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

doi:10.3233/978-1-61499-830-3-1075

Predicting Harm Scores from Patient Safety Event Reports

Chen Liang[a], Yang Gong[b]

[a] *Louisiana Tech University, Ruston, Louisiana, USA*
[b] *School of Biomedical Informatics, University of Texas Health Science Center, Houston, Texas, USA*

Abstract

The identification of the severity of patient safety events promotes prioritized safety analysis and intervention. The Harm Scale developed by the Agency for Healthcare Research and Quality is widely used in the US hospitals. However, recent studies have indicated a moderate to poor inter-rater reliability of the Harm Scale across a number of US hospitals. Although the reasons are multi-folded, biased human judgments are recognized as a prominent factor. We proposed that key information to identify and refine the severity of harm is contained in the narrative data in patient safety reports. Using automated text classification to categorize harm scores is intended to provide reduced subjective judgments and much improved efficiency. We evaluated different types of classification algorithms using a corpus of patient safety reports from a US health care system. The results demonstrate the effectiveness and efficiency of the proposed methods. Accordingly, human biases on the application of harm scores are expected to be largely reduced. Our finding holds promise to serve as a semi-supervised tool during the process of manually reviewing and analyzing patient safety events.

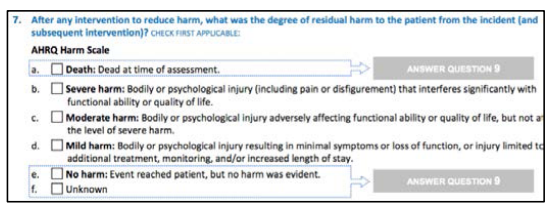
Keywords:
Patient Safety, Patient Harm, Data Mining

Introduction

Harm Classification and Scales

Reducing patient harm is a top priority of US hospitals and health care organizations. During the past two decades, researchers have been focusing on compiling patient safety events and detecting errors through nationwide patient safety reporting [1,2]. Event reporting at all levels has shown remarkable advantages to gather concurrent and retrospective events, including patient harms, near misses, and unsafe conditions in a timely-manner. Most importantly, it enables a close analysis on aggregate data, which increases the chance of disclosing vulnerability of health care systems. To accommodate safety event reporting at federal level, the Patient Safety Organization (PSO) has employed standardized event reporting formats (a.k.a., the Common Formats) to collect and classify reports [3].

Severity of safety events is an influential factor that can be identified by using the Common Formats. This piece of information plays a crucial role in triggering intervention actions and prioritizing limited resources of root cause analysis. In the Common Formats, Harm Scale is used to describe the degree of harm by assigning each event a harm score. The latest version of Harm Scale (v1.2) released in 2012 consists of a 5-point scale of severity of harm and a 2-point scale of anticipated duration of the harm (see Figure 1). In the meantime, a number of health care organizations have also participated in developing harm scales from different perspectives. The World Health Organization (WHO) developed a five-point harm scale, consisting of 'no harm to death', 'mild', 'moderate', 'severe', and 'death' [4]. This scale is centered on the patient harms arising from the provision of care. The National Coordinating Council for Medication Error Reporting and Prevention (NCC MERP) developed a scale that takes duration and permanency of harm into account. The Institute for Healthcare Improvement (IHI) also developed Global Trigger Tools to measure the severity of patient harm [5].


7. After any intervention to reduce harm, what was the degree of residual harm to the patient from the incident (and subsequent intervention)? CHECK FIRST APPLICABLE:

AHRQ Harm Scale

a. ☐ **Death:** Dead at time of assessment. → ANSWER QUESTION 9

b. ☐ **Severe harm:** Bodily or psychological injury (including pain or disfigurement) that interferes significantly with functional ability or quality of life.

c. ☐ **Moderate harm:** Bodily or psychological injury adversely affecting functional ability or quality of life, but not at the level of severe harm.

d. ☐ **Mild harm:** Bodily or psychological injury resulting in minimal symptoms or loss of function, or injury limited to additional treatment, monitoring, and/or increased length of stay.

e. ☐ **No harm:** Event reached patient, but no harm was evident. → ANSWER QUESTION 9

f. ☐ Unknown

Figure 1 – A Harm Scale Screenshot from the Common Formats, Patient Information Form (v1.2).

Harm Scale Reliability

One of the most significant challenges of using harm scale is reliability, which is the deviation between reporters' judgment about the type and severity of harm [6-8]. In practice, the deviation not only influences the classification of harm but also relates to the determination of intervention actions. For example, if an event is determined at a certain level that is preventable, it is likely that significant analysis and intervention efforts will be assigned. Otherwise, the complication of the care occurred in the event is likely to be labeled as a risk factor.

The reliability of harm scale, especially standard scales used at a national level, is not as high as expected in the practice. A recent survey study on the reliability of the Common Formats Harm Scale across different roles of clinicians and different settings yielded an overall moderate level of reliability [8]. The findings in this study show that some levels of harm are difficult to distinguish from neighbor levels. This problem is most significant for the moderate severity levels of harm. Another study that is performed on a relatively smaller size of data showed similar findings [7].

The deviations may be caused by several reasons. Firstly, the reporters vary in background. Reporting is open to clinicians in the hospitals, including physicians, nurses, pharmacists, etc. Nurses and pharmacists are reported to be more active in the reporting because they witness errors more frequently during the course of care [9]. For example, they have more chances to witness and report medication errors. When they do, they are likely to assign medications errors with a lower harm score

compared to other clinicians [10,11]. Secondly, reporters' understanding of the harm scale exerts an influence on the rating [10]. Studies have suggested an important role of education and training in the reporting [12]. Besides, biased harm scores can be a result of unclear guidelines and information representation, such as the definitions [7] and knowledge structure [13]. Thirdly, the way in which events are reported may influence the reliability of harm scores. This argument is mainly centered on the capability of hospitals to discover and make adjustments of potentially biased scores. Compared to paper-based reporting, web-based reporting holds potential to disclose biased harm scores, as it is advanced in viewing aggregate data and trends.

Predicting Harm Scores from Patient Safety Reports

An alternative of calibrating biased harm score is to develop a mechanism of predicting harm score from patient safety reports. During the reporting, the decision of assigning a harm score to an event is made by reporters' understanding of the event, their experience, and perception of environment. While components such as experience and perception are subjective, the event itself is relatively more objective. As such, decisions that are purely based on the events are likely to reduce the bias caused by human. In most of the hospital reporting systems, events are storytelling-like and recorded in a text format, namely patient safety reports. These reports contain substantial and essential information to make judgment of harm scores. Most importantly, informatics techniques are available to extract information without human biases. Text classification is a candidate technique that purports to predict classes of text based on statistical regulations of term distribution in the text. To perform text classification, a statistical model is trained through learning term frequency from a set of categorized documents. The trained model is then capable of predicting un-categorized homogeneous documents with correct classes. This method has been broadly used in biomedical domain to reduce manual production time [14].

We propose that text classification can be used to predict harm scores based on patient safety reports. In this study, we will train classifiers from a set of reports that are assigned with harm scores according to the Common Formats Harm Scale (v1.2). The classifiers will predict harms scores of unlabeled reports where the performance of classifiers will be evaluated. From a practical perspective, the classification results are expected to eliminate potentially biased harm scores based on stored reports.

Table 1 – Distribution of harm scores among 2919 reports.

Harm Score	Meaning	Frequency
a	Death	11
b	Severe harm	36
c	Moderate harm	144
d	Mild harm	336
e	No harm	626
f	Unknown	1766

Methods

Data

The dataset consists of a corpus of 2919 de-identified patient safety reports from a university health care system. The reports cover a range of incident types that are labeled by reporters (see Figure 2).

The reports have been cross-validated by a group of domain experts, assigned with harm scores using the Common Formats Harm Scale (see Table 1). In the text classification task, the assigned harm scores serve as the gold standard to be compared with machine prediction.

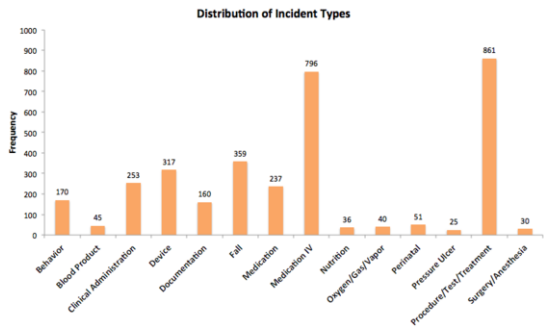


Figure 2 – Distribution of incident types.

Procedure

Environment

We provided a 64-bit OS system with a processing power of 2.2 GHz with 4 Cores 8 Threads and a memory of 8 GB RAM to perform the experiments. The classification experiment was performed on WEKA 3.6 [15].

Text Processing

To extract term frequency information from the raw data, we followed a serial of procedures to prepare the data. (1) Snowball stemmer was used to reduce inflected terms to theirs root form [16]. (2) Rainbow list was used to remove stop words [17]. (3) Alphabetic tokenizer was used to break a string of text into terms. (4) Lower case token was applied to all the terms. (5) TF-IDF (term frequency-inverse document frequency) was used in the transformation of documents into a bag-of-words (BOW) matrix keeping 1000 unique terms [18].

Text Classification Algorithms

We selected three types of algorithms that are well documented in text processing and biomedical application. They include decision tree algorithm, lazy algorithm, probabilistic algorithm, and support vector machine (SVM) [19]. For the decision tree, we employed C4.5 since it is reported effective in processing text [20]. For the lazy algorithm, we employed k-Nearest Neighbor (kNN) [21] for its well-balanced efficiency and predictive performance in medical text [22]. For the probabilistic algorithm, we employed Naïve Bayesian [23]. See Table 2 for a list of algorithms we used. A benchmark comparison is performed among these algorithms.

Table 2 – A list of selected algorithms.

Algorithm	Implementation	Parameter
Decision Tree	C4.5	
k-Nearest Neighbor	IBk	k = 1
Naïve Bayes	NaiveBayes	
Support Vector Machine	LibSVM	Linear SVM

Evaluation

For all the six harm scores and four algorithms, we employed a 10-fold cross validation to compare between algorithm performance. Each round of evaluation ran 10 times, which produced a total of 2,400 results. Performance was measured by F measure, which is a weighted average between precision and recall. The generic F measure is given as

$$F\ Measure = \frac{2 \times Precision \times Recall}{Precision + Recall} \quad (1)$$

In a classification task, precision is the fraction of retrieved documents that are relevant to a given label. It measures the ability of a classifier not to label a document as relevant when it is not.

$$Precission = \frac{y \cap \hat{y}}{\hat{y}} \quad (2)$$

where $\hat{Y}$ denotes the predicted set of labels; Y denotes the exact set of labels.

Recall is the fraction of relevant documents that are retrieved. It measures the ability of a classifier to retrieve as more relevant documents as possible.

$$Recall = \frac{y \cap \hat{y}}{y} \quad (3)$$

In addition, we provided an estimate of receiver operating characteristic curve (ROC) as a metric for assessing the trade- off between true positive and false positive.

Results

Table 3 shows a ranking between the six tasks of classifying each harm score. The numbers indicate the number of wins or losses (negative number) of any task against the other tasks. On the metrics of precision, recall, and F measure, the task becomes more difficult if the harm score becomes smaller. This is probably because the narratives in the mild-harm or unknown-harm events contain less significant term frequency information that distinguishes the events from severe events. However, the ROC shows that tasks of classifying score 0 (unknown) and score d (mild harm) outperform others, indicating a well-controlled false positive.

Table 3 – Ranking test for classification tasks. ($p < .05$)

	Precision	Recall	F measure	ROC
f	-20	-18	-20	11
e	-12	-12	-12	-7
d	-4	-1	-2	11
c	4	0	2	0
b	13	14	12	-7
a	19	17	20	-8

Table 4 shows the results of ranking test for different algorithms. Naïve Bayesian outperformed in the ranking of precision and ROC. C4.5 is ranked the best algorithm on Recall and F measure.

Table 4 – Ranking test for different algorithms. ($p < .05$)

	Precision	Recall	F measure	ROC
C4.5	1	6	8	-3
kNN	-9	6	4	-9
Naïve Bayesian	9	-11	-10	18
SVM	-1	-1	-2	-6

Figure 3 shows the benchmark comparison on precision. Naïve Bayesian (precision = 0.88) and C4.5 (precision = 0.88) outperformed kNN (precision = 0.85) and SVM (precision = 0.87) in the comparison across all the six classification tasks. Paired *t* test shows that Naïve Bayesian performed better than C4.5 on tasks of classifying score e, d, c, and b, respectively ($p < .05$). But for the tasks of classifying score f and a, C4.5 performed better ($p < .05$).

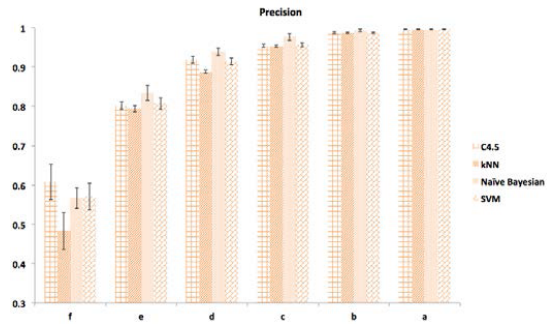


Figure 3 – Precision of algorithms by different harm scores.

Figure 4 shows the benchmark comparison on recall. C4.5 (recall = 0.90) ranks the best algorithm compared to kNN (recall = 0.88), Naïve Bayesian (recall = 0.83), and SVM (recall = 0.87).

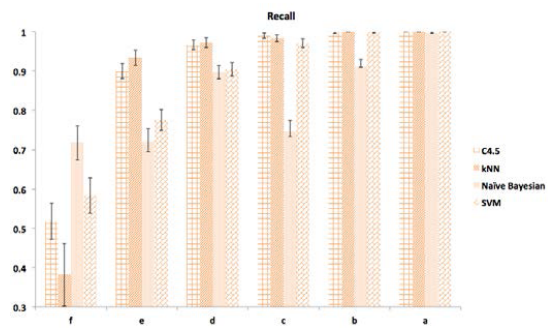


Figure 4 – Recall of algorithms by different harm scores.

Figure 5 shows the benchmark comparison on F measure. C4.5 (F = 0.89) ranks the best algorithm against kNN (F = 0.86), Naïve Bayesian (F = 0.85), and SVM (F = 0.87) overall.

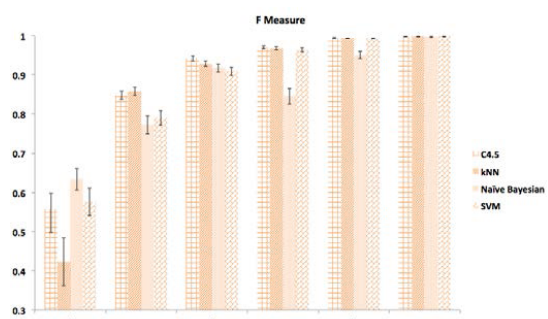


Figure 5 – F measure of algorithms by different harm scores.

Figure 6 shows the benchmark comparison on the area under ROC curve. Naïve Bayesian outperformed C4.5 (ROC = 0.76), kNN (ROC = 0.59), and SVM (ROC = 0.56) at all the classification tasks ($p < .05$).

Concerning the efficiency of algorithm, kNN used an average of 0.15 seconds of model training time, recognized the most efficient algorithm ($p < .05$) compared to C4.5 (time = 22.48 seconds), Naïve Bayesian (time = 0.92 seconds), and SVM (time = 1.70 seconds). An interaction of efficiency and performance is observed on Recall and F measure only, indicating that more time is needed for better performed algorithms (see Figure 7).

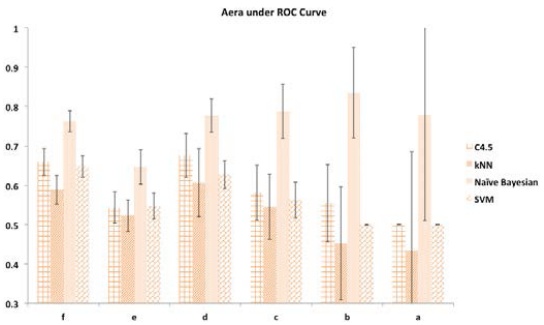


Figure 6 – Area under ROC curve of algorithms by different harm scores.

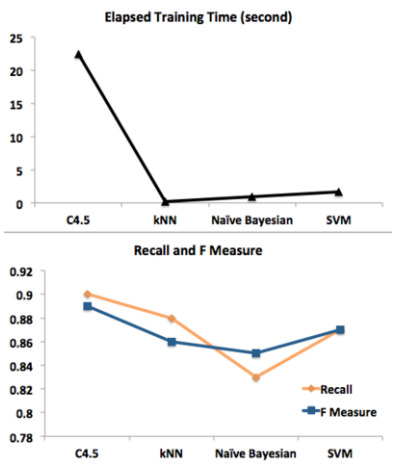


Figure 7 – Interaction between model training time and, recall and F measure, respectively.

Discussion

Experimental Findings

Our findings demonstrate the effectiveness of applying text classification to identify harm scores. The narratives of patient safety reports contain sizable information to determine the harm scores of events. In the experiment, the four types of text classification were well performed. However, the classification performance varied by the tasks of identifying different harm scores. There may be two factors that influenced the results. Firstly, the amount of reports of score a and b is significantly smaller, which might cause a label imbalance problem. This problem states that a classifier may be under trained because there are not sufficient reports in the minority labels. As an example, the classification tasks of identifying score a and b show outstanding performances on precision and recall but much worse ones on ROC. Secondly, levels of difficulties may vary by tasks. The task of identifying score f (unknown harm) is worst performed in terms of precision, recall, and F measure, but not ROC. Intuitively, the term frequency information in the reports of score f is sparse compared to the others.

Predicting harm scores based on patient safety reports is significantly efficient. Our findings confirmed a small computational cost of building the four types of classifiers but revealed some differences between these classifiers. Rule-based classifiers, i.e., C4.5 in our study, demand more time and computational resources. On the contrary, lazy classifiers and probabilistic classifiers use comparatively less time and computational resources, indicating a much improved efficiency. However, there was a tradeoff between efficiency and predictive power. Although, C4.5, for example, is most time consuming in our experiment, it showed best performance on Recall and F measure. This effect implies that C4.5 has the best capability to identify as more true positive reports as possible, regardless of identifying false positive ones. Moreover, this capacity is still dominant on the comparison of F measure, which is a combined metric of precision and recall.

Selecting a suitable classifier is task dependent. If the task concerns more about false alarm, i.e. mistakenly assigning a report to an irrelevant harm score, C4.5 and Naïve Bayesian are better. If the task concerns more about retrieving more reports that belong to a given harm score, C4.5 is preferred. When it takes both factors into account, C 4.5 is recommended. Because it still won on the F measure by showing a statistically significant difference. In the practice, however, it is mostly concerned to enlarge the true positive rate and reduce the false positive rate. Thus, Naïve Bayesian is the best classifier.

Clinical Implementation

We envision that the automated classification of harm scores could assist in calibrating any human biases raised in the process of reporting. The proposed methods are implementable to most of the existing web-based reporting systems. In our view, text classification takes advantages of the existing reporting systems from three aspects.

Firstly, text classification corresponding to the narrative data that are commonly used in the reporting. Narrative data makes patient safety reports different from many other clinical data. A patient safety event is mostly encoded in a story-telling fashion since the elements of health care are complex and temporally organized. It is less likely to include many detailed and critical information in structured data, such as numeric or categorical data. This fact hinders many analytical methods that are applicable for generic clinical data.

Secondly, text classification promotes aggregate analysis and reporting. The Patient Safety and Quality Improvement Act of 2005 has called on to build a national mechanism of error reporting and analysis. Over the past decade, the rapidly increased volume of reported data has shown a quantitative improvement as well as a technique bottleneck of timely processing such a huge amount of data. Without a feasible solution, the analysis at a national level is of less practical value. The efficiency of our method suggests a practical use to be implemented in clinic, which is promising in largely reducing the demand of human labor.

Thirdly, the automated classification of harm scores is controllable because it is semi-supervised in practice. Caution should be taken when we apply automated method to medical decisions. We noted that all the evaluated classifiers are at a certain rate of error. Though this is not unique in our case, such an error should be controlled at a reasonable level. It is highly recommended to perform the classification under human supervision, as it is the case for most of the informatics tools implemented in medicine.

Limitation and Future Direction

We highlighted a need for creating an objective mechanism of overseeing human biases of categorizing harm scores. However, our study is of less value without a discussion of limitations. One limitation is the relatively small sample size used for the experiment. This problem may cause imbalanced labels,

which further harm the performance. In addition, a small sample limits the possibility to evaluate rare cases that may hold important clinical value but have limited distribution. Therefore, we will enrich the sample size and include a broader and more representative dataset.

The other limitation includes less consideration of label correlation, which may cast a crucial influence on the classification performance. The label correlation may not only be limited within the different harm scores but also with a number of categories such as contributing factors, settings, and procedures. In the future, we will evaluate such relational information by experimenting a multi-label classification.

Last but not least, the present study did not consider categorizing harm scores on temporal information. For example, the Common Formats Harm Scale (v1.2) consists of a two-point scale assessing the duration of harm. The extraction of temporal information from medical (e.g., clinical notes) has not been adequately investigated but has been on our research agenda.

Conclusion

This study aims to apply automated text classification to categorize harm scores from patient safety reports. This method is applicable to discover and calibrate potential classification biases caused by human judgment. Four types of classification were evaluated in the experiment, yielding effectiveness and efficiency of the proposed approach. In addition, this approach holds promise in facilitating the labor intense analysis of large volume patient safety reports which is in accordance with the goal of establishing a nationwide safety reporting mechanism.

Acknowledgements

The study is supported by a research grant (1R01HS022895) from the Agency of Healthcare Research and Quality, and University of Texas System Grants Program (#156374).

References

[1] P.J. Pronovost, L.L. Morlock, J.B. Sexton, M.R. Miller, C.G. Holzmueller, D.A. Thompson, L.H. Lubomski, and A.W. Wu, Improving the value of patient safety reporting systems, *Advances in patient safety: new directions and alternative approaches* **1** (2008).
[2] L.T. Kohn, J.M. Corrigan, and M.S. Donaldson, *To Err Is Human: Building a Safer Health System*, National Academies Press, 2000.
[3] C.M. Clancy, Common formats allow uniform collection and reporting of patient safety data by patient safety organizations, *American Journal of Medical Quality* **25** (2010), 73-75.
[4] H. Sherman, G. Castro, M. Fletcher, 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, *International Journal for Quality in Health Care* **21** (2009), 2-8.
[5] D.C. Classen, R. Resar, F. Griffin, F. Federico, T. Frankel, N. Kimmel, J.C. Whittington, A. Frankel, A. Seger, and B.C. James, 'Global trigger tool' shows that adverse events in hospitals may be ten times greater than previously measured, *Health Affairs* **30** (2011), 581-589.
[6] M. Hoppes and J. Mitchell, Serious Safety Events: A Focus on Harm Classification: Deviation in Care as Link, 2014.
[7] T. Abbasi, D. Adornetto-Garcia, P.A. Johnston, J.H. Segovia, and B. Summers, Accuracy of Harm Scores Entered Into an Event Reporting System, *Journal of Nursing Administration* **45** (2015), 218-225.
[8] T. Williams, M. Szekendi, S. Pavkovic, W. Clevenger, and J. Cerese, The reliability of AHRQ common format harm scales in rating patient safety events, *Journal of patient safety* **11** (2015), 52-59.
[9] S.M. Evans, J.G. Berry, B.J. Smith, A. Esterman, P. Selim, J. O'Shaughnessy, and M. DeWit, Attitudes and barriers to incident reporting: a collaborative hospital study., *Quality & safety in health care* **15** (2006), 39-43.
[10] S.D. Williams and D.M. Ashcroft, Medication errors: how reliable are the severity ratings reported to the national reporting and learning system?, *International Journal for Quality in Health Care* **21** (2009), 316-320.
[11] P.J. Pronovost, C.G. Holzmueller, J. Young, P. Whitney, A.W. Wu, D.A. Thompson, L.H. Lubomski, and L.L. Morlock, Using incident reporting to improve patient safety: a conceptual model, *Journal of patient safety* **3** (2007), 27-33.
[12] A.M. Mayo and D. Duncan, Nurse perceptions of medication errors: what we need to know for patient safety, *Journal of nursing care quality* **19** (2004), 209-217.
[13] C. Liang and Y. Gong, Knowledge Representation in Patient Safety Reporting: An Ontological Approach, *Journal of Data and Information Science* **1** (2016), 75-91.
[14] A.M. Cohen and W.R. Hersh, A survey of current work in biomedical text mining, *Briefings in bioinformatics* **6** (2005), 57-71.
[15] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I.H. Witten, The WEKA data mining software: an update, *ACM SIGKDD explorations newsletter* **11** (2009), 10-18.
[16] M.F. Porter, Snowball: A language for stemming algorithms, 2001.
[17] A. McCallum, Rainbow project, 1998.
[18] G. Salton and C. Buckley, Term-weighting approaches in automatic text retrieval, *Information processing & management* **24** (1988), 513-523.
[19] C.C. Chang and C.J. Lin, LIBSVM: a library for support vector machines, in: *ACM Transactions on Intelligent Systems and Technology (TIST)*, ACM, 2011, p. 27.
[20] J.R. Quinlan, *C4. 5: programs for machine learning*, Elsevier, 2014.
[21] J. Wang and J.D. Zucker, Solving multiple-instance problem: A lazy learning approach, in: *Proc. 17th International Conf. on Machine Learning*, Morgan Kaufmann, 2000.
[22] C. Liang and Y. Gong, Enhancing Patient Safety Event Reporting by K-nearest Neighbor Classifier. Context Sensitive Health Informatics, in: *Context Sensitive Health Informatics*, 2015.
[23] G.H. John and P. Langley, Estimating continuous distributions in Bayesian classifiers, in: *Proceedings of the Eleventh conference on Uncertainty in artificial intelligence*, 1995, pp. 338-345.

Address for correspondence

Yang Gong, MD, PhD
7000 Fannin St. Suite 600, Houston 77030 Texas
Yang.Gong@uth.tmc.edu
Tel: +1 713 500 3547