Original Article

Health Informatics Journal

Health Informatics Journal
1–17

DOI: 10.1177/1460458214555040
jhi.sagepub.com
SAGE

Statistical classification of drug incidents due to look-alike sound-alike mix-ups

Zoie Shui Yee Wong
City University of Hong Kong, Hong Kong

Abstract

It has been recognised that medication names that look or sound similar are a cause of medication errors. This study builds statistical classifiers for identifying medication incidents due to look-alike sound-alike mix-ups. A total of 227 patient safety incident advisories related to medication were obtained from the Canadian Patient Safety Institute's Global Patient Safety Alerts system. Eight feature selection strategies based on frequent terms, frequent drug terms and constituent terms were performed. Statistical text classifiers based on logistic regression, support vector machines with linear, polynomial, radial-basis and sigmoid kernels and decision tree were trained and tested. The models developed achieved an average accuracy of above 0.8 across all the model settings. The receiver operating characteristic curves indicated the classifiers performed reasonably well. The results obtained in this study suggest that statistical text classification can be a feasible method for identifying medication incidents due to look-alike sound-alike mix-ups based on a database of advisories from Global Patient Safety Alerts.

Keywords

International Classification for Patient Safety, look-alike sound-alike mix-ups, patient safety, statistical classifiers, text mining

Introduction

Medication error has been regarded as a major type of healthcare mishap putting patients at risk of harmful outcomes.[1] According to a study of medical incidents by the Japan Council for Quality Health Care (JCQHC), drug errors contributed to more than 70 per cent of the medical incidents that occurred from 2009 to 2011.[2] Medication names that look-alike sound-alike (LASA) are the most common cause contributing to medication errors.[3] A cause of a problem is the basic contributory factor that can be reasonably identified and prevented. By its nature, LASA mix-up is a cause that leads to medication error such that if it had not occurred, the adverse event would not have taken place.

Corresponding author:
Zoie Shui Yee Wong, Centre for Systems Informatics Engineering, City University of Hong Kong, Tat Chee Avenue, Kowloon Tong, Hong Kong.
Email: zoiewong@cityu.edu.hk; zoiesywong@gmail.com

There are six main types of medication errors found in pharmacological and pharmaceutical patient care; they are prescribing faults, prescription errors, transcription errors, dispensing errors, administration errors and 'across settings' errors.[4] Among all the causes, LASA mix-ups are regarded as one of the major causes of dispensing errors.[5–8] Common causes usually repeat themselves as error incidents that happen if not corrected properly. To deal with this, once the key contributory factor for incidents has been understood, management can employ administrative means to avoid the occurrence of the contributory factors and reduce the likelihood of the event recurring. This study investigates the potential feature sets associated with drug incidents due to LASA mix-ups using Global Patient Safety Alerts (GPSA). The findings may be beneficial to automate the identification of similar mediation-related harms using other incident reports in the future and may provide insights into design patient safety taxonomy on medication-related harm.

Background

Drug safety initiatives

The United States Pharmacopeia (USP) is a global cross-disciplinary platform that aims to promote higher quality, appropriate and reliable over-the-counter and prescription medicines.[1] According to the eighth annual MEDMARX data report released by the USP, more than 1400 commonly used drugs are associated with errors related to LASA drug names, and it has identified more than 3000 confusing drug name pairs. LASA drug mix-ups have drawn a lot of attention as it has been found that 1.4 per cent of errors due to LASA drug mix-ups have resulted in adverse and harmful patient outcomes.[9,10] The USP's Drug Error Finder provides a free searchable database of more than 1400 drugs involved in LASA errors with information on the potential severity of each error.

In addition, there is a reporting mechanism for events involving LASA medication incidents, namely, the Institute for Safe Medication Practices' (ISMP) National Medication Errors Reporting Program (MERP).[11] The ISMP has published a list of common confusing drug names on the Internet. These resources allow healthcare policy makers to devise strategies to reduce the risk of errors caused by LASA drugs. The LASA name-pair list is used to inform healthcare professionals which medications require special attention to reduce the risk of errors.

Currently, the Institute for Healthcare Improvement (IHI) has established adverse drug event trigger tools and the medication module of the Global Trigger Tool for detecting medication-related harm using sentinel words or 'trigger' for daily clinical operation.[12] A study has investigated the performance and effectiveness of the methods as various criticisms were widely reported One of the problems of this method is that the 'trigger' or sentinel words were selected by expert reviewers using the adverse drug event chart review sheet, and its inclusion may not be comprehensive enough. The study showed that the performance of the tool is enhanced when hospitals implement a refined list of 'triggers'. The study provided comments on further research directions to refine existing and explore new 'trigger' events.[13] Many healthcare facilities have been collecting textual incident reports on adverse events and near-misses, preparing patient safety incident advisories, alerts and recommendations to document medical incidents better. We believe that these resources can provide valuable information to refine a list of sentinel words in an evidence-based manner.

Reporting for patient safety

According to Heinrich's Law, for every accident that causes a major loss, there are 29 other accidents that result in minor injuries (incidents) as well as 300 that cause no injuries

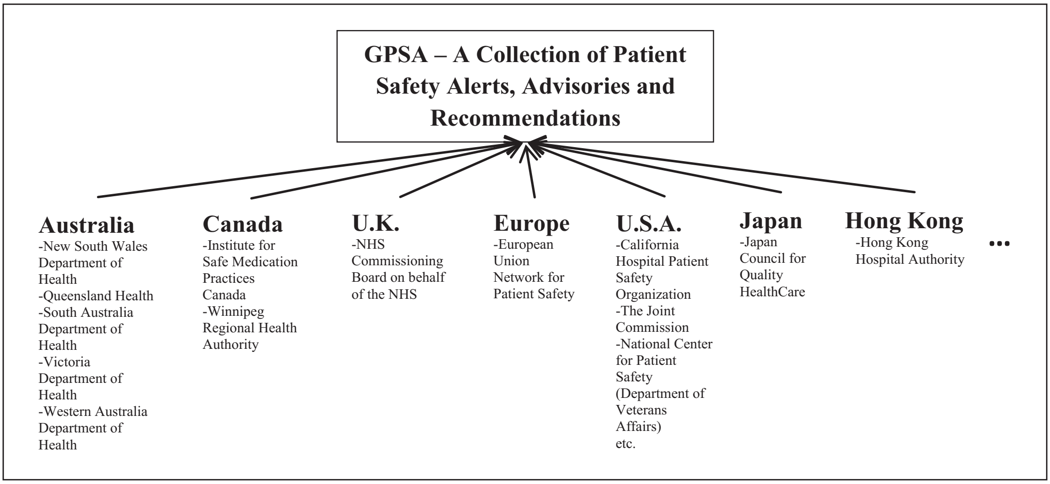

GPSA – A Collection of Patient Safety Alerts, Advisories and Recommendations
Australia
-New South Wales Department of Health
-Queensland Health
-South Australia Department of Health
-Victoria Department of Health
-Western Australia Department of Health
Canada
-Institute for Safe Medication Practices Canada
-Winnipeg Regional Health Authority
U.K.
-NHS Commissioning Board on behalf of the NHS
Europe
-European Union Network for Patient Safety
U.S.A.
-California Hospital Patient Safety Organization
-The Joint Commission
-National Center for Patient Safety (Department of Veterans Affairs)
etc.
Japan
-Japan Council for Quality HealthCare
Hong Kong
-Hong Kong Hospital Authority
•••

Figure 1. GPSA and its contributing organisations in the world.
GPSA: Global Patient Safety Alerts; NHS: National Health Service.

(near-misses). Analysing routine incidents and near-misses can prevent accidents that result in such injuries. Independent patient safety systems and reporting mechanisms for patient safety incidents have been established in the United Kingdom, Canada, Australia, Japan, Hong Kong and so on to capture the cause, progression and responses to various incidents and near-misses. For instance, the Advanced Incident Management System (AIMS) is an incident reporting system commonly used by many Australian public hospitals. The Project to Collect Medical Near-miss/Adverse Event Information is introduced by the Division of Adverse Event Prevention of the JCQHC documents near-miss incidents and adverse events by cooperating with Japanese medical institutions.

In Hong Kong, the Hospital Authority has implemented the Advanced Incident Reporting System (AIRS) to collect, manage, classify, analyse and monitor medical incidents for both adverse events and near-misses. In addition, many hospitals worldwide independently operate their own systems for voluntary reporting of patient safety incidents for clinical operations improvement and documentation purposes. To date, there are no mandates, controls or standards for reporting medical incidents across hospitals worldwide. The World Health Organization (WHO) Member States developed a World Health Assembly resolution on patient safety in 2002. The WHO workgroup on patient safety has been investigating a proper set of terminologies to increase the coherence and expressiveness of patient safety classification. A study has been carried out to model patient safety texts to understand the categorical structure and terminological systems.[14]

The GPSA system developed by the Canadian Patient Safety Institute with the support of the WHO is a global platform for sharing patient safety information with frontline healthcare providers and healthcare organisations worldwide. The web-based system contains a comprehensive collection of patient safety alerts, advisories and recommendations from the contributing organisations in Japan, Hong Kong, Australia, Canada, Demark, the United States and so on, as shown in Figure 1. It provides access to more than 1000 patient safety incident advisories, alerts and recommendations and groups the patient safety advisories into more than 20 topics, including medications, devices, surgery, care management, suicide, blood products/transfusions and so on. Among these, the highest number of alerts is medication errors.

A quantitative approach for situation awareness

There are a number of patient safety–related studies carried out using information technology–based incident report systems, while millions of incident reports are captured every day in hospitals from all over the world. However, many patient safety studies are still using manual review methods to screen out incident reports with respect to their own topic of interest for further study. The critical weakness of manually reviewing incident reports is that it is subjective, expensive, time-consuming and demanding in terms of human resources.

Text classification and statistical modelling have already been successfully applied to identify safety problems from aviation incident reports.[15] Automated classification of text in patient safety–related information is a new area.[16,17] Most of the available research on patient safety has been conducted in a qualitative manner. The majority of incident reports are not relevant for machine learning analysis in the first place, so a great deal of effort to blend know-how in natural language processing and relevant domain knowledge is required to extract signals from noise in medical text. Currently, most of the event detection studies have been carried out at term level, although a few studies have overcome the challenges of complex sentence identification and detect events at sentence level.[18] It is still challenging and barely sufficient to justify the correct classification of medical events from unstructured texts/records.[19] In reality, effective use of health informatics techniques for patient safety research is relatively undeveloped because such successful implementation requires access to relevant data (in most cases, incident reports are not publicly available) and extensive cross-disciplinary know-how, including machine learning, text mining, statistical modelling, ontology and domain clinical knowledge – traditionally, these are distinctly different disciplines.

Franklin et al.[20] used free text searching to screen potential reports related to the use of intravenous vinca alkaloids from 9 million reports for further study. Some recent studies have also detected adverse events or disorders using the texts of clinical reports.[17,21,22] Others have demonstrated that text classification methodologies can be used to automate the detection of extreme risk events in medical incident reports.[23,24] In the area of text mining related to patient safety subjects, some pioneer studies have shown that statistical text classifiers can be feasible for categorising clinical incident reports[23] and detecting extreme risk events[24] and health information technology incidents.[25]

While LASA drug mix-ups are one of the major causes of medications errors, it has yet to be determined if statistical classifiers can be built to efficiently detect cases with the same contributing factors. Also, the current GPSA patient safety-sharing platform cannot identify which advisories are associated with which particular cause. It is believed that screening medical incidents based on underlying causes may be helpful to healthcare professionals to access incidents with a similar cause, and the selected features for such classification may be insightful to develop patient safety taxonomies and refine ‘triggers’ for medication-related harm. This study attempts to build statistical classifiers for detecting incidents due to LASA mix-ups from the medication errors recorded in the GPSA system. The findings demonstrate the possibility of applying text-mining and statistical modelling techniques to automate the classification of medical errors with particular causes based on incident text data. The author believes that this study can be insightful for identifying specific sentinel words contributing to LASA medication-related harm.

Method

A statistical text classifier for identifying medication incidents due to LASA mix-ups from the GPSA data was developed using various modelling methods. The precise procedure and

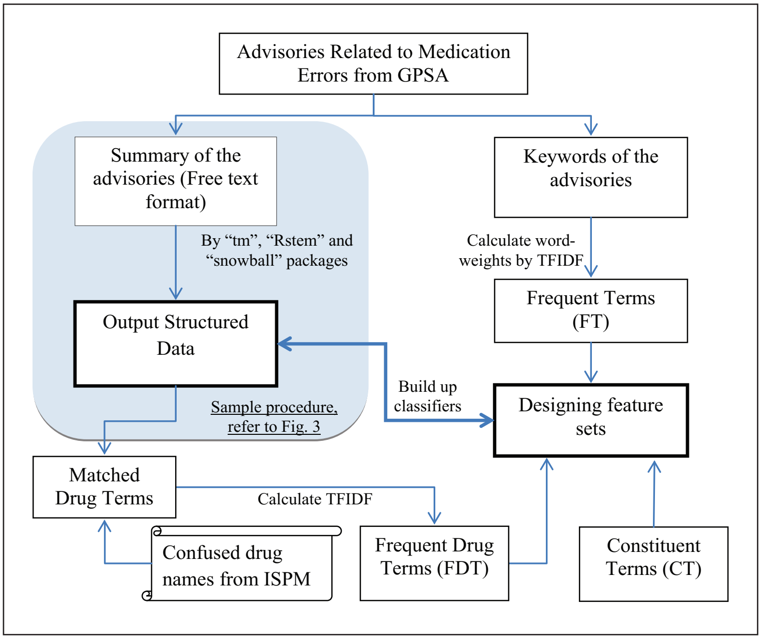

Advisories Related to Medication Errors from GPSA
Summary of the advisories (Free text format)
Keywords of the advisories
By "tm", "Rstem" and "snowball" packages
Calculate word-weights by TFIDF
Frequent Terms (FT)
Output Structured Data
Build up classifiers
Designing feature sets
Sample procedure, refer to Fig. 3
Matched Drug Terms
Calculate TFIDF
Confused drug names from ISPM
Frequent Drug Terms (FDT)
Constituent Terms (CT)

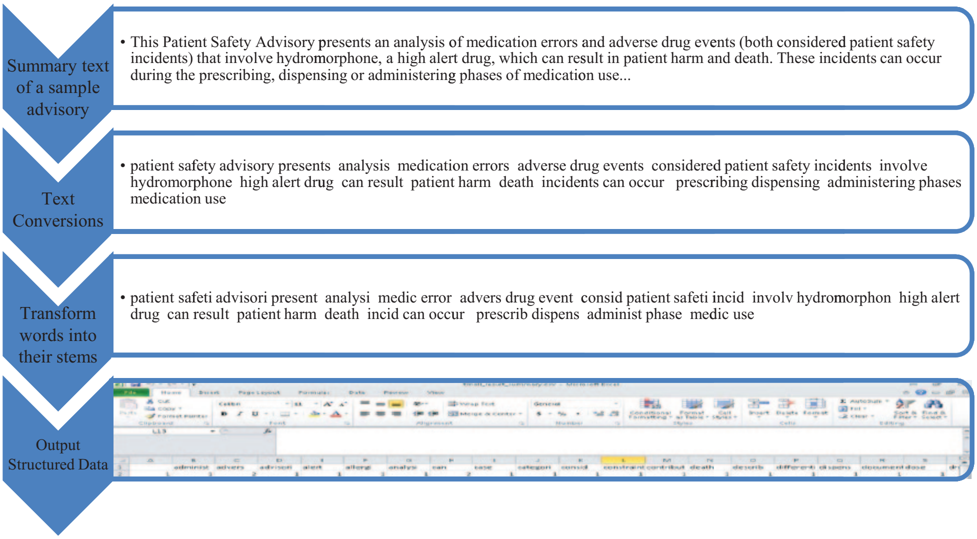
Figure 2. Procedure of structured data preparation.
GPSA: Global Patient Safety Alerts; TFIDF: term frequency–inverse document frequency; ISPM: Institute of Social and Preventive Medicine.

implementation is addressed in the following sections. It includes structuring unstructured text data, selection of classifiers, feature selection and performance evaluation metrics.

Structuring unstructured free texts

A total of 227 patient safety incident advisories related to medications were extracted from the GPSA of the Canadian Patient Safety Institute in October 2012. The advisories contained basic incident information such as the title, keywords, country and contributing organisation. The descriptive text content included a summary and actions for reducing risk in terms of communication, training, environment, equipment, rules/policies/procedures and fatigue/scheduling. Most of the cases contained text documents drawn from the contributing organisations, although not all of the advisories contained all of the information listed above. We applied text-mining and statistical modelling methods to the summary text of the advisories. All missing information was classified as 'void'.

Figure 2 demonstrates the procedure of text data preparation that was used. A transformation was performed on all documents in the corpus, and the free text summary of the advisories was processed using a standard text-mining procedure. The original paragraphs were processed by converting to lower case, trimming word-spacing and removing English stop-words, numbers and punctuation. As the same word may appear in different forms, a stemming process was applied to transform words into their stems. The presence of a particular stemmed term was classified as dichotomous, that is, present (1) or absent (0). A structured text outcome summarising the term document matrix was then ready for further use as the source of the potential predictor variables for the classifier. The text-mining analysis requires R packages, namely, 'tm', 'snowball' and 'Rstem'. Figure 3 illustrates the phases of text conversion using a sample incident text.


Summary text of a sample advisory
Text Conversions
Transform words into their stems
Output Structured Data

- This Patient Safety Advisory presents an analysis of medication errors and adverse drug events (both considered patient safety incidents) that involve hydromorphone, a high alert drug, which can result in patient harm and death. These incidents can occur during the prescribing, dispensing or administering phases of medication use...

- patient safety advisory presents analysis medication errors adverse drug events considered patient safety incidents involve hydromorphone high alert drug can result patient harm death incidents can occur prescribing dispensing administering phases medication use

- patient safeti advisori present analysi medic error advers drug event consid patient safeti incid involv hydromorphon high alert drug can result patient harm death incid can occur prescrib dispens administ phase medic use

Figure 3. Sample advisory and text-mining procedure.

Meanwhile, we created the structured text outcome for the keywords of the advisories using the above-mentioned text-mining procedure. As some words are more common than others, a weighting approach was used to understand the properties of the corpus. The weights were determined by term frequency–inverse document frequency (TFIDF), a normalised word frequency statistic indicates how important a word is to a document in the corpus. TFIDF word weighting of each term was computed and a group of frequent terms (FT) was generated as one feature set (further details can be found in the section 'Feature selection').

Cross-checking with Institute of Social and Preventive Medicine (ISPM) drug terms, the feature set of frequent drug terms (FDT) was selected. The procedure aims to develop a set of matched drug terms that were both listed in ISPM[11] and found in the structured text outcome of the summary of the advisories. Similarly, TFIDF among the matched drug terms were calculated to check how frequently LASA drug terms occur in the corpus. The top TFIDF terms were selected during the feature selection stage. The above algorithms are available upon request. We further elaborate the design of the feature set in the section 'Feature selection'.

Classifiers

Statistical text classifiers based on logistic regression (LR), support vector machines with linear (L.SVM), polynomial (P.SVM), radial-basis (R.SVM) and sigmoid (S.SVM) kernels, and decision tree (DT) were trained and tested on the GPSA data to automate the detection of LASA-related advisories. A binary label where one observation can be either positive (LASA cases) or negative (non-LASA cases) was designed. The outcome (response) variable is binary, that is, it is either a LASA case (1) or not (0). 'Target' means the class label of a case, and the objective of classification is to predict the target class for a case. The dichotomous outcome variable of each advisory was determined by careful manual review of all of the advisories by an experienced patient safety researcher with a PhD degree in the relevant discipline, more than 3 years' experience in researching

Table 1. Classifier outcomes compared to manual review.

Class	Assignment	
	+	−
+	True positive (TP)	False negative (FN)
−	False positive (FP)	True negative (TN)

patient safety–related texts and has extensive experience in public health research. This baseline was determined by two sets of reviews of the entire document set. At the first stage, the researcher scrutinised the entire document set and marked the tabbed LASA cases as the first identifiers. Another inspection was carried out 6 months after the first review by the same researcher. A second identifier was checked in relation to the identified LASA cases. The two identifiers were then cross-checked in order to ensure a consistent representation of the class label of LASA cases. In the case of inconsistencies, the researcher carried out additional detailed inspection of the documents. The labels for both review class and assignment are listed in Table 1.

Several classifiers with empirically good performance in text-mining and classification research were selected in this study.[23–26] LR is a popular and powerful discriminative classifier for modelling binary response data. Support vector machines have been shown to be powerful tools for modelling complex and real-world problems in text and image classification, handwriting recognition, and bio-sequence analysis.

Various high-dimensional feature spaces were considered in the support vector machine analyses using linear, polynomial, radial-basis and sigmoid kernels. To construct a DT classifier, each leaf node is a classification decision, and non-terminal nodes contain tests that separate observations based on attribute values. In this study, LR, L.SVM, P.SVM, R.SVM, S.SVM and DT were performed using R packages.

Feature selection

When building a text classifier with too many candidate variables, it is important to observe the relevant features to create a reasonably good model. By incorporating such additional knowledge on the characteristics of LASA incidents, the feature selection process can help improve the performance of statistical learning models by alleviating the effect of dimensionality, improving generalisation by reducing over-fitting, reducing the learning time and enhancing model interpretation.

The selection of features to create this predictive model was especially challenging due to the abundance of available terms that can be treated as candidate variables. Selecting a set of relevant features to build a robust statistical learning model is a complicated process. The text-mining results often contain a large number of possible parameters, that is, a large number of potential variables with relatively few observations. Too many features relative to the number of observations may result in over-fitting. In this study, the feature selection strategy was developed based on the characteristics of the LASA mix-up incidents. The output terms were categorised into several feature sets, including the following:

- *Constituent terms (CT)*. A source of potential variables was identified through the constituents of LASA. As a result, the terms ‘look’, ‘alik(e)’ and ‘sound’ were selected.
- *FT*. A list of potential variables was selected based on analysing the keywords of the advisories by word weights. The top 10 TFIDF terms from this analysis were considered in the variable selection process.

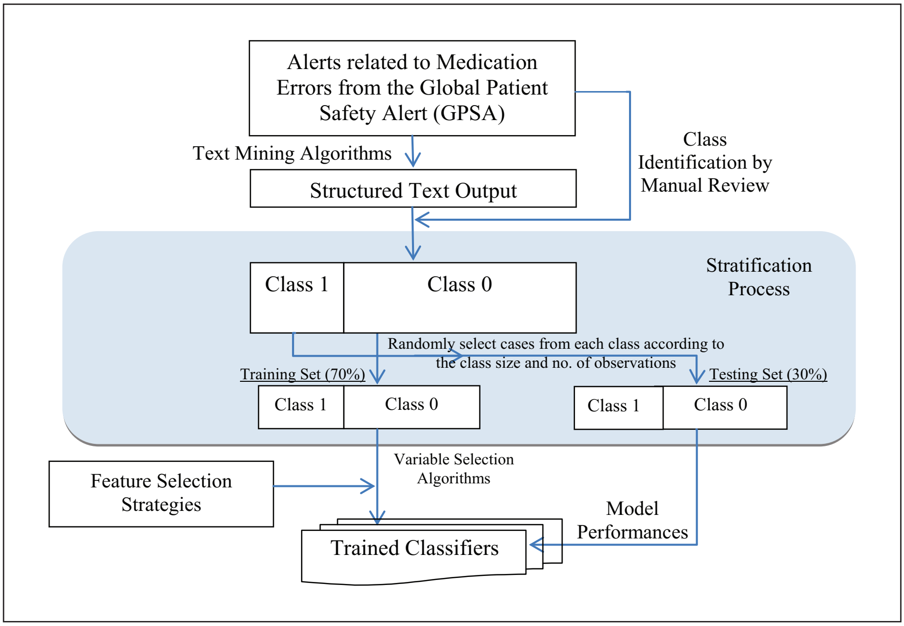

Alerts related to Medication Errors from the Global Patient Safety Alert (GPSA)
Text Mining Algorithms
Structured Text Output
Class Identification by Manual Review
Stratification Process
Class 1
Class 0
Randomly select cases from each class according to the class size and no. of observations
Training Set (70%)
Testing Set (30%)
Class 1
Class 0
Class 1
Class 0
Variable Selection Algorithms
Feature Selection Strategies
Model Performances
Trained Classifiers

Figure 4. Flowchart of training and testing classifiers.

- *FDT*. A list of common LASA drug names, which were generated by fully cross-checking drug names in the corpus with the name pairs in the ISMP's list of confusing drug names,[11] was considered. The top 10 TFIDF matched drug names were regarded as FDT.

In this study, we tested for all the seven combinations of the three feature sets. One additional feature set was developed using the full feature combination with the Stepwise Akaike Information Criterion (AIC) approach to search for an optimal model that best fit the data according to the minimal AIC value.

Performance evaluation

Stratification was shown to be successful in achieving better performance of the classifiers when identifying health information technology incidents in the past.[25] This study adopted stratified (10-fold) cross-validation. The available data were split into two non-overlapping parts. In order to minimise loss of information and avoid over-fitting using this imbalanced dataset (only 21% of them related to LASA cases), we selected random samples of 70 per cent training and 30 per cent testing datasets from each class based on a number proportional to the class size compared with the total number of observations; the stratification process is illustrated graphically in Figure 4. The above procedure was repeated 10 times in each model setting, and we calculated the overall model performance measures based on the average of the model's results.

Precision, which is also known as positive predictive value, represents the fraction of retrieved instances that are relevant, whereas *recall* (or sensitivity, or true-positive rate (TPR)) is the fraction of actual positives which are correctly identified. *F-score*, which is expressed as the harmonic mean of precision and recall, is also a measure of performance of the classifier. In binary classification, *accuracy* is a statistical measure that expresses how well a classifier correctly identifies a condition. The mathematical expressions of the evaluation metrics are as follows

$$\text{Precision} = \frac{\text{TP}}{(\text{TP} + \text{FP})}$$

$$\text{Recall} = \frac{\text{TP}}{(\text{TP} + \text{FN})}$$

$$\text{F-score} = \frac{(2\text{PR})}{(\text{P} + \text{R})}$$

$$\text{Accuracy} = \frac{(\text{TP} + \text{TN})}{(\text{TP} + \text{TN} + \text{FP} + \text{FN})}$$

A receiver operating characteristic (ROC) curve is a graphical method to explore the changes of sensitivity and specificity as the test thresholds vary. The entire analysis was conducted on a R v2.14.0 (64-bit) platform using the 'tm', 'Rstem', 'snowball', 'epicalc', 'MASS', 'e1071', 'ROCR', 'rpart', 'Rcmda', 'klaR' and 'caret' packages.[27]

Results

In this section, the results of the statistical text classifier are presented in two sub-sections: structured text outputs and classification performance.

Structured text outputs

Among the 227 advisories, 48 were regarded as LASA cases after careful manual review and double-checking. From the text-mining results of the summary of advisories, 2004 structured terms were obtained. As the predictor variables of interest might occur in the extracted wordings from the summary, we attempted to select the relevant features in predicting the reporting of LASA errors based on the strategies of using FT, FDTs and CT. The ISMP's list of confusing drug names was carefully cross-checked against the drug names that appeared in the corpus, and 52 matched drug names were found. For those drug names that appeared in two or more individual words, the frequency of occurrence of the mined outcomes against the original corpus was confirmed.

The top 50 TFIDF terms identified by analysing the keywords of the advisories and top 50 TFIDF matched drug names identified by analysing the summary of the advisories are displayed in Figure 5. It was found that the TFIDF scores for both series dropped dramatically within 10 top score terms, indicating that a limited number of terms were used more frequently than others. The lines gradually decreased and levelled off, indicating that the used terms were more and more diverse. The top 10 TFIDF terms identified by analysing the keywords of the advisories were 'label', 'dose', 'wrong', 'storag', 'dispens', 'drug', 'medic', intraven', 'packag' and 'differenti'. We found that the top 10 TFIDF matched drug names were 'heparin', 'fentanyl', 'epinephrin', 'morphin', 'hydromorhon', 'immunoglobulin', 'amphotericin', 'salin', 'codein' and 'ephedrine'.

All combinations of feature sets using FT (top 10 TFIDF terms identified by analysing the keywords of the advisories and the summary of the advisories), FDT (top 10 TFIDF terms identified by analysing cross-checked drug terms with ISMP's list of confusing drug names) and CT ('look', 'alik(e)' and 'sound') were considered. Variable selection using Stepwise AIC by all feature sets was undertaken, and a minimal AIC value of 80.32 was achieved. The selected variables were used to carry out the feature set of 'CT+FT+FDT (StepAIC)'.

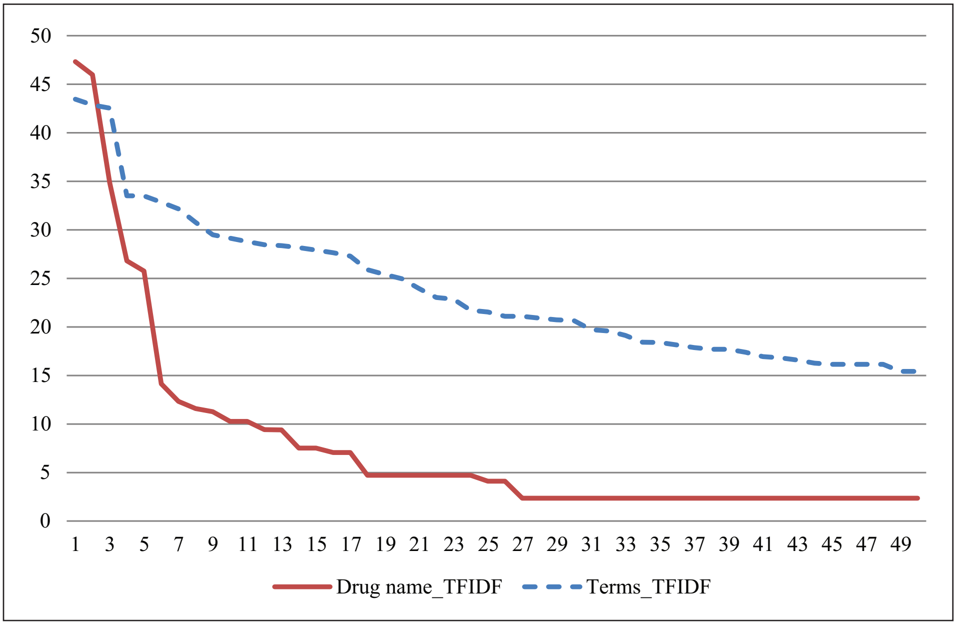


Figure 5. TFIDF scores for the top 50 TFIDF terms (*y*-axis: TFIDF score and *x*-axis: terms).
TFIDF: term frequency–inverse document frequency.

Classification performance

Eight feature strategies were tested using LR, L.SVM, P.SVM, R.SVM, S.SVM and DT. Table 2 and Figure 6 illustrate the overall performance measure results.

From the above results, it was found that the developed models achieved an average accuracy of 0.78 or above across all the model settings. The best average precision of 0.895 and accuracy of 0.861 were achieved by the LR model with a CT feature set. The LR model using selected variables based on all feature sets achieved the highest average recall (also known as sensitivity) of 0.529 and F-score of 0.575, while its average precision and accuracy were found to be 0.644 and 0.841, respectively. This model setting achieved improved rates of recall of 40 per cent and F-score of 8 per cent compared with model setting by ‘CT’. The best average TPR measured by recall was achieved by the ‘CT+FT+FDT (StepAIC)’ feature set using the LR method, whereas the best average true-negative rate (TNR) measured by specificity (0.986) was achieved by the ‘CT’ feature set using the LR method.

Due to the skewed data property discussed previously, we checked the ROC curve in addition to the performance measure metrics. We included the best performing classifiers based on model selection criteria of top two recall, precision, F-score and accuracy values and created the ROC curve to illustrate the performance of the classifiers for different discrimination thresholds. The models that we took into consideration were as follows: the ‘CT + FT + FDT (StepAIC)’ feature set using the LR method, ‘CT’ feature set using the LR and R.SVM methods and ‘CT + FT + FDT’ feature set using the LR method. Figure 7 shows the ROC curve of the best four classifiers plotted on the same graph. In this kind of classification, it is preferable to acquire models that obtain better sensitivity, that is, are able to correctly identify LASA cases as positive. However, while the TPR

Table 2. Experiment results.

Methods	CT						FT					
	LR	L.SVM	DT	P.SVM	R.SVM	S.SVM	LR	L.SVM	DT	P.SVM	R.SVM	S.SVM
Recall	0.379	0.343	0.343	0.343	0.393	0.374	0.230	0.234	0.336	0.264	0.164	0.137
Precision	0.895	0.894	0.894	0.894	0.897	0.894	0.517	0.589	0.538	0.585	0.629	0.453
F-score	0.533	0.496	0.496	0.496	0.547	0.527	0.318	0.335	0.414	0.364	0.260	0.210
Accuracy	0.861	0.854	0.855	0.854	0.864	0.860	0.793	0.805	0.795	0.804	0.801	0.784
	FDT						CT + FT					
	LR	L.SVM	DT	P.SVM	R.SVM	S.SVM	LR	L.SVM	DT	P.SVM	R.SVM	S.SVM
Recall	0.223	0.196	0.083	0.101	0.089	0.073	0.396	0.341	0.370	0.311	0.344	0.329
Precision	0.661	0.676	0.428	0.618	0.562	0.389	0.697	0.785	0.764	0.822	0.806	0.665
F-score	0.533	0.496	0.496	0.496	0.547	0.527	0.318	0.335	0.414	0.364	0.260	0.210
Accuracy	0.813	0.810	0.791	0.801	0.800	0.787	0.836	0.841	0.837	0.842	0.844	0.825
	FT + FDT						CT + FDT					
	LR	L.SVM	DT	P.SVM	R.SVM	S.SVM	LR	L.SVM	DT	P.SVM	R.SVM	S.SVM
Recall	0.399	0.353	0.306	0.144	0.126	0.130	0.458	0.417	0.343	0.201	0.360	0.349
Precision	0.610	0.596	0.502	0.675	0.644	0.515	0.770	0.751	0.782	0.821	0.817	0.798
F-score	0.533	0.496	0.496	0.496	0.547	0.527	0.563	0.526	0.464	0.515	0.488	0.461
Accuracy	0.816	0.811	0.787	0.808	0.800	0.793	0.858	0.849	0.843	0.825	0.848	0.841
	CT + FT + FDT						CT + FT + FDT (StepAIC)					
	LR	L.SVM	DT	P.SVM	R.SVM	S.SVM	LR	L.SVM	DT	P.SVM	R.SVM	S.SVM
Recall	0.511	0.437	0.354	0.240	0.350	0.281	0.529	0.441	0.364	0.191	0.349	0.240
Precision	0.633	0.680	0.755	0.813	0.835	0.798	0.644	0.715	0.757	0.776	0.839	0.799
F-score	0.533	0.496	0.496	0.496	0.547	0.527	0.575	0.545	0.477	0.509	0.468	0.399
Accuracy	0.835	0.836	0.839	0.829	0.843	0.830	0.841	0.845	0.839	0.821	0.845	0.824

FT: frequent terms; LR: logistic regression; CT: constituent terms; DT: decision tree; FDT: frequent drug terms.

increases, the false-positive rate inevitably decreases. From Figure 7, the ROC curves indicate that these classifiers performed reasonably well.

Discussion

Different medical incident reporting systems exist among various health organisations and hospitals, and medical incident text data have not been adequately researched. The GPSA system is a global platform for sharing patient safety with frontline healthcare providers, healthcare organisations and healthcare policy makers, and it can be regarded as providing a reference structure and format for reporting medical incidents. Healthcare professionals from various disciplines from all over the world can access the alerts and advisories freely. The current categorisation method of the GPSA system cannot determine which advisories are associated with what kind of causes of the medical errors, even though the advisories are communicated to healthcare professionals and organisations around the world that may report or manage medical incidents. Manual reviews of incidents are

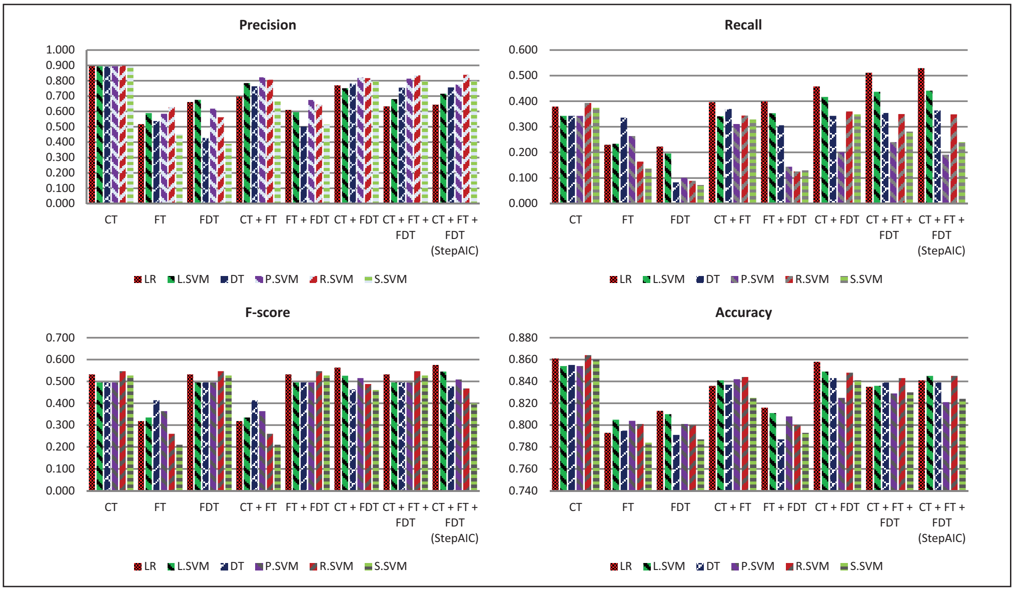


Figure 6. Graphical display of experiment results.
FT: frequent terms; CT: constituent term; FDT: frequent drug term; LR: logistic regression; L.SVM: support vector machines with linear kernel (L.SVM), polynomial (P.SVM), radial-basis (R.SVM) and sigmoid (S.SVM) kernels; P.SVM: support vector machines with linear polynomial kernel; R.SVM: support vector machines with radial-basis kernels; S.SVM: support vector machines with sigmoid kernel; DT: decision tree.

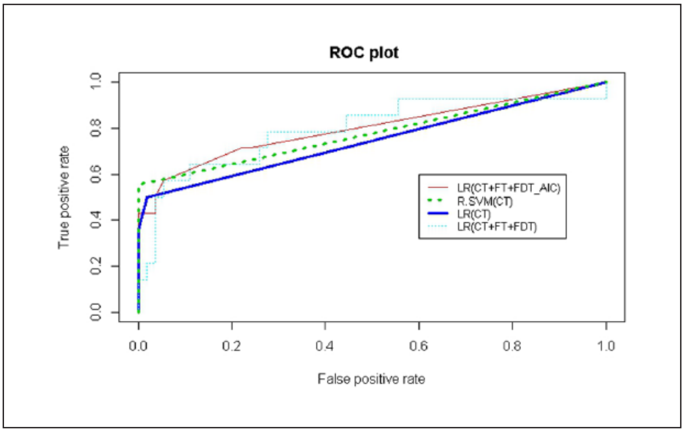


Figure 7. ROC curve plot.
ROC: receiver operating characteristic; FT: frequent terms; LR: logistic regression; CT: constituent terms; DT: decision tree; FDT: frequent drug term.

time-consuming and non-exhaustive, and there is no comprehensive solution to automate learning from reports. Understanding the causes of an incident and its ability to recognise cases with related causes are essential to healthcare system improvement. A statistical text classifier built under this platform will be able to create access to patient safety alerts with certain contributing causes.

In this study, statistical classifiers to identify incidents due to LASA mix-ups using structured text outputs from the GPSA system were developed. The developed models using LR, L.SVM, P.SVM, R.SVM, S.SVM and DT achieved an average accuracy of more than 0.8 across all the model settings. The LR model using selected variables based on all feature sets improved the recall rate and F-score but reduced the precision and accuracy. In general, the classifiers achieved a good TNR (the mean of specificities among all classifiers was 0.962) and only a moderate TPR was found using LR with all feature sets ('CT + FT + FDT') and selected variables based on all the feature sets ('CT + FT + FDT (StepAIC)') (the mean of sensitivities under the two cases was 0.520). The developed models achieved 81 per cent accuracy overall using the training and testing datasets, and the ROC curve indicated that the four best models are satisfactory.

The results of this study indicate that statistical text classification can be a feasible method for identifying medication incidents due LASA mix-ups based on the database of advisories from GPSA. The selected feature set can be used as a reference in the future for other categorisation of incidents due to LASA mix-ups and should enable healthcare professionals to scrutinise the occurrence of similar incidents together, thus enabling a comprehensive understanding of the causes of and responses to incidents due to such cause.

Limitations, implications and directions

In this study, the data came from a relatively small number of medical incident contributions from a number of medical incident databases from all over the world. It should be noted that, in accord with common practice,[23] the outcome variable was manually determined by the researcher, which may potentially result in input errors on the binary class label. Moreover, the data were skewed as they unequally represented classes of LASA and non-LASA cases. The number of LASA cases was relatively small (21%) compared with non-LASA cases. The small number of observations compared with the large number of potential variables, diverse sources of alerts and the imbalanced property of the dataset increased the challenge of developing accurate text classifiers that identify rare events. There is a large number of types of medical incident in healthcare settings, and many of them do not occur or get detected frequently.[25] It is also common to find imbalanced data in medical incident reports. Future directions of study could include exploring new statistical methods that focus on imbalanced classification problems to improve the performance of rare event classification.[28,29] Besides, unsupervised learning methods[30] for analysing incident-free text can be promising, while most of the incident reports are written in a freestyle format, that is, non-standardised and current report platforms usually fail to categorise them systematically and properly.

The International Classification for Patient Safety (ICPS) developed by the WHO in 2009 is a conceptual framework for the standardised classification of key patient safety concepts.[31] The goal of the ICPS is to enable the methodical categorisation of patient safety information using widely agreed definitions, preferred terms and a distinct ontology.[14] Serving as the basis for the Information Model for Patient Safety, some studies[32,33] have adopted a knowledge engineering approach to develop a semantic model that underpins the concepts of the existing patient safety framework. The proposed methodology and reference models for classifying medication incidents due to LASA provide insights that will facilitate automatic categorisation using alternative textual medical incident databases. Together with the ontological approach, such as using WordNet to trace metonyms, synonyms and hypernyms, further studies can be carried out to standardise patient safety taxonomy for medication-related errors due to LASA.

The selection of variables is challenging when thousands of mined terms are available. In this study, reasonably good predictive models were obtained because relevant features related to the subject were considered, that is, CT and the frequent matched LASA drug name terms. It is recommended that causes of incidents, relevant medication names and other effective identifiers be included in the future reporting of incidents relating to medication errors. Research efforts should focus on investigating a well-structured and standardised method and procedure for medical incident reporting to facilitate effective collective knowledge concerning medical incidents.

Based on various discussions with patient safety practitioners and experts, it is understood that the current medical incident reporting systems group incidents in relatively brief and generalised categories, such as medication incident, patient falls and so on (some systems provide different reporting templates). It is known that healthcare is complicated, and incidents may occur due to various root causes and progression of events. Important common grounds and details usually are not recorded structurally in many reporting systems. The recent minimal information model, which is being examined by the WHO, aims to simplify the reporting process while meaningful data categories are captured.[34] To date, free texts still dominate the report contents and inevitably increase the difficulties of identifying and scrutinising related cases together. The ability to identify particular cause of errors (such as LASA) can be a meaningful step that facilitates a more focused investigation of the same type of medical incidents[20] to design effective intervention strategies.

This study also demonstrates the feasibility of carrying out text-mining and statistical modelling techniques to categorise medical errors caused by LASA automatically. Furthermore, it is believed that the statistical text classification method will be extensible to other causes contributing to medical incidents and transferable to other medical incident datasets. The concepts can be applied to other types of medical incidents of a similar nature. Incorporating additional relevant features that are closely associated with the incident properties is essential for developing similar kinds of robust medical incident classifiers.

The third version of AIRS operated by Hong Kong Hospital Authority was launched at the beginning of 2014. Since the first introduction of AIRS in Hong Kong, the system has accumulated more than 10,000 free text reports in a year. Statistical learning using these incident-free text is challenging, and, due to technical difficulties and lack of demonstrated outcome, analytics using a quantitative approach (such as text-mining and modelling methods) has never been done on such a free text dataset.

This centralised patient safety information collection system includes several major incidents' templates and collects both clinical near-miss and clinical incident occurrences in the public hospitals in Hong Kong. The current system has been collecting free text related to medication incidents – which can have the most severe consequences, comparing with other types of incidents. Although there are some structured items captured in the system, those categorised items are far too abstract/rough to single out every incident with many possible causes and progression of events, and thus provide very few insights in root cause analysis. For instance, the system can only distinguish three types of events: *prescribing, dispensing* and *administration*. Some important event types, such as LASA cases and self-medication cases, cannot be identified from the system as those class labels are not recorded structurally. Currently, all the incident reports have to be reviewed and screened manually by quality and safety staff. Manual review has been criticised in that it is subjective, expensive, time-consuming and human resources demanding. With such review workload and unstructured free text format, the analysis of past similar incident records over time becomes extremely difficult. Building on this work and related studies, in the future, these demonstrated outcomes can be extended to practical patient safety free text analysis using real text data. Future research will be of high significance and establish a better position from which to single out cases of a similar nature (such as LASA cases), understand the reporting patterns from healthcare professionals, figure out a systematic free text reports review method and collect the most relevant information related to incidents. Further developments based on the outcomes of this pilot study may give insights into future development of minimal information models for incident reporting and effective incident template design so as to increase patient safety situational awareness.

There are various patient safety reporting systems in place in different developed countries/territories, such as Australia, the United Kingdom, Japan and Hong Kong. However, healthcare professionals may exhibit huge differences in terms of reporting style and terminology used, while their national and cultural contexts and English language proficiency may be greatly different. The systematic approach suggested here will be able to help single out similar cases and set up a draft standard for guiding an appropriate reporting standard/terminology for similar types of events.

Text mining and machine learning for analysing medical incident texts can provide insights that can help derive a taxonomy of medical incidents. Currently, some medical incident–related texts are publicly available but most are private. When the data confidentiality issue is resolved, it is envisioned that in the future, frontline healthcare providers, health products manufacturers (including drug firms), healthcare organisations and healthcare policy makers will be in a better position to understand the characteristics of medical incidents and structure incident reporting of related incidents. A cross-disciplinary research effort involving close collaboration between text miners, statisticians, ontologists and practitioners is essential if this is to be achieved.

Conclusion

This research successfully built statistical text classifiers to distinguish incidents due to LASA mix-ups from GPSA. The results demonstrate the feasibility of applying text-mining and statistical learning techniques to automate the classification of the causes of medical errors based on medical incident texts.

Acknowledgements

The author would like to thank the two anonymous reviewers for their helpful and constructive comments that contributed to improve the final version of the article and also thank the editor for his support during the review process. The author thanks Dr Abhishek K. Shrivastava, Assistant Professor at Florida A&M University – Florida State University for his helpful comments on a draft article.

Funding

The work described in this article was supported by the Japan Society for the Promotion of Science (JSPS) – FY2011 JSPS Postdoctoral Fellowship for Foreign Researchers (ID no. P11411).

References

1. Thompson CA. USP says thousands of drug names look or sound alike. *Am J Health Syst Pharm* 2008; 65(5): 386–388.
2. Wong ZSY, Fujita K and Akiyama M. Prioritization of type of medical incidents: a preliminary result. In: *The 32nd joint conference on medical informatics*, Niigata, Japan, 15–17 November 2012, p. 234.
3. McCoy LK. Look-alike, sound-alike drugs review: include look-alike packaging as an additional safety check. *Joint Comm J Qual Patient Saf* 2005; 31(1): 47–53.
4. Van den Bemt PMLA and Egberts ACG. Drug related problems: definitions and classification. *Eur J Hosp Pharm Pract* 2007; 13: 62–64.
5. Beso A, Franklin BD and Barber N. The frequency and potential causes of dispensing errors in a hospital pharmacy. *Pharm World Sci* 2005; 27: 182–190.
6. Knudsen P, Herborg H, Mortensen AR, et al. Preventing medication errors in community pharmacy: root-cause analysis of transcription errors. *Qual Saf Health Care* 2007; 16: 285–290.
7. Cheung K-C, Bouvy1 ML and De Smet PAGM. Medication errors: the importance of safe dispensing. *Br J Clin Pharmacol* 2009; 67(6): 676–680.

8. James KL, Barlow D, McArtney R, et al. Incidence, type and causes of dispensing errors: a review of the literature. *Int J Pharm Pract* 2009; 17(1): 9–30.
9. Kim S. US Pharmacopeia 8th Annual MEDMARX(R) Report indicates look-alike/sound-alike drugs, 2008, http://us.vocuspr.com/Newsroom/ViewAttachment.aspx?SiteName=USPharm&Entity=PRAsset &AttachmentType=F&EntityID=105435&AttachmentID=6b770787-571e-4dc1-ba3b-eb21d7bac147
10. Kaufman MB. Preventable medication errors: look-alike/sound-alike mix-ups. Formulary (serial on the Internet), 2011, http://formularyjournal.modernmedicine.com/formulary/article/articleDetail.jsp?id=579387 (accessed 12 December 2012).
11. Institute for Safe Medication Practices. List of confused drug names, 2011, https://www.ismp.org/tools/confuseddrugnames.pdf (accessed 12 December 2012).
12. Rozich JD, Haraden CR and Resar RK. Adverse drug event trigger tool: a practical methodology for measuring medication related harm. *Qual Saf Health Care* 2003; 12: 194–200.
13. Carnevali L, Krug B, Amant F, et al. Performance of the adverse drug event trigger tool and the global trigger tool for identifying adverse drug events: experience in a Belgian hospital. *Ann Pharmacother* 2013; 47(11): 1414–1419.
14. Souvignet J, Bousquet C, Lewalle P, et al. Modeling patient safety incidents knowledge with the categorical structure method. *AMIA Annu Symp Proc* 2011; 2011: 1300–1308.
15. Wong ZSY and Akiyama M. Patient safety systems between Japan and Hong Kong. In: *ANQ congress 2012: program book* (ed ATC Wong and KS Chin), Hong Kong, 30 July–3 August 2012. Hong Kong: Hong Kong Society for Quality, p. 118.
16. Wong ZSY and Akiyama M. Statistical text classifier to detect specific type of medical incidents. In: *Medinfo 2013. Proceedings of the 14th world congress on medical and health informatics* (eds Lehmann CU, Ammenwerth E and Nøhr C), Copenhagen, August 2013. DOI: 10.3233/978-1-61499-289-9-1053.
17. 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: 160–164.
18. Naughton M, Stokes N and Carthy J. Sentence-level event classification in unstructured texts. *Inform Retrieval* 2010; 13(2): 132–156.
19. Cogley J, Stokes N, Carthy J, et al. Analyzing patient records to establish if and when a patient suffered from a medical condition. In: *Proceedings of the 2012 Workshop on Biomedical Natural Language Processing (BioNLP 2012)*, Montreal, Canada, June 8, 2012, pp. 38–46. Association for Computational Linguistics.
20. Franklin BD, Panesar SS, Vincent C, et al. Identifying systems failures in the pathway to a catastrophic event: an analysis of national incident report data relating to vinca alkaloids. *BMJ Qual Saf* 2014; 23(9): 765–772.
21. 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.
22. Melton GB and Hripcsak G. Automated detection of adverse events using natural language processing of discharge summaries. *J Am Med Inform Assoc* 2005; 12: 448–457.
23. Ong MS, Magrabi F and Coiera E. Automated categorization of clinical incident reports using statistical text classification. *Qual Saf Health Care* 2010; 19(6): e55.
24. Ong MS, Magrabi F and Coiera E. Automated identification of extreme-risk events in clinical incident reports. *J Am Med Inform Assoc* 2012; 19(e1): e110–e118.
25. Chai KEK, Anthony S, Coiera E, et al. Using statistical text classification to identify health information technology incidents. *J Am Med Inform Assoc* 2013; 20: 980–985.
26. Wei Q and Collier N. Towards classifying species in systems biology papers using text mining. *BMC Res Notes* 2011; 4(32): e1–e8.
27. R-Project. Contributed packages, 2014, http://cran.r-project.org/web/packages/ (accessed 17 May 2014).
28. Zhao Y and Shrivastava AK. Combating sub-clusters effect in imbalanced classification. In: *Proceedings of the 2013 IEEE 13th international conference on data mining (ICDM) 2013*, Dallas, TX, 7–10 December 2013, pp. 1295–1300. New York: IEEE.

29. Zhao Y, Tsui KL, Shrivastava AK, et al. Decomposition based Logistic regression on Methicillin-resistant Staphylococcus Aureus (MRSA) patient prognosis. In: *Spring research conference on systems engineering and management science 2014 (SRC-SEMS2014)*, Shenzhen, China, 16–17 May 2014.
30. Wong ZSY. Text mining of medication incidents using Topic models. In: *AMIA 2013 annual symposium*, Washington, DC, 16–20 November 2013.
31. World Health Organization (WHO). The conceptual framework for the International Classification for Patient Safety (v.1.1). Final technical report and technical annexes, 2009, http://www.who.int/patientsafety/implementation/taxonomy/icps_download/en/index.html (accessed 20 December 2012).
32. Souvignet J, Bousquet C and Lewalle P. Modeling patient safety incidents knowledge with the Categorical Structure method. *AMIA Annu Symp Proc* 2011; 2011: 1300–1308.
33. Rodrigues JM, Larizgoitia I, Hansen J, et al. International information model for patient safety (2IMPS): enhancing care, patient safety and outcome. In: *Medinfo 2013*, Copenhagen, 21 August 2013.
34. World Health Organization (WHO). Information Model for Patient Safety (IMPS), 2014, http://www.who.int/patientsafety/implementation/information_model/en/ (accessed 21 July 2014).