MEDINFO 2017: Precision Healthcare through Informatics
A.V. Gundlapalli et al. (Eds.)

doi:10.3233/978-1-61499-830-3-609

Automating the Identification of Patient Safety Incident Reports Using Multi-Label Classification

Ying Wang[a], Enrico Coiera[a], William Runciman[b, c], Farah Magrabi[a]

[a] Centre for Health Informatics, Australian Institute of Health Innovation, Macquarie University, Australia
[b] Centre for Population Health Research, School of Health Sciences, University of South Australia, Australia
[c] Australian Patient Safety Foundation, Adelaide, Australia

Abstract

Automated identification provides an efficient way to categorize patient safety incidents. Previous studies have focused on identifying single incident types relating to a specific patient safety problem, e.g., clinical handover. In reality, there are multiple types of incidents reflecting the breadth of patient safety problems and a single report may describe multiple problems, i.e., it can be assigned multiple type labels. This study evaluated the abilty of multi-label classification methods to identify multiple incident types in single reports. Three multi-label methods were evaluated: binary relevance, classifier chains and ensemble of classifier chains. We found that an ensemble of classifier chains was the most effective method using binary Support Vector Machines with radial basis function kernel and bag-of-words feature extraction, performing equally well on balanced and stratified datasets, (F-score: 73.7% vs. 74.7%). Classifiers were able to identify six common incident types: falls, medications, pressure injury, aggression, documentation problems and others.

Keywords:

Machine Learning; Risk Management; Patient Safety

Introduction

Approximately 10% of admissions to acute-care hospitals are associated with unnecessary harm to patients [1, 2]. Events that could have resulted, or did result in unnecessary harm are called *patient safety incidents* [3]. The reporting and rapid analysis of patient safety incidents is important to prevent similar events from occurring in the future [4]. However, the volume of incident reports has dramatically increased over the last 20 years with wide implementation of incident monitoring systems [5]. Retrospective review of these incident reports by human experts is highly resource intensive and can no longer keep up with the growing volume of incidents being reported. To facilitate timely analysis, many incident monitoring systems now ask reporters to assign incident types so that reports of a specific type can be easily grouped for detailed classification by experts. However, incidents are reported by many different groups of health professionals who generally have limited expertise in classification [3, 6, 7].

One way of improving the efficiency of identifying incidents of a specific type is to automatically classify reports using text classification techniques. We have previously shown the feasibility of using statistical text classification to identify reports about three types of incidents: patient identification [8], clinical handover [8] and health information technology [9] using binary classifiers based on Naïve Bayes, logistic regression and Support Vector Machines (SVM). We have subsequently shown that extreme-risk events could be identified using a similar approach [10]. However, these studies have focused on distinguishing specific incident types from all others, e.g., patient identification. In reality, there are multiple types of incidents reflecting the breadth of problems in patient safety. Moreover, a single report can describe problems in more than one patient safety area, i.e., it can be assigned to multiple incident types [3, 11]. For example, "*Episode label A for patient X was placed incorrectly onto the specimen that belongs to patient Y,*" describes an error in patient identification that also relates to documentation.

The use of automated methods to identify multiple incident types remains largely unexplored. Some studies have sought to apply topic modelling [12, 13]. However, the mapping between topics and incident types is not straightforward. In this paper, we evaluated the feasibility of using multi-label classification to automate the identification of two labels or two incident types per report. Such multi-label problems are typically decomposed into one or more binary problems to simplify decision boundaries [14, 15]. In this study, classifiers were trained, validated and tested on balanced datasets. We then examined generalizability by applying the classifiers to imbalanced, i.e., stratified datasets which represented the real-world distribution of incidents.

Methods

Multi-label classification was decomposed into a number of binary classification problems, one for each label. We sought to evaluate the performance of three decomposition methods:
1. *Binary Relevance* (BR) is a widely used decomposition method which is theoretically simple and intuitive. For a given set of labels *L*, BR learns *L* binary classifiers (e.g., SVM). Each classifier is trained independently for a specific label against the rest of labels (one-versus-rest scheme) [15]. However, it does not preserve dependencies between labels. |
2. *Classifier Chains* (CC) is another decomposition method which is based on the BR but considers dependencies between labels [16]. CC learns *L* binary classifiers like the BR method. However, these binary classifiers are linked in a chain through a feature space where additional features with binary values indicate which other labels are assigned to a report. Testing begins with the first classifier and processes to the *L*th classifier by passing label information between classifiers through this feature space. Hence, the inter-label dependency is preserved. The performance of CC is very sensitive to the choice of label order in the chain, as the label dependency is demonstrated by label order in the feature space.

3. Ensemble of Classifier Chains (ECC) were introduced to reduce the influence of label order and improve classification performance [16]. Multiple CCs can be trained using different random chain ordering (determined by the order of labels) on a random subset of data. The final decision is the average of the multi-label predictions of CC. In general, the performance of ECC tends to improve and converge by combining more CCs with diverse label structures. However with increasing ensemble sizes, ECC becomes uncessarily large, and imposes an extra computational cost. Although there are several variants of CC which consider complex structures for label dependence and use different random search methods [17], we took a more practical solution by using a subset of CCs in the ensemble, which also improved classification performance.

Database

We used 6,000 randomly selected reports from 137,522 submitted to the Advanced Incident Management System (AIMS) across an Australian state between January and December 2011 [4]. Ethical approval was obtained from university committees as well as a committee governing the hospital and state datasets. Reports were de-identified and then labelled by three experts in the classification of patient safety incident reports. Experts provided a primary label and where applicable a secondary label was also given. These labels were used as a "gold standard" for classifier training and testing. Only descriptive narratives in reports were retained for experiments including incident description, patient outcome, actions taken, prevention steps, investigation findings and results. All system-specific codes, punctuation and non-alphanumerical characters were removed and text was converted to lower case.

The distribution of incident types from AIMS is imbalanced i.e. "stratified", and their real-world ratio is shown in Table 1.

Table 1: The composition of incident types in balanced and stratified datasets. N1 is the number of reports based on primary label. N2 is the number of reports by considering both labels.

	Training: balanced datasets		Testing: stratified datasets		
Incident type	N1	N2	N1	N2	%
Falls	260	261	90	91	20
Medications	260	304	68	74	15
Pressure injury	260	264	37	38	8
Aggression	260	271	49	57	11
Documentation	260	589	26	67	6
Blood product	260	273	5	6	1
Patient identification	260	337	7	8	2
Infection	260	274	6	6	1
Clinical handover	260	301	7	8	2
Deteriorating patient	260	264	1	2	<1
Others	260	689	148	173	33
Total	**2860**	**3827**	**444**	**530**	

To build reliable classifiers which capture characteristics of rare incident types, for training we used balanced AIMS datasets where each of the 10 types were evenly distributed (Table 1). Applicability to real-world conditions was then examined by testing on stratified data. The distribution of balanced and stratified datasets was based on primary labels.

Experimental workflow

Figure 1 shows an overview of our approach. Datasets were first decoded into 11 subsets according to one-vs-rest ensemble schemes [18, 25]. For the BR, 11 base classifiers were trained for each two-class subset involving feature extraction, classifier training and cross validation. A threshold-based decision-making scheme was applied to identify testing reports by combining the predicted probabilities from all base classifiers. With CC, the feature space was extended to represent label connections using binary values that indicated label co-occurrence [18, 19]. With ECC, incident types were randomly reordered and several ensembles of CC were generated. The hard decision for each report was made based on the probability values provided by all CCs.

1. ***Feature extraction***: To provide informative features for classification, removal of stop words and short words with fewer than two characters, stemming and lemmatization were applied to reports [18]. The bag-of-words model, commonly used in document classification, was adopted to extract features [19]. Irrespective of grammar, incident narratives were represented as an unordered collection of words and unique words were used as features. The bag of words was then transformed into a numeric representation interpretable by classifiers. A binary count, transforming the bag of words representation into 1 or 0 corresponding to word occurrences, was used, as it was the most effective feature representation along with the SVM classifier in practice [9, 10].

2. ***Feature space extension for CC and ECC***: The feature space of each link in the classifier chain was extended with the binary label associations of all links in the training procedure [16]. The order of the chain itself affects accuracy.

3. ***Base binary classifier training and validation***: For base classifiers, we chose discriminative classifier of SVM with radial-basis function kernel, as the SVM-based methods perform better for smaller datasets with a large feature space [20, 21]. Especially in text classification, documents are typically represented as a bag-of-words, where each feature captures crucial information but in a very sparse format [20]. Furthermore, in our previous incident classification work, SVM outperformed other binary classifiers, such as a logistic regression model [21].

To train the base classifiers, a ten-fold repeated random sub-sampling, cross-validation method was used to assign reports

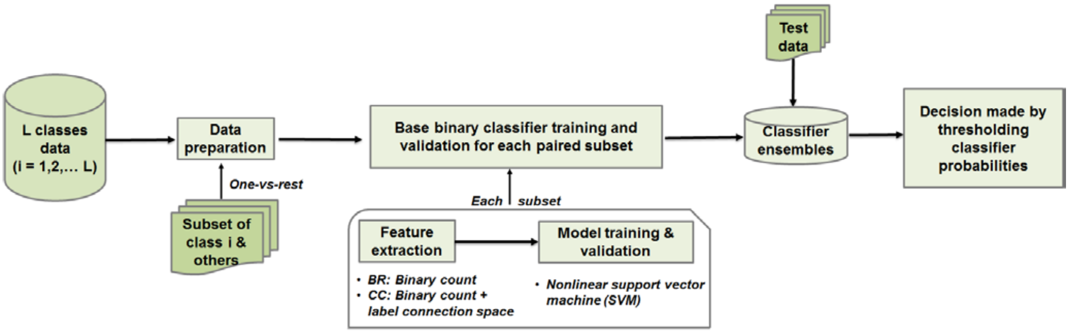
L classes data (i = 1,2,... L)
Data preparation
One-vs-rest
Subset of class i & others
Base binary classifier training and validation for each paired subset
Each subset
Feature extraction
Model training & validation
• BR: Binary count
• CC: Binary count + label connection space
• Nonlinear support vector machine (SVM)
Test data
Classifier ensembles
Decision made by thresholding classifier probabilities

Figure 1: Experimental workflow to train and evaluate methods of binary relevance (BR) and classifier chains (CC) to identify reports. Ensemble classifier chains (ECC) is a repetition of CC using randomly ordered labels.

to training (80%), validation (10%) and testing (10%) sets. Given that, with random assignment, a testing report for one base binary classifier might be used in the training set of another base classifier. For example, a report about a fall incident which was used for training a base classifier (fall vs. medications), may also be assigned to the test set of another base classifier (fall vs. blood product). To avoid potential overlaps between training and testing sets, we first randomly selected 10% of reports for each type and set them aside for testing. Then we created the folds using repeated random subsampling for traning and validation sets. Classifiers which outperformed others by achieving higher classification accuracy were adopted for testing.

4. ***Probability threshold based decision-making scheme***: For ECC, the hard decision was made by averaging multiple predictions from individual CC. Two labels were predicted if the averaged classfiication probabilites exceeded a predefined threshold. If the threshold is reached only for one label then the report was assigned to a single incident type.

5. ***Performance evaluation:*** Evaluation metrics for multi-label classification performance are inherently different from those used in multi-class or binary classification, due to the additional degrees of freedom that the multi-label setting introduces. Overall performance was examined using two types of evaluation measures, example-based and label-based measures [20]. Example-based measures are based on the average difference of the true and the predicted sets of labels over all testing reports. On the other hand, label-based measures evaluate classification performance separately for each incident type and then average the performance over all types. We used six example-based measures including [20]:

1. *Hamming loss* evaluates how many times an incident is misclassified, i.e. label not belonging to the report is predicted or a label belonging to the report is not predicated. This is a loss function, so the optimal value is zero:

$$Hamming\ loss\ = \frac{1}{N}\sum_{i=1}^{N}\frac{\sum_{j=1}^{L} xor(T_{i,j},P_{i,j}\ )}{L} \qquad (1)$$

Where L is the total number of incident types and N is the number of testing reports. X_{or} means the symmetric difference between two sets. $T_{i,j}$ denotes the set of true labels and $P_{i,j}$ denotes the set of predicted labels.

2. *Accuracy* is micro-averaged across all reports and is defined as the number of correct labels divided by the union of predicted and true labels.

$$Accuracy = \frac{1}{N}\sum_{i=1}^{N}\frac{|T_i\cap P_i|}{|T_i\cup P_i|} \qquad (2)$$

Note that any predicted label matches the real labels, count 1; so the size of correct labels for a report could be 0, 1, and 2 if there are two real labels. However, the number of correct labels in this way does not consider label orders.

3. *Exact match score* (0/1 loss) is a very strict evaluation measure as it requires the predicted set of labels to be an exact match of the true set of labels.

4. *Precision* is the correct labels divided by the number of number of predicted labels.

$$Precision = \frac{1}{N}\sum_{i=1}^{N}\frac{|T_i\cap P_i|}{|P_i|} \qquad (3)$$

5. *Recall* is defined as the correct labels divided by the number of number of true labels.

$$Recall = \frac{1}{N}\sum_{i=1}^{N}\frac{|T_i\cap P_i|}{|T_i|} \qquad (4)$$

6. *F-score* is the harmonic mean between precision and recall.

$$Fscore = (1+\beta^2)\frac{precision*recall}{(\beta^2*precision)+recall} \qquad (5)$$

Where β enables the F-score to favor either precision or recall. Precision and recall are given equal weight by setting β to 1.

We also used six label-based measures, which are defined below. Given two labels per report, the classification performance of individual types was evaluated based on OR logic when matching the predicted and true set of labels per report.

1. *Macro-precision* is defined as precision averaged across all labels:

$$Macro-precision\ = \frac{1}{L}\sum_{j=1}^{L}\frac{|tp_j|}{|tp_j+fp_j|} \qquad (6)$$

Where $tp_j$and fp_j are the number of true positives and false positives for the label j considered as a binary class.

2. Macro-recall is defined as recall averaged across all labels:

$$Macro-recall\ = \frac{1}{L}\sum_{j=1}^{L}\frac{|tp_j|}{|tp_j+fn_j|} \qquad (7)$$

Where fn_j is the number of false negatives for the label j which is considered as a binary class.

3. *Macro-F-score* is the harmonic mean between precision and recall where the average is calculated per label and then averaged across all labels.

$$Macro-F-score\ = \frac{1}{L}\sum_{j=1}^{L}\frac{2*precision_j*recall_j}{precision_j+recall_j} \qquad (8)$$

Micro-precision and micro-recall are defined as general definitions of precision and recall but the number of true positives, false positives and false negatives are averaged over all the labels, respectively. Micro-F-score is the harmonic mean of micro-precision and micro-recall, defined as below.

$$Micro-precision = \frac{\sum_{j=1}^{L} tp_j}{\sum_{j=1}^{L} tp_j+\sum_{j=1}^{L} fp_j} \qquad (9)$$

$$Micro-recall = \frac{\sum_{j=1}^{L} tp_j}{\sum_{j=1}^{L} tp_j+\sum_{j=1}^{L} fn_j} \qquad (10)$$

$$Micro-score = \frac{2 * micro_{precison} * micro_{recall}}{micro_{precison} + micro_{recall}} \qquad (11)$$

Results

We examined the performance of three multi-label classification methods on balanced and stratified datasets. Using the most effective classifiers, we then evaluated performance when identifying different incident types.

ECC performance with increasing ensemble sizes

Using SVM RBF with binary count we examined the ECCs using different ensemble sizes ranging from three to forty. We found that six classifier chains obtained much smaller ensembles while achieving better or at least comparable performance on balanced and stratified datasets (Table 2, columns 3-7).

Performance of BR, CC and ECC

We compared BR, CC and ECC with six CCs (Table 2). Performance improved from BR to CC to ECC with both balanced and stratified datasets. Considering exact matching between the predicted and true sets of incident types, ECC (balanced/stratified: 39.9%, 44.4%) outperformed BR (35.7%, 39.0%) and CC (36.7%, 37.8%), indicating that better label dependency was maintained by ECC. Hamming loss value also showed that ECC was more efficient than BR. Example-based F-scores and label-based Macro-F-scores and Micro-F-scores increased slightly from BR to ECC with both testing datasets, for instance, F-scores of BR: 73.7%, CC: 74.2% and ECC: 74.7% on the stratified datasets. The most effective ECC performed equally well on both balanced and stratified datasets, considering example-based measures and Micro-F-score, indicating that this approach is generalizable. However, it obtained a relatively worse Macro-F-score of 59.2% on stratified datasets (balanced: 73.7%).

Table 2: Classifier performance of BR, CC, and ECC with different ensemble sizes on balanced and stratified AIMS datasets, where hamming loss, accuracy, recall, precision, F-score, extract match score, Macro/Micro-averaged measures were considered. Base binary classifier was SVM RBF with binary count representation of bag-of-word features.

Measures (%)	***BR***	***CC***	***ECC***				
			3CC	**6CC**	**11CC**	**20CC**	**40CC**
Hamming Loss	8.4	8.3	8.1	7.8	7.8	7.9	7.7
	8.0	8.2	7.6	7.2	7.5	7.5	7.5
Accuracy	62.6	63.5	64.4	64.4	65.1	64.2	65.3
	65.4	65.6	67.1	68.0	66.8	66.7	66.6
Exact match score	35.7	36.7	39.9	39.9	40.6	40.2	40.6
	39.0	37.8	43.0	44.4	44.4	43.5	43.5
Precision	67.7	68.4	70.3	70.6	71.3	70.5	71.9
	68.9	69.3	71.1	72.9	71.7	71.6	71.6
Recall	80.1	80.4	77.1	77.1	77.8	76.7	77.8
	79.1	80.0	76.6	76.6	75.5	75.9	75.6
F-score	73.3	73.9	73.5	73.7	74.4	73.5	74.7
	73.6	74.2	73.7	74.7	73.6	73.7	74.4
Macro-precision	67.2	66.2	69.0	69.7	69.9	70.0	70.0
	50.0	49.0	51.0	52.4	52.3	52.2	51.9
Macro-recall	80.7	80.0	78.9	79.0	79.6	79.2	79.5
	80.4	79.7	77.5	77.4	77.6	77.4	77.3
Macro-F-score	72.9	72.2	73.2	73.7	74.0	73.9	74.0
	57.6	56.7	57.9	59.2	59.0	58.9	58.6
Micro-precision	63.3	63.3	66.1	67.1	67.1	67.3	67.4
	62.8	62.3	66.0	67.1	67.0	66.9	66.6
Micro-recall	73.4	73.4	71.9	71.9	72.6	71.9	72.3
	71.5	72.2	71.2	70.6	70.7	70.5	70.4
Micro-F-score	68.0	68.0	68.9	69.4	69.7	69.5	69.7
	66.9	66.9	68.5	68.8	68.8	68.7	68.4

** measures from stratified datasets are shaded.*

Performance on identifying individual incident types

The most effective ECC based on six CCs achieved high F-scores above 81% on balanced datasets when identifying falls, pressure injury, infection and deteriorating patients (Table 3). This ECC trained on balanced datasets was very robust in identifying six types of incidents (falls, medication, pressure injury, aggression, documents and others), achieving similar F-scores on stratified datasets. For blood products, patient identification, infection, clinical handover and deteriorating patients, recall was similar on both balanced and stratified datasets, but precision was lower on stratified datasets.

Table 3: Individual incident type classification performance (recall, precision and F-score) on balanced (B) and stratified (S) datasets using 6 CC of binary SVM RBF with binary count

Incident type	**Recall (%)**		**Precision (%)**		**F-score (%)**	
	B	S	B	S	B	S
Falls	96.3	96.8	81.3	83.5	88.1	89.7
Medication	73.8	77.9	62.0	74.7	67.4	76.3
Pressure injury	90.0	90.5	93.1	80.9	91.5	85.4
Aggression	87.1	80.3	64.3	60.6	74.0	69.1
Documentation	62.2	62.6	62.2	59.8	62.2	61.2
Blood products	81.8	85.7	62.8	30.0	71.1	44.4
Patient identification	73.9	72.7	59.6	22.2	66.0	34.0
Infection	83.3	85.7	78.9	31.6	81.1	46.2
Clinical handover	76.9	72.7	52.6	20.5	62.5	32.0
Deteriorating patient	96.3	66.7	83.9	33.3	89.7	44.4
Others	53.9	59.6	67.6	78.9	60.0	67.9

Discussion

We evaluated three multi-label text classification methods using binary classifier ensemble on both balanced and stratified datasets from a state-wide incident reporting system. The most effective classifiers performed equally well in identifying reports from six common types including falls, pressure injury, aggression, documents and others. These types made up 93% of all reports (Table 1). Even so, it should be emphasized that automated identification of incident reports is not intended as a replacement for expert review but as a first step in grouping incidents and identifying clusters when human resources are lacking [22].

Comparison between BR, CC and ECC

Although the ECC achieved the best performance, the performance difference between BR, CC and ECC was relatively small (Table 2). Compared to CC and ECC, BR is more computationally efficient as it requires less dimensionalities of features and no classifier chain for training. However the BR completely ignores the possible correlations among labels, so the binary classifiers make decisions independently from each other. On the other hand, it makes BR very flexible in practice, e.g., evolving more incident types, because it can add labels without affecting the rest of classifiers. Thus, BR is more suitable for problems with only a small number of labels associating with each other.

The CC method can be seen as a direct extension of the BR, capable of exploiting label dependencies. Similar to BR, CC involves training a group of L binary classifiers. Nevertheless, instead of being kept isolated from each other, these L classifiers are linked in a chain structure, which allows each one to pass their predictions to the other binary classifiers connected in the chain. Compared to BR, CC did not improve much of classification performance in this study. BR was even slightly better than CC according to hamming loss and exact matching score on the stratified datasets. This might be due to the original text feature space being huge and the label relationship in such a sparse space not being captured well by classifiers with limited training datasets.

ECC keeps relationship of patient safety types in reports by building multiple CCs using random label orders. Obviously, the larger ensemble size leads to higher computational cost. We tried to reduce the ensemble size whilst keeping or improving the performance. In training, we built several ECCs with ensemble sizes ranging from three to forty. We found that the classification performance became stable when the size of CC was greater than six (Table 2). Thus the most effective ECC used six CCs was used for testing.

Identification of common and rare incident types

An important finding of this study is that the most effective ECC trained on balanced datasets appears generalizable to unseen datasets. The ECC achieved similar F-scores on stratified datasets in identifying the common incident types including falls, medication, pressure injury, aggression, documentation and others (4-33% Table 1, 93% of reported incidents in total). Especially for falls and pressure injury, the ECC achievied high F-scores (above 85.4%). However the ECC tended to be weaker when identifying rare types in stratified datasets such as patient identification, infection, clinical handover and deteriorating patient (<2% Table 1, 7% of all reported incidents). This explains why similar example-based measures were achieved on balanced and stratified datasets because the common types dominate the performance measures. On the other hand, the label-based measures showed relatively worse performance on stratified datasets, as the averaged measures were based on individual performance of each type. Given the imbalanced nature of incident distribution, automated identification can reduce the effort spent in identifying common types and provide small volumes of like incident reports for further investigation by experts.

Limitations and future work

There are several limitations. First, we used datasets from one Australian state. Our classifiers may not be generalizable to other regions with different linguistic styles and terminology. Secondly, we use the balanced dataset for classifier training because a limited number of incidents had been reported over a 12-month period. Given the imbalanced nature of multiple incident types, a stratified training set may work better in real-world conditions. To improve the identification of rare classes, one solution might be to review rare classes flagged by classifiers, which is practical because overall volumes in real-world datasets will be low. Another possible way is to use rule-based methods that involve expert knowledge and incorporate specific criteria for identifying incidents.

Conclusions

The use of text-based ECC is a feasible approach for automatically identifying multiple incident types. Evaluation of BR, CC, and ECC using binary classifiers of SVM RBF with binary count feature extraction, showed that the most effective combination was the six ECC of binary SVMs. Despite the limitations listed above, automated identification can provide a more efficient way to categorize the common incident reports, so that human resources can be redirected to detailed classification, allowing remedial actions to be triggered more quickly to respond to emerging safety issues.

Acknowledgements

We thank B Shumack, K Pappas and D Arachi for assisting with the data extraction and, A Deakin, A Agers and S Suffolk for their assistance with labeling the incident reports.

References

[1] N. Rafter, A. Hickey, S. Condell, R. Conroy, P. O'Connor, D. Vaughan, and D. Williams, Adverse events in healthcare: learning from mistakes, *QJM* **108** (2015), 273-277.

[2] W.B. Runciman, R.K. Webb, S.C. Helps, E.J. Thomas, E.J. Sexton, D.M. Studdert, and T.A. Brennan, A comparison of iatrogenic injury studies in Australia and the USA. II: Reviewer behaviour and quality of care, *Int J Qual Health Care* **12** (2000), 379-388.

[3] G. World Alliance For Patient Safety Drafting, H. Sherman, G. Castro, M. Fletcher, S. World Alliance for Patient, M. Hatlie, P. Hibbert, R. Jakob, R. Koss, P. Lewalle, J. Loeb, T. Perneger, W. Runciman, R. Thomson, T. Van Der Schaaf, and M. Virtanen, Towards an International Classification for Patient Safety: the conceptual framework, *Int J Qual Health Care* **21** (2009), 2-8.

[4] W.B. Runciman, J.A. Williamson, A. Deakin, K.A. Benveniste, K. Bannon, and P.D. Hibbert, An integrated framework for safety, quality and risk management: an information and incident management system based on a universal patient safety classification, *Qual Saf Health Care* **15 Suppl 1** (2006), i82-90.

[5] NRLS Quarterly Data Workbook up to September 2015, in, 2016, p. NRLS Quarterly Data Workbook updates analysis of the NRLS patient safety incidents reported by NHS organisation in England and Wales to the National Reporting and Learning System up to September 2015.

[6] J.F. Travaglia, M.T. Westbrook, and J. Braithwaite, Implementation of a patient safety incident management system as viewed by doctors, nurses and allied health professionals, *Health (London)* **13** (2009), 277-296.

[7] J.Westbrook, L. Li, E.Lehnbom, M.Baysari, J. Braithwaite, R. Burke, C. Conn, and R. Day, What are incident reports telling us? A comparative study at two Australian hospitals of medication errors identified at audit, detected by staff and reported to an incident system, *International Journal for Quality in Health Care* **27** (2015), 1-9.

[8] M.S. Ong, F. Magrabi, and E. Coiera, Automated categorisation of clinical incident reports using statistical text classification, *Qual Saf Health Care* **19** (2010), e55.

[9] K.E.K. Chai, S. Anthony, E. Coiera, and F. Magrabi, Using statistical text classification to identify health information technology incidents, *Journal of the American Medical Informatics Association* **20** (2013), 980-985.

[10] M.S. Ong, F. Magrabi, and E. Coiera, Automated identification of extreme-risk events in clinical incident reports, *J Am Med Inform Assoc* **19** (2012), e110-118.

[11] National Safety and Quality Health Service Standards Report, *Australia Commission on Safety and Quality in Health Care* (2012).

[12] A. Fong, A. Hettinger, and R.Ratwani, Exploring methods for identifying related patient safety events using structured and unstructured data, *J Biomed Inform* **58** (2015), 89-95.

[13] A. Fong and R. Ratwani, An Evaluation of Patient Safety Event Report Categories Using Unsupervised Topic Modeling, *Methods Inf Med* **54** (2015), 338-345.

[14] J. Read, B. Pfahringer, G. Holmes, and E. Frank, Classifier chains for multi-label classification, *Machine Learning* **85** (2011), 333-359.

[15] G. Tsoumakas and I. Katakis, Multi-label classification: An overview, *Dept. of Informatics, Aristotle University of Thessaloniki, Greece* (2006).

[16] J. Read, B. Pfahringer, G. Holmes, and E. Frank, Classifier Chains for Multi-label Classification, *Machine Learning and Knowledge Discovery in Databases, Pt Ii* **5782** (2009), 254-269.

[17] E.C. Goncalves, A. Plastino, and A.A. Freitas, Simpler is Better: a Novel Genetic Algorithm to Induce Compact Multi-label Chain Classifiers, *Proceedings of the 2015 Genetic and Evolutionary Computation Conference* (2015), 559-566.

[18] T. Korenius, J. Laurikkala, K. Järvelin, and M. Juhola, Stemming and lemmatization in the clustering of finnish text documents, in: *Proceedings of the thirteenth ACM international conference on Information and knowledge management*, ACM, 2004, pp. 625-633.

[19] J. Sivic and A. Zisserman, Efficient Visual Search of Videos Cast as Text Retrieval, *Ieee Transactions on Pattern Analysis and Machine Intelligence* **31** (2009), 591-606.

[20] G. Madjarov, D. Kocev, D. Gjorgjevikj, and S. Dzeroski, An extensive experimental comparison of methods for multi-label learning, *Pattern Recognition* **45** (2012), 3084-3104.

[21] C. Cortes and V. Vapnik, Support-Vector Networks, *Machine Learning* **20** (1995), 273-297.

[22] C.P. Friedman, A "fundamental theorem" of biomedical informatics, *J Am Med Inform Assoc* **16** (2009), 169-170.

Address for correspondence

Ying Wang, Email: ying.wang@mq.edu.au.