Using statistical and machine learning to help institutions detect suspicious access to electronic health records

Aziz A Boxwala,[1] Jihoon Kim,[1] Janice M Grillo,[2] Lucila Ohno-Machado[1]

► Additional materials are published online only. To view these files please visit the journal online (www.jamia.org).

[1]Division of Biomedical Informatics, University of California San Diego, La Jolla, California, USA
[2]IS Clinical Systems, Partners Healthcare System, Boston, Massachusetts, USA

Correspondence to
Aziz A Boxwala, Division of Biomedical Informatics, UCSD School of Medicine, 9500 Gilman Drive #0728, La Jolla, CA 92093-0728, USA; aboxwala@ucsd.edu

Received 28 February 2011
Accepted 2 May 2011

ABSTRACT

Objective To determine whether statistical and machine-learning methods, when applied to electronic health record (EHR) access data, could help identify suspicious (ie, potentially inappropriate) access to EHRs.

Methods From EHR access logs and other organizational data collected over a 2-month period, the authors extracted 26 features likely to be useful in detecting suspicious accesses. Selected events were marked as either suspicious or appropriate by privacy officers, and served as the gold standard set for model evaluation. The authors trained logistic regression (LR) and support vector machine (SVM) models on 10-fold cross-validation sets of 1291 labeled events. The authors evaluated the sensitivity of final models on an external set of 58 events that were identified as truly inappropriate and investigated independently from this study using standard operating procedures.

Results The area under the receiver operating characteristic curve of the models on the whole data set of 1291 events was 0.91 for LR, and 0.95 for SVM. The sensitivity of the baseline model on this set was 0.8. When the final models were evaluated on the set of 58 investigated events, all of which were determined as truly inappropriate, the sensitivity was 0 for the baseline method, 0.76 for LR, and 0.79 for SVM.

Limitations The LR and SVM models may not generalize because of interinstitutional differences in organizational structures, applications, and workflows. Nevertheless, our approach for constructing the models using statistical and machine-learning techniques can be generalized. An important limitation is the relatively small sample used for the training set due to the effort required for its construction.

Conclusion The results suggest that statistical and machine-learning methods can play an important role in helping privacy officers detect suspicious accesses to EHRs.

INTRODUCTION

With health records becoming computerized, patients are becoming increasingly concerned about the privacy and security of their health information.[1 2] Health-information privacy and the loss of such information are covered in the USA by federal laws (Health Insurance Portability and Accountability Act and Health Information Technology for Economic and Clinical Health Act)[3 4] and myriad state regulations. Breaches of privacy require notification to the affected individuals, to government agencies, and the media. The cost to the healthcare organization of the loss of information during security breaches is one of the highest of any industry.[5] The notifications and adverse news reports also lead to the loss of trust of patients.[6 7]

Much work has been done by healthcare organizations to keep patient data in electronic health records (EHRs) secure and private. Among common security mechanisms are secure networks with firewalls, encrypted devices and messages, strong user passwords, auditing of access logs of clinical systems, and device timeouts. Despite these security measures, there remains a problem of how to allow the 'right' users to access the 'right' patient records, while preventing inappropriate accesses by authorized users of the system. Organizational polices allow access to records strictly for treatment, payment, and healthcare operations (TPO) reasons. Authorized users are typically granted limited access to those functions in the EHR needed to perform their jobs. This is in contrast to restricting access to records by patients. Such restrictions are more difficult to regulate and implement, since it is difficult to predict accurately the records for which a provider will need access. Thus, one form of privacy breach occurs when staff members of a healthcare organization access records without TPO reasons. Among the reasons for inappropriate access is snooping of records of celebrities, neighbors, coworkers, and family members.[6 7] At other times, inappropriate access has been associated with criminal activity.[8] In order to counter such breaches, some institutions, such as the one providing data for this study, have implemented additional safeguards, including:

- annual privacy training and annual confidentiality agreements for all employees;
- access rules, either prohibiting access or requiring users to enter a reason for access, when the patient being accessed is not registered at the same site as the user;
- an annual assessment process where individual managers review privileges for their own staff members and can either approve continued access or remove access that is no longer required; and
- employee self-audit, a tool that allows individuals to see who has looked at their record.

Investigations of breach complaints are done by privacy officers in the Health Information Systems (HIS) or Medical Records departments. Users found to have violated the policy are subject to disciplinary action, which may include termination.

RESTRICTING ACCESS TO RECORDS

Ferreira *et al* performed a comprehensive review on access control models for EHRs over 10 years

UNLOCKED

This paper is freely available online under the BMJ Journals unlocked scheme, see http://jamia.bmj.com/site/about/unlocked.xhtml

(1996–2006) and found 59 studies.[9] Of these, only one of them had been implemented in a real scenario; the others were described in a theoretical framework or as prototypes. Later, Role-Based Access Control (RBAC), proposed by National Institute of Standards and Technology[10] and the most commonly used access control model among the above 59 studies, was further extended to a contextual-RBAC and Situation-Based Access Control (SBAC).[11 12] Unfortunately, the feasibility of the proposed models was seldom tested on an operational clinical system. Several factors make such access control approaches challenging:

- unpredictable and dynamic care patterns including scheduled and unscheduled inpatient, outpatient, and emergency department visits;
- varied workflow: providers may fill-in in unexpected areas;
- mobile workforce: access may be needed at unexpected locations and times;
- collaborative nature of clinical work and teaching environments;
- large number of users and job titles/roles (the organization in this study has approximately 40 000 authorized users with over 1700 unique job titles); and
- user job titles that do not always translate directly into a list of patient(s) whose records would be appropriate to be accessed at any point in time.

AUDITING ACCESS LOGS

The current practice of auditing access logs involves identifying suspicious accesses to records based on known and simple patterns, such as accesses to records of VIP or celebrity patients, or accesses involving last-name matches between the patient and the user suggesting access to a family member's record. This use of simple patterns could lead to a large number of false-positive cases. Furthermore, these simple patterns cannot produce ranked lists of cases, thus not providing a means for prioritizing which cases should be investigated further. One of our objectives was to create a system that could identify suspicious accesses to EHRs based on certain patterns known to us. However, the system should score each access for appropriateness, so that the top scoring cases can be prioritized for further investigation by privacy officers. This would allow us to accelerate the process of detecting suspicious accesses, that is, accesses that are similar to patterns of breaches known to us. The approach we have taken in this study is based on the assumption that a vast majority of authorized users are accessing records only of those patients for whom they have a valid TPO reason. Further, many users have roles and responsibilities that are consistent and repeatable, and their accesses can be tracked over time to define patterns of access. Thus, we used statistical and machine-learning techniques on EHR access logs to detect rarely occurring suspicious accesses.

Other investigators have studied the use of audit logs and other databases to identify appropriate and suspicious accesses to EHRs. As described in the next few paragraphs, the breadth of this research includes the development of algorithms, modeling and definition of information sources used for determining appropriateness of access, architectures for auditing systems, and the application of business intelligence platforms.

Salazar-Kish *et al* defined patterns and a heuristic algorithm for determining the patient–user match from data including the patient–PCP relationship, patient appointment information, and user belonging to a clinic or department.[13] They studied the impact of the algorithm to assess how often a warning might need to be issued to a user attempting to access a record for which a patient–user relationship was not found by the algorithm. Over 25% of accesses would result in a warning in their cases, which would impose a tremendous burden on the users. Our work uses similar types of patterns to define user–patient relationships for creating a training set of data. However, we go further in that we attempt to identify, using statistical and machine-learning methods, patterns for suspicious access to records.

Asaro and colleagues created a taxonomy of 'indicators' for EMR breaches.[14] The top-level categories of this taxonomy included patient characteristics (eg, VIP patient, patient with sensitive data such as HIV test results), patient–user interactions (eg, coworkers, spouse going through divorce), and session characteristics (eg, access to a large number of records). They describe positive and negative characteristics, where the latter are similar to the patient–user match described in the previous study. The authors also created a taxonomy of information requirements to perform an audit to discover breaches; some of this information is available within the healthcare organization such as appointment information, while some information such as legal proceedings between spouses is not available. In an article that extends this work, Herting defines internal and external data sources classified by the repository type (eg, relational database, organizational directory, internet resource) that can be used to augment information on individuals.[15] They provide high-level guidelines on how one could score the reliability of information from different sources and how to combine information from different sources. The authors describe the partial implementation of an auditing system based on the above design ideas. They also highlight the potential for disclosure of private information during auditing in trying to access external information sources to learn about the user–patient relationship. In this article, we describe how we bring together data from many of these information sources to build a system for detecting suspicious accesses.

Malin proposes a novel protocol, using cryptography, that allows an EMR access auditor to obtain information from other systems (eg, human-resource database) without revealing to the custodians of those systems, the identity of those being investigated.[16] This can be useful when the third-party system was not within the same security and privacy purview as the access logs. In our study, we were able to integrate several of the databases from our institution including the human-resources database into a single audit database, thus precluding the need to create the type of architecture described by Herting and Asaro, and to use external information sources during the audit process. Zhou, Liu, and colleagues designed a system architecture and created a prototype implementation of an auditing system that collects access logs from picture archiving and communication systems (PACS) and other data with the objective of identifying appropriateness of access.[17] One component is an audit analysis tool to automatically analyze logs and discover inappropriate accesses. This component, which was not implemented by the authors, is the focus of our research. We have implemented a system that largely automates the data collection and, following a manual effort to train the system, can automatically analyze logs.

Coleman implemented a commercially available business intelligence system that combines access logs with other institutional data in a dimensional data warehouse to assist users in auditing the logs.[18] The business intelligence tool allows users to rapidly drill down the data to manually identify accesses that seem inappropriate. The authors conclude that, even with such a system, detecting inappropriate accesses is difficult and time-consuming. They identify a challenge in automating the analysis

as that of defining the rules to detect inappropriate access. The statistical and machine learning approaches we utilized in this study created patterns of suspicious access that can be readily implemented in these data warehouses. Furthermore, the database we constructed for this project enables privacy officers to review suspected cases, by drilling down into accesses by the user, and comparing patterns of access within a department or across role-types.

None of the above studies investigated the use of statistical and machine-learning techniques to support the detection of suspicious access to EHRs in a real clinical system. These techniques have been used to detect fraud in financial reporting for an audit client,[19] to detect fraud in credit card transaction data,[20] to construct a spam email detector,[21] and to solve a fraud-detection problem at a car-insurance company.[22] In this article, we demonstrate a novel application of these techniques in the analysis of EHR access logs.

METHODS

Overview

The data flow in our study is summarized in figure 1. We constructed an Event Data Mart (see Appendix 1 online at www.jamia.org) of record access events (RAEs) from the institution's operational databases (DBs) as a start-up set. The Event Data Mart was populated with data from a 2-month period from December 1, 2007 to January 31, 2008. Then, we built a Feature DB—a collection of useful features for detection of suspicious access to the EHR, derived from the Event Data Mart. Some events were selected as a training set and were sent to privacy officers at the participating institutions. Lastly, we built a classifier for identification of suspicious access to EHRs.

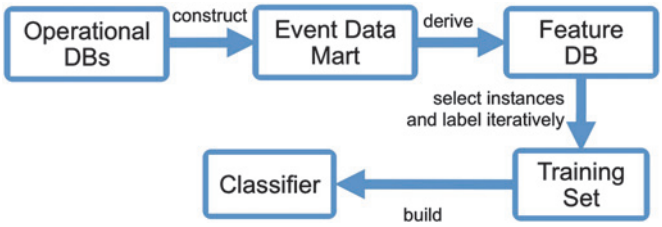

Figure 1 Overview of the system to detect suspicious record access events (RAE). DB, database.

Constructing the Feature DB

We defined a set of features considered likely to help in the detection of suspicious RAEs based on our review of previously reported breach cases. The role-based and situation-based access control models also were considered in the feature set definition. These features were derived from the Event Data Mart and inserted into the Feature DB (figure 1). For example, the feature 'Is provider' was coded as '1' (for 'yes') or '0' (for 'no') based on the patient's primary care physician information in the patient's registration record. Twenty-six features were constructed into the Feature DB. These can be categorized into five classes: user-related, patient-related, RAE-related, encounter-related, and user—patient-relationship-related (box 1).

Preparing the training set

The next step was to create a data set to be used for training statistical and machine-learning algorithms. We created the data set iteratively in collaboration with privacy officers from the HIS departments of the study sites. These officers labeled a subset of

Box 1 List of variables in a feature database derived from the Event Data Mart in figure 1

Record access event features
Time of access, Application*, Over 200 accesses in a day†, Month of access, Day of month, Day of week, Hour of day, Access after clinic hours, Access on clinic day

User features
Integrated delivery network employee, Hospital #1 employee, Hospital #2 employee, Nurse, Researcher, Is provider‡

Patient features
VIP§, Is employee¶, Had recent visit**

Encounter type features
Encounter location type††

Patient—user relationship features
Same street address, Same city, Same zip code, Accessing own record, Care unit visit match, Clinic visit match, Work in same department, Patient registered in user site‡‡, Same family name.

The following explains variables whose meaning may not be clear from the label.
*Software program used to access the record.
†Accessed over 200 patient records in the same calendar day. (We selected a threshold of 200 accesses because in our experience most providers would access well below this number of patients in a day for TPO reasons. A lower threshold might create many false positives, and a higher threshold might lead to false negatives. A possible reason for such large number of accesses might be a researcher trying to find patients for a study. Without approval from an institutional review board, this type of access is inappropriate.)
‡User is listed as patient's PCP in the patient registration system.
§Patient marked as VIP in the registration system.
¶Patient is an employee of the integrated delivery network or one of its sites.
**Patient had visit in the past 60 days at any site in the integrated delivery network.
††Visit to inpatient, outpatient, or emergency room locations.
‡‡Patient has an entry in the registration system at the site at which the user is an employee.

the RAEs as being either suspicious or appropriate —'positive' or 'negative' respectively in the context of machine learning—after a thorough review of each event. They defined a suspicious RAE as one that appeared on initial review to be 'inappropriate' or 'should prompt for a reason for access.' An inappropriate access is one where the user has no TPO reason for accessing the record. The privacy officers were making their judgment without conducting a full investigation that would involve interviewing the user, and which might reveal an RAE to be appropriate. Therefore, such RAEs were marked as suspicious rather than inappropriate. This approach was desirable, since our end objective is to construct an autonomous system to identify suspicious cases either for further investigation or potentially for the EHR user interface to request a reason for access.

The study used RAEs only where the user accessing the record was an employee of one of three sites: the integrated delivery network (IDN) corporate entity itself, or two of the hospitals belonging to the IDN. Medical residents and most of the Information Systems staff are employees of the IDN corporate entity. Clinicians and administrative support staff are employees of the hospital. Each RAE added to the labeled data set was reviewed by a privacy officer from the respective site within the IDN. The review team utilized the Event Data Mart and external information sources, such as patient records and an internal audit tool, to make their decisions. The audit tool contained subset of the information in our Event Data Mart. It included details of the user actions within the patient's record (eg, signing an order which would strongly indicate appropriate access) that were not present within the Event Data Mart. These details were not included in the Event Data Mart in this phase of the study because of the very large volume of data. For this study, the review team did not interview the user, or their supervisors, which would be part of a complete investigation in the case of a suspected breach. The review of each RAE in this study took on the average approximately 10 min per reviewer.

In order to select a mix of RAEs that represented various patterns of appropriate and suspicious events, the labeling was performed through several iterations, as illustrated in figure 2. In the first round, we selected RAEs based on patterns of features (scenarios, see Appendix 2 online at www.jamia.org) that were, in our experience, illustrative of appropriate or suspicious/inappropriate accesses. Since such types of rule-based patterns were used to detect suspicious accesses at the IDN, we defined this collection of seven patterns as the baseline method. We selected a set of 505 RAEs from the Feature DB by stratified sampling, where a stratum corresponded to one of the scenarios above. The privacy officers were not told which pattern was used to select the case. Of the 505 RAEs selected for review, 216 were labeled suspicious and 289 appropriate.

We built a logistic regression (LR) model from this labeled training set of 505 RAEs. Then, we applied the fitted model to the 10.5 million RAEs in the Feature DB to assign LR prediction probabilities to all. Next, we used the LR probabilities to select unlabeled events. RAEs were randomly selected from the top, middle, and bottom terciles. The selected RAEs were labeled by the privacy officers, who were blinded to the prediction probabilities of the selected RAEs. In the next round, we built a new LR model using labeled RAEs from the previous rounds. This 'Build-Predict-Select-Label' procedure was repeated three times.

Since suspicious RAEs are very rare compared to appropriate RAEs, these data were subject to a class imbalance problem, that is, in other words, the significant difference in prior probabilities of these classes of RAEs could cause degradation in the performance of the classifiers we were creating. In order to overcome

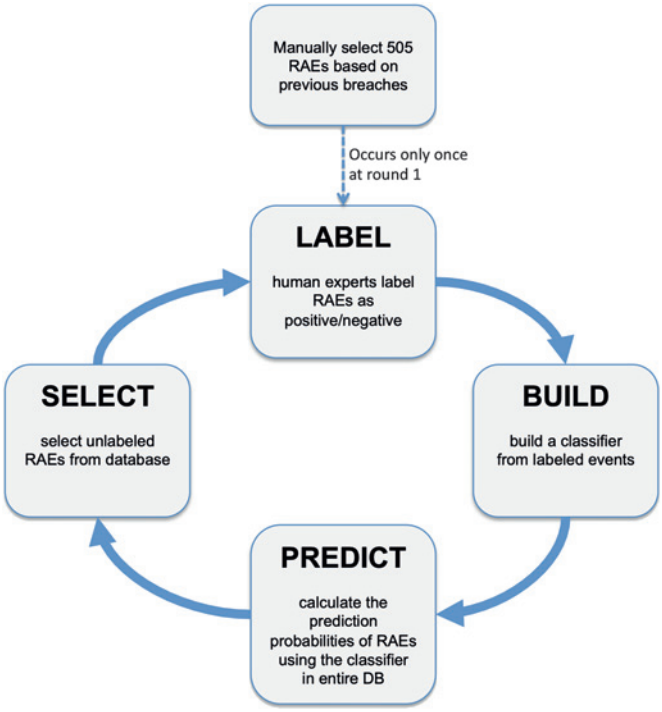

Figure 2 Construction of the training set. After the initialization set was obtained, three cycles of 'BUILD—PREDICT—SELECT—LABEL' were conducted. RAE, record access events. DB, database.

this problem, we oversampled the suspicious RAEs to have the balanced sets, similar to a sampling process commonly used in active learning.[23 24] To further reduce the effect of the problem, in each round described above and illustrated in figure 2, prediction scores based on the previous round's model were used as prior probabilities to select new RAEs to be labeled.[25] Oversampling positive cases has the consequence that the model's estimates cannot be used as proxies for absolute risk. However, proper ranking of the events was the focus of our interest, as this determines which cases need follow-up by privacy officers, and this ranking did not change with oversampling. The prevalence of suspicious access as well as the resources available to investigate them varies from institution to institution. Hence, the important aspect of this type our model is the ability to correctly rank the events, and not the ability to produce an accurate estimate of absolute risk for each individual event.

At the end of this process, the final training set contained 1291 labeled RAEs. Of these, 643 RAEs were marked as suspicious, and 648 RAEs were marked as appropriate.

Building a classifier

We built classifiers on the training set using logistic regression (LR) and a support vector machine (SVM). Since our training set was relatively small (0.09% of the total available set), we adopted a 10-fold cross validation approach.[26] The whole training set was first partitioned into 10 near-equal parts. Then, 10 iterations of training and validation were performed, such that, within each iteration a model was trained on nine parts, and then the fitted model was applied to the held-out part. The area under the receiver operating characteristic curve (AUC) was calculated on the held-out part. The AUC was used as a discrimination measure for the binary classification; a value for AUC close to 1 indicates that the model discriminates well between the positive and negative accesses, and a value close to 0.5 indicates poor discrimination.[27] Ten AUC values were

obtained per technique. To minimize overfitting of the model to the data, the one with the median AUC on the held-out test set was chosen as the final model.

Evaluation on the study set

We measured the performance of the classifiers by applying the final models to the entire training set. We used the Wilcoxon rank sum test to see whether there was any difference in performance between the two techniques.

Evaluation on the investigated events set

We measured the performance of our models on an external set of cases that were investigated for inappropriate access. Note that the label in this data set is not suspicious, but rather truly inappropriate. During the 2-month data-collection period for this study, the privacy officers of the three sites had investigated several events. Of these 58 events (involving five users and 11 patient records) events were determined by them to be inappropriate accesses. The investigation of these events were triggered and conducted independently from our study. Our final models were applied to this external set of 58 investigated events. Since all events were true positives, we used sensitivity, but not specificity, as the primary measure of performance.

RESULTS

In the 2-month study period, a total of 10.5 million RAEs occurred. There were over 781 000 unique patients whose records were accessed. There were over 75 000 users on the network, and more than 40 000 had access to at least one clinical system.

Each of three sites—the IDN, Hospital 1 and Hospital 2—had one dedicated privacy officer per site. The labeling result of one site was reviewed by the privacy officers of the other two sites. Then, an average of two agreement rates of labeling were obtained per site. For example, the average agreement rate of IDN was 92%, calculated from 94% for 'IDN versus Hospital 1' and 90% for 'IDN versus Hospital 2.' The average agreement rates for Hospital 1 and Hospital 2 were 90% and 83% respectively.

Performance by cross-validation in the training set

We also applied the rule-based classification technique, consisting of the seven scenarios of suspicious access described in Appendix 2 online (www.jamia.org), as a baseline method to the 1291 RAEs in the training set. Unlike LR or SVM, this rule-based method produced only binary prediction labels, suspicious or appropriate without prediction scores. The results are shown in the confusion matrix (table 1). As can be seen, the baseline method had a sensitivity of 80% for this data set.

Univariate analysis results of the labeled 1291 RAEs are displayed in the online appendix, table A1 (www.jamia.org).

Figure 3 is a box plot of the AUC values from the 10-fold cross-validation procedure used to construct the SVM and LR models. Although the median AUC for the SVM (=0.909) was larger than that for the LR (=0.885) model, the difference was not significant (p=0.684, Wilcoxon rank sum test).

Table 1 Confusion matrix depicting the performance of baseline method in the training set

		Gold standard*		
		Suspicious	Appropriate	
Prediction	Suspicious	514	0	514
	Appropriate	129	648	777
		643	648	1291

*The gold standard was determined by security officers.

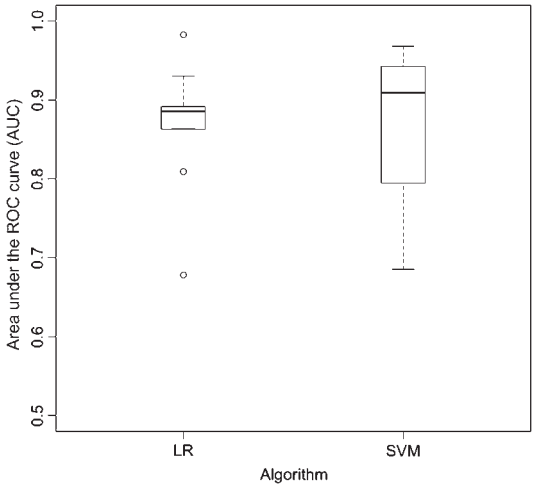

Figure 3 Area under the receiver operating characteristic (ROC) curve (AUC) estimates from 10 held-out test sets using cross-validation drawn as box plots. Each box extends from the 25% percentile to the 75% percentile of AUCs. The 50% percentile (median) is shown with a bold line. LR, logistic regression; SVM, support vector machine.

The final model within each algorithm was chosen as the one with the median AUC among the 10 models obtained from cross-validation. This model was applied to the whole set of 1291 events.[28] The performances of the two classifiers are shown in figure 4. The AUC for the SVM is 0.949, and that for the LR model is 0.911. The 95% CI of the AUC was (0.894 to 0.927) for the LR and (0.937 to 0.962) for the SVM model, which suggests a significant difference in the performance of the two techniques.[29]

The result of the final LR model is shown in table 2. Predictors are sorted by descending order of the OR estimates. All of the 18 predictors (including interactions) were deemed significant.

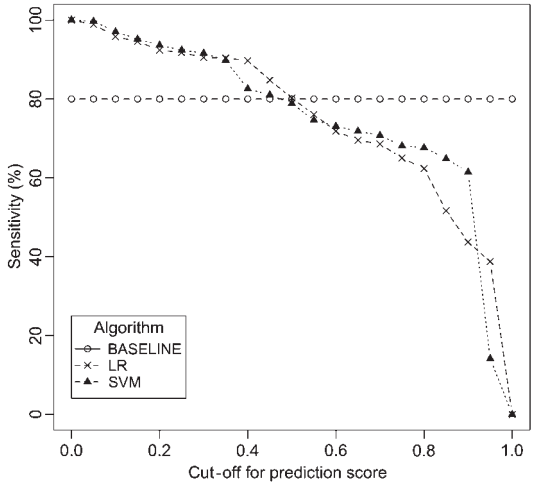

Figure 4 Sensitivity of the final model applied to the training set (n=1291). The performance of the baseline technique is depicted as a horizontal line. Since the baseline technique is a rule-based method, its outputs are binary (suspicious/appropriate), and its sensitivity is a single number unlike the statistical and machine-learning methods, for which we can obtain different sensitivities depending on the threshold. LR, logistic regression; SVM, support vector machine.

Table 2 Logistic regression model

Feature	Coefficient	OR (95% CI)
Works in same department	3.158	23.524 (2.450 to 225.843)
Same street address	2.599	13.450 (2.212 to 81.790)
Same family name	2.340	10.381 (6.653 to 16.199)
Over 200 accesses in day	1.300	3.669 (1.534 to 8.778)
VIP patient	1.175	3.238 (1.049 to 9.994)
Same city	1.047	2.849 (1.449 to 5.602)
Access on clinic day	0.869	2.385 (1.421 to 4.001)
Hospital 1 employee	−0.434	0.648 (0.426 to 0.986)
Hospital 2 employee	−0.448	0.639 (0.423 to 0.966)
Integrated delivery network employee	−0.531	0.588 (0.363 to 0.952)
Had recent visit	−0.616	0.540 (0.302 to 0.965)
Same zip code	−1.465	0.231 (0.111 to 0.479)
Researcher	−1.475	0.229 (0.114 to 0.458)
Had recent visit×VIP patient	−1.610	0.200 (0.061 to 0.652)
Is provider	−2.324	0.098 (0.050 to 0.190)
Care unit visit match	−3.124	0.044 (0.022 to 0.088)
Had recent visit×works in same department	−3.397	0.033 (0.003 to 0.347)
Works in same department× same family name	−9.343	0.000 (0.000 to 0.002)

Table 3 shows the number of RAEs that matches the different patterns for suspicious access in the baseline method for 10.5 million RAEs.

Comparison of sensitivities in an external set of inappropriate accesses

The baseline method, and the final LR and SVM models were applied to the set of 58 inappropriate events. The baseline method failed to detect any of these events. Figure 5 is a plot of sensitivity as a function of the cut-off for the predicted probability. The higher cut-off value will result in the smaller sensitivity. With a cut-off value of 0.5, LR correctly detected 75.8%, and SVM detected 79.3% of these events.

The following scenarios illustrate the investigated events that the models failed to detect:

- The models failed to detect inappropriate access by an employee into a coworker's record because the coworker/patient was not marked as an 'employee' in the operational database.
- One event in which a mother accessed her child's record was missed, since the mother and her child had different last names. The models infer family association based on the last names.

Table 3 Frequency of record access events (RAEs) from the 10.5 million RAEs matching each of the baseline patterns for suspicious access

Baseline method pattern	Frequency
Accessing a coworker's record	975
Accessing a VIP patient's record	1487
Accessing the record of a patient with no recent visit	1738
Accessing a family member's record	9202
Researcher accessing over 200 records in a day	12 439
Accessing a neighbor's record	58 152
Accessing the record of a patient who did not have a medical record number at the user's site*	1 097 496

*The large number of RAEs in this row is an artifact of the workflows and organizational structures at this integrated delivery network. For example, over 400 000 of these RAEs are accesses by resident physicians who are employees of the integrated delivery network and not of any clinical site.

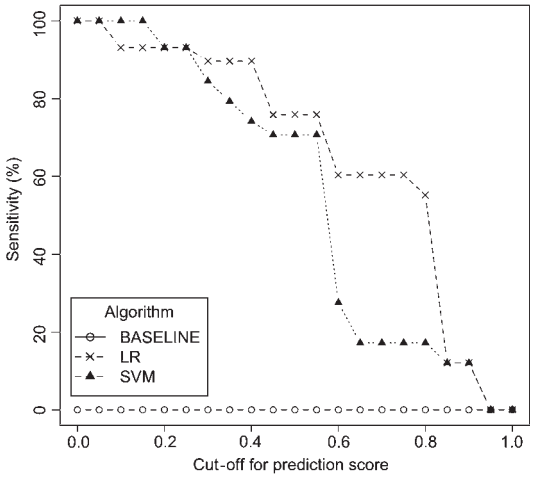

Figure 5 Sensitivity of the final model applied to the external validation set (n=58). The performance of the baseline technique is drawn as a reference. LR, logistic regression; SVM, support vector machine.

- In one event, the user self-reported that she had accessed the record by mistake.

DISCUSSION

This pilot study suggests that statistical and machine-learning techniques can help automate the process of detecting suspicious RAEs. The performance of the LR and SVM models, as measured by the AUCs on the study set (>0.90) and sensitivity on the investigated events set (>0.75), were very promising. This study, unlike other approaches discussed earlier, addresses the problem of detecting suspicious access to electronic medical records using a scalable approach based on statistical and machine-learning techniques. Inappropriate access to medical records, while occurring rarely, is a serious data-security issue for healthcare organizations. Due to the large number of record access events and the challenges of limiting access of providers to records in a healthcare setting, automated approaches to identifying suspicious access can be a powerful tool for improving data security. Highly suspicious cases can be investigated by privacy officers to determine whether they were appropriate or inappropriate accesses. While there is an initial manual effort in constructing a training set and a predictive model using our approach, subsequent use of this model to rank suspicious accesses can be automated. During the data-collection period of the study, the integration of data from Operational DBs into the Event Data Mart was partly automated. We subsequently have completely automated the integration of data, thus allowing us to execute the SVM and LR models on the data set automatically.

An advantage of using the statistical and machine-learning models is that they produce scores indicating suspiciousness of access. The scores can be used to prioritize the investigations of events. In contrast, using simple rule-based patterns, such as in the baseline method, results in a very large unranked list of suspicious events (table 2) that have to be investigated by the privacy officers. Investigating tens of thousands of events every months is far beyond the resource capabilities of any healthcare organization.

In this study, we retrospectively analyzed data in an offline setting to create models for detecting suspicious access. These models can be implemented in an operational clinical setting by integrating the data sources used to construct the model. These

models can be used to flag RAEs suspected of being inappropriate. Such knowledge could be used prospectively to prompt for a reason for access to a medical record. It could also be used to improve the yield of retrospective security audits by identifying RAEs that need further investigation and hence utilizing the limited time of the privacy officers to address the cases that most warrant an investigation. A more ambitious future objective will be to detect highly suspicious access attempts in real-time, and prevent such access.

This preliminary study has limitations. The prediction models that we created may not generalize to other institutions because of differences in organizational structures, workflows, and software systems. Nevertheless, we expect that our approach of constructing the models should be generalizable. By showing that the process can be automated, we hope to encourage other institutions to create models that facilitate the process of identifying inappropriate accesses.

The procedure used to construct the training set relies on the types of inappropriate RAEs known to the privacy officers and the research team that might bias the models. We used our knowledge of inappropriate types of accesses in selecting initial cases for the training set. The privacy officers used their knowledge to decide on appropriateness of accesses. Thus, the model that is constructed might be biased toward detecting the types of inappropriate accesses that are known to us. For example, we assume that any staff member in a clinic in which a patient has an appointment has a valid reason to access a patient's record. However, a particular staff member within a clinic might not be involved in a patient's care and might still access the record inappropriately.

The baseline method, which had a sensitivity of 80% with the training set, could not detect a single event in the external set of 58 inappropriate events. This difference in performance is of concern. It calls into question the validity of the baseline method, which could be overfitting the training data. The rules in the baseline method were derived from experts who also helped label the training set; hence this could be a possible explanation. The performance difference could indicate a limitation of systems that have to produce binary outputs instead of continuous ones. The external set consisted only of truly inappropriate accesses, which are just a small subset of the suspicious cases. Due to the binary nature of the baseline method, it had to declare each case as appropriate or not, instead of assigning a continuous estimate. What makes a suspicious case also an inappropriate one appears not to be well captured in the baseline method, and the small sample size calls for more investigations. In contrast, LR and SVM produce continuous estimates indicating degrees of appropriateness. Thus, their performance on the external set, though degraded, retains a high sensitivity.

We used supervised machine-learning methods to build prediction models for rare events, which led to challenges. First, the construction of the training set is an effort-intensive activity. Each RAE had to be reviewed by an expert, a privacy officer, to label the cases for the training set. Thus, constructing a sufficiently large training set comes at a cost. Furthermore, due to time-consuming manual review required by privacy officers, it was not possible to build a large training set. Since the inappropriate events are rare, we chose to represent them in a disproportionately larger number in the training set, so that a sufficiently diverse pattern of events were included in the training set. One problem with this representation of events in the training set is that the estimates produced by the models cannot be considered true probabilities. However, the models are still appropriate to rank cases in descending order of suspiciousness to target the investigation by privacy officers.

The set of features we used in this study was influenced by the selection of operational data. We did not use clinical data such as patient problems to determine appropriateness of access. For example, in a patient with respiratory insufficiency, access by a respiratory therapist might likely be appropriate. We used some social associations between users and patients, such as being department colleagues or neighbors at home. However, there might be relationships that our database is not aware of, such as high-school friendships that might lead to inappropriate access. These kinds of data would be difficult to impossible to incorporate in the database.

A challenge in operationalizing the use of statistical and machine-learning techniques for detecting suspicious access is to be able to create larger training sets. This will require systematic improvements to the 'BUILD–PREDICT–SELECT–LABEL' procedure shown in figure 2. We are currently working on optimized selection of events to be reviewed and automating a review process for privacy officers as well as collecting a larger data set of labeled examples.

CONCLUSIONS

We developed an approach to utilizing statistical and machine-learning methods to identify suspicious accesses to electronic medical records. The approach automates the integration of data from disparate sources in the enterprise into a single data mart. The integrated data source is used to construct predictive models for identifying suspicious access to electronic health records. The results of our study indicate that these methods show promise. However, methodological refinements and further evaluation are necessary, particularly in constructing training sets, to put these methods into operational use in diverse clinical settings.

Acknowledgments We acknowledge K Grant, P Rubalcaba, D Mikels, C Spurr, D Adair, J Raymond, J Einbinder, H Shea, M Hamel, C Griffin, J Saunders, M Miga, M Muriph, A Kennedy, and C Soran for their contributions to this research.

Funding This study was funded in part by PHS Information Systems Research Council (ISRC) and in part by NIH grants 1R01LM009520 and U54HL108460.

Provenance and peer review Not commissioned; externally peer reviewed.

REFERENCES

1. **Connecting for Health.** *The personal health working group final report*. New York, NY: Markle Foundation, 2003.
2. **Kaelber DC,** Jha AK, Johnston D, *et al*. A research agenda for personal health records (PHRs). *J Am Med Inform Assoc* 2008;**15**:729–36.
3. **HIPAA Privacy Rule.** *Standards for Privacy of Individually Identifiable Health Information*. US Dep. Health Hum. Serv. 67 Fed. Reg. 157, 2002:53181–273.
4. *Health Information Technology for Economic and Clinical Health Act*. US Dep. Health Hum. Serv. 74 Fed. Reg. 160, 164, 2009.
5. **Ponemon L.** *Fourth Annual US Cost Of Data Breach Security*. Traverse City, MI: Ponemon Institute, 2008.
6. **Ornstein C.** *Ex-Worker Indicted In Celebrity Patient Leaks, La Times*. (Archived by WebCite®). http://www.webcitation.org/5oKaFQsP9 (accessed 24 May 2010).
7. **Arkansas News Bureau.** *Three Sentenced For Privacy Violations In Pressly Case*. Arkansas News. (Archived by WebCite®). http://www.webcitation.org/5oKacngTT (accessed 24 May 2010).
8. **Kerr J.** *Walter Reed: Data breach at Military Hospitals, Army News*. (Archived by WebCite®). http://www.webcitation.org/5oKatgFD5 (accessed 24 May 2010).
9. **Ferreira A,** Cruz-Correia R, Antunes L, *et al*. Access control: how can it improve patients' healthcare? *Stud Health Technol Inform* 2007;**127**:65–76.
10. **Ferraiolo DF,** Sandhu R, Gavrila S, *et al*. Proposed NIST standard for role-based access control. *ACM Trans Info Syst Security* 2001;**4**:224–74.
11. **Motta GH,** Furuie SS. A contextual role-based access control authorization model for electronic patient record. *IEEE Trans Inf Technol Biomed* 2003;**7**:202–7.
12. **Peleg M,** Beimel D, Dori D, *et al*. Situation-based access control: privacy management via modeling of patient data access scenarios. *J Biomed Inform* 2008;**41**:1028–40.
13. **Salazar-Kish J,** Tate D, Hall PD, *et al*. Development of CPR security using impact analysis. *Proc AMIA Symp* 2000:749–53.

14. **Asaro PV,** Herting RL Jr, Roth AC, *et al*. Effective audit trails—a taxonomy for determination of information requirements. *Proc AMIA Symp* 1999:663—5.
15. **Herting RL Jr,** Asaro PV, Roth AC, *et al*. Using external data sources to improve audit trail analysis. *Proc AMIA Symp* 1999:795—9.
16. **Malin B,** Airoldi E. Confidentiality preserving audits of electronic medical record access. *Stud Health Technol Inform* 2007;**129**:320—4.
17. **Zhou Z,** Liu BJ. HIPAA compliant auditing system for medical images. *Comput Med Imaging Graph* 2005;**29**:235—41.
18. **Coleman RM,** Ralston MD, Szafran A, *et al*. Multidimensional analysis: a management tool for monitoring HIPAA compliance and departmental performance. *J Digit Imaging* 2004;**17**:196—204.
19. **Bell T,** Carcello J. A decision aid for assessing the likelihood of fraudulent financial reporting. *Audit J Pract Theor* 2000;**9**:169—78.
20. **Brause R,** Langsdorf T, Hepp M. Neural data mining for credit card fraud detection. *Proceedings of the 11th International Conference on Tools with Artificial Intelligence*. 1999:103—6.
21. **Yoshida K,** Adachi F, Washio T, *et al*. *Density-Based Spam Detector. Proc SIGKDD04*. New York, NY: ACM Press, 2004:486—93.
22. **Perez JM,** Muguerza J, Arbelaitz O, *et al*. Consolidated tree classifier learning in a car insurance fraud detection domain with class imbalance. *LNCS* 2005;**3686**:381—9.
23. **Chawla NV,** Bowyer KW, Hall LO, *et al*. SMOTE: synthetic minority over-sampling technique. *J Artif Intell Res* 2002;**16**:341—78.
24. **Ertekin S,** Huan J, Giles CL. Active learning for class imbalance problem. *Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* 2007:823—4.
25. **Chawla NV.** C4. 5 and imbalanced data sets: investigating the effect of sampling method, probabilistic estimate, and decision tree structure. *Workshop on Learning from Imbalanced Datasets II. Proceedings of the International Conference on Machine Learning*, 21—24 August, 2003.
26. **Hastie T,** Tibshirani R, Friedman J. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. 2nd edn. New York: Springer, 2009.
27. **Metz CE.** Basic principles of ROC analysis. *Semin Nucl Med* 1978;**8**:283—98.
28. **Nisbet R,** Elder J, Miner G. *Handbook of Statistical Analysis and Data Mining Applications*. Burlington, MA: Academic Press, 2009.
29. **Hanley JA,** McNeil BJ. A method of comparison of the areas under the receiver operating characteristic curves derived from the same cases. *Radiology* 1983;**148**:839—43.