Initializing and Growing a Database of Health Information Technology (HIT) Events by Using TF-IDF and Biterm Topic Modeling

Hong Kang, Zhiguo Yu, Yang Gong
School of Biomedical Informatics, the University of Texas Health Science Center at Houston, Houston, TX, USA

Abstract

Health information technology (HIT) events were listed in the top 10 technology-related hazards since one in six patient safety events (PSE) is related to HIT. Although it becomes a common sense that event reporting is an effective way to accumulate typical cases for learning, the lack of HIT event databases remains a challenge. Aiming to retrieve HIT events from millions of event reports related to medical devices in FDA Manufacturer and User Facility Device Experience (MAUDE) database, we proposed a novel identification strategy composed of a structured data-based filter and an unstructured data-based classifier using both TF-IDF and biterm topic. A dataset with 97% HIT events was retrieved from the raw database of 2015 FDA MAUDE, which contains approximately 0.4~0.9% HIT events. This strategy holds promise of initializing and growing an HIT database to meet the challenges of collecting, analyzing, sharing, and learning from HIT events at an aggregated level.

Introduction

Patient safety event (PSE), defined as any process, act of omission, or commission that results in hazardous healthcare conditions and/or unintended harm to the patient[1], is the third leading cause of death in the United States[2-4]. PSEs are complex and difficult to control because they are related to healthcare systems, operations, drug administration, or any clinical aspect of patient care[5]. Due to the wide application of health information technology (HIT) in clinical settings, HIT events become a key component of PSEs. HIT includes hardware or software that is used to electronically create, maintain, analyze, store, receive (information), or otherwise aid in the diagnosis, cure, mitigation, treatment, or prevention of disease, and that is not an integral part of an implantable device or medical equipment[6]. The positive impacts of HIT[7, 8] include cost savings and improved patient outcomes[9, 10], decreased occurrence of medication errors[11], and improved healthcare process measures across diverse settings[12]. However, if HIT is poorly designed or implemented, it poses a risk to patient safety[13-15]. For instance, an anesthesiologist did not know his patient had taken oxycodone because of the lack of interoperability between the office-based medical record platform and the inpatient system; the patient became somnolent upon receiving morphine after the oxycodone[16]. HIT events were listed in the top 10 technology-related hazards identified by the Emergency Care Research Institute among a range of common problems in 2015[17]. One in six PSEs is related to HIT, such as medical devices, EHRs, and CPOE[18]. Therefore, HIT related events pose a major threat and barrier toward a safer healthcare system, and the large number of HIT related events must be addressed to reduce patient harm.

Safety event reporting has been proven effective in many high-risk industries[19-21], for improving safety and enhancing organizational learning from errors. Healthcare systems have adopted event reporting since the 1999 Institute of Medicine (IOM) report[22] that greatly raised the public awareness of patient safety issues. Through collecting reports of adverse events and near misses in healthcare, reporting systems would enable safety specialists to analyze events, identify underlying factors, and generate actionable knowledge to mitigate risks[23]. In the U.S., the IOM recommended using patient safety reporting systems[24, 25] to determine why patients are sometimes harmed during medical care. The Agency for Healthcare Research and Quality (AHRQ) created the Common Formats (CF), common definitions and reporting formats[21] to help healthcare providers uniformly report PSEs.

However, collecting data on HIT-related PSEs for learning purposes is challenging owing to the lack of HIT reporting forms or platforms. Although CF forms are commonly used in Patient Safety Organizations (PSO), the CF form containing HIT event categories is defined at a very high level and embedded with the category of *device or medical/surgical supply*. This makes it difficult for reporters to recognize the proper categories when reporting, and as a result, reporters often leave fields blank rather than responding to the prepopulated questions. In the 2015 annual report of a PSO institute, the Missouri Center for Patient Safety, only one PSE was identified in the original report as an HIT event. Therefore, the scarcity of HIT event-exclusive databases and event reporting systems indicates the challenge of identifying the HIT events from existing resources.

EHR seems to be a potential resource to extract HIT events because EHRs are real-time, patient-centered records that make information available instantly and securely to authorized users. While an EHR does contain the medical and treatment histories of patients, an EHR system is built to go beyond standard clinical data collected in a provider's office and can be inclusive of a broader view of a patient's care[26]. However, EHR itself belongs to HIT, which may hide a large amount of EHR related events that happened during data entry and data transfer. In addition, there is no structured field that could help extract HIT events in an EHR that is not designed for such a purpose.

The U.S. FDA Manufacturer and User Facility Device Experience (MAUDE) database[27] is a rich and publicly accessible resource with the potential of extracting HIT events. Different from EHRs, the FDA MAUDE database focuses on the reports of events involving medical devices, voluntary reports of medical device malfunction, and reports of problems leading to serious injury and death since June 1993[28]. The database houses medical device reporting submitted to the FDA by mandatory reporters (manufacturers, importers and device user facilities) and voluntary reporters (healthcare professionals, patients and consumers). The FDA MAUDE is updated weekly and searchable online. As of February 2017, MAUDE had more than 6 million reports. The challenge of identifying HIT events from the FDA MAUDE is that only 0.1% of reports are related to HIT, and are mixed with other reports regarding equipment failures and hazards[29]. Due to the enormous size of the FDA MAUDE and the small percentage of HIT events, directly identifying and extracting all HIT events from the FDA MAUDE is almost impossible using a straightforward strategy. The classification of data with imbalanced class distribution has encountered a significant drawback of the performance attainable by most standard classifier learning algorithms[30]. In addition, the rapid evolvement of topics and free text regarding HIT events presents another challenge for distinguishing these events from others. The initialization of HIT event database based on the FDA MAUDE database requires an up-to-date, efficient, and effective strategy.

A MAUDE report is composed of structured fields (device and patient information) and unstructured fields (textual information). In the preliminary study, we developed an HIT filter based on the generic name and manufacturer name, two structured fields of the 2015 MAUDE database, enabling us to screen the FDA MAUDE and consequently create a report subset with more than 50% HIT events[31]. We sampled and reviewed the filtered reports which resulted in an estimate that 0.4~0.9% FDA MAUDE reports were HIT events. Although this ratio has been significantly enhanced, it is still far away from establishing an HIT dataset. Fortunately, a classifier based on unstructured field may be feasible to further identify HIT events, since the information from unstructured data provides an integrated and comprehensive view. In addition, the ratio of the HIT events in filtered data is perfect for classifier training and evaluation.

In this study, six popular classification algorithms were compared on 2015 MAUDE database by using term frequency-inverse document frequency (TF-IDF)[32] as the feature of each single word in the unstructured fields. Then biterm topic model (BTM)[33], a word co-occurrence based topic model that learns information from word combinations, was applied to further improve the best classifier of the six. After discussing the balance of precision and recall, according to the learning requirement, a final model that can provide a dataset with more than 97% HIT events was proposed. This model offers the probability of initializing and growing an HIT database from the FDA MAUDE and other potential resources. The outcome will be a timely reflection of the evolvement of HIT events and will be helpful for enriching HIT knowledge and better using the historic reports toward an overall understanding and analysis of the characteristics, occurrence, observation, and description of HIT events. In addition, this project provides an intelligent strategy to connect structured data with unstructured data and holds promise in triggering a revolution of data management and analysis in healthcare and other industries.

Methods

Improve the filter for the structured fields of FDA MAUDE

Most of the 45 structured fields of the FDA MAUDE database are either left blank by reporters or are of little use for the purpose of identifying HIT related events. The only two fields have the greatest potential in identifying HIT related events are *generic name* and *manufacturer name*. Therefore, a filter based on these two fields was established in our preliminary study[31].

We started with a set of common computer hardware and software related keywords that had been previously identified[34]. The starting keyword list was expanded by the addition of several generic terms such as "software," "program," and "hardware" and several more modern terms such as "electronic medical record" and "portal technology." Then all of the generic names from the FDA MAUDE database starting from Jan 2010 to Dec 2015 were extracted, yielding a total of 60,000 unique generic names. Next, through partial keyword matching of the generic names to the keywords list, a subset of generic names appearing HIT-related was extracted. The subset was further

analyzed by domain experts to determine which generic names were actually linked to HIT reports by using a small portion of the 2015 FDA MAUDE database. A similar approach was utilized in selecting a list of manufacturers for the filter. We started with a list of seven popular HIT manufacturers[34] and then added 347 manufacturers of HIT software. The reports from the 2015 FDA MAUDE database related to those manufacturers were extracted and checked to determine the most relevant manufacturers.

To evaluate and improve the HIT filter, the filter was first applied to the 2015 FDA MAUDE database. Then a subset (10%) of the reports screened by the filter was manually reviewed by two domain experts. The experts labeled each report with one of three labels: HIT, Not HIT, or Unsure. The reports with disagreements among the reviewers were resolved through group discussion. During the review, our team narrowed the HIT definition to identify the most clinically relevant and consequential HIT devices. Under our current understanding, an HIT device is any device that utilizes both hardware and software to facilitate health information exchange in order to aid in the diagnosis, treatment, or prevention of disease. Using this definition, priority is given to HIT systems that focus on information exchange such as electronic health records, computerized physician order entry, and picture archiving communications systems. Implantable devices, glucose monitors, defibrillators, and similar devices are excluded under this definition, as they do not actively facilitate health information exchange.

Compare six classification algorithms based on the unstructured fields of FDA MAUDE

Six popular machine learning classifiers including logistic regression, support vector machine (SVM), naïve Bayes, decision tree, JRip rules, and random forest were constructed using the unstructured data of the reviewed reports. Weka 3.8.1[35] was applied for training and validating the classifiers. The unstructured data of 2015 FDA MAUDE was filtered by removing the words in the Rainbow stoplist[36]. Then each report in the labeled training set was treated as a vector of words and was weighted by a TF-IDF technique. The effectiveness of this word-based information retrieval approach has been proven effective in searches of patient records, which are often corrupted by misspelled words and conventional graphs or abbreviations[37, 38]. Eq.1 shows how to calculate the weight of term x within document y using TF-IDF ($tf_{x,y}$ is the frequency of term x in document y, df_x is the number of documents containing term x, N is the total number of documents). The six classifiers were evaluated using leave-one-out cross-validation (LOOCV) and their performances were weighted based on both their F-scores and ROC areas.

$$w_{x,y} = tf_{x,y} \times \log(\frac{N}{df_x}) \quad \text{(Eq.1)}$$

Improve HIT classifier by using BTM

TF-IDF only considers the information from a single word and its distribution across the whole corpus, but cannot capture the semantics information among the dataset. Topic models[39] are a class of statistical machine learning algorithms that extract the semantic themes (topics) from a corpus of documents. These topics generated based on the document-level word co-occurrence patterns describe the thematic composition of each document and can thus capture the semantic similarity between topics and documents. Moreover, those generated topics have been widely used as features to improve document classification and information retrieval performance[40, 41].

The BTM[33] that we used in this work to improve the HIT classifier is specifically designed for short text documents. In general, an HIT report is too short to provide enough word counts for conventional topic models, like PLSA[42] and LDA[43], to know how words are related, comparing to lengthy documents[44]. Also, the limited contexts make it very difficult for topic models to identify the sense of ambiguous words in short documents. BTM overcomes those short text limitations by assuming that all the biterms (i.e. unordered co-occurring word pairs) are generated by a corpus level topic distribution to benefit from the global word co-occurring patterns. Informally, the "generative story" for BTM is as follows. Firstly, each topic's word distribution is drawn over the vocabulary of the whole corpus. Then a corpus-level topic distribution is drawn to describe this dataset. To generate each biterm in the biterm set extracted from the whole corpus, one draws a topic from the corpus-level topic distribution and subsequently selects two words from this distribution across all vocabulary of the whole corpus corresponding to this topic. BTM uses this generative process to model the co-occurrence of a pair of words over the whole corpus to reveal the topics than just the occurrence of a single word, and then enhance the learning of topics. Specifically, it estimates the parameters that define the topic mixture over the whole corpus and the conditional probability of each word given each topic. The topic distribution of each document can then be naturally derived based on the learned model. Parameter estimation is done via Gibbs sampling approach.

The number of topics generated by BTM must be pre-specified. Determining the "right" number of topics for different data sets remains a challenge. When the number of topics increases, redundant and nonsense topics may be generated.

In this study, we conducted experiments with various numbers of topics ranging from 5 to100. The performance in the classification tasks was quite consistent after a particular topic number threshold. Hence, it is likely that a BTM with a relatively large number of topics will capture all the key themes over this corpus and in further to improve the document classification task.

Results

HIT event filter based on structured data

Based on the analysis of pre-viewed FDA MAUDE reports and the discussion with two healthcare professionals, 58 keywords from *generic names* (39 software and 19 hardware keywords respectively) and 16 keywords from *manufacturer names* were determined to compose the filter for HIT related reports, as shown in Table 1. The filter was first applied to the 2015 FDA MAUDE database including 860,915 reports. 4871 reports (2479 software and 2392 hardware reports) were initially found. 490 reports (10%) were randomly selected according to the keyword distribution for expert review and labeling. 289 reports were identified as HIT related by experts, which means the filter can generate a report subset from original 2015 MAUDE database with about 50-60% HIT related reports. This proportion is much more practicable in terms of classifier training than that on the entire MAUDE database (0.4~0.9%)[29].

Table 1. Keywords of the HIT filter in alphabetical order

Generic Names				Manufacturer Names
Hardware related:	Telemetry Monitor,	Dispensing System,	Management System,	Allscripts,
Anesthesia Monitor,	Telemetry Transmitter,	Dose Suggestion,	Monitoring System,	Centricity,
Apnea Monitor,	Vital Sign Monitor,	Downloader,	Network,	Cerner,
Arterial Monitor,	Workstation	Drug Suggestion,	Order Entry,	Epic,
Atlas Monitor,		EHR,	PACS,	GE Healthcare,
Blood Pressure Monitor,	***Software related:***	Electronic Heath,	Picture Archiving,	Hass,
Central Monitor,	ADC,	Electronic Medical,	Portal,	Healthtronics,
Computer,	Alert,	Electronic Patient,	Powerchart,	Henry Schein,
Console Monitor,	Automated Dispensing Cabinet,	EMR,	Program,	Homer,
Drug Screen,	Communication Device,	ICT,	Server,	Isite,
Fetal Monitor,	Communication System,	Imaging System,	Soarian,	iSOFT,
Patient Monitor,	CPOE,	Information System,	Telemetry,	Kestral,
Physiological Monitor,	Data Backup,	Internet,	Trima Accel Platelet,	McKesson,
Pressure Monitor,	Database,	Invision,	Web	Medical Director,
PT Monitor,	Decision Support,	LIS,		MedPro,
Safety Monitor,	Digital,			Oasis

HIT Classifiers using TF-IDF model

The manually labeled reports (289 HIT and 376 non-HIT reports) from 2015 MAUDE database were applied as the training set of the classifiers. 1,541 words extracted from the unstructured fields were fixed as the feature set after removing stopwords. TF-IDF was calculated for each word in each report. A comparison was made among the classifiers trained using six popular classification algorithms: logistic regression, random forest, naïve Bayes, SVM, decision tree (J48), and JRip rules. As shown in Table 2 and Figure 1, random forest has the best performance in terms of accuracy, ROC area, and F-score.

Table 2. Performances of the six HIT classifiers

Method	Accuracy	F-score	ROC
Random Forest	0.809	0.807	0.900
Logistic Regression	0.802	0.801	0.871
Naïve Bayes	0.786	0.786	0.853
SVM	0.792	0.792	0.787
Decision Tree (J48)	0.737	0.737	0.718
JRip Rules	0.716	0.715	0.716

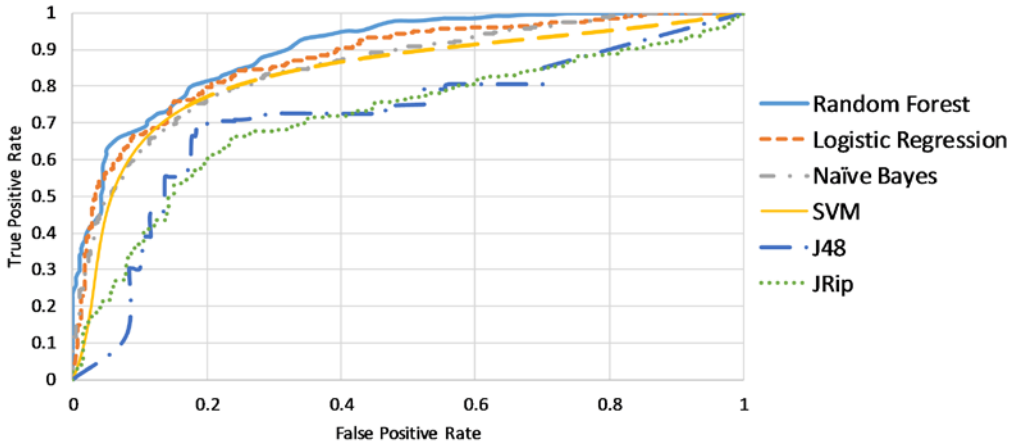

Figure 1. ROC areas of six HIT classifiers by using TF-IDF as feature (word) values

HIT Classifiers using BTM

BTM was implemented to analyze the words of the unstructured fields of the training set and to discover the topics without any prior annotations or labeling of the reports. This statistical model reflects the intuition that narrative reports exhibit multiple topics. Each report exhibits the topics in different proportions; each word in a report is drawn from one of the topics, where the selected topic is chosen from the per-document distribution over topics. All the reports in the collection share the same set of topics, but each report exhibits those topics in different proportions. We applied the topics as the feature set and the weights of the topics as the features' values. To figure out the best topic number, BTM was run for 20 times by setting topic numbers from 5 to 100 with an interval 5. Random forest classifiers were also trained for 20 times as the topic number changed. As shown in Figure 2, most BTM-based classifiers have higher F-scores than the TF-IDF-based classifier, while the ROC areas are not improved by BTM. The BTM-based classifiers with the best F-score and ROC were obtained when setting the topic number at 70 and 80. Therefore, we combined the features of BTM and TF-IDF at each of the two topic numbers and trained two additional classifiers respectively (the dots in Figure 2). The classifier combining TF-IDF features with 70 topics from BTM has the best performance with ROC area 0.920 and F-score 0.834, which are significant improvements comparing those of the original TF-IDF-based classifier.

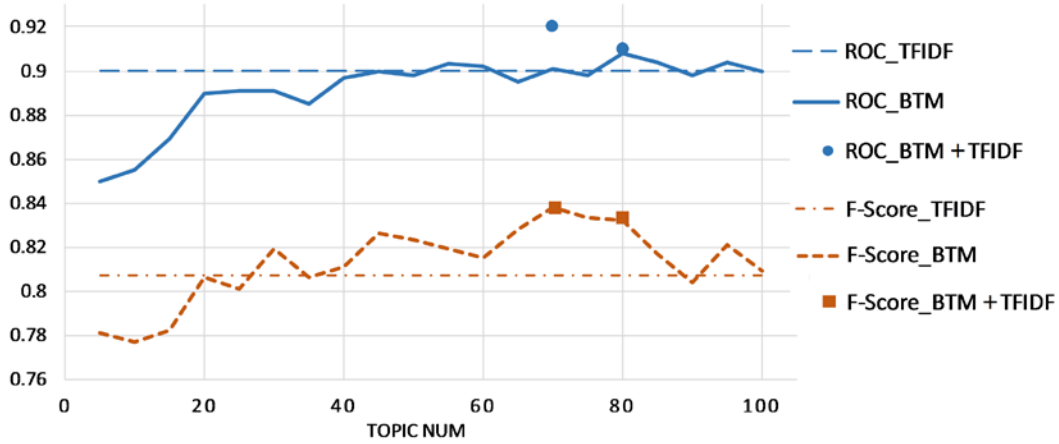

Figure 2. BTM improves TF-IDF model for HIT classification

A dataset with 97% HIT events

The original purpose of building the HIT filter and classifiers was not to identify all HIT related reports from the FDA MAUDE. Instead, we need an HIT exclusive database or a report dataset with a large proportion of HIT related events. In other word, we can tolerate the loss of recall to obtain a higher precision. Random forests build an ensemble of tree classification predictors using bagging. Each node of the trees only considers a small subset of features for the split. The classification is done by voting, which means a probability is calculated based on the voting result of all predictors.

Figure 3 shows the changes of true positive and false positive as the threshold of the probability changes. Based on the case review, we found that one third of the HIT related events are typical enough for learning purposes. Therefore, if setting the threshold at 0.7, we could get a high precision 0.97 and adequate true positive cases. By using this strategy, a report dataset with 97% HIT related events was derived from the original 2015 FDA MAUDE, which contains up to 0.9% HIT related events.

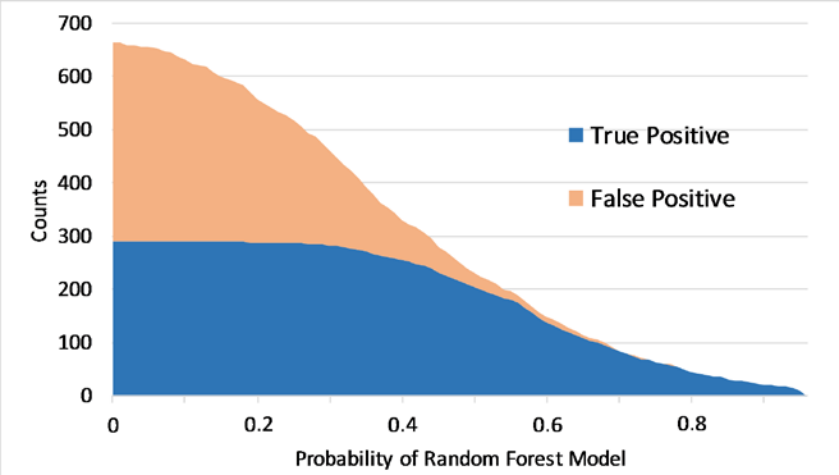

Figure 3. The effects of random forest probability

Discussion

Initialize and grow an HIT database

Learning directly from the AHRQ CF and FDA MAUDE is challenging since the two reporting forms are not HIT exclusive. This study provides an effective way to extract HIT related events from the FDA MAUDE and initialize an HIT event database. To establish and grow the database, meanwhile, further improve the HIT filter and classifier, we will apply the HIT event filter and classifier on the 2014 dataset. Only the cases that are identified as HIT will be manually reviewed. All the reviewed cases will be added to the training set, which will consequently help build a new classifier. The same process will, in turn, be retrospectively applied on the datasets from 2013, 2012, 2011, 2010, and beyond. At the end of each iteration, the four training methods will be re-evaluated based on the corresponding manual review, and the best method will be used in the classifier. The classifier is expected to be improved as more labeled cases are included. As the manually reviewed HIT events are accumulated, an HIT event database will be established and keep growing.

In our preliminary study, we prototyped a user-centered PSE reporting system based on a PSE knowledge base, which includes PSE reports, solutions, and their connections[45]. Patient falls, another important PSE subtype, were applied to test the system. Users can either choose an existing fall case or report a new case, then the system retrieves similar cases and customized solutions based on the query and the reporter's role (e.g., manager, clinician, staff, patient). The user preference may be diverse for different purposes. The system also allows the user to click the feedback button to indicate their preferences to a certain similar case or solution. All feedback will be returned to the algorithm implementation step in order to update the weights of similarity matrices and dynamically upgrade the system performance. This mechanism, similar to the ranking strategy of the Google search engine, will gradually stabilize the similarity matrices, making them more convincing as the feedback increases. The HIT identification strategy together with the TF-IDF and BTM features will be incorporated in this system to grow the database.

Explore contributing factors and connections of HIT events for shared learning

Intuitively, finding a way to compare two relevant HIT cases will be beneficial for learning from previous cases. The FDA MAUDE database is rich, broad and unique because it is a collection of errors of many HIT products from prevailing vendors such as Epic, Cerner, GE Healthcare, Allscripts, and McKesson; the connections among the reported errors will be helpful for their users and developers alike. The benefit once perceived would attract more HIT vendors and users to join the reporting and promote shared learning. The BTM can map the HIT events to the topic space that each event has an HIT topic rank according to the relationship between this event and each topic. We can observe how those topics changed over time and how they are connected to each other. Rather than finding reports

through keyword searches alone, we can find a small group of central topics as contributing factors and then examine the reports related to the topics. More importantly, the topics can help compare and measure the vector-based similarity between HIT reports, which could provide more targeted knowledge support to the reported cases in comparison with similar or relevant cases. We can also extract the common characteristics of the HIT cases within a certain similarity range, which will help us track the changes of HIT events and further improve the HIT identification strategy and benefit healthcare professionals for shared learning.

Reduce the human labor

Manually reviewing all cases in the FDA MAUDE database was simply infeasible and that machine learning was likely to be the only viable approach. Traditionally, the two paradigms of machine learning have been supervised (all labeled data) and unsupervised learning (all unlabeled data). However, recently, much attention has been placed on semi-supervised learning for its ability to utilize only a small amount of labeled cases combined with a large amount of unlabeled cases to improve classification accuracy. In the case of HIT, this approach seemed to be well-suited as the cost of labeling narrative text by manual review is quite high, while the cost of obtaining unlabeled reports is minuscule in comparison. One of the simplest methods within semi-supervised learning is self-training. In self-training, predictions made with high confidence by the classifier are added back to the labeled data, so that the classifier keeps being updated and improved after each iteration. This approach seems to a viable method to extract and classify HIT reports from large databases such as the FDA MAUDE.

Connecting structured and unstructured data is essential for patient safety data retrieval

The variety of reporting formats is the primary challenge to improving quality and connection of PSE reports. Our preliminary work suggests that communication among PSEs can be established if we extract all the necessary information and properly annotate it to the same hierarchical feature structure[46]. However, when the quality of the structured features is unsatisfactory (e.g., HIT related reports in AHRQ CF), we need to find more information in the unstructured data (narrative reports) to ensure a comprehensive view. An analogy can be made with the U.S. National Library of Medicine's PubMed literature filtering, in which a controlled vocabulary, Medical Subject Headings (MeSH) was developed to index MEDLINE articles. MeSH has been successfully used to improve PubMed query results, information retrieval, document clustering, and query refinement in "downstream" applications that use PubMed abstracts[47, 48]. Similarly, a MeSH-like patient safety vocabulary holds promise in connecting structured and unstructured features and in becoming the framework for the PSE knowledge base, the collection of PSE reports, solutions, and the potential connections among them. The TF-IDF and BTM features in this study have the potential to explore the "MeSH" in patient safety domain and standardize PSE reporting and management.

Benefits of shared learning and reporting systems

While all technology is fallible, technology-induced errors in healthcare may have far more serious repercussions than those in other fields. Even with extensive testing, HIT may be especially vulnerable to failure due to the fast pace and complexity of the healthcare system. As an example, many EHR systems perform well in testing when used by only one user of each user category in an office setting. However, when released to the public, the EHR systems may malfunction as a group of physicians from all areas of medicine may use them in a variety of settings. Manufacturers of technology often do not have the resources to test their products on such a large scale and must instead rely on user generated reports to inform them of possible defects with their products. Perhaps, most well-known, Microsoft's error reporting system (codenamed Watson) has allowed the company to collect and learn from billions of error reports all across the world. Integrated into Windows, Watson automatically collects information on program crashes and sends it to Microsoft. Microsoft analyzes each crash report, and if available, sends a solution back to the user. Watson has played an integral role in helping Microsoft identify errors and has also helped the company prioritize its debugging efforts.

Implementation of a similar system for HIT events could yield tremendous benefits. With the knowledge gained from a shared learning system, patient safety experts could analyze the distribution of HIT events and identify common underlying factors among the reports. Safety experts could then prioritize their efforts to generate actionable solutions for the most important and pressing patient safety events. While healthcare systems have utilized event reporting in the last decade to increase awareness of patient safety issues, several barriers still exist that prevent patient safety reporting systems from reaching their full potential. These include the fact that many reporting systems are still paper based, which prevents data sharing; many of the user reports may be hard to understand due to ambiguous and vague language used; and much of the patient safety data is scattered across many reporting systems, which are challenging to aggregate.

Barriers of voluntary reporting systems

The proposed HIT database was built on the voluntary reports generated by clinicians, patients, or consumers. Such reports are indispensable for preventing future occurrences of these adverse events and help foster a healthcare culture driven toward a safer healthcare system. Unfortunately, according to the US Department of Health and Human Services, 86% of medical errors go unreported[49]. Many barriers still exist that prevent safety reporting systems from reaching their full potential. Perhaps a primary reason safety events go unreported, regardless of the field they occur in, is that people feel their reports will not make a difference. The origins of their feelings are completely understandable. In a well-publicized incident, Toyota was fined 1.2 billion dollars for negligence in addressing known safety issues that caused unintended acceleration in their vehicles[50]. Prior safety reports were not enough to cause the company to fix the safety issue, resulting in a series of tragedies. In one tragedy, Officer Mark Saylor was driving his car with his family when the accelerator became stuck. Unable to free the accelerator, Mr. Saylor called 911 to narrate his terrifying ordeal. However, it was too late and Mr. Saylor and his family died after crashing into an SUV[51].

In a high stakes field such as healthcare, it is critical that errors are reported and that manufacturers are held accountable for their products. In regards to HIT, the challenge to get healthcare providers to report these events may be especially high. With the massive spending and efforts taken to adopt HIT by hospitals and the government, there may be a positive bias toward reporting the benefits of HIT and a negative bias toward reporting adverse events caused by HIT. To make matters worse, even when safety events are reported, manufacturers may not have enough resources allocated toward diagnosing and fixing those issues. While the staggering pace of technology has driven much innovation in healthcare, the need to take a careful look at the adverse events caused by HIT has never been greater. Doing so may ultimately keep HIT on the right track toward becoming a safe and integral part of the healthcare system.

Limitations

There is no exact figure about the proportion of HIT related reports in the FDA MAUDE, because the proportion was only an estimation based on limited sampling and reviewing[29]. After applying the HIT filter on the 2015 FDA MAUDE, a subset (10%) of the reports of the screened data was selected for manual review to save human labor. Thus, the ratio 50-60% representing the HIT related reports among the filtered data was estimated based on the 10% random chosen reports. There might be a bias when using the subset's profile to predict the profile of the whole dataset. An effective approach to reduce the bias could be utilizing a semi-supervised learning to grow the HIT database. The classifier can be trained using an iterative approach, and cases that are labeled by the classifier with a high probability (e.g., >0.90) of being HIT related are added directly to the training set. Using this self-training semi-supervised learning approach, it is possible to introduce biases and for early mistakes to become compounded through subsequent rounds of training. Despite several rounds of error analysis can be performed to prevent such an outcome, selection bias toward certain HIT events cannot be completely eliminated.

Conclusion

A strategy to initialize and grow a database for HIT related event reports from the FDA MAUDE database was proposed by retrieving the information from both structured and unstructured fields. The strategy helps us connect up to 0.9% FDA MAUDE reports with HIT events. The creation of this database holds promise in aiding the understanding, characterization, discovery, and reporting of HIT related events.

Acknowledgement

This project is supported by UTHealth Innovation for Cancer Prevention Research Training Program Post-Doctoral Fellow-ship (Cancer Prevention and Research Institute of Texas grant #RP160015), Agency for Healthcare Research & Quality (1R01HS022895), and University of Texas System Grants Program (#156374).

References

1. International Classification for Patient Safety: World Health Organization. Available from: http://www.who.int/patientsafety/implementation/taxonomy/development_site/en/.
2. Barach P, Small SD. Reporting and preventing medical mishaps: lessons from non-medical near miss reporting systems. BMJ. 2000;320(7237):759-63.
3. Cha AE. Researchers: Medical errors now third leading cause of death in United States. The Washington Post. 2016 May 3.
4. James JT. A new, evidence-based estimate of patient harms associated with hospital care. J Patient Saf. 2013;9(3):122-8.

5. Patient safety event report: Center for Leadership, Innovation and Research in EMS. Available from: http://event.clirems.org/Patient-Safety-Event.
6. Common Formats: Agency for Healthcare Research and Quality. Available from: https://pso.ahrq.gov/common.
7. Chaudhry B, Wang J, Wu S, Maglione M, Mojica W, Roth E, et al. Systematic review: impact of health information technology on quality, efficiency, and costs of medical care. Ann Intern Med. 2006;144(10):742-52.
8. Goldzweig CL, Towfigh A, Maglione M, Shekelle PG. Costs and benefits of health information technology: new trends from the literature. Health Aff (Millwood). 2009;28(2):w282-93.
9. Garg AX, Adhikari NK, McDonald H, Rosas-Arellano MP, Devereaux PJ, Beyene J, et al. Effects of computerized clinical decision support systems on practitioner performance and patient outcomes: a systematic review. JAMA. 2005;293(10):1223-38.
10. Amarasingham R, Plantinga L, Diener-West M, Gaskin DJ, Powe NR. Clinical information technologies and inpatient outcomes: a multiple hospital study. Arch Intern Med. 2009;169(2):108-14.
11. Kaushal R, Shojania KG, Bates DW. Effects of computerized physician order entry and clinical decision support systems on medication safety: a systematic review. Arch Intern Med. 2003;163(12):1409-16.
12. Bright TJ, Wong A, Dhurjati R, Bristow E, Bastian L, Coeytaux RR, et al. Effect of clinical decision-support systems: a systematic review. Ann Intern Med. 2012;157(1):29-43.
13. Ammenwerth E, Schnell-Inderst P, Machan C, Siebert U. The effect of electronic prescribing on medication errors and adverse drug events: a systematic review. J Am Med Inform Assoc. 2008;15(5):585-600.
14. 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(2):104-12.
15. Coiera E, Westbrook J, Wyatt J. The safety and quality of decision support systems. Yearb Med Inform. 2006:20-5.
16. Adler-Milstein J. Case 382: Falling Between the Cracks in the Software: AHRQ; 2016 [cited 2016 Aug 8]. Available from: https://psnet.ahrq.gov/webmm/case/382.
17. Top 10 health technology hazards for 2015. Emergency Care Research Institute (ECRI), November 2014.
18. Cheung KC, van der Veen W, Bouvy ML, Wensing M, van den Bemt PM, de Smet PA. Classification of medication incidents associated with information technology. J Am Med Inform Assoc. 2014;21(e1):e63-70.
19. Billings CE. Some hopes and concerns regarding medical event-reporting systems: lessons from the NASA Aviation Safety Reporting System. Arch Pathol Lab Med. 1998;122(3):214-5.
20. Kivlahan C, Sangster W, Nelson K, Buddenbaum J, Lobenstein K. Developing a comprehensive electronic adverse event reporting system in an academic health center. Jt Comm J Qual Improv. 2002;28(11):583-94.
21. Jha AK, Prasopa-Plaizier N, Larizgoitia I, Bates DW, Research Priority Setting Working Group of the WHOWAfPS. Patient safety research: an overview of the global evidence. Qual Saf Health Care. 2010;19(1):42-7.
22. Kohn LT, Corrigan JM, Donaldson MS. To err is human: building a safer health system. U.S. Institute of Medicine, 1999.
23. Shaw R, Drever F, Hughes H, Osborn S, Williams S. Adverse events and near miss reporting in the NHS. Qual Saf Health Care. 2005;14(4):279-83.
24. Holzmueller CG, Pronovost PJ, Dickman F, Thompson DA, Wu AW, Lubomski LH, et al. Creating the web-based intensive care unit safety reporting system. J Am Med Inform Assoc. 2005;12(2):130-9.
25. Edwards SA. Computer-based management gaming: a method of executive development. Hospitals. 1965;39(24):59-60.
26. What is an electronic health record (EHR)? : HealthIT.gov. Available from: https://www.healthit.gov/providers-professionals/faqs/what-electronic-health-record-ehr.
27. MAUDE - Manufacturer and User Facility Device Experience [Internet]. U.S. Food & Drug Administration. Available from: https://www.accessdata.fda.gov/scripts/cdrh/cfdocs/cfmaude/search.cfm.
28. Tamura N, Terashita T, Ogasawara K. [Development of an attitude-measurement questionnaire using the semantic differential technique: defining the attitudes of radiological technology students toward X-ray examination]. Nihon Hoshasen Gijutsu Gakkai Zasshi. 2014;70(3):206-12.
29. Magrabi F, Ong MS, Runciman W, Coiera E. Using FDA reports to inform a classification for health information technology safety problems. J Am Med Inform Assoc. 2012;19(1):45-53.
30. Mistry M, Pavlidis P. Gene Ontology term overlap as a measure of gene functional similarity. BMC Bioinformatics. 2008;9:327.
31. Kang H, Wang F, Zhou S, Miao Q, Gong Y. Identifying and Synchronizing Health Information Technology (HIT) Events from FDA Medical Device Reports. MedInfo 2017 (Accepted).
32. Salton G, McGill MJ. Introduction to modern information retrieval. New York: McGraw-Hill; 1983.

33. Yan X, Guo J, Lan Y, Cheng X. A Biterm Topic Model For Short Text. WWW2013; Rio de Janeiro, Brazil2013.
34. Magrabi F, Ong MS, Runciman W, Coiera E. An analysis of computer-related patient safety incidents to inform the development of a classification. J Am Med Inform Assoc. 2010;17(6):663-70.
35. Weka 3: Data Mining Software in Java: the University of Waikato. Available from: http://www.cs.waikato.ac.nz/ml/weka/.
36. Rainbow stoplist. Available from: http://www.cs.cmu.edu/~mccallum/bow/rainbow/.
37. Ruch P, Baud R, Geissbuhler A. Evaluating and reducing the effect of data corruption when applying bag of words approaches to medical records. Int J Med Inform. 2002;67(1-3):75-83.
38. Luo G, Tang C, Yang H, Wei X, editors. MedSearch: a specialized search engine for medical information retrieval. CIKM '08 Proceedings of the 17th ACM conference on Information and knowledge management; 2008; New York: ACM.
39. Blei D, Carin L, Dunson D. Probabilistic Topic Models: A focus on graphical model design and applications to document and image analysis. IEEE Signal Process Mag. 2010;27(6):55-65.
40. Yu Z, Bernstam E, Cohen T, Wallace BC, Johnson TR. Improving the utility of MeSH(R) terms using the TopicalMeSH representation. J Biomed Inform. 2016;61:77-86.
41. Rubin TN, Chambers A, Smyth P, Steyvers M. Statistical topic models for multi-label document classification. Machine learning. 2012;88(1-2):157-208.
42. Hofmann T. Probabilistic latent semantic indexing. The 22nd annual international ACM SIGIR conference on Research and development in information retrieval; Berkeley, California: ACM; 1999.
43. Blei DM, Ng AY, Jordan. MI. Latent dirichlet allocation. Journal of machine Learning research. 2003;Jan(2003):993-1022.
44. Hong L, Davison BD. Empirical study of topic modeling in twitter. The first workshop on social media analytics; Washington D.C.: ACM; 2010.
45. Kang H, Gong Y. Design of a user-centered voluntary reporting system for patient safety events. MedInfo 2017 (Accepted).
46. Kang H, Gong Y. A Novel Schema to Enhance Data Quality of Patient Safety Event Reports. AMIA Annu Symp Proc. 2016;2016:1840-9.
47. Lu Z, Kim W, Wilbur WJ. Evaluation of Query Expansion Using MeSH in PubMed. Inf Retr Boston. 2009;12(1):69-80.
48. Richter RR, Austin TM. Using MeSH (medical subject headings) to enhance PubMed search strategies for evidence-based practice in physical therapy. Phys Ther. 2012;92(1):124-32.
49. Levinson D. Hospital Incident Reporting Systems Do Not Capture Most Patient Harm. Washington, DC: US Department of Health and Human Services, 2012.
50. Justice Department Announces Criminal Charge Against Toyota Motor Corporation and Deferred Prosecution Agreement with $1.2 Billion Financial Penalty: U.S. Department of Justice; 2014. Available from: https://www.justice.gov/opa/pr/justice-department-announces-criminal-charge-against-toyota-motor-corporation-and-deferred.
51. Devine R, Payton M, Stickney R. CHP Officer, Family Killed in Crash 2009. Available from: http://www.nbcsandiego.com/news/local/CHP-Officer-Family-Killed-in-Crash-56629472.html.