Using Active Learning to Identify Health Information Technology Related Patient Safety Events

Allan Fong[1]; Jessica L. Howe[1]; Katharine T. Adams[1]; Raj M. Ratwani[1,2]

[1]MedStar Institute for Innovation – National Center for Human Factors in Healthcare, 3007 Tilden St. NW, Suite 7M, Washington, D.C. 20008, USA;
[2]Georgetown University Medical Center, 3800 Reservoir Rd NW, Washington, DC 20007

Keywords
Patient safety, health information technology, active learning, human-in-the-loop, machine learning, patient safety event reports

Summary
The widespread adoption of health information technology (HIT) has led to new patient safety hazards that are often difficult to identify. Patient safety event reports, which are self-reported descriptions of safety hazards, provide one view of potential HIT-related safety events. However, identifying HIT-related reports can be challenging as they are often categorized under other more predominate clinical categories. This challenge of identifying HIT-related reports is exacerbated by the increasing number and complexity of reports which pose challenges to human annotators that must manually review reports. In this paper, we apply active learning techniques to support classification of patient safety event reports as HIT-related. We evaluated different strategies and demonstrated a 30% increase in average precision of a confirmatory sampling strategy over a baseline no active learning approach after 10 learning iterations.

Correspondence to:
Allan Fong
National Center for Human Factors in Healthcare
3007 Tilden St. NW, Suite 7M
Washington, D.C. 20008, USA
202–244–9807
Email: allan.fong@medicalhfe.org

Appl Clin Inform 2017; 8: 35–46
https://doi.org/10.4338/ACI-2016-09-CR-0148
received: September 1, 2016
accepted: November 9, 2016
published: January 18, 2017
Citation: Fong A, Howe JL, Adams KT, Ratwani RM. Using active learning to identify health information technology related patient safety events. Appl Clin Inform 2017; 8: 35–46
https://doi.org/10.4338/ACI-2016-09-CR-0148
Funding
This project was funded under contract/grant number Grant R01 HS023701–02 from the Agency for Healthcare Research and Quality (AHRQ), U.S. Department of Health and Human Services. The opinions expressed in this document are those of the authors and do not reflect the official position of AHRQ or the U.S. Department of Health and Human Services.

Introduction

Over 90% of hospitals in the United States have adopted basic electronic health record technology which includes functions such as clinical information, computerized provider order entry (CPOE), results management, and decision support [1]. These health information technology (HIT) systems have a tremendous impact on clinical care processes. While HIT can improve care processes, it can also pose serious safety hazards if not designed, developed, implemented, and used effectively [2]. In order to optimize HIT and reduce safety hazards, it is imperative that we develop an in-depth understanding of HIT system-induced errors during the clinical care process.

Patient safety event (PSE) reports, which are descriptions of safety hazards or errors reported and initially categorized by frontline staff, offer one view of the impact of HIT on safety. Most hospitals have a PSE reporting system in place and frontline clinicians are encouraged to report events. PSE reports generally consist of structured and unstructured data elements. The structured data elements generally require the reporter to select a predefined category that best represents the nature of the event report (fall, medication, laboratory, etc.). The unstructured data consists of a free-text narrative describing the hazard [3]. Free-text narratives can be a rich source of information, particularly for complex events. In most hospitals, the PSE reports are then reviewed by a patient safety officer or other analysts to identify trends.

Often the predefined category from the structured data is misleading for several reasons. First, many safety hazards cannot be classified into one specific category. This is particularly true for HIT-related events. For example, consider an entry error while ordering a medication using CPOE. This event may be categorized as a medication-related event or a HIT-related event. Most PSE reporting systems do not allow two primary structured categories to be selected. Second, frontline staff may not have the time to determine the specific category that is the best fit for the event being entered given other clinical duties he or she must perform. Third, frontline staff may not know how each category is specifically defined by the analyst that reviews PSE reports. Below are examples of PSE free-text descriptions that were categorized by the frontline staff into a general predefined category of "Other" and "Diagnosis/Treatment" (respectively). Both events have a clear HIT contributing factor:

> "Admission weight of 75 kg entered into EHR in pound field as 75 lbs. This weight was automatically converted to 34.02 kg. This caused the drip calculations to be based on weight of 34.02 kg."
> "Pt [Patient] on electronic board was automatically removed after system reset at midnight. Pt missed medication at appropriate time."

To identify HIT-related events, which is a critical step to improve HIT systems and improve safety, PSE events must be manually reviewed by patient safety officers or other analysts to identify those events that have a HIT component. Accurately identifying HIT-related events from PSE databases containing tens of thousands of reports is labor intensive. Consequently, the manual review process is a serious limiting factor in identifying HIT hazards from PSE data.

To improve HIT safety event analysis and the categorization of PSE events as being HIT-related, we examine the utility of applying an active learning (AL) approach to identify HIT-related PSE reports. We implement and evaluate different AL strategies, describe a system for annotation, and discuss insights gained from this real-world application.

Background

Although an agreed upon definition of HIT safety hazards is lacking, there has been tremendous interest in better understanding these hazards to reduce patient harm [4–7]. For the purposes of this study, we utilized the Agency for Healthcare Research and Quality's (AHRQ) definition of a HIT hazard to identify HIT-related events [8]. "A [HIT] hazard is a characteristic of any [HIT] application or its interactions with any other health care system (e.g. the people, equipment and work spaces of an ICU) that increases the risk that care processes will be compromised and patients harmed… hazards may arise due to the inadequacies in the design, manufacture, implementation,

or maintenance of [HIT]. They may also arise in the interactions between [HIT] and other complex health care systems… [and] combine with other characteristics of the care system to overwhelm the vigilance and skill of the health care team..." [8].

Most research seeking to understand HIT safety hazards has focused on developing taxonomies and analysis tools, including computational approaches, to identify HIT-related safety event reports. However, these models have relied on static, manually annotated datasets which are time consuming, labor intensive, and costly [4, 9–11]. Active learning offers a potential solution to this resource intensive manual coding process.

Active learning

Active learning (AL) is a semi-supervised human-in-the-loop machine learning approach that leverages both human insight and natural language processing algorithms to annotate data [12, 13]. The motivation behind AL is that human annotation is resource intensive and, strategically, data points that would add the most value to a machine learning model should be annotated. While the benefits of AL have been demonstrated in several research areas including medical and clinical applications [12, 14–16], most previous research has used simulated human annotators as opposed to actual human annotators [17, 18]. The reliance on simulated humans can be a major weakness when this approach is applied to PSE data given the difficulty in interpreting PSE reports. In addition, the datasets previously used have been well annotated and cleaned making it difficult to translate the methods to real-world problem areas like HIT safety hazard identification. In this paper we build upon previous research by comparing three different AL strategies integrated with machine learning models to identify HIT-related PSE reports from a large set of diverse reports [9]. We compare the three different AL strategies to a support vector machine (SVM) model with no AL and examine model evolution characteristics and outputs.

Methods

We first describe the dataset used and the AL workflow, ▶Figure 1. We then discuss three AL query strategies and the evaluation metrics.

Dataset

Data were comprised of a selection of anonymous PSE reports from the Institute for Safe Medication Practices (ISMP) between 2007 and 2016. As a Patient Safety Organization (PSO), ISMP serves as a safe harbor for all PSE reports from hospitals in Pennsylvania, US. Although ISMP as an organization is focused on improving medication safety, the dataset that was analyzed contains all categories of PSE reports (fall, medication, surgery, etc.) and is not limited to medication reports. Data were divided into a set of 252 labeled reports and 5,000 unlabeled reports randomly selected from the ISMP dataset. Labeled reports were annotated in the context of HIT as either "Likely" (120/252), "Unlikely" (96/252), or "Need More Information" (36/252). "Likely" refers to HIT likely being a contributing factor, based on the HIT hazard definition, the PSE free-text narrative, and reasonable assumptions about clinical practice [8]. Examples of "Likely" reports include: usability design issues (e.g. information hard to find or difficult data entry) that result in order or administration error, downtime occurrences, and system interaction inaccuracies. "Unlikely" refers to HIT unlikely being a contributing factor of the PSE given the narrative and reasonable assumptions about clinical practice. Reports lacking sufficient detail to make the above distinction were coded as "Need More Information." The labeled reports were annotated by three annotators [RR, KA, JH] with expertise in human factors and HIT. Inter-rater reliability (IRR) between the three annotators was assessed for every set of 50 reports; differences were reconciled through discussion. IRR resulted in a final Fleiss' kappa of 0.81 for the three-way labeling task. This study was approved by the MedStar Health Research Institute Institutional Review Board (protocol #2014–101).

Active Learning Workflow

Reports were preprocessed by removing punctuations, numbers, and common stop words. Reports on average contained 27 words with a standard deviation of 33. Words were converted to lower case, stemmed, and a term frequency-inverse document frequency (tf-idf) vector for each report was generated. These tf-idf feature vectors were used for the training and model development in this bag-of-words approach and there were approximately 700 features. We initialized an SVM learning model with a radial basis function trained on an approximately balanced dataset of 216 manually annotated reports (120 "Likely"). This approach has been shown to perform well in similar tasks [19, 20].

The resulting SVM was then applied to 5,000 unlabeled reports. Five reports from the unlabeled set were then selected for human annotation using one of the query strategies described below. The annotated reports were then added to the labeled reports and removed from the unlabeled set. The model was retrained (one learning iteration) and the human was presented with five new reports. We developed a system with a minimalistic user interface that allows the human to annotate each report, providing feedback to the underlying SVM model. While this batch learning approach might converge slower than an incremental approach, we believe reviewing reports in batches is more reflective of an actual human annotation process [14]. In addition, to reduce Type II errors, reports annotated as "Need More Information" were excluded from the training process.

Stopping criteria

To better evaluate and compare the incremental benefits of the AL strategies, we set the number of learning iterations to ten (stopping criteria). We examined the results after each iteration to understand the evolution of the models.

Query strategies – sample reports

We utilized the LIBSVM probability estimate extension for SVM class labels in our AL workflow [21]. For this binary classification task ("Likely" or "Unlikely"), probabilities associated with SVM predictions is a fitted logistic distribution using maximum likelihood to the decision values [21]. As a result, each SVM prediction has an associated likelihood probability (ranging from 0.5 to 1). For example, a model is more confident in a "Likely" prediction with 0.98 probability compared to another "Likely" prediction with a probability of 0.75. In addition, previous research has identified the importance of minimizing AL iteration time when involving human annotators [17]. We used these concepts to select the three different AL query strategies below.

Uncertainty Sampling – Baseline (US_B)

In US_B, reports with the lowest likelihood probability, regardless of their predicted classification, were selected for annotation. This approach parallels other uncertainty sampling approaches in that the highest uncertain, or least confident, model predictions are selected for annotation [22]. There was no specific cut-off threshold used for sampling. Instead the associated classification probability was used to rank the results. The top 100 results were evaluated, a typical approach for information retrieval tasks [23].

Uncertainty Sampling – Likely (US_L)

This strategy modifies the US_B approach in that only uncertain reports tending towards "Likely" were selected, ▶Figure 2. This approach is motivated by decreasing the occurrence of false positives.

Confirmatory Sampling – Likely (CS_L)

This approach requires annotators to confirm reports predicted as "Likely" with high probability [16]. The motivation for this last strategy is to reduce false positives and thus improve the precision of the model early in its development.

Evaluation

We evaluated and compared four model conditions: three different AL models and a control SVM without AL: US_B, US_L, CS_L, and SVM. Each AL condition was randomly assigned to a different annotator. To equalize the amount of training data across conditions, we added an additional 50 annotated reports (reports were randomly selected from the initial unlabeled reports) to the training data in the control SVM condition. Each model was evaluated based on precision, or positive predictive value (PPV), at K, where K is a threshold parameter. This is a common approach for evaluating information retrieval search results when global sensitivity and specificity metrics are difficult to assess [23]. For example, it is very difficult to fully evaluate the number of false positives or false negatives in a Google search. Instead only the top K retrieved search items are reviewed rather than doing an exhaustive review. In addition to assessing high probability K (where K = 100) reports, we also reviewed K reports predicted as "Likely" with low probability. This evaluation was done through manual review and was to assess the performance at model boundaries. Lastly, we reported on the evolution of the AL model predictions.

Results

Model classifications

Each model had high precision when evaluating high probability predictions, with the Uncertainty Sampling - Likely strategy, US_L, (0.85) slightly outperforming other models, ▶Table 1. This demonstrates there was little performance difference in sampling strategy for reports that can be easily identified as HIT-related. However, Confirmatory Sampling - Likely (CS_L) had the highest precision for lower bound reports (a ten-fold increase and a 30% average precision increase over the SVM control). This difference in performance suggests CS_L may be better at discerning or identifying more complex reports, specifically with earlier detection and removal of false positives. Being able to identify HIT-related reports after fewer iterations is beneficial for annotators in that their time would be spent reading more relevant reports. CS_L also had the fewest "Likely" categorized reports after ten iterations which could also be reflective of more efficient filtering of false positives. Having a lower number of "Likely" labeled reports while still maintaining reasonable precision suggests better overall model performance, although this requires additional investigation to confirm.

Model evolution

We investigated the evolution of predictions after each AL iteration, ▶Figure 3. The number of reports where the predicted label switched from "Likely" to "Unlikely" (left sub-graph) and from "Unlikely" to "Likely" (right sub-graph) are shown for each AL strategy. The results after each learning iteration are shown on the vertical axis. Confirmatory Sampling - Likely (CS_L) had the least label prediction fluctuations suggesting more stable and gradual model changes through each iteration. US_B had the most oscillation in predictions through each iteration which could be reflective of sampling at decision boundaries. Furthermore, the variability with US_L tended to have reports relabeled as "Likely." This may be due to better identification of false positives which is also reflected in less "Likely" labeled reports after ten iterations.

Discussion

The active learning (AL) approach offers a potentially resource efficient alternative approach to manual review of PSE reports to identify reports that are HIT-related. Instead of randomly sampling reports to review, AL leverages human insight encoded in the iterative training process to more intelligently sample reports. We demonstrated that a confirmatory sampling strategy has on average higher precision at identifying HIT-related events. The success of the confirmatory sampling strategy suggests that earlier removal of false positives and giving humans more useful reports to anno-

tate is more effective than other strategies. At a practical level, patient safety analysts might utilize AL with a confirmatory sampling strategy when reviewing PSE reports to develop categorization models and this would result in a less resource intensive review process. Patient safety analysts may be able to review fewer reports and therefore spend time on other important activities to improve safety. Furthermore, the degree of HIT literacy and insight of the reporters are examples of report variability that make AL useful in this real-world application. This method is also preferred over iterative development of keyword search phrases because AL utilizes advanced statistical techniques to identify relevant reports.

Applications of Active Learning

Our work suggests that AL has real-world applicability. It can be used to identify HIT-related events from large datasets where human annotation is a major barrier to understanding trends and patterns in the data. Our AL approach required the development of a system that was used by the human annotators to rapidly review data and annotate the data for further model refinement. The system we built required several important design decisions to fit the needs of the human annotators. First, query strategies and the sampling size were selected to reduce the model iteration time to 1–2 seconds. Second, for this initial evaluation, we only provided the free-text of the reports to the annotators. Future research could investigate additional design decisions such as the impact of showing annotators model predictions and confidence levels on the annotation process.

Limitations

This analysis is limited by the integrity and quality of voluntary, self-reported PSE data. The data used in our analysis stemmed from one source, the Institute for Safe Medication Practices, which sources data from hospitals in Pennsylvania, US. To improve the generalizability of the results, additional data sources must be examined. In addition, we excluded the reports classified as "Need More Information" to simplify the modeling approach. Future work might use the approach of assigning annotator weights or confidence assumptions to reports in this category [24]. Only 200 output reports from each condition were evaluated. Additional evaluation techniques could be developed to address this type of real-world evaluation challenge. We tested ten AL iterations with three annotators. It would be interesting to compare these results with an approach that uses more learning iterations and annotators with different domain expertise. Lastly, model learning time and complexity will increase with the number of reports. Different techniques, such as staged learning increments and parallel distributed machine learning systems, are needed to maintain a short iteration time and acceptable user experience [25, 26].

Conclusion

PSE reports offer a lens to better understand how HIT impacts patient safety and patient care. To improve the analysis of patient safety event reports to identify HIT-related events, we demonstrate the benefits of an AL approach with confirmatory sampling over a SVM model with no AL, helping to focus and increase the value of human annotators.

Question

Which metric is generally more useful when evaluating the relevance of the first 100 retrieve reports from an active learning approach?

A – Specificity
B – Sensitivity
C – Positive Predictive Value
D – Negative Predictive Value

Correct answer is C: Positive Predictive Value, or precision, is a standard metric used for evaluating information retrieval tasks and is more appropriate than other metrics because such metrics cannot be accurately or meaningfully calculated without a fully annotated dataset (Manning 2008). For example, it is burdensome to fully evaluate the number of false positives or false negatives in a Google search. Instead, one typically only reviews the top few search returns.

Conflict Of Interest
The authors have no conflict of interest.

Human Subjects Protection
No human subjects were involved in the project.

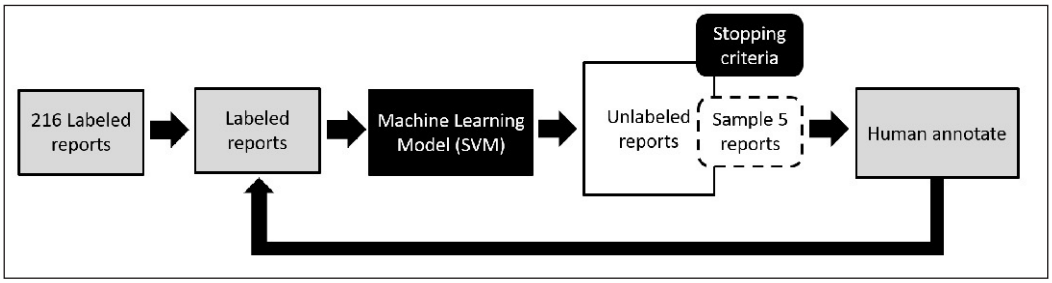

Fig. 1 Active learning experiment workflow shows an iterative loop requiring human annotation feedback to update a machine learning model until an externally defined stopping criteria is met.

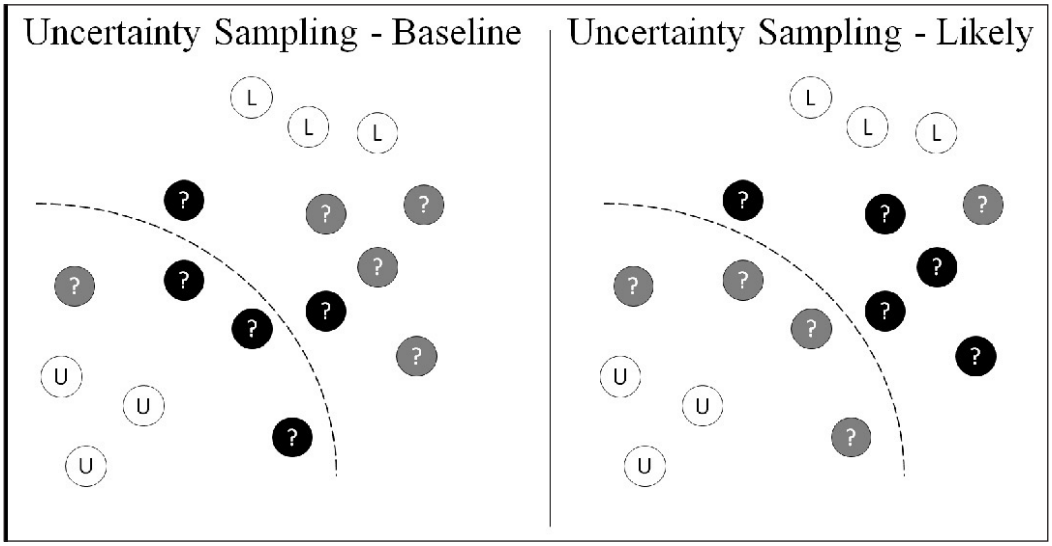

Fig. 2 Black solid circles indicate the unlabeled (circles with "?") samples selected under the two different uncertainty sampling strategies for a binary classification task ("L" circles as "Likely" and "U" circles as "Unlikely").

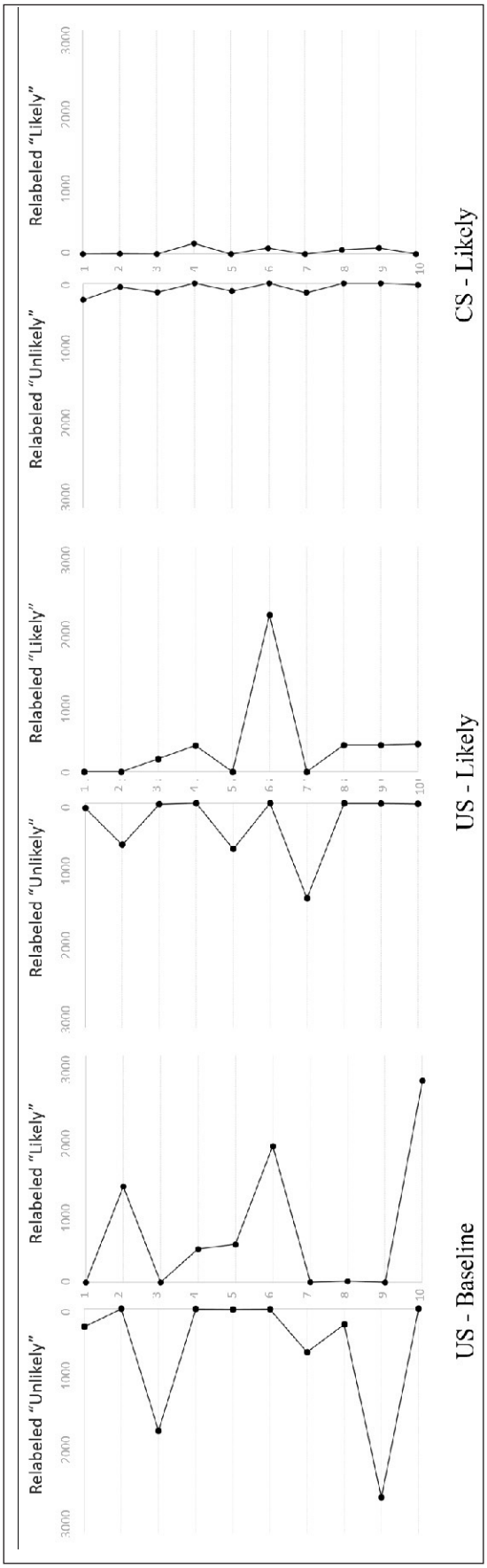

Fig. 3 After each iteration (vertical axis), the number of reports where the predicted label switched from "Likely" to "Unlikely" (left sub-graph) and from "Unlikely" to "Likely" (right sub-graph) for each AL strategy.

Table 1 Comparison of model performance evaluated for precision, or positive predictive value, at K=100, for both high and low probability classifications as well as the percent of "Likely" categorized reports.

	Precision (K=100)			Label "Likely"
	High Probability	Low Probability	Average	
Uncertainty Sampling – Baseline (USB)	0.83	0.02	0.43	68% (3364/4950)
Uncertainty Sampling – Likely (USL)	0.85	0.02	0.44	61% (3009/4950)
Confirmatory Sampling – Likely (CSL)	0.84	0.28	0.56	34% (1678/4950)
Control SVM (SVM)	0.83	0.02	0.43	74% (3661/4950)

References

1. Charles D, Gabriel M, Searcy T, Carolina N, Carolina S. Adoption of Electronic Health Record Systems among U . S . Non – Federal Acute Care Hospitals: 2008–2014. 2015.
2. Karsh BT, Weinger MB, Abbott PA, Wears RL. Health information technology: fallacies and sober realities. J Am Med Informatics Assoc 2010; 17(6): 617–623. Available from: http://www.ncbi.nlm.nih.gov/pubmed/20962121
3. Fong A, Ratwani RM. An Evaluation of Patient Safety Event Report Categories Using Unsupervised Topic Modeling. Methods Inf Med 2015; 54(4): 338–345.
4. Magrabi F, Ong M-S, Runciman W, Coiera E. Using FDA reports to inform a classification for health information technology safety problems. J Am Med Informatics Assoc 2012; 19(1): 45–53. Available from: http://www.ncbi.nlm.nih.gov/pubmed/21903979
5. Borycki EM, Kushniruk a W. Towards an integrative cognitive-socio-technical approach in health informatics: analyzing technology-induced error involving health information systems to improve patient safety. Open Med Inform J 2010; 4: 181–187.
6. Walker JM, Hassol A, Bradshaw B, Rezaee ME. Health IT Hazard Manager Beta-Test. Rockville, MD; 2012. Available from: https://healthit.ahrq.gov/sites/default/files/docs/citation/HealthITHazardManager-FinalReport.pdf
7. Meeks DW, Takian A, Sittig DF, Singh H, Barber N. Exploring the sociotechnical intersection of patient safety and electronic health record implementation. J Am Med Inform Assoc 2014; 21(e1): e28-e34. Available from: http://dx.doi.org/10.1136/amiajnl-2013–001762\nhttp://www.ncbi.nlm.nih.gov/pubmed/24052536
8. Walker JM, Hassol A, Bradshaw B, Rezaee ME. Health IT Hazard Manager Beta-Test. Rockville, MD; 2012.
9. Chai KEK, Anthony S, Coiera E, Magrabi F. Using statistical text classification to identify health information technology incidents. J Am Med Informatics Assoc 2013; 20(5): 1–6. Available from: http://www.ncbi.nlm.nih.gov/pubmed/23666777
10. Singh H, Sittig DF. Measuring and improving patient safety through health information technology: The Health IT Safety Framework. BMJ Qual Saf 2015; (0): 1–7.
11. Amato MG, Salazar A, Hickman TT, Quist AJL, Volk LA, Wright A, McEvoy D, Galanter WL, Koppel R, Loudin B, Adelman J, McGreevey JD, Smith DH, Bates DW, Schiff GD. Computerized prescriber order entry – related patient safety reports: analysis of 2522 medication errors. J Am Med Informatics Assoc 2016; 0: 1–6.
12. Settles B. Active Learning. Synthesis Lectures on Artificial Intelligence and Machine Learning. Morgan & Claypool; 2012.
13. Settles B. Active Learning Literature Survey. Madison (WI): University of Wisconsin-Madison. 2010 Jan 16. Computer Sciences Technical Report 1648.
14. Kholghi M, Sitbon L, Zuccon G, Nguyen A. Active learning: a step towards automating medical concept extraction. J Am Med Informatics Assoc 2015; 0: 1–9. Available from: http://jamia.oxfordjournals.org/lookup/doi/10.1093/jamia/ocv069
15. Zhang H, Huang M-L, Zhu X-Y. A Unified Active Learning Framework for Biomedical Relation Extraction. J Comput Sci Technol 2012; 27(6): 1302–1313.
16. Boström H, Dalianis H. De-identifying health records by means of active learning. In: Recall (micro). 2012; 90–97.
17. Clancy S, Bayer S, Kozierok R. Active Learning with a Human In The Loop. Bedford, MA; 2012.
18. Settles B, Craven M, Friedland L. Active Learning with Real Annotation Costs. Proceedings of the NIPS Workshop on Cost-Sensitive Learning 2008.
19. Ong M-S, Magrabi F, Coiera E. Automated categorisation of clinical incident reports using statistical text classification. Qual Saf Health Care 2010; 19(6): e55. Available from: http://www.ncbi.nlm.nih.gov/pubmed/20724392
20. Chai KEK, Anthony S, Coiera E, Magrabi F. Using statistical text classification to identify health information technology incidents. J Am Med Inform Assoc 2013; 20(5): 980–985. Available from: http://jamia.bmj.com/cgi/doi/10.1136/amiajnl-2012–001409
21. Chang C, Lin C. LIBSVM: A Library for Support Vector Machines. ACM Trans Intell Syst Technol 2011; 2(3): 27.
22. Lewis DD, Catlett J. Heterogeneous Uncertainty Sampling for Supervised Learning. In: Machine Learning: Proceedings of the Eleventh International Conference. 1994; 148–156.
23. Manning CD, Raghavan P, Schütze H. Introduction to Information Retrieval. Cambridge University Press; 2008.

24. Donmez P, Carbonell JG. Proactive Learning: Cost-Sensitive Active Learning with Multiple Imperfect Oracles. In: Proceedings of the 17th ACM conference on Information and knowledge management. ACM 2008; 619–628.
25. Meng X, Bradley J, Yavuz B, Sparks E, Venkataraman S, Liu D, Freeman J, Tsai D, Amde M, Owen S, Xin D, Xin R, Franklin M, Zadeh R, Zaharia M, Talwalkar A. MLlib: Machine Learning in Apache Spark. In: CoRR 2015.
26. Kraska T, Talwalkar A, Duchi JC, Griffith R, Franklin MJ, Jordan MI. MLbase: A Distributed Machine-learning System. In: CIDR 2013.