Using statistical text classification to identify health information technology incidents

Kevin E K Chai,[1] Stephen Anthony,[2] Enrico Coiera,[1] Farah Magrabi[1]

▶ Additional material is published online only. To view please visit the journal online (http://dx.doi.org/10.1136/amiajnl-2012-001409).

[1]Centre for Health Informatics, Australian Institute for Health Innovation, The University of New South Wales, Sydney, Australia
[2]The Kirby Institute for Infection and Immunity in Society, The University of New South Wales, Sydney, Australia

Correspondence to
Dr Farah Magrabi, Centre for Health Informatics, Australian Institute for Health Innovation, The University of New South Wales, Sydney, NSW 2052, Australia; f.magrabi@unsw.edu.au

Received 10 October 2012
Revised 4 April 2013
Accepted 14 April 2013
Published Online First
10 May 2013

To cite: Chai KEK, Anthony S, Coiera E, *et al*. *J Am Med Inform Assoc* 2013;**20**:980–985.

ABSTRACT

Objective To examine the feasibility of using statistical text classification to automatically identify health information technology (HIT) incidents in the USA Food and Drug Administration (FDA) Manufacturer and User Facility Device Experience (MAUDE) database.

Design We used a subset of 570 272 incidents including 1534 HIT incidents reported to MAUDE between 1 January 2008 and 1 July 2010. Text classifiers using regularized logistic regression were evaluated with both 'balanced' (50% HIT) and 'stratified' (0.297% HIT) datasets for training, validation, and testing. Dataset preparation, feature extraction, feature selection, cross-validation, classification, performance evaluation, and error analysis were performed iteratively to further improve the classifiers. Feature-selection techniques such as removing short words and stop words, stemming, lemmatization, and principal component analysis were examined.

Measurements κ statistic, F1 score, precision and recall.

Results Classification performance was similar on both the stratified (0.954 F1 score) and balanced (0.995 F1 score) datasets. Stemming was the most effective technique, reducing the feature set size to 79% while maintaining comparable performance. Training with balanced datasets improved recall (0.989) but reduced precision (0.165).

Conclusions Statistical text classification appears to be a feasible method for identifying HIT reports within large databases of incidents. Automated identification should enable more HIT problems to be detected, analyzed, and addressed in a timely manner. Semi-supervised learning may be necessary when applying machine learning to big data analysis of patient safety incidents and requires further investigation.

INTRODUCTION

Information technology has many benefits for improving the quality and safety of healthcare. However, there are also inherent risks to patient safety when health information technology (HIT) is poorly designed, implemented, and maintained.[1–5] For example, serious incidents have occurred when incorrect medication dosage and treatment have been prescribed as the result of software errors.[6] One of the main reasons for these incidents is that designing and maintaining safe HIT systems is both complex and difficult.[7 8] Moreover problems with HIT may only emerge after systems are deployed in real-world settings.[6] Therefore, it is important that incidents with the potential to harm patients be reported, so that risks can be understood and addressed in a timely fashion.

HIT reported amongst medical device incidents

An important source of HIT incidents is the USA Food and Drug Administration (FDA) Manufacturer and User Facility Device Experience (MAUDE) database. The MAUDE database contains reports of events involving medical devices.[9] As part of FDA regulatory requirements, manufacturers in the USA are required to report medical device malfunction and problems that lead to serious injury and death. At present, there is considerable debate about the FDA's role in regulating HIT.[10] Under the Federal, Food, Drug, and Cosmetic Act, HIT is a medical device.[1] However, the FDA does not currently enforce its regulatory requirements with respect to HIT. Nevertheless, some manufacturers have voluntarily listed their systems, and in our previous work we analyzed HIT incidents reported to the FDA.[11] The new Patient Safety Action and Surveillance Plan from the Office of the National Coordinator proposes to monitor HIT adverse event reports in MAUDE to identify HIT patient safety risks.[12]

Identifying HIT incidents

A major obstacle in addressing HIT risks is the difficulty in identifying the small numbers of reports about HIT within large databases of incidents. The MAUDE database currently contains over two million incidents reported since 1991, with 358 229 incidents reported in 2011 alone. Manual review of such a large database on an ongoing basis is not feasible. Another approach is to perform keyword searches and eliminate non-HIT incidents through manual reviews of smaller subsets of reports.[11] On the basis of this approach, we previously searched almost 900 000 MAUDE reports to identify over 400 relating to HIT. A similar approach was used to identify 99 HIT reports from 42 616 patient safety incidents in an Australian incident-monitoring database.[13] However, this method is both time consuming and not exhaustive unless all HIT-relevant keywords are known in advance.

In the present study, we set out to evaluate the feasibility of using statistical text classification to automatically identify HIT-related incidents within MAUDE. HIT was broadly defined to include computer hardware and software used by health professionals to support patient care. To the best of our knowledge, the present study, focusing on automatically identifying HIT incidents within MAUDE, is novel. We have previously demonstrated the feasibility of using text classification to identify incident reports relating to clinical handover, patient identification, and extreme risk events.[14 15] Other studies have applied text mining methods to electronic

patient records,[16] radiology reports,[17 18] pathology reports,[19] and clinical notes,[20] but not incident reports. The identification of HIT incidents within MAUDE may be more challenging than for other types of incidents (eg, patient identification and clinical handover) because the specific terms used for HIT incidents may be quite similar to the language used to describe incidents involving medical devices. For example, problems involving software embedded in medical devices are not considered HIT incidents, but might be described using HIT-related words such as 'programming' or 'systems'.

BACKGROUND

Training text classifiers on imbalanced datasets

The class imbalance between HIT and non-HIT incidents in the MAUDE database presents a challenge for developing accurate text classifiers. A dataset is said to be 'imbalanced' when the number of examples from each class are not equal. Training statistical text classifiers on imbalanced datasets, particularly those with rare classes, can negatively affect their performance.[21 22] Our previous study performed text classification on balanced datasets of patient safety incidents with promising results.[14 15] However, these experimental results did not explore performance on imbalanced ('stratified') datasets, which represent real-world conditions. Therefore, in the present study, we focus on the more difficult problem of building and validating classifiers with a stratified subset of MAUDE.

Identifying rare classes

HIT incidents are considered rare classes because they amount to <1% of incidents in MAUDE. Seiffert and colleagues[21] showed that data-sampling approaches can increase classification performance when rare classes comprise 0.1–1.6% of a dataset. Other suitable approaches include oversampling, undersampling, cost-sensitive learning, ensemble methods, and constructing *k* neural networks.[21 22] In the present study, we experiment with an undersampling approach by training classifiers on a dataset containing a balanced number of HIT and non-HIT incidents and then evaluating performance on stratified validation and test datasets. This approach allows us to evaluate the feasibility of our previous work[14 15] where only balanced datasets were used to train and test classifiers.

Thus the aims of this study were to examine the feasibility of using statistical text classification to identify HIT incidents. We specifically sought to evaluate the performance of classifiers with combinations of balanced and stratified (real-world distribution) datasets for training, validation, and testing. In addition, feature-selection techniques such as removing short words and stop words in addition to stemming, lemmatization, and principal component analysis were evaluated.

METHODS

Dataset

Our study used 570 272 incidents extracted from the MAUDE database between 1 January 2008 and 1 July 2010. After data cleansing and preparation, the experimental dataset was reduced to 515 897 incidents (see online supplementary appendix A). The study dataset included a subset of 405 HIT incidents (0.079%; box 1) that had been labeled in our previous study.[11 13]

Experimental setup

We performed a set of 'preliminary experiments' followed by a set of 'relabeled data experiments'. Preliminary experiments were performed using the 405 known HIT incidents taken from our previous study.[11] In these early experiments, we determined that incidents involving handheld devices were causing the classifiers to mislabel events. These handheld device events within MAUDE had not been detected in our earlier study, and so we elected to enlarge the HIT incident dataset by including search terms for handheld devices. Two annotators labeled these new incidents that were retrieved from MAUDE using our HIT classification scheme.[11] The annotators achieved a high level of consensus, agreeing on 1145 incidents and disagreeing on 66 (which were then resolved by discussions). However, we achieved only a moderate inter-rater agreement of 0.579 (κ statistic). This was due to the sensitivity of the κ statistic to class imbalance,[23] as there were 1129 HIT and 82 non-HIT examples in the labeling set. After expanding the MAUDE search and annotating the new candidate incidents, we added an additional 1129 HIT incidents to the dataset. In our relabeled data experiments, the preliminary experiments were repeated on the relabeled data, now containing 1534 HIT incidents (original 405 +1129; 0.297%).

The experiments involved building text classifiers using different combinations of balanced and stratified datasets (table 1). Firstly, datasets with an equal number of HIT and non-HIT incidents were used to generate benchmark results. Secondly, classifiers were built using stratified datasets with HIT populations of 0.297%. Thirdly, classifiers were built using a balanced training dataset in an attempt to improve the performance of identifying rare HIT incidents within stratified validation and test datasets. Experiments with stratified datasets were initially performed with 10% of the dataset to test the classifiers and then scaled up to 100%. Details of how the datasets were partitioned are provided in online supplementary appendix B.

Experimental workflow

Each experiment comprised seven main tasks (figure 1): dataset preparation, feature extraction, feature selection, cross-validation, classification, performance evaluation, and error analysis.

Box 1 Health information technology incident example

It was reported that, while viewing an examination on a PACS (Picture Archiving and Communications System) workstation, the site reported that, in 20 examples throughout the year, the CT exam reports are being assigned to incorrect exams. The report info (information) may belong to other pts (patients) info leading to a misdiagnosis. This event is not isolated to a single workstation, radiologist or particular event. It does not occur all the time and is not immediately obvious to the caregiver. However, the event did not cause any harm or potential injury to a pt (patient) during this instance.

Table 1 Experimental setup

	Validation and test dataset	
	Balanced	**Stratified**
Training dataset		
Balanced	1. Benchmark	3. Rare class
Stratified		2. Original dataset

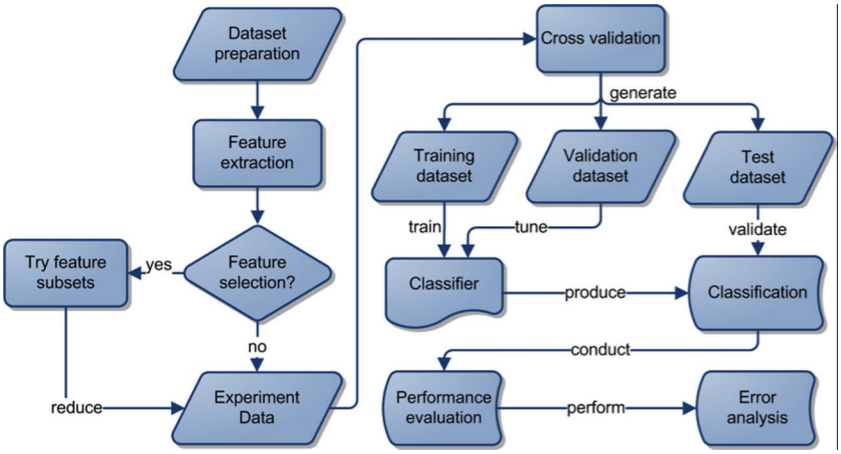

Figure 1 Experimental workflow.

Feature extraction

A bag-of-words model was used to represent incidents, treating them as an unordered collection of words, and each unique word was used as a feature. We performed data preprocessing before extracting words from the dataset such as removing punctuation and non-alphanumerical characters as well as converting numbers into the token, 'number'. In addition, an unknown-word feature was incorporated into the feature set to capture the incidence of new words not found in the training dataset vocabulary.

Feature selection

Common feature-selection tasks such as removing short words with fewer than two characters, removing stop words, stemming, lemmatization, and principal component analysis[24] were used in our experiments. The goal of feature selection is to reduce the number of features in the model while maintaining low classification error. We used feature selection to reduce memory usage and processing time rather than improve performance because regularized logistic regression (LR) allows the use of the full feature set while avoiding model overfit.[25]

Cross-validation

A 10-fold repeated random sub-sampling cross-validation method was used to assign incidents to training (60%), validation (20%), and test (20%) sets.[26] This random assignment was performed 10 times to generate 10 different training, validation, and test datasets. Therefore, 10 different classifiers were built for each dataset, and their results were averaged to measure overall performance.

Classification

We used LR over Naïve Bayes and Support Vector Machines (SVMs) used in our previous work for a number of reasons.[14 15] First, discriminant classifiers (eg, LR, SVMs) have been shown to generally outperform generative classifiers (eg, Naïve Bayes) on large datasets.[27] Therefore, we favor discriminative classifiers, as MAUDE contains millions of incidents. Second, LR has been shown to be more accurate and faster than SVMs on large datasets in certain studies.[28 29] In particular, we used the conjugate gradient descent algorithm with LR to reduce the classifier training time over large datasets.[30] Third, kernel functions were used by SVMs to transform the training data (input space) into a feature space for constructing decision boundaries.[26] However, these kernels increase memory and processing costs for datasets where the number of training examples, m, is larger than the number of features, n, as with MAUDE —that is, LR can be trained using a $m \times n$ matrix, whereas SVMs require a larger $m \times m$ matrix to represent the feature space. Furthermore, the feature space matrix cannot be represented efficiently as a sparse matrix as with the input space for LR.

Regularization was used to avoid overfitting (see online supplementary appendix C). Regularization is used when training a classifier on a small number of examples or learning from a large number of features.[25] Overfitting occurs when a classifier is tuned too finely to training data and then performs poorly on unseen test data. Text classification tasks performed on large datasets often involve training classifiers with thousands to hundreds of thousands of features. For example, the dataset used in our experiments contains up to 85 560 features. The features were extracted from the training set, and incidents were then encoded with this feature set. The classifier then learns incidents in the training set and is tuned on the validation set to find the best regularization parameter, λ, by maximizing the F1 score using a grid search algorithm.[31] The classifier was retrained with the selected λ and used to classify incidents in the unseen test set.

Performance evaluation

Precision, recall, and F1 score metrics were used to evaluate the performance of the classifiers (see online supplementary appendix D). Precision measures the percentage of incidents predicted as HIT that are actually HIT. Recall calculates the percentage of HIT incidents that were successfully identified out of all HIT incidents in the dataset. The F1 score is the harmonic mean of precision and recall and measures performance on imbalanced datasets more effectively than alternative metrics such as classification accuracy and the receiver operating characteristic.[32]

Error analysis

Error analysis involves analyzing incidents that were incorrectly classified to gain insights for improving the classifier. Learning curves were constructed on the basis of classification error with the training and validation datasets. These curves allowed us to identify bias and variances that can be addressed to improve performance. In addition, the learning curves allowed us to gage the number of incidents required to sufficiently train the classifier to achieve low classification error.

RESULTS

The results of the experiment are shown in table 2. The classifier built and tested with 100% of the original stratified dataset

Table 2 Experiment results

Datasets	Features*	F1 score	Precision	Recall
1. *Benchmark (balanced† training, validation, and test)*				
3068 incidents	8020	0.994	0.998	0.990
Stop words	99%	**0.995****	0.998	0.991
<2 characters	97%	0.994	0.998	0.990
Lemmatize	94%	0.993	0.996	0.989
Stem	73%	0.991	0.959	0.987
PCA	15%	**0.995**	0.998	0.992
2. *Original dataset (stratified‡ training, validation, and test)*				
10%	31011	0.945	0.977	0.916
Stop words	99%	0.946	0.977	0.919
<2 characters	98%	0.943	0.974	0.916
Lemmatize	94%	0.951	0.987	0.919
Stem	75%	**0.953**	0.990	0.919
100%	85560	0.953	0.966	0.941
Stop words	99%	**0.954**	0.966	0.943
<2 characters	99%	**0.954**	0.967	0.942
Lemmatize	95%	0.952	0.966	0.939
Stem	79%	0.945	0.959	0.932
3. *Rare classes (balanced training, stratified validation, and test)*				
10%	2606	0.274	0.165	0.929
100%	7957	**0.283**	0.165	0.989

*The number of features from the best performing classifier is shown; percentages represent the reduced feature set size.
**Bold figures represent the best performing classifiers based on F1-score for each dataset e.g. the classifier using stop words and PCA for the 1. Benchmark dataset achieved the top F1-scores of 0.995.
†Balanced dataset=50% health information technology (HIT) and 50% non-HIT.
‡Stratified dataset=0.297% HIT and 99.703% non-HIT.
PCA, principal component analysis.

achieved an F1 score of 0.954 along with high precision (0.953 and 0.954) and recall (0.919 and 0.943). The performance of classifiers using balanced training sets performed poorly (0.274 and 0.283 F1 scores) compared with those using stratified training, validation, and test sets. Feature-selection techniques were able to achieve comparable performance when evaluating all features. This experiment was initially performed on the original 405 HIT incidents (see online supplementary appendix E), and improved performance was achieved in these experiments with relabeled HIT data.

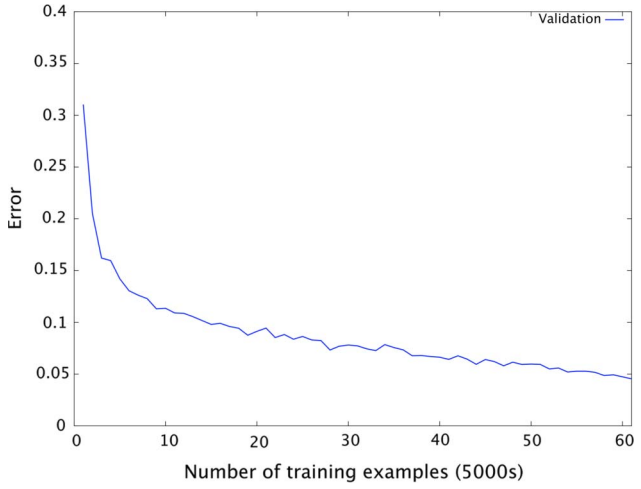

Figure 2 Learning curve of the best classifier on the full dataset.

The best performing classifier built and tested with 100% of the original stratified dataset achieved an F1 score of 0.972 with a selected λ value of 0.3 (figure 2). The classifier is trained in increments of 5000 randomly selected training examples, and the overall average training error is 0.02%.

Tokens for identifying HIT events

We identified the most important tokens for classifying HIT incidents by ranking their coefficients as shown in table 3. Intuitively, tokens such as 'handhelds', 'cport', and 'software/firmware' appear useful for identifying HIT incidents, while 'ceroplasm' and 'laboratory/diagnostic' may be useful for distinguishing non-HIT incidents. Tokens such as 'ra(number)q' are created during data preprocessing, and capture words such as 'ra1028q' are used in MAUDE incident reports.

DISCUSSION

Analyses of patient safety incidents involving HIT allow emerging problems to be identified and addressed in a timely manner. However, current methods for identifying these incidents are time consuming or non-exhaustive in discovering all HIT incidents. We evaluated text classification methods and showed that the approach is feasible for automatically identifying HIT incidents from large collections such as the FDA's MAUDE dataset. Classifiers based on a stratified training set achieved an average F1 score of 0.953 in identifying HIT incidents within a representative test dataset.

Semi-supervised learning

An important finding of this study is that semi-supervised learning may be necessary to apply machine learning techniques to big data analysis of patient safety incidents. The iterative workflow of classifying, analyzing, and retraining used in this study resembles a manual implementation of a semi-supervised learning approach. More specifically, a small number of labeled incidents (405 HIT) from MAUDE were initially used for training. We discovered that our classifiers identified many non-HIT incidents as HIT with a high probability, as many non-HIT had been incorrectly labeled. We retrained the classifiers on relabeled data and improved performance over our initial experiments. Semi-supervised learning can be used to perform these steps in an automated fashion. For example, incidents classified as HIT above a probability threshold (eg, 95%) can be automatically relabeled as HIT to retrain the classifier without human annotation. The improvements gained by performing this method manually provide evidence that the approach is suitable for MAUDE. In addition, the classifier learning curve in figure 2

Table 3 Top classification tokens

Token	Coefficient
handhelds	9.56
pacsic	3.49
cport	2.32
software/firmware	2.30
ra(number)q	2.18
ceroplasm	2.16
imagesanti	2.05
centrifued	2.01
millennium	1.89
laboratory/diagnostic	1.79

indicates that classification error may continue to decline if trained on more labeled incidents. Recent studies have used semi-supervised learning in text classification of medical data including patient discharge summaries and clinical reports.[20]

Classification performance

The high performance of our classifiers was achieved largely from conducting error analysis on the preliminary experiment results (0.953 F1 score; see online supplementary appendix E). More specifically, we rectified a number of mislabeled incidents for the relabeled data experiments that improved performance on 100% of the original stratified dataset from 0.759 to 0.954. Overall the performance of classifiers built and tested with 10% and 100% of the original stratified dataset (0.953 and 0.954 F1 score) was comparable to those built and tested with balanced datasets (0.995 F1 score). In addition, classifiers trained on stratified datasets outperformed those trained on balanced datasets in identifying HIT incidents within a representative test dataset. This poor performance can be attributed to the low precision of classifiers when trained on balanced data (0.165).

Compared with our previous studies in building and testing classifiers with a balanced dataset, we found that classifiers for HIT incidents performed better (0.995 F1 score) than our previous studies that used a similar approach for classifying up to 600 clinical handover incidents (0.84–0.92 F1 score) and up to 500 patient identification incidents (0.91–0.98 F1 score).[14] More advanced text classification methods such as n-grams where n>1, part-of-speech tagging, semantic role labeling, and relationship extraction can be investigated to further improve performance. However, the simple unigram features achieved an F1 score of 0.953, and advanced techniques may increase computational complexity for minor performance gains.

Feature selection

The use of feature-selection techniques did not significantly reduce the feature set size with the exception of stemming (73–79%) and principal component analysis (15%). However, stemming produced similar performance with an average F1 score of 0.945 compared with 0.953 when all features were used. These results suggest that stemming is a useful feature-selection technique for identifying HIT incidents in MAUDE.

Training classifiers with balanced datasets

We discovered that classifiers trained and tested with balanced datasets (0.994) marginally outperformed the classifiers trained and tested with 10% (0.945) and 100% (0.953) of the original dataset. These results are in accordance with studies that have shown that training classifiers on imbalanced data can degrade performance.[21 22] In addition, we experimented with training classifiers on balanced datasets but to validate and test on the stratified datasets. Our results show that classifiers trained, validated, and tested on balanced datasets (0.953 F1 score) overestimate classification performance compared with testing them on real-world stratified data (0.274–0.283 F1 scores).

Identifying rare HIT incidents

An undersampling approach was adopted by using balanced training datasets to improve the performance of classifying rare HIT incidents (0.297%) in MAUDE. We observed that this approach increases recall but significantly degrades precision and F1 score. For example, recall (0.929–0.989) is improved compared with the original stratified dataset (0.916–0.943). However, it could be argued that more weighting should be applied to recall than precision for calculating the F-score in a real-world context. For example, healthcare professionals will likely undertake detailed analysis of the identified HIT in order to learn and address problems. Therefore, it may be more important to identify potential HIT incidents even if many non-HIT incidents are incorrectly flagged (false positives) since they can be quickly discarded. In addition, other useful techniques such as ensemble methods for improving the rare class detection were not evaluated in this study and could be tested in future work.

Cost of classifying HIT incidents

The learning curve of the best performing classifier on the full dataset is displayed in figure 2. This curve illustrates that classification error may continue to decline if more incidents (ie, 300 000+) can be used for training or through incorporation of other types of text classification features. However, there are costs involved in acquiring human expertise to review and label potential HIT incidents retrieved from the initial keyword searches used in our previous work.[11] The learning curve indicates that convergence is not achieved, so we are unable to determine how many incidents need to be labeled to achieve stable performance. Additional costs incurred involve performing time-consuming but useful error analyses of classification results.

Big data

The number of incidents in MAUDE and other incident-monitoring systems is growing. For example, the subset of MAUDE used in this experiment contained 515 897 incidents, with 85 560 generated features that are represented as a matrix with approximately 44.1 billion elements. Therefore, we selected computationally efficient classification techniques and algorithms in this study that can scale to large datasets. Improving efficiency affords many opportunities, including the use of richer features such as semantic role labels in large-scale experiments and performing near-real-time classification of HIT incidents in MAUDE. However, there are other techniques not used in this study that can be evaluated in future work. For example, applying mini-batch gradient descent with a distributed and parallelized framework such as MapReduce[33] can reduce training time in addition to applying efficient feature-reduction techniques such as feature hashing.[34]

Limitations

We exclusively evaluated LR classifiers to identify rare HIT incidents using different combinations of balanced and stratified datasets and feature-selection techniques. The HIT incidents we used were voluntarily reported by vendors and users to MAUDE. We used incidents ranging from 1 January 2008 to 1 July 2010 and not the complete dataset starting from 1991. Therefore, it is unlikely that our dataset is representative of all types of HIT incidents reported to MAUDE and other incident-monitoring databases. It is likely that there are more HIT incidents that have not yet been discovered by our labeling. The labeling of training data is critical in developing an effective classifier. This study clearly shows how problems with the training data in the preliminary experiments can impair the development of an effective classifier. The original 405 HIT incidents were labeled in our previous work, using keyword search and then manual reviews. This may have introduced a selection bias and the classifiers are likely to overfit the data by identifying these keywords as the most effective features, which may not reveal the true difficulty of the problem. This was evident in the relabeled data experiments where handheld-related incidents

were relabeled as HIT, and the token with the highest coefficient from our new classifiers was 'handhelds'. Unfortunately, there is no other practical method of identifying HIT incidents from such large datasets without manual inspection of every incident, which is not feasible. In addition, there may be bias toward improved results because of the repeated use of the same dataset. Although we took great care to randomly select multiple training, validation, and test subsets, the fact that they were repeatedly sampled from the same dataset may have resulted in an overstatement of the classification performance. If the best classifier from this study were tested on new incident reports, its performance may be lower.

CONCLUSION

Statistical text classification is a feasible approach for automatically identifying HIT incidents from large collections of incidents. The use of automated methods can reduce the time for HIT problems to be identified and therefore analyzed and addressed by healthcare professionals. Semi-supervised learning may be necessary to apply machine learning techniques to big data analysis of medical incidents.

Acknowledgements The authors wish to thank Ms D Arachi for assistance with labeling incidents.

Contributors FM and EC: conceived the study. KEKC and SA: designed, implemented, and refined the text classifiers; KEKC, FM and SA: undertook the analysis and interpretation of findings; KEKC and FM: drafted the initial version of the manuscript; all authors contributed to subsequent versions and gave final approval.

Funding This research is supported in part by grants from the Australian National Health & Medical Research Council Centre for Research Excellence in e-Health 1032664 and Project Grant 1022964.

Competing interests None.

Provenance and peer review Not commissioned; externally peer reviewed.

Data sharing statement The data used in this study is freely available from the FDA Manufacturer and User Facility Device Experience (MAUDE) website.

REFERENCES

1 US Office of the National Coordinator for Health IT, HIT Policy Committee. Adoption/Certification Workgroup meeting 25 February 2010. http://healthit.hhs.gov/ (accessed Apr 2010).
2 US Joint Commission on Accreditation of Healthcare Organizations. 2008. http://jointcommission.org/SentinelEvents/SentinelEventAlert/sea_42.htm (accessed Dec 2008).
3 Ammenwerth E, Schnell-Inderst P, Machan C, *et al*. The effect of electronic prescribing on medication errors and adverse drug events: a systematic review. *J Am Med Inform Assoc* 2008;15:585–600.
4 Ash JS, Berg M, Coiera E. Some unintended consequences of information technology in health care: the nature of patient care information system-related errors. *J Am Med Inform Assoc* 2004;11:104–12.
5 Coiera E, Westbrook J, Wyatt J. The safety and quality of decision support systems. *Methods Inf Med* 2006;45(Suppl 1):20–5.
6 Coiera E, Aarts J, Kulikowski C. The dangerous decade. *J Am Med Inform Assoc* 2012;19:2–5.
7 Bloomrosen M, Starren J, Lorenzi NM, *et al*. Anticipating and addressing the unintended consequences of health IT and policy: a report from the AMIA 2009 Health Policy Meeting. *J Am Med Inform Assoc* 2011;18:82–90.
8 Karsh BT, Weinger MB, Abbott PA, *et al*. Health information technology: fallacies and sober realities. *J Am Med Inform Assoc* 2010;17:617–23.
9 Runciman WB, Edmonds MJ, Pradhan M. Setting priorities for patient safety. *Qual Saf Health Care* 2002;11:224–9.
10 Goodman KW, Berner ES, Dente MA, *et al*. Challenges in ethics, safety, best practices, and oversight regarding HIT vendors, their customers, and patients: a report of an AMIA special task force. *J Am Med Inform Assoc* 2011;18:77.
11 Magrabi F, Ong MS, Runciman W, *et al*. Using FDA reports to inform a classification for health information technology safety problems. *J Am Med Inform Assoc* 2012;19:45–53.
12 The Office of the National Coordinator for Health Information Technology, Health Information Technology Patient Safety Action & Surveillance Plan for Public Comment. 2012. http://www.healthit.gov/sites/default/files/safetyplanhhspubliccomment.pdf (accessed Jan 2013).
13 Magrabi F, Ong MS, Runciman W, *et al*. An analysis of computer-related patient safety incidents to inform the development of a classification. *J Am Med Inform Assoc* 2010;17:663–70.
14 Ong MS, Magrabi F, Coiera E. Automated categorisation of clinical incident reports using statistical text classification. *Qual Safety Health Care* 2010;19:1–7.
15 Ong MS, Magrabi F, Coiera E. Automated identification of extreme-risk events in clinical incident reports. *J Am Med Inform Assoc* 2012;19:110–18.
16 Savova GK, Olson JE, Murphy SP, *et al*. Automated discovery of drug treatment patterns for endocrine therapy of breast cancer within an electronic medical record. *J Am Med Inform Assoc* 2011;19:83–9.
17 Percha B, Nassif H, Lipson J, *et al*. Automatic classification of mammography reports by BI-RADS breast tissue composition class. *J Am Med Inform Assoc* 2012;19:913–16.
18 Xu Y, Tsujii J, Eric I, *et al*. Named entity recognition of follow-up and time information in 20 000 radiology reports. *J Am Med Inform Assoc* 2012;19:792–99.
19 Wagholikar KB, MacLaughlin KL, Henry MR, *et al*. Clinical decision support with automated text processing for cervical cancer screening. *J Am Med Inform Assoc* 2012;19:833–39.
20 Figueroa RL, Zeng-Treitler Q, Ngo LH, *et al*. Active learning for clinical text classification: is it better than random sampling? *J Am Med Inform Assoc* 2012;19:833–39.
21 Seiffert C, Khoshgoftaar TM, Van Hulse J, *et al*. Mining data with rare events: a case study. IEEE International Conference on Tools with Artificial Intelligence, Patras, Greece, 2007;2:132–39.
22 Weiss GM. Mining with rarity: a unifying framework. *SIGKDD Explorations* 2004;6:7–19.
23 Gwet K. Inter-rater reliability: dependency on trait prevalence and marginal homogeneity. *Statistical Methods Inter-rater Reliability Assessment* 2002;2:1–9.
24 Jolliffe IT. *Principal component analysis*. Berlin: Springer, 1986.
25 Lee S-i, Lee H, Abbeel P, *et al*. Efficient L1 Regularized Logistic Regression. The 21st AAAI Conference on Artificial Intelligence, Boston, USA 2006.
26 Liu B. *Web data mining: exploring hyperlinks, contents, and usage data*. Secaucus, USA: Springer Verlag, 2007.
27 Ng A, Jordan M. On discriminative vs. generative classifiers: a comparison of logistic regression and naive bayes. *Adv Neural Inf Process Syst* 2002;14:841.
28 Komarek P, Moore A. Fast robust logistic regression for large sparse datasets with binary outputs. *Artif Intell Stat* 2003;83.
29 Komarek P, Moore AW. Making logistic regression a core data mining tool with tr-irls. Fifth IEEE International Conference on Data Mining, Houston, USA, 2005:4.
30 Hestenes MR, Stiefel E. Methods of conjugate gradients for solving linear systems. *J Res Natl Bur Stand* 1952;49:409–36.
31 Hsu CW, Chang CC, Lin CJ. A practical guide to support vector classification. 2003. http://www.csie.ntu.edu.tw/~cjlin/papers/guide/guide.pdf (accessed 10 Oct 2012).
32 Davis J, Goadrich M. The relationship between Precision-Recall and ROC curves. *Proceedings of the 23rd International Conference on Machine Learning (ICML)*, Pittsburgh, USA, 2006:233–40.
33 Chu CT, Kim SK, Lin YA, *et al*. Map-reduce for machine learning on multicore. *Adv Neural Inf Process Sys* 2007;19:281.
34 Weinberger K, Dasgupta A, Langford J, *et al*. Feature hashing for large scale multitask learning. Proceedings of the 26th Annual International Conference on Machine Learning (ICML), Montreal, Canada, 2009:1113–20.

Using statistical text classification to identify health information technology incidents

Kevin E K Chai, Stephen Anthony, Enrico Coiera, et al.

J Am Med Inform Assoc 2013 20: 980-985 originally published online May 10, 2013
doi: 10.1136/amiajnl-2012-001409