Automated identification of extreme-risk events in clinical incident reports

Mei-Sing Ong, Farah Magrabi, Enrico Coiera

▸ Additional appendices are published online only. To view these files please visit the journal online (www.jamia.bmj.com/content/19/e1.toc).

Centre for Health Informatics, University of New South Wales, Sydney, Australia

Correspondence to
Dr Mei-Sing Ong, Centre for Health Informatics, University of New South Wales, Sydney 2052, Australia;
m.ong@unsw.edu.au

Received 22 August 2011
Accepted 16 December 2011
Published Online First
11 January 2012

ABSTRACT

Objectives To explore the feasibility of using statistical text classification to automatically detect extreme-risk events in clinical incident reports.

Methods Statistical text classifiers based on Naïve Bayes and Support Vector Machine (SVM) algorithms were trained and tested on clinical incident reports to automatically detect extreme-risk events, defined by incidents that satisfy the criteria of Severity Assessment Code (SAC) level 1. For this purpose, incident reports submitted to the Advanced Incident Management System by public hospitals from one Australian region were used. The classifiers were evaluated on two datasets: (1) a set of reports with diverse incident types (n=120); (2) a set of reports associated with patient misidentification (n=166). Results were assessed using accuracy, precision, recall, F-measure, and area under the curve (AUC) of receiver operating characteristic curves.

Results The classifiers performed well on both datasets. In the multi-type dataset, SVM with a linear kernel performed best, identifying 85.8% of SAC level 1 incidents (precision=0.88, recall=0.83, F-measure=0.86, AUC=0.92). In the patient misidentification dataset, 96.4% of SAC level 1 incidents were detected when SVM with linear, polynomial or radial-basis function kernel was used (precision=0.99, recall=0.94, F-measure=0.96, AUC=0.98). Naïve Bayes showed reasonable performance, detecting 80.8% of SAC level 1 incidents in the multi-type dataset and 89.8% of SAC level 1 patient misidentification incidents. Overall, higher prediction accuracy was attained on the specialized dataset, compared with the multi-type dataset.

Conclusion Text classification techniques can be applied effectively to automate the detection of extreme-risk events in clinical incident reports.

INTRODUCTION

Detecting adverse events is pivotal for measuring and improving patient safety. Voluntary reporting of patient safety incidents in healthcare provides a means for adverse events to be reported, so that lessons may be learned from them.[1] Aggregation of incidents provides an indication of the nature and extent of errors in healthcare. Several studies implicated that incident reporting is capable of identifying more preventable problems, and provides more useful information about the context in which adverse events occur, compared with traditional methods such as medical chart review.[2 3] However, manual review of incident reports is both costly and resource intensive. As a result, most reports submitted are not given due attention, and thus opportunity to learn from them is lost.

Most reporting systems apply a severity rating system to grade the seriousness of an incident, so that investigation of high-risk events can be prioritized. This is critical in ensuring that major events with significant consequences that are likely to recur are followed-up immediately. One such rating system is the Severity Assessment Code (SAC), which was first developed by the US Veteran Administration to determine whether a reported event merits a dedicated root cause analysis.[4] The system has since been adopted by many reporting systems worldwide, including the Advanced Incident Management System (AIMS), an incident reporting system used by many public hospitals in Australia. Using the SAC system, the reporter of an incident rates the incident on the basis of its actual or potential severity and likelihood of recurrence. This rating then determines how the incident is managed.

The use of a severity rating system facilitates consistent handling of reports and prevents potential downplay of the importance of an event.[4] However, a rating system works only if reporters are knowledgeable about the system and are able to apply it consistently. In practice, clinicians are often uncertain about how to use incident reporting systems.[5] A study carried out to assess the reliability of the severity rating scale used by the National Reporting and Learning System in England and Wales for medication errors showed that there were marked differences in the severity ratings graded between different health professional groups, and at different times rated by the same individuals.[6] Studies evaluating the SAC system have also found poor agreement between raters.[7 8] Severity rating was highly subjective, and reporters were generally hesitant in ranking incidents with a high severity score.[7] It was further found that the rating of severity was determined by criteria other than the stated policy, and resource limitations are a major influence.[8] Clinicians may report an incident at a lower severity level so as to avoid the disruption of formal investigation for themselves and their colleagues. Conversely, clinicians may assign low-severity incidents a higher severity score, in an attempt to escalate the matter to management, rather than managing it at the local level.

Thus the assignment of severity rating to incident reports is highly variable. Current methods, which rely on retrospective manual review of incident reports, do not facilitate timely detection of safety problems. In one Australian state alone, 62 369 patient safety incidents were reported between January and June 2009.[9] This is equivalent to 20 incidents per 1000 bed days, or 8.2% of admissions. Manual reviews are not only time consuming, they also suffer from a lack of consistency as to what constitutes a high-risk event, and there is no agreement on how to classify or prioritize harm.

Text classification using machine learning has been used successfully to identify safety issues in aviation incident reports.[10–12] In healthcare, automated text classification has also been applied in detecting adverse events or disorders in clinical text,[13] including automatic identification and classification of surgical margin status from pathology reports,[14] automatic detection of disorders in chest radiograph reports,[15 16] and detection of adverse events in discharge summaries.[17 18] However, automated classification of the severity of adverse events in clinical incident reports remains relatively unexplored. A recent study examined the use of text classification to detect anaphylaxis incident reports.[19] In our previous work, we investigated the feasibility of applying statistical text classification as an alternative to manual review of clinical incident reports. Classifiers based on Naïve Bayes algorithm and Support Vector Machine (SVM) were constructed to automatically identify groups of related incidents from reporting systems.[20] These techniques performed well in identifying incident types such as patient misidentification and handover failure. In the present study, we turned our attention to the recognition of severity of harm across incident types, as this is the most common driver for institutional investigations. We hypothesize that the techniques used to predict incident types can also be applied to identification of incident severity. Specifically, our study objective was to determine if automated text classification can (1) perform equally well or better than a human classifier and (2) identify high-severity cases in reports that were either not initially classified or misclassified by the reporter. Our ultimate goal was to demonstrate that automated text classification is sufficiently robust to automatically monitor large volumes of incident reports.

BACKGROUND

The Advanced Incident Management System (AIMS)

The AIMS reporting system was set up by the Australian Patient Safety Foundation in the late 1980s as an initiative to reduce adverse events in anesthetic practice.[21] Participating anesthetists were invited to report, on an anonymous and voluntary basis, any unintended incident that reduced or could have reduced the safety margin for a patient. Later the system was extended beyond anesthesia to other specialties.[22] The system is increasingly being adopted by hospitals across Australia.

AIMS provides an in-built classification system consisting of 21 healthcare incident types (HITs) (online appendix A). An HIT is a descriptive term for a category made up of incidents of a common nature, grouped because of shared, agreed features.[23] Each HIT contains further sub-categories. On submission, a reporter assigns an incident to one or more HITs, and their sub-categories. The ‘clinical management’ HIT describes incidents relating to the clinical management of patients. One of its sub-categories is ‘wrong patient’, which describes incidents where care was given to the incorrect patient. An AIMS report also contains a number of unstructured free-text fields, including incident description, outcome for the subject, initial action taken, prevention steps, and results of incidents (online appendix B).

Severity rating system in AIMS

AIMS uses the SAC system to rate the severity of incidents. The rating system comprises four numerical ratings, scaling from 1 (most severe) to 4 (least severe). SAC level 1 represents extreme-risk incidents. SAC levels 2, 3, and 4 are assigned to high-risk incidents, medium-risk incidents, and low-risk incidents, respectively. Rating is assigned by the reporter on incident submission, based on the severity of the event and its likelihood to recur. Once the severity and likelihood categories are determined, a SAC score can be assigned by consulting the SAC Matrix (online appendix C). Appropriate management of the incident will depend on the level of risk that the incident poses to patients and the healthcare organization (online appendix C).

METHODS

Statistical text classifiers using Naïve Bayes and SVM were developed to identify extreme-risk events based on the free-text description of the incident reports. Naïve Bayes and SVM were selected because they are commonly used in classification of documents.[24 25] The text classifiers were first trained and tested on a dataset comprising diverse types of incidents retrieved from the AIMS incident reporting database. The same process was then repeated on a subset of incidents relating to patient misidentification. The ability of the classifiers to detect extreme-risk events in both datasets was then evaluated. In this study, extreme-risk events were defined as events that satisfy SAC level 1 criteria, in accordance with the AIMS Incident Management Policy Directive.[26]

Data source

The datasets consisted of incident reports submitted to the AIMS database by public hospitals from one Australian region between March 2004 and July 2009.

Dataset

For the purpose of this study, we extracted a non-exhaustive set of incidents from the AIMS database. Two sets of incidents were randomly retrieved: (1) reports classified under any incident type; and (2) reports classified under the ‘clinical management/wrong patient’ subcategory. For each dataset, incidents ranked as SAC level 1 by the reporters were set aside as the positive dataset. An equal number of incidents not ranked as SAC level 1 were selected to form the negative dataset. A balanced set of positive and negative datasets was required, since SVM is known to perform poorly when trained on unbalanced datasets.[27] A stratified random sampling approach was used to ensure that the negative dataset contained incidents of SAC levels that were representative of the AIMS database (online appendix D). Reports were manually reviewed to ensure correctness of classification. The reviewer was blinded to the initial incident severity score associated with the report. Discrepancies between the initial classification and the reviewer were arbitrated by a second independent reviewer. Inter-rater reliability for selecting SAC level 1 incidents was good for both datasets (κ=0.81, 95% CI 0.68 to 0.94 for the multi-type dataset; κ=0.94, 95% CI 0.82 to 1.0 for the patient misidentification dataset). Disagreements were resolved by discussion. Both reviewers were health informatics researchers who have received training in incident classification.

Training methodology

Feature extraction

In order to construct classifiers based on text description of the incident reports alone, all AIMS-specific codes were removed, including the reporter-classified SAC score, retaining only descriptive narratives in the following fields: incident description, outcome for the subject, initial action taken, prevention steps, and results of incident. Punctuation was removed, and text was converted into lower case. All narratives were then processed into an unordered collection of words, known as a 'bag of words', which formed the feature set.

Input representation

The bag of words was then transformed into a numeric representation interpretable by the classifiers. Several common input representation methods were tested: (1) binary—transforming input features into 1 or 0, where 1 indicates the occurrence of a word in the report, and 0 indicates non-occurrence; (2) term frequency—converting input features into the actual frequency of occurrence for each word in the report; (3) thresholding—converting input features into one of three values (0 if a word did not occur, 1 if it occurred once, and 2 if it occurred two or more times); and (4) term frequency-inverse document frequency (tf-idf)—transforming input features into the term frequency of each word multiplied by the inverse document frequency. In tf-idf, weightings for words were computed, depending on the importance of the words.[28] When a word appeared in many reports, it was considered less important. When the word was relatively unique and appeared in few reports, a greater weighting was assigned. In all cases, the input space was normalized.

Feature selection

Several feature selection strategies were tested to see if they enhance the quality of processed reports. Strategies to reduce the dimension of the feature space included: (1) excluding words that occurred in similar frequency (with a difference of <2) in both the positive and negative datasets; (2) excluding determiners, prepositions, pronouns, conjunctions, particles and misspelt words; (3) stemming, which involves regularizing grammatical variants such as singular/plural, present/past; (4) the addition of bigrams (word pairs that co-occur in text) selected through manual inspection of reports (online appendix E). Term frequency was used as a criterion for selecting bigrams, and bigrams were restricted to word sequences of length two, as bigrams with longer sequences were known to degrade classification performance.[29]

Training and validating the classifiers

Classifiers were trained using Naïve Bayes and SVM. The kernel function in an SVM plays the central role of implicitly mapping the input vector into a high-dimensional feature space. In this study, we considered three kernel types: linear, polynomial, and radial-basis function (RBF). Kernel parameters (degree for a polynomial kernel, and γ for an RBF kernel) and the trade-off parameter (C) were tuned to optimize an SVM classifier.[30]

Data mining software known as WEKA was used to train the classifiers.[31] A stratified 10-fold cross-validation was applied to evaluate the classifiers.

Evaluating performance

Results were assessed using the following measures: (1) accuracy—the percentage of incidents classified correctly to be in a given category in relation to the total number of incidents tested; (2) precision—the percentage of true positives detected in relation to the total number of incidents classified for a category; (3) recall—the percentage of true positives detected in relation to the actual number of incidents in a category; (4) F-measure—the harmonic mean of precision and recall; (5) area under the curve (AUC) of the receiver operating characteristic curve; (4) κ statistics—a chance-corrected measure of agreement between the classifications and the true classes; and (5) mean absolute error.

The classifiers were evaluated in three phases: (1) the impact of input representations on the prediction accuracy of each classifier was evaluated; (2) using the best-performing input representation for each classifier, feature selection strategies were applied and evaluated; and finally (3) using the input representation and feature selection combination that produced the highest accuracy, the classifiers were optimized to achieve the best F-measure. Since disproportionate stratified sampling was used to select reports (with SAC level 1 incidents being oversampled, and the other severity categories being undersampled), the outcome of the classifiers may not reflect their actual performance on the target population. To address this, the performance metrics of the classifiers were adjusted using post-stratification weighting, so that the results from each severity category carry as much weight in the calculation of the overall results as its share in the target distribution.[32] A qualitative error analysis was also performed to assess textual patterns that were most likely to contribute to misclassifications.

RESULTS

Seven thousand incident reports between January 2009 and July 2009 were extracted randomly from the AIMS database. Of these, 0.86% incidents were rated SAC level 1, 2.73% were rated SAC level 2, 36.94% were rated SAC level 3, and 51.66% were rated SAC level 4. In the remaining 7.81% reports, a SAC level was not assigned. Querying the AIMS database for reports specifically related to patient misidentification between March 2004 and July 2008 yielded 1083 incidents. Of these, 10.43% incidents were rated SAC level 1, 1.48% were rated SAC level 2, 25.12% were rated SAC level 3, and 55.77% were rated SAC level 4. SAC classification was not provided in the remaining 7.2% reports. Of all the incident reports retrieved, there were in total 625 (7.7%) that did not have a severity rating assigned.

Manual review of the incidents showed that 62.5% of SAC level 1 incidents in the multi-type dataset were correctly classified. The remaining incidents were either misclassified (34.4%) or not classified (3.1%). Of the SAC level 1 patient misidentification incidents, 83.8% were correctly classified by the reporters; the remaining incidents (16.2%) were misclassified.

Examples of incidents for both datasets are provided in box 1, including incidents of different SAC levels and SAC level 1 incidents that were not classified and misclassified by the reporter.

The performance of the classifiers

Feature extraction

There were 2803 unique words in the multi-type dataset, and 2624 unique words in the patient misidentification dataset.

Input representation

Table 1 shows the effects of different input representations on the performance of the classifiers. In the multi-type dataset, tf-idf appeared to facilitate the best prediction accuracy for Naive Bayes and SVM with a linear kernel. SVM with polynomial and RBF kernels performed best when simple binary representation was used. In the patient misidentification dataset, term frequency clearly stood out to be the best method for SVM, and tf-idf performed best with the Naïve Bayes classifier. Overall, SVM was less sensitive than Naïve Bayes to the types of input representation.

Box 1 Examples of incidents

Incidents from the multi-type dataset

SAC level 1

Example 1: A patient was found dead at the psychiatry ward, suspected suicide (overdose).

Example 2: Patient transferred to ward in an alert state at 19:45. Patient received oxygen via humidifier at 40%. At 02:00, patient's observations were attended. No oxygen saturations were attended. Patient became agitated, removing oxygen. No action was taken by the nursing staff member. No medical officer notified. 04:00, patient was found deceased.

Example 3: Lumbar puncture done at low thoracic instead of lumbar level, resulting in probably laceration of radicular vessel and subarachnoid spinal bleeding, and associated injury to distal cord/conus.

Example 4: Patient was transferred from Hospital A with a 6-month history of dysplasia. Patient had a biopsy and was diagnosed with moderate to well differentiated adenocarcinoma. Patient had esophagectomy, but subsequent esophagectomy specimen showed no evidence of carcinoma.

SAC level 2

Example 5: Woman came to ward from recovery without prophylactic 40 U synotocinon infusion. 450 ml blood loss in OT, proceeded to lose a further 1000 mls, MET team called, BP 80/40.

SAC level 3

Example 6: Patient developed a pressure ulcer on the left heel, after having a total knee replacement.

SAC level 4

Example 7: Staff went to administer charted medication to a patient, however the patient was asleep. Patient's mother requested medication not be given until the patient woke up. Medication had been signed for by staff prior to being given. The patient's mother stated that the medication was never given to patient when she woke up, and since it was already signed for prior to being given, it was hard to determine if it was actually given.

Incidents from the patient misidentification dataset

SAC level I

Example 8: Wrong patient underwent venesection as a result of incorrect documentation. Medical order for venesection written in wrong patient's chart.

Example 9: Wrong patient was brought from ward to Nuclear Medicine for a bone scan, and was given intravenous injection of 814MBq Tc99m-MDP before the error was detected.

SAC level 2

Example 10: Patient arrived from the nursing home, with the wrong patient notes and medication chart. Patient had a cardiac arrest, the wrong family was notified and end-of-life decisions were being made for the wrong patient. Family identified that it was the wrong person when the nursing home was contacted. The correct family was contacted and arrived prior to patient's death.

SAC level 3

Example 11: Mislabeled specimen, blood collected from the wrong patient.

SAC level 4

Example 12: The wrong patient admission was filed with another patient's medical record.

SAC level 1 incidents that were not classified by reporter

Example 13: Patient presented to abortion clinic stating she was 5 weeks pregnant. Ultrasound performed in abortion clinic stated she was 5 weeks pregnant. IV sedation given and cervical cannula inserted by doctors, noted small amount of clear fluid. Procedure terminated and patient transferred to delivery suite via ambulance. Mother transferred to S4EA, delivered 22+ week fetus. Baby died in NICU shortly afterwards from extreme prematurity and sepsis.

Example 14: Patient was admitted for Primary PTCA June 27, 2008. Coronary artery dissected in the process of the PTCA, but coronary blood flow OK. Heparinized overnight. Review by Nurses 9:00 showed bruised groin. Ultrasound showed small false aneurysm. 'Femstop' applied and i.v. heparin ceased. Nurses noted that they were awaiting review at 14:45. Patient developed Chest Pain and Ac Myocardial Infarction at 17:40. Cardiothoracic Team contacted and CABG performed using veins rather than arterial conduits and in the emergent situation. Myocardial Infarction occurred which had previously been avoided. The main problem was that the decision to stop the Heparin was made and review was awaited. However, it did not happen, and the consultants involved were not contacted.

SAC level 1 incidents that were misclassified by the reporter

Example 15: Patient admitted to medical ward with psychiatrist's name on the paperwork as the main treating doctor, when the psychiatrist had not agreed to accept care. As a consequence, the patient was not seen by a doctor until 4 days later, despite deterioration in her status. She died from a gastric hemorrhage.

Example 16: Patient did not receive a MET call after his condition deteriorated, due to confusion regarding 'Not For Resuscitation' order. In this case a MET call was warranted since the patient was not preterminal. Many of the nursing staff wrongly interpret NFR as pre-terminal, and therefore palliation.

Table 1 Prediction accuracy for Severity Assessment Code level 1 incidents with different input representations. The highest prediction accuracy achieved by each algorithm is highlighted in grey

Dataset	Algorithm	Input representation Binary	 Thresholding	 Term frequency	 tf-idf
All incident types (n=120) Positive dataset (n=60) Negative dataset (n=60)	Naive Bayes	67.5%	69.2%	73.3%	80.8%
	SVM linear (C=1)	82.5%	81.7%	79.2%	83.3%
	SVM linear (C=2)	82.5%	80.8%	79.2%	82.5%
	SVM linear (C=3)	81.7%	81.7%	78.3%	83.3%
	SVM linear (C=4)	81.7%	80.8%	79.2%	81.7%
	SVM linear (C=5)	82.5%	80.8%	79.2%	81.7%
	SVM polynomial (degree=2)	82.5%	80.8%	79.2%	82.5%
	SVM polynomial (degree=3)	83.3%	80.8%	79.2%	83.3%
	SVM polynomial (degree=4)	84.2%	80.8%	80.0%	83.3%
	SVM polynomial (degree=5)	83.3%	81.7%	79.2%	83.3%
	SVM polynomial (degree=6)	83.3%	81.7%	79.2%	83.3%
	SVM RBF (γ=0.0001)	81.7%	80.8%	79.2%	80.8%
	SVM RBF (γ=0.0002)	81.7%	79.2%	78.3%	83.3%
	SVM RBF (γ=0.0003)	84.2%	80.8%	79.2%	81.7%
	SVM RBF (γ=0.0004)	82.5%	81.7%	78.3%	82.5%
	SVM RBF (γ=0.0005)	83.3%	80.0%	77.5%	81.7%
Patient misidentification incidents (n=166) Positive dataset (n=83) Negative dataset (n=83)	Naive Bayes	82.5%	88.0%	86.7%	89.2%
	SVM linear (C=1)	92.2%	91.6%	95.2%	92.2%
	SVM linear (C=2)	92.2%	91.0%	95.8%	92.2%
	SVM linear (C=3)	92.2%	91.0%	96.4%	92.2%
	SVM linear (C=4)	91.6%	91.6%	96.4%	92.2%
	SVM linear (C=5)	92.2%	91.6%	95.8%	92.2%
	SVM polynomial (degree=2)	92.2%	91.0%	96.4%	92.2%
	SVM polynomial (degree=3)	92.2%	91.6%	96.4%	91.6%
	SVM polynomial (degree=4)	92.2%	91.0%	96.4%	92.2%
	SVM polynomial (degree=5)	92.2%	91.0%	96.4%	92.2%
	SVM polynomial (degree=6)	92.2%	91.0%	96.4%	92.2%
	SVM RBF (γ=0.0001)	92.2%	91.6%	96.4%	92.2%
	SVM RBF (γ=0.0002)	92.2%	91.6%	95.8%	92.2%
	SVM RBF (γ=0.0003)	92.2%	91.0%	95.2%	92.2%
	SVM RBF (γ=0.0004)	92.2%	91.6%	95.8%	92.2%
	SVM RBF (γ=0.0005)	92.2%	91.0%	95.2%	92.2%

RBF, radial-basis function; SVM, Support Vector Machine; tf-idf, term frequency-inverse document frequency.

Feature selection

Table 2 shows the effects of different feature selection strategies on the performance of the classifiers. Overall, most classifiers were relatively unaffected by the feature selection strategies. In the multi-type dataset, SVM with a linear kernel improved slightly with all feature selection strategies, except for stemming. SVM with an RBF kernel also improved slightly when words with difference in frequency between bags <2 were excluded from the feature set. In the patient misidentification dataset, the prediction accuracy of Naïve Bayes improved very slightly when determiners, prepositions, pronouns, conjunctions, particles, and misspelt words were excluded. In most cases, classification performance degraded when stemming was applied.

Best performance

Table 3 shows the best performance achieved by each classifier with input representation, feature selection strategies, and SVM parameters tuned to optimize for F-measure. In the multi-type dataset, SVM with a linear kernel performed best, identifying 85.8% (95% CI 79.6% to 92.1%) of SAC level 1 incidents (precision=0.88, recall=0.83, F-measure=0.86, AUC=0.92). Adjusting the performance metrics to account for the oversampling of SAC level 1 incidents gives an accuracy of 86.7% (precision=0.85, recall=0.87, F-measure=0.86, AUC=0.91). In the patient misidentification dataset, 96.4% (95% CI 93.6% to 99.2%) of SAC level 1 incidents were detected when SVM with linear, polynomial or RBF kernel was used (precision=0.99, recall=0.94, F-measure=0.96, AUC=0.98). Post-stratification accuracy of the classifiers was 98.3% (precision=0.95, recall=0.98, F-measure=0.96, AUC=0.98). The automated classifiers outperformed reporter classification by 23% in the multi-type dataset and 13% in the specialized dataset. This improved performance was found to be statistically significant (p<0.001). Naïve Bayes showed reasonable performance, detecting 80.8% of SAC level 1 incidents in the multi-type dataset and 89.8% of SAC level 1 patient misidentification incidents. Overall, higher prediction accuracy was attained on the specialized dataset than on the multi-type dataset.

Error analysis

Features associated with accurate predictions

Examination of misclassified incidents reveals several distinct patterns. Overall, incidents containing the words 'death', 'deceased', 'died', 'adverse', or 'suicide' were more likely to be classified as SAC level 1. Further, patient misidentification incidents containing the phrases 'wrong patient' or 'wrong procedure' were accurately identified as SAC level 1. All cases of documentation errors were correctly identified as true negatives. Common words relating to these errors included 'document', 'form', 'label', 'charts', and 'records'.

Features associated with false positives

Features that were associated with true positives were also found in many false-positive predictions. About 90% of false

Table 2 Text classification accuracy with different feature selection strategies. The best-performing input representation for individual classifiers (table 1) was used to allow comparison. Cases where feature selection resulted in improvement in prediction accuracy are highlighted in grey

Dataset	Algorithm	Feature selection strategies			
		Exclude words with difference in frequency between bags less than 2	Exclude determiners, prepositions, pronouns, conjunctions, particles, misspelt words	Stemming	Bigrams
All incident types (n=120)	Naive Bayes	79.2%	79.2%	77.5%	80.8%
	SVM linear (C=1)	82.5%	85.0%	81.7%	81.7%
Positive dataset (n=60)	SVM linear (C=2)	83.3%	82.5%	80.8%	82.5%
Negative dataset (n=60)	SVM linear (C=3)	82.5%	83.3%	82.5%	84.2%
	SVM linear (C=4)	84.2%	85.8%	82.5%	82.5%
	SVM linear (C=5)	84.2%	84.2%	83.3%	84.2%
	SVM polynomial (degree=2)	84.2%	81.7%	83.3%	84.2%
	SVM polynomial (degree=3)	84.2%	83.3%	84.2%	84.2%
	SVM polynomial (degree=4)	84.2%	84.2%	83.3%	81.7%
	SVM polynomial (degree=5)	82.5%	84.2%	83.3%	83.3%
	SVM polynomial (degree=6)	82.5%	83.3%	83.3%	81.7%
	SVM RBF (γ=0.0001)	82.5%	83.3%	84.2%	83.3%
	SVM RBF (γ=0.0002)	82.5%	82.5%	84.2%	83.3%
	SVM RBF (γ=0.0003)	83.3%	84.2%	82.5%	83.3%
	SVM RBF (γ=0.0004)	83.3%	81.7%	82.5%	82.5%
	SVM RBF (γ=0.0005)	85.0%	82.5%	82.5%	82.5%
Patient misidentification incidents (n=166)	Naive Bayes	89.2%	89.8%	88.6%	89.2%
	SVM linear (C=1)	95.8%	94.0%	91.0%	96.4%
	SVM linear (C=2)	95.8%	94.0%	91.6%	95.8%
Positive dataset (n=83)	SVM linear (C=3)	96.4%	94.6%	92.2%	95.8%
Negative dataset (n=83)	SVM linear (C=4)	96.4%	95.2%	91.6%	96.4%
	SVM linear (C=5)	96.4%	95.2%	91.6%	96.4%
	SVM polynomial (degree=2)	96.4%	94.6%	92.2%	96.4%
	SVM polynomial (degree=3)	96.4%	94.6%	92.2%	96.4%
	SVM polynomial (degree=4)	95.8%	94.0%	91.6%	95.8%
	SVM polynomial (degree=5)	96.4%	94.6%	92.2%	96.4%
	SVM polynomial (degree=6)	96.4%	94.6%	91.6%	96.4%
	SVM RBF (γ=0.0001)	96.4%	94.6%	91.6%	95.8%
	SVM RBF (γ=0.0002)	95.2%	95.8%	91.6%	95.8%
	SVM RBF (γ=0.0003)	95.8%	96.4%	92.2%	95.8%
	SVM RBF (γ=0.0004)	95.8%	95.8%	92.8%	95.2%
	SVM RBF (γ=0.0005)	95.8%	96.4%	92.2%	95.2%

RBF, radial-basis function; SVM, Support Vector Machine.

positives were observed in incidents involving a near-miss, where potential adverse outcomes were described but were avoided.

Features associated with false negatives

In both datasets, prediction accuracy was affected by spelling errors in the reports. About 70% of false negatives were a result of misspellings. In particular, the word 'misidentification', often used to describe patient misidentification incidents, was commonly misspelt. As this word was critical to the classification task, error in spelling the word may have resulted in the incidents being missed by the classifiers. Other critical words that were commonly misspelt included 'deceased' and 'deterioration'. In the multi-type dataset, incident types that occurred less frequently in our dataset were more likely to be misclassified. These included incidents relating to blood transfusion and drug overdose.

DISCUSSION

Classification performance

Machine learning techniques appear able to detect extreme-risk events from the free text of incident reports. The performance of SVM is particularly encouraging, achieving high accuracies even on a small sample size. This finding is in agreement with our previous experiment, where classifiers trained on SVM achieved an accuracy of above 80% in predicting incident types when the training set was as small as 100 samples.[20] Post-stratification adjustment of the performance measures indicates that both Naive Bayes and SVM will perform well in the real world. The adjusted accuracy of Naive Bayes is slightly poorer than the unadjusted accuracy, as the algorithm performs better in identifying SAC level 1 incidents, which occur less frequently in a real-world dataset than other SAC categories. Conversely, the adjusted accuracy of SVM is an improvement from the unadjusted measure, as the algorithm performs better in identifying other SAC categories.

It also appears likely that our techniques will be adaptable to allow detection of a broad array of adverse events. The classifiers were capable of identifying SAC level 1 incidents when tested on the dataset containing a diverse range of incident types, indicating that the method is generalizable across different types of incidents. However, both Naïve Bayes and SVM performed better on the specialized dataset consisting of incidents relating to patient misidentification only. A plausible explanation is the relative homogeneity of the data in the specialized dataset compared with the multi-type dataset. The specialized dataset

Table 3 Classifiers with best performance, with input representation, feature-selection strategies, and SVM parameters tuned to optimize for F-measure

		Performance measures							Adjusted performance measures*				
Dataset	Algorithm	Accuracy (95% CI)	Precision	Recall	F-measure	AUC	κ	Mean absolute error	Accuracy (95% CI)	Precision	Recall	F-measure	AUC
All incident types (n=120)	Naive Bayes Input representation: tf-idf Feature selection: none	80.8% (73.8 to 87.9)	0.80	0.82	0.81	0.87	0.62	0.20	80.0% (72.8 to 87.2)	0.81	0.80	0.81	0.88
	SVM linear Input representation: tf-idf Feature selection: exclude determiners, prepositions, pronouns, conjunctions, particles, misspelt words Parameters: C=4	85.8% (79.6 to 92.1)	0.88	0.83	0.86	0.92	0.72	0.26	86.7% (80.6 to 92.8)	0.85	0.87	0.86	0.91
	SVM polynomial Input representation: binary Feature selection: none Parameters: degree=5, C=1	84.2% (77.7 to 90.7)	0.85	0.83	0.84	0.91	0.68	0.27	85.0% (78.6 to 90.6)	0.84	0.85	0.84	0.91
	SVM RBF Input representation: binary Feature selection: exclude words with difference in frequency between bags <2 Parameters: γ=0.0005, C=1	85.0% (78.6 to 90.6)	0.88	0.82	0.85	0.91	0.70	0.28	89.9% (84.5 to 95.3)	0.82	0.90	0.86	0.90
Patient misidentification incidents (n=166)	Naive Bayes Input representation: tf-idf Feature selection: exclude determiners, prepositions, pronouns, conjunctions, particles, misspelt words	89.8% (85.2 to 94.4)	0.88	0.93	0.90	0.97	0.80	0.10	87.3% (81.3 to 93.3)	0.92	0.87	0.89	0.96
	SVM linear Input representation: term frequency Feature selection: none Parameters: C=4	96.4% (93.6 to 99.2)	0.99	0.94	0.96	0.98	0.93	0.11	98.3% (96.0 to 100.0)	0.95	0.98	0.96	0.98
	SVM polynomial Input representation: term frequency Feature selection: none Parameters: degree=3, C=1	96.4% (93.6 to 99.2)	0.99	0.94	0.96	0.98	0.93	0.11	98.3% (96.0 to 100.0)	0.95	0.98	0.96	0.98
	SVM RBF Input representation: term frequency Feature selection: none Parameters: γ=0.0001, C=1	96.4% (93.6 to 99.2)	0.99	0.94	0.96	0.98	0.93	0.11	98.3% (96.0 to 100.0)	0.95	0.98	0.96	0.98

*Performance metrics adjusted to account for the disproportionate stratified random sampling of the data source.

AUC, area under the curve; RBF, radial-basis function; SVM, Support Vector Machine; tf-idf, term frequency-inverse document frequency.

consisted of a single narrow subject with limited vocabulary. The multi-type dataset, however, encompasses a wider range of incident types and diverse subject matters with overlapping vocabularies. Thus training the classifiers became slightly more challenging. Box 1 provides examples of SAC level 1 incidents from both datasets. The diversity of the multi-type dataset is clear from these examples.

Also evident from the experiment was that, with careful selection of input representation, feature selection strategies, and classifier parameters, classification performance could be significantly improved. The choice of input representation appeared to have a greater impact than feature selection. Feature selection is a technique commonly used in selecting a subset of relevant features for building robust learning models. By removing most irrelevant and redundant features from data, feature selection helps improve the classification performance. A major characteristic, or difficulty, of text classification problems is the high dimensionality of the feature space. The feature space consists of the unique terms (words or phrases) that occur in documents, and there can be tens of thousands of terms for a moderate-sized text collection. This is prohibitively high for many learning algorithms, because of the training cost it incurs. Thus it is highly desirable to reduce the input space. However, in our study, the number of unique words in the dataset was <3000. Thus the role of dimension reduction through feature selection was less important. Conversely, removing features from the dataset may have the effect of removing useful data, resulting in poorer classification performance. Further, SVM tends to generalize well in high-dimensional feature spaces, thus eliminating the need for feature selection.[25]

As with any classification problem, the strategies that would yield the best classification performance are dependent on the dataset. The techniques that worked best on the multi-type dataset performed differently on the patient misidentification dataset, and vice versa. The same also holds for the choice of kernel and parameters for SVM classifiers. There is no special kernel, which has the best generalization performance for all kinds of problem domains. Different kernel functions and combinations of SVM parameters would need to be tested to ascertain the most appropriate classifier.

Unreliability of reporters' classification

Our analysis of the incident reports confirms existing literature that the severity rating given by the reporters was highly unreliable. Many reports involving severe patient harm events did not have a severity rating assigned or were incorrectly given a lower severity rating. As a result, these incidents were unlikely to have been sufficiently addressed, and therefore the opportunity to learn from them is missed. Our analysis underscores the inadequacy of relying on reporters' ratings to detect high-risk events. The potential benefits that can be realized from using automated detection techniques are evident. Since 0.86% of incidents were SAC level 1, and 37.5% were either misclassified or not classified by human raters, and we can detect 86% of these, then our system would identify about three serious events that would otherwise have been missed per 1000 incident reports.

Applications of machine learning

While we have demonstrated the feasibility of using machine learning techniques to automate the detection of SAC level 1 incidents, the same techniques can be applied to detect incidents of different severities. Further, the approach is generalizable across different reporting systems and reporting languages. To apply the technique, classifiers must be trained on datasets specific to a particular setting and language. This is because terminology, reporting, and linguistic styles may differ between reporting systems, hospitals, and countries. Once developed, the classifiers can be integrated into the reporting system, so that extreme-risk events can be automatically detected on submission. Additional software can be implemented to send electronic alerts to the appropriate parties when an incident is submitted. This can potentially provide a scalable and effective means of managing incidents.

It is important to note that, as goals and priorities for patient safety change through time, retraining of the classifiers is critical to ensure that they are consistent with the most recent policies. Further, as reporting systems and reporting culture can evolve through time, it is necessary to retrain the classifiers on a regular basis for optimal performance.

Although automated text classification is a powerful tool, it should be emphasized that it is not a replacement for manual review. Manual analysis provides insights that cannot be captured by any automated methods. However, when human resources are lacking, automated classifiers can significantly reduce the effort spent in identifying incidents.

Limitations

As with any statistical machine learning technique, the performance of our classifiers is only as good as the training data. The poor quality of some reports may have hampered their classification. Further, we were limited by the number of incident reports available for training and testing. Thus the incident classes tested here may not be representative of all incident types. Nevertheless, for our results to be generalizable, it is important that our set of incidents were representative of their class, but they need not be exhaustive.

Despite these limitations, the potential benefits that can be realized from automating detection of high-risk incidents are clear. It has been widely recognized that incident reporting systems fall short of meeting their objectives of improving patient safety. The success of the Aviation Safety Reporting System has been attributed to three factors: reporting is safe (anonymous), simple (a one-page report), and worthwhile (timely analysis and feedback).[33] Current incident reporting systems are often complex, and timely feedback is a challenge because of limited human resources. Thus the systems are underutilized. We believe that an automated approach can help simplify the systems and allow more expedient feedback. In addition, miscategorised incidents can be detected, providing a more reliable system.

CONCLUSIONS

Machine learning techniques appear to be a viable enhancement to manual detection of extreme-risk events in clinical incident reports. In this study, classifiers were trained successfully to identify SAC level 1 incidents reported to the AIMS database. The same techniques could be extended to other incident reporting systems and for identifying incidents of different severity levels.

This study is a proof-of-concept, and only simple methods were used. Further enhancements to the algorithms may improve performance efficiency.

Acknowledgments We thank Ms D Arachi for assistance with classifying the incidents, Dr S Anthony for sharing expertise, and Ms S Michael for assistance with searching the AIMS database. We are grateful to the healthcare region for providing access to the incident reports.

Funding This research is supported by NHMRC Program Grant 568612 and the Australian Research Council (ARC) LP0775532. The funding sources played no role in the design, conduct, or reporting of the study.

Competing interests None.

Contributors MSO developed the classifiers, analyzed the data, and wrote the manuscript. FM and EC analyzed the data and reviewed the manuscript.

Provenance and peer review Not commissioned; externally peer reviewed.

REFERENCES

1. **Kohn L,** Corrigan J, Donaldson M. *TO err is Human: Building a safer Health System. Committee on Quality of Health Care in America*. Washing DC: US Institute of Medicine, National Academy Press, 1999:86–101.
2. **Beckmann U,** Bohringer C, Carless R, *et al*. Evaluation of two methods for quality improvement in intensive care: facilitated incident monitoring and retrospective medical chart review. *Crit Care Med* 2003;**31**:1006–11.
3. **O'Neil AC,** Petersen LA, Cook EF, *et al*. Physician reporting compared with medical-record review to identify adverse medical events. *Ann Intern Med* 1993;**119**:370–6.
4. **Bagian JP,** Lee C, Gosbee J, *et al*. Developing and deploying a patient safety program in a large health care delivery system: you can't fix what you don't know about. *Jt Comm J Qual Improv* 2001;**27**:522–32.
5. **Vincent C,** Stanhope N, Crowley-Murphy M. Reasons for not reporting adverse incidents: an empirical study. *J Eval Clin Pract* 1999;**5**:13–21.
6. **Williams SD,** Ashcroft DM. Medication errors: how reliable are the severity ratings reported to the national reporting and learning system? *Int J Qual Health Care* 2009;**21**:316–20.
7. **University Centre for Clinical Governance Research in Health.** *Evaluation of the safety improvement program in New South Wales: Study no 6 report on program outcomes*. Kensington: Centre for Clinical Governance Research, University of New South Wales. http://www.med.unsw.edu.au/medweb.nsf/resources/SIP2/$file/Final+SIP+Study+6.pdf (accessed 6 Apr 2011).
8. **Hor SY,** Iedema R, Williams K, *et al*. Multiple accountabilities in incident reporting and management. *Qual Health Res* 2010;**20**:1091–100.
9. **NSW Health.** *Clinical Incident Management in the NSW public health system, January–June 2009*. http://www.health.nsw.gov.au/resources/quality/incidentmgt/pdf/incident_management_2009_01to06.pdf (accessed 6 Apr 2011).
10. *Megaputer Intelligence Application of Polyanalyst to Flight Safety Data at Southwest Airlines—Proof-of-concept Demonstration of Data and Text Mining*. http://flightsafety.org/files/flight_safety_data_sw.pdf (accessed 6 Jun 2011).
11. **Srivastava AN,** Akella R, Cruz S. *Enabling Discovery of Recurring Anomalies in Aerospace Problem Reports Using High-Dimensional Clustering Techniques*. http://ti.arc.nasa.gov/m/pub-archive/archive/IEEE_TextMiningPaper_vrs4.pdf (accessed 6 Jun 2011).
12. **Castle JP,** Stutz JC, McIntosh DM. Automatic discovery of anomalies reported in aerospace systems health and safety documents. *American Institute of Aeronautics and Astronautics 2007 Conference and Exhibit*. Rohnert Park, California: American Institute of Aeronautics and Astronautics, 2007. http://ti.arc.nasa.gov/m/pub-archive/1322h/1322%20(Castle).pdf (accessed 5 Jan 2012).
13. **Govindan M,** Van Citters AD, Nelson EC, *et al*. Automated detection of harm in healthcare with information technology: a systematic review. *Qual Saf Health Care* 2010;**19**:e11.
14. **D'Avolio LW,** Litwin MS, Rogers SO Jr, *et al*. Automatic identification and classification of surgical margin status from pathology reports following prostate cancer surgery. *AMIA Annu Symp Proc* 2007;**11**:160–4.
15. **Webber WC,** Cooper GF, Hanbury P, *et al*. Creating a text classifier to detect radiology reports describing mediastinal findings associated with inhalation anthrax and other disorders. *J Am Med Inform Assoc* 2003;**10**:494–503.
16. **Hripcsak G,** Austin JH, Alderson PO, *et al*. Use of natural language processing to translate clinical information from a database of 889,921 chest radiographic reports. *Radiology* 2002;**224**:157–63.
17. **Visweswaran S,** Hanbury P, Saul M, *et al*. Detecting adverse events in discharge summaries using variations on the simple bayes model. *AMIA Annu Symp Proc* 2003:689–93.
18. **Melton GB,** Hripcsak G. Automated detection of adverse events using natural language processing of discharge summaries. *J Am Med Inform Assoc* 2005;**12**:448–57.
19. **Botsis T,** Nguyen MD, Woo EJ, *et al*. Text mining for the Vaccine Adverse Event Reporting System: medical text classification using informative feature selection. *J Am Med Inform Assoc* 2011;**18**:631–8.
20. **Ong MS,** Magrabi F, Coiera E. Automated categorisation of clinical incident reports using statistical text classification. *Qual Saf Health Care* 2010;**19**:e55.
21. **Runciman WB.** The Australian patient safety Foundation. *Anaesth Intensive Care* 1988;**16**:114–16.
22. **Runciman WB.** Lessons from the Australian Patient Safety Foundation: setting up a national patient safety surveillance system—is this the right model? *Qual Saf Health Care* 2002;**11**:246–51.
23. **Runciman W,** Hibbert P, Thomson R, *et al*. Towards an international classification for patient safety: key concepts and terms. *Int J Qual Health Care* 2009;**21**:18–26.
24. **Sebastiani F.** Machine learning in automated text categorization. *ACM Computing Surveys* 2002;**34**:1–47.
25. **Joachims T.** Text categorization with support vector machines: learning with many relevant features. Machine Learning: ECML-98. *Lect Notes Comput Sci* 1998;**1398**:137–42.
26. **NSW Health.** *Severity Assessment Code (SAC) November 2005*. http://www.health.nsw.gov.au/pubs/2005/pdf/sac_matrix.pdf (accessed 6 Apr 2011).
27. **Akbani R,** Kwek S, Japkowicz N. Applying support vector machines to imbalanced datasets. *Lect Notes Comput Sci* 2004;**3201**:39–50.
28. **Weiss SM,** Indurkhya N, Zhang T, *et al*. *Text mining: Predictive Methods for Analyzing Unstructured Information*. Spinger, 2005:30.
29. **Fürnkranz J.** A study using n-gram features for text categorization. *Technical Report OEFAI-TR-98-30*. Wien, Austria: Austrian Research Institute for Artificial Intelligence, 1998. http://en.scientificcommons.org/43006206 (accessed 5 Jan 2012).
30. **Scholkopf B,** Smola AJ. *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond*. MIT Press, 2001:17.
31. *Weka 3: Data Mining Software*. http://www.cs.waikato.ac.nz/ml/weka (accessed 6 Apr 2011).
32. **Kish L.** *Statistical Design for Research*. John Wiley and Sons, 2004:113–14.
33. **Billings CE.** The NASA Aviation Safety Reporting System: lessons learned from voluntary incident reporting. In: *Proceedings of Enhancing Patient Safety and Reducing Errors in Health Care*. Chicago: National Patient Safety Foundation, 1999:97–100.

Automated identification of extreme-risk events in clinical incident reports

Mei-Sing Ong, Farah Magrabi and Enrico Coiera

J Am Med Inform Assoc 2012 19: e110-e118 originally published online January 11, 2012
doi: 10.1136/amiajnl-2011-000562