or evaluate for these more serious sequelae. As such, radiology reports offer a basis for our first pass at identifying the cohort of interest. Moreover, radiology reports have a more restricted vocabulary and are fewer in number relative to, for example, progress notes.

This study is significant because it identifies a targeted patient event, specifically a “fall,” from radiology reports when structured information (eg, E-codes) in the EHR is missing or insufficient.

Machine learning has been used to identify instances of patients having fallen (McCart et al.⁴) and patients who incurred fall-related injuries (Tremblay et al.⁵) in the Veterans Health Administration’s EHR. Both of these studies represent a clinical document as a “bag of words” and then apply several machine learning classifiers to label whether a given note in the clinical document indicates the occurrence of a fall. Our proposed method in this study applies machine learning to identify falls as well; however, we have additionally enriched our
feature space by matching text phrases to Metathesaurus concepts in the Unified Medical Language System (UMLS). Specifically, we represent each radiology report as a “bag of features,” including both words and UMLS concepts. Moreover, we evaluate whether applying a filter or the embedded feature selection technique enhances the performance of the classifier, and we also consider the effect of using different misclassification costs per class (“fall”/“not fall”).

BACKGROUND

McCarter et al.5 used supervised feature selection and classification to identify instances of patients having fallen. For classification, the authors evaluated logistic regression, support vector machine (SVM), and cost-sensitive SVM (SVM-cost) classifiers. All three classifiers were found to have comparable and acceptable performances, although the performance of SVM-cost was marginally better. Although the authors found the “bag of words” and text mining approach to yield acceptable results, they proposed that inputting more complex language features to the classifier may reduce the context-dependent errors found in their error analysis.

To mine instances of patient falls from EHR notes, Tremblay et al.5 employed both supervised and unsupervised learning techniques. For the former, the authors used information gain for feature selection and logistic regression for classification. For the latter, the authors applied entropy and clustering. They concluded that using supervised classification produced encouraging initial results and that unsupervised techniques alone “did not produce viable results.”5 Moreover, similar classification methods have been applied to radiology reports to identify fractures and abnormalities,7 reportable cancers,8 and acute wrist fractures.9 Shivade et al.10 reviewed automated methods for identifying cohorts with specific phenotypes. For an overview of clinical natural language processing (NLP), we refer to Demner-Fushman et al.,11 and, for an overview of some popular NLP systems, we refer to Doan et al.12

METHODS

Creating a Reference Standard

To train and test the classifier, we created a reference standard of labeled radiology reports (N = 8288). Two groups of patients were included. The first group comprised all patients with a fall-related E-code recorded in their EHR in 2009. The E-codes we included are: E817, E824, E880.X-E888.X, E927.X. We identified all radiology reports that were generated within 30 days of the E-code date. We picked 30 days because we considered this time to be enough time for the patient’s radiology reports to be recorded in their EHR. The second group comprised patients at risk for a fall but whose EHR did not include a fall-related E-code. Conditions that were associated with risk for a fall included: congestive heart failure, convulsions/seizures, impaired coordination, dementia, diabetes, gait abnormalities, peripheral neuropathy, peripheral arterial disease, syncope, traumatic brain injury, dementia/Alzheimer’s disease, and cerebrovascular disease/stroke. We randomly selected 20% of all the radiology reports of patients with these diagnoses generated in 2009 for inclusion in the reference standard. There were 168 patients in the first group and 108 patients in the second group. Once the radiology reports were identified, two domain experts labeled each report as “fall” or “not fall,” according to the contents of the report (whether or not they indicated that the patient had suffered from a fall) (Table 1). The inter-rater agreement had a Cohen’s kappa value of 0.927. The reference standard was divided into a training set and test set. To prevent information leakage, we stratified the reference standard by patient into two-thirds training patients and one-third test patients.

Applying Machine Learning

Once all the radiology reports were represented in the bag of words and concepts model, we trained a linear SVM to classify each report as either “fall” or “not fall.” From the training set, the classifier “learns” the region of feature space (ie, combinations of features) that corresponds to “fall.” The boundary of this region is known as the decision boundary. The linear SVM maps each feature vector to a decision value, which is the signed distance from the feature vector to the decision boundary. For example, decision values of 10, 1, and -1 would correspond to “fall,” less certain “fall,” and “not fall,” respectively.

Several SVM classifiers were trained for comparison. The baseline classifier was an L2-regularized soft margin SVM with one model hyper-parameter, the misclassification cost ("Cost"). We chose the “Cost” parameter that maximized the F1-score through 10 runs of 10-fold cross validation on the training set, which ranged in powers of 10 from 10^-6 to 10^6. After noting the class size imbalance in the training set (Table 1), we expected the baseline SVM to learn to favor the “not fall” label.16 Thus, to reduce the potential for bias from the training set, we also trained an SVM classifier analogously to the baseline, with the difference that each class had its own misclassification cost, inversely proportional to the class size.16,17

Feature Selection

Feature selection is a recommended step when training a classifier on data with many variables, because, for instance, doing so improves classifier performance and reduces classification time.18 We considered both variable ranking feature selection and embedded feature selection methods. Table 1: Radiology Reports Reference Standard

<table>
<thead>
<tr>
<th>Set</th>
<th>Not Fall</th>
<th>Fall</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Training</td>
<td>4953</td>
<td>396</td>
<td>5349</td>
</tr>
<tr>
<td>Test</td>
<td>2794</td>
<td>145</td>
<td>2939</td>
</tr>
<tr>
<td>Total</td>
<td>7747</td>
<td>541</td>
<td>8288</td>
</tr>
</tbody>
</table>

Bag of Words and Concepts Model

We used YTEX,13 an NLP tool built on top of Apache clinical Text Analysis and Knowledge Extraction System (cTAKES),14 to extract features from each radiology report. In particular, the lowercased output of the tokenizer and the concept unique identifiers of the named entity recognition components were combined to create a feature index, and each report was represented by a binary feature vector in this index (Figure 1). We used the out-of-the-box pipeline of annotators for YTEX (v0.8), which comprises sentence splitting, tokenization, part-of-speech tagging, shallow parsing, named entity recognition, and storage of all annotations in a database. The named entity recognition component maps spans of text to a dictionary lookup table, which must be prepopulated with UMLS concepts. The out-of-the-box dictionary lookup table is populated via inclusion of the following type unique identifiers: diseases and disorders, signs and symptoms, anatomical sites, medications and drugs, procedures, device, and laboratory.15 In addition to these identifiers, we included in the dictionary lookup table any UMLS concepts with partial matches to any of the following words/phrases: blackout, black out, blanked out, dizzy, dizziness, drop attack, fall, fell, lightheaded, loss of consciousness, LOC, orthostatic, passing out, passed out, pass out, seizure, slid, slip, stumble, syncopy, syncope, syncopal, trip, vasovagal.
selection. In the first approach, we used mutual information (MI) to rank features according to their predictive power for the target “fall” value. We selected those features with the highest MI with respect to the training labels. Specifically, using only the training set, the 100 features with the highest MI for the “fall” label and the 100 features with the highest MI for the “not fall” label were chosen. These top-ranked features were then used to train two SVM classifiers, one analogous to the baseline classifier and one with class size-adjusted misclassification costs analogous to the second classifier.

In the second approach, we used an embedded feature selection method.18 Embedded feature selection refers to an implicit mechanism in some machine learning algorithms that reduces the feature space as part of the classifier training. It is desirable to select features that complement the particular learning algorithm. In the context of linear SVMs, certain training algorithms for the SVM classifier \( f(x) = w^T x + b \) force zero entries in the weight vector \( w = (w_1, w_2, \ldots, w_n) \). Nonzero weights \( w_i \) then correspond to the selected features, hence embedded feature selection. The sparseness of the weight vector \( w \) is implicitly determined by what is known as the “regularization penalty.” Traditionally, the regularization penalty in the soft margin SVM optimization problem uses the L2 norm, which is sometimes called the “ridge penalty.”19 However, L2-regularization does not generally shrink the SVM coefficients to zero and, hence, does not have a selection mechanism. Several other regularization penalties for SVM that encourage sparsity in the SVM coefficients have been proposed. Becker et al.20 compared several such penalties, namely: Lasso (L1), smoothly clipped absolute deviation (SCAD), elastic net, and elastic SCAD. In a test of mean misclassification rates for these penalties on simulated data, SCAD-SVM had the best performance when there were few predictive variables. Because we expected the proportion of fall-related terms in the corpus vocabulary to be very small, we considered the linear SCAD-SVM classifier as an alternative to the feature selection approach. The linear SCAD-SVM classifier has one cost hyper-parameter that was chosen by cross-validation on the training set, analogously to the other SVM models.

We implemented the various classifiers using LIBSVM,17 for L2-regularized SVM, and the penalizedSVM package21 in R, for SCAD-SVM.

RESULTS

Performance on the Test Set

After training the classifiers, we evaluated their performance on the test set. The patients in the test set were distinct from the patients in the training set; thus, the test set is representative of a novel cohort. The area under the curve score, F1-score, positive predictive value, and the sensitivity of the classifiers are provided in Table 2. For these metrics, a “positive” value corresponds to a “fall.”

Classification of VACS-VC and Comparison with E-codes

After comparing the performances (Table 2) of the various SVM classifiers on the test set, we concluded that MI filtering and class size-adjusted misclassification cost should be used for the final classifier. Using all the labeled data, we trained a final linear L2-regularized SVM classifier with these options. We then ran this classifier on all of the unlabeled patient radiology reports \( (N = 2,977,739) \) in VACS-VC.

A total of 80,416 (2.7%) of the unlabeled patient radiology reports in VACS-VC were classified as positive for a fall. For comparison, we counted how many radiology reports that were positive for a fall were not generated within the 30-day window of the date that a fall-related E-code was input to the patient’s EHR that we had chosen for this analysis. Of the 80,416 radiology reports that were positive for a fall, 11,484 were associated with a fall-related E-code and 69,932 were not, corresponding to 29,280 patients who had incurred potential fall-related injuries and could not have been identified using E-codes alone.

For further evaluation of the classifier, we reviewed 100 reports that were randomly sampled across the range of assigned SVM decision values. Specifically, we divided the range of reports classified as positive for a fall into 50 intervals and uniformly drew a report from each interval, then did the same for reports classified as “not fall,” for a total of 100 reports. Out of the 100 reports sampled across the SVM decision values, there was one false positive and no false negatives.

DISCUSSION

The performance of all the classifiers we considered in our analyses, including the baseline L2-regularized SVM classifier, was acceptable. By adjusting the misclassification costs per class, an interesting jump in sensitivity occurred. This seems to confirm that the baseline classifier is biased towards the most prevalent class. Note that this adjustment also increased the positive predictive value. Next, feature selection clearly increased the classification performance of the model across all metrics. Interestingly, there is no difference in the performance of the three classifiers that apply feature selection. Furthermore, feature selection appears to overcome the bias due to imbalanced classes, because the performance of both the uniform and
adjusted misclassification costs classifiers is the same when MI ranking is used.

To compare the feature selection approaches, we considered the variables selected by the training algorithms. The magnitude of the feature weights (ie, SVM coefficients) for the linear SVM classifier provides a measure of significance for each feature. The top 50 features by weight for the baseline and final classifier (L2-SVM with MI and size-adjusted cost) are shown in Figure 2.

The SCAD-SVM classifier selected exactly two features: “fall” and “(fall).” Obviously, the classification power of such a simple model is as good as that of any other model. However, the option to choose a given number of features in the MI ranking approach allows us to capture clinically relevant contextual information about falls. In particular, noting, by comparison with cosine similarity, that the SVM coefficients in the linear case may be viewed as the feature vector of an ideal fall-positive radiology report, adding more features while maintaining performance provides insight into the context in which injurious falls occur. This more nuanced report may also be of use to clinicians and clinical researchers. In particular, looking at the weights of features in Figure 2, additional information surrounding the falls, such as sequelae (“dislocation”, “fracture”) as well as the most common sites of injury (“vertebral,” “rib,” “(hip),” “shoulder”) are included in the report and can be used to guide patient education and falls prevention efforts.

CONCLUSION
We used machine learning to build a classifier to identify fall events in VACS-VC. Starting with a corpus of unlabeled radiology reports from a carefully chosen patient population, we created a labeled reference standard to train our machine learning algorithms. The NLP tool YTEX was used to extract token and UMLS concept features from each report. We then applied machine learning methodology to create a “fall” event classifier, first training a simple classifier to provide baseline performance scores and then testing several possible classifier refinements. Feature selection was found to be crucial for increasing the performance of the algorithms, which yielded classifiers with area under the curve scores of 97.04 on the test set. Ultimately, the linear SVM classifier with features selected using MI and with class size-adjusted misclassification costs was singled out to be the final classifier to run on the full set of 2,977,739 patient radiology reports in VACS-VC.
Excluding the patients whose EHRs included fall-related E-codes, a total of 29,280 patients were identified as potentially having fallen. We have demonstrated the use of two feature selection methods of classifying radiology reports. Filter feature selection methods seem to be more commonly used in the clinical text mining community (for example, in Tremblay et al., McCarron et al., and the other literature we refer to in the background section of this article), and embedded feature selection methods appear to be less commonly used. We have contrasted our two feature selection approaches to highlight the potential advantages and disadvantages of each. Finally, we have observed that feature selection, independently of the approach one chooses, can significantly improve the performance of the classifier. All of this information will be combined with the goal of identifying falls as a patient-level variable. We will then build predictive models to identify fall risk factors, using this variable as well as fall-related E-codes as our outcome measures.

CONTRIBUTORS
J.B., S.F., J.W., and C.B. contributed to the design of the study; management of data, analysis, and interpretation of results; drafting; and revision of the paper. J.W. and C.B. annotated the reference standard. J.B. and S.F. implemented the machine learning methods. All authors read and approved the final manuscript.

FUNDING
This work was supported by the Yale Center for Clinical Investigation and the Clinical and Translational Science Award grant number UL1 RR024139 from the National Center for Research Resources; National Institute of Nursing Research grant number K01 NR013437; National Library of Medicine University Biomedical Informatics Research Training Award grant number 5 T15 LM007056. The data collection, maintenance, and coordinating center for the VACS Virtual Cohort were supported by awards U10 AA013566, U24 AA020794, and U01 AA020790.

COMPETING INTERESTS
None.

ACKNOWLEDGEMENTS
The authors wish to thank the reviewers for their careful reading of the manuscript and for their many helpful comments. The views expressed here are those of the authors and do not necessarily reflect the position or policy of the Department of Veterans Affairs or the United States government.

REFERENCES

AUTHOR AFFILIATIONS
1Yale School of Medicine, New Haven, CT
2Yale School of Nursing, West Haven, CT
3VA Connecticut Healthcare System, West Haven, CT