Flagging and Ranking Suspicious Accesses in Electronic Health Record Systems

By

Monica Satyanarayan Hedda

Thesis

Submitted to the Faculty of the

Graduate School of Vanderbilt University

in partial fulfillment of the requirements

for the degree of

MASTER OF SCIENCE

in

COMPUTER SCIENCE

MAY 11, 2018

Nashville, Tennessee

Approved:

Daniel Fabbri, PhD

Bradley A. Malin, PhD

To My Grandmother.

ACKNOWLEDGMENTS

I would like to thank my advisor, Dr. Daniel Fabbri, for his guidance and support. I am grateful, for the opportunity he provided to a novice researcher like me and for helping me develop an insight into the problem which has been instrumental in the success of this work. He taught me not to get overwhelmed by a complex problem and how to systematically break it down into simpler ones.

I would also like to thank Dr. Bradley Malin for reviewing this work and providing valuable feedback which has helped in improving the quality of this work tremendously. Group discussions with Dr. Fabbri, Dr. Malin, Dr. Eugene Vorobeychik and Chao Yan have been very helpful in broadening my understanding of this domain. Thanks to Joseph Coco for his help with preparation of the data required for this work.

I would like to extend my thanks to Dr. Janos Sztipanovits and Dr. Sandeep Neema for getting me started in the field of research. I am thankful for the help of the faculty and staff of Computer Science and Biomedical Informatics department throughout my time at Vanderbilt.

I want to express my deepest gratitude to my parents, I would not have come this far without their love and support. My thanks go to all my friends and family for their support and help over the years. Finally, special thanks to my husband, Ashish Tapdiya, for encouraging me to go back to school and join the Graduate program. He has been my constant source of inspiration in this pursuit of learning.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

Chapter 1

Introduction

Electronic health record (EHR) systems can improve the quality of patient care, safety and education, while reducing costs and enabling research [1]. To encourage the adoption and use of EHR systems by healthcare providers, the US government passed the Health Information Technology for Economic and Clinical Health (HITECH) Act of 2009 [2] and established incentives for healthcare providers that demonstrate meaningful use of EHR system to provide better patient care [3]. As a result, integration of EHR systems into healthcare organizations (HCOs) has continually increased. However, the increased accessibility of protected health information (PHI) in EHR systems leads to a greater potential for misuse and abuse by the authorized users. Such events can result in penalties levied by federal and state regulators.

An EHR is fundamentally a collaborative information system, which, traditionally, is protected through proactive strategies, such as fine-grained access control technologies [4]. Such technology is often integrated into EHR systems; however, the dynamics of patient care, in combination with the difficulty in predicting who needs access to a patient's medical record when, make it challenging to deploy such fine-grained control schema without triggering a substantial quantity of false alerts and slowing care workflows [5]. Despite acknowledging the potential for insider threats, HCOs typically do not instantiate fine-grained controls [6]. This implicitly suggests that HCOs deem the losses associated with impacts on workflow and care to be greater than those brought about by employees who misuse or abuse their privileges.

Still, HCOs do not neglect insider threats entirely. In lieu of fine-grained proactive protections, HCOs tend to rely upon retrospective mechanisms, such as auditing and investigation. In the United States, the Security Rule of the Health Information Portability and

Accountability of 1996 (HIPAA) requires that all HCOs maintain audit logs, analyze them for inappropriate use and report misuse [7]. Hospitals maintain audit log of all accesses to PHI and the audit log is often reviewed by administrative officers to detect inappropriate access. However, the sheer volume of accesses documented by large HCOs makes manual review infeasible. The number of access transactions is often over one million per day [8], while officers have only one or two people at their disposal (often allocating only a portion of their time) to run investigations. As a result, many HCOs prioritize their investigations by monitoring patient records deemed to be very important persons (VIPs) [9] or upon patient complaints [10]. In the latter scenario, compliance officers investigate the accesses to patient records after a complaint has been registered.

More recently, there has been a push to (semi-)automate the auditing process. However, there are many challenges an HCO faces to do so. For instance, the information often required to determine if an access is inappropriate is not stored in the audit log [11], [12]. As a consequence, HCOs have deployed rule-based methods [10], [13] to capture high-risk behavior and promote them to compliance officers for review. Figure 1.1 shows the process associated with such a traditional rules-based auditing system. Unfortunately, rule-based flagging systems can result in high false positives [9]. For example, a typical rule is to flag when an employee accesses an EHR of a patient with the same last name. Yet, for individuals with a common name, clearly this rule will trigger an excessive amount of alerts.

Given the state of affairs, we set out to, assess the validity of rules for auditing accesses made in EHR systems. Our goal is to test, through simulation and theoretical analyses, if these flags occur at a higher rate than expected, and therefore serve as a valid means to detect inappropriate behavior. In order to achieve this goal, we solve three sub-problems that are listed below.

1. **Investigate the difference between observed and expected high-risk accesses.** We introduce an approach to investigate the difference between the observed and ex-

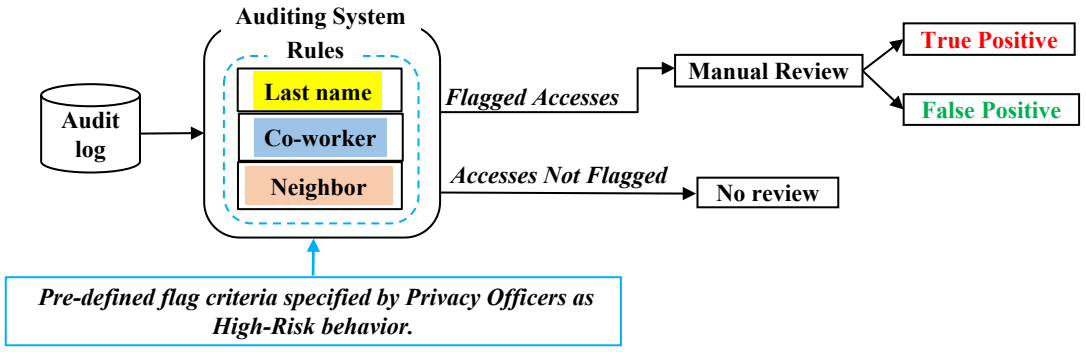


Figure 1.1: The process by which an HCO investigates accesses to EHRs deemed to be of high-risk.

pected rate of high-risk accesses in EHR systems for typical expert-specified rules. If a rule holds merit, we anticipate that the observed rate of high-risk accesses will be higher than the expected rate of high-risk accesses to the EHR system. Using one week of data from Vanderbilt University Medical Center (VUMC), we show that there are many rules for which this difference is statistically significant.

2. **Select and prioritize rules based on deviation between observed and expected.** We introduce an approach for selection and prioritization of the high-risk rules. This approach is based on the magnitude of the deviation between the observed and expected frequency of high-risk accesses for each rule.

3. **Prioritize flagged high-risk accesses for investigation.** To improve the manageability of a manual review process in resource constrained environments, explanation-based filtering [10] can be utilized to prioritize the flagged accesses for manual review. Note that rule-based flagging and explanation-based filtering are complementary approaches to detect inappropriate behavior. While rules capture the high-risk behavior, explanations reduce the set of accesses that need to be investigated to a set of un-explained accesses. We find synergy between these two auditing approaches and introduce an explanation-based mechanism to prioritize high-risk ac-

cesses flagged by rules for manual investigation. We show that many, though not all, of the high-risk accesses can be explained away with clinically justifiable reasons.

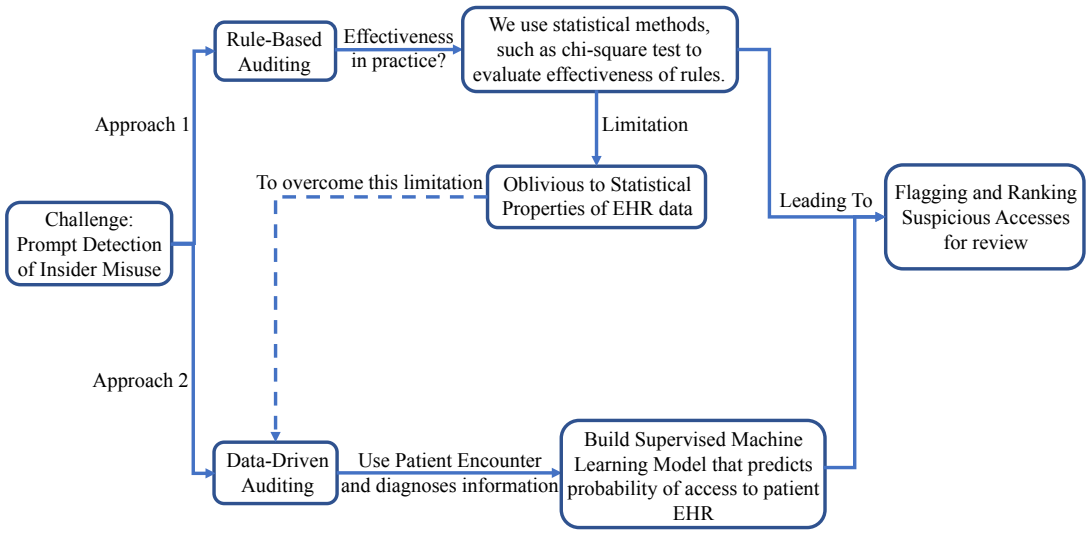


Figure 1.2: Summary of challenges, approaches and contributions in this work.

Despite its potential, rule-based auditing system is inherently limited by its reliance on predefined rules, which themselves are often based on domain expertise. However, in practice, there are many possible reasons for inappropriate access. As a result, the access coverage (i.e., proportion of accesses effectively monitored by high-risk rules) is low. Another limitation of rule-based auditing system is that the rule-based flagging of high-risk accesses is dependent on the correctness and completeness of EHR data. Incorrect EHR data (e.g., the wrong patient's last name is entered into the EHR system) leads to gaps in identifying a potential high-risk access. Similarly, an incomplete address or a P.O. Box would lead to gaps in flagging a potentially high-risk access according to the Residential Proximity or Residential Street rule.

To this end, we propose an automated suspicious access detection system which can overcome above mentioned limitations of rule based auditing system. We hypothesize that the department from which an employee will access a patient's EHR and the time of this access depends upon the patient's clinical encounter times and diagnosis. We believe that the patient clinical encounter and diagnosis information can be used to predict which hospital

department's employee will access patient's EHR and when will the access occur. We propose an automated suspicious access detection system based on supervised machine learning technique that learns access patterns of various hospital departments using the audit log of accesses and corresponding patient's clinical encounters and diagnosis. This system predicts the probability of access of a patient's record on a given date by an employee of a given hospital department. The value of probability of occurrence for each access can be used to identify suspicious accesses in the audit log, which according to the system are the accesses predicted to have zero/low probability of occurrence. This automated suspicious access detection process can reduce the time and manual effort in identifying suspicious accesses to EHR. The suspicious accesses detected by our system can be investigated by administrative officers to identify if the suspicious access is in fact an inappropriate access. This fast detection of insider misuse can reduce further harm to the sensitive patient health information. Figure 1.2 summarizes challenges, approaches and contributions in this work.

Chapter 2

Background

In this chapter, we first provide an overview of the different frameworks and techniques we utilize in this thesis. Next, we review works related to the Electronic Health Record (EHR) auditing systems.

2.1 Background

2.1.1 Supervised Machine Learning

Machine learning is the process of analyzing the data by establishing the relationship between multiple features in the data to solve various problems including classification, prediction etc. Machine learning algorithms are classified into supervised and unsupervised algorithms. A supervised machine learning algorithm learns from a dataset of training instances that are represented using same set of features with known associated labels. It applies the learned relationships to predict labels for future new instances of data [14]. For example, given a training set of patient disease diagnoses mapped to a department that treats the diseases, supervised machine learning learns from this dataset and builds a function to best determine the department from the disease diagnoses. Then, given a set of disease diagnoses withheld from the training dataset, the function will predict a department that treats the diseases.

Random Forest Classifier. Random forest classifier is an ensemble learning algorithm for building a predictor with a set of decision trees which grow in random subset of data [15]. Random forest fits the set of decision tree classifiers on the various subsets and utilizes averaging to achieve optimal predictive accuracy and controls over-fitting of data.

Classifier Performance. The performance of random classifier is evaluated using area under the receiver operating characteristics curve (AUC ROC) metric [16]. ROC curve is a

graph of true positive rate vs false positive rate of a classifier.

2.2 Related Work

Various auditing strategies have been proposed to detect inappropriate insider accesses in EHRs. Boxwala et al. have introduced an automation strategy based on statistical and supervised machine-learning techniques to detect suspicious accesses to EHRs [9]. This strategy uses audit logs and EHR data to construct features to learn predictive models that rank suspicious and non-suspicious accesses according to their risk. The features constructed in this technique are user-related, patient-related, record-access-events-related, encounter-related, and user-patient-relationship-related. The encounter-related features in this technique only cover Encounter location type and patient visits. Our technique focuses on different types of encounters and encounter times. Also, their technique includes Patient features based on patient type such as whether patient is VIP or patient is also an employee, it does not use patient's diagnosis for feature construction.

Recognizing that not all suspicious accesses are affiliated with a specific pattern, a variety of frameworks have been developed to detect anomalous accesses based on deviations from expected behavior [17], [18].

Fabbri et al. have proposed notion of an explanation-based auditing system (EBAS) considering that most accesses to EHRs occur for a valid clinical or operational reason. EBAS works by filtering out accesses to the EHR according to explanations generated automatically from the data by a mining algorithm [10]. EBAS is also equipped to explain reasons for an access to EHR based on diagnosis information [19].

While all of these auditing strategies offer certain benefits over the simple rule-based auditing system, currently approaches based on the latter are in common use by HCOs. Hence, we first evaluate the effectiveness of the auditing rules in identifying inappropriate accesses to EHR and propose methods to improve effective use of auditing rules. Recognizing the inherent limitations of auditing rules which lead to low access coverage, we

propose an automated auditing system that utilizes patient clinical encounter and diagnosis information to automatically detect suspicious accesses to EHR, which has not been investigated in the previous works on EHR auditing.

Chapter 3

Flagging and Ranking Suspicious Accesses in Electronic Health Record Systems Using Auditing Rules

Healthcare organizations (HCOs) often deploy rule-based auditing systems to detect insider threats to sensitive patient health information in electronic health record (EHR) systems. These rule-based systems define behavior deemed to be high-risk a priori (e.g., family member, co-worker access). While such rules seem logical, there has been little scientific investigation into the effectiveness of these auditing rules in identifying inappropriate behavior. Thus, in this work, we introduce an approach to evaluate the effectiveness of individual high-risk rules and rank them according to their potential risk. We investigate the rate of high-risk access patterns and minimum rate of high-risk accesses that can be explained with appropriate clinical reasons in a large EHR system. An analysis of 8M accesses from one-week of data shows that specific high-risk flags occur more frequently than theoretically expected and the rate at which accesses can be explained away with five simple reasons is 16 - 43%.

3.1 Methodology

We hypothesize that a high-risk audit rule holds merit when the observed frequency at which it fires is higher than what would occur due to routine daily behavior. To test this hypothesis, we compare the observed frequency of high-risk accesses in a large EHR audit log with what one might expect to observe at random. We apply a goodness of fit test to determine if there is a significant difference between the observed and expected frequencies. We further examine the observed high-risk accesses flagged by each rule to determine the minimum rate in that these accesses can be explained with clinical reasons.

3.1.1 Data overview

The data investigated in this study is drawn from the VUMC EHR system. Table 3.1 depicts the data investigated in this study. These data are an integration of EHR audit log, employee personal information, patient personal information (e.g., names, dates of birth, and residential addresses) with information about the department for which the employee is affiliated (e.g., the Anesthesiology department).

Table 3.1: Summary of the VUMC data used in this investigation.

Total Accesses	7.5M
Repeat Accesses	6.9M
Self-Accesses	21K
Unique Non-Self Accesses (L_{EP})	710K
Unique Employees (E)	13K
Unique Patients (P)	152K
Unique Departments	2.1K

We designate an access as a *Self-access* when the employee has accessed his/her own record. We assume this occurs when the first name, last name and date of birth of the employee and patient in the access are the same. We designate an access as a *Repeat access* when the employee accesses the record of the same patient earlier in the week. All of the accesses except the first access are considered as *Repeat accesses*.

3.1.2 Types of high-risk behavior

While there are many types of high-risk behavior, we selected the following types for our experiments through background analysis. Specifically, we investigate five high-risk rules in this study:

1. **Co-Worker:** The EHR user and patient are both employees of the VUMC.
2. **Department Co-Worker:** The EHR user and patient work in the same VUMC department.

3. **Last Name:** The EHR user and patient have the same last name.

4. **Geographic Proximity:** The EHR user lives within 0.25 miles of the patient.

5. **Residential Street:** The EHR user lives on the same street as the patient. In addition, we added one rule to ascertain if the results of our experiments are merely an artifact of the data or if they are indicative of suspicious behavior:

6. **First Name:** The EHR user and patient have the same first name.

3.1.3 Method overview

In this section, we provide an overview of the method to test our hypothesis and to determine the minimum rate of high-risk accesses explained with a clinical reason, as depicted in Figure 3.1.

The steps in this method are defined broadly as follows:

1. Determine the observed frequency of the high-risk accesses.

2. Determine the expected frequency of high-risk accesses by:

   (a) Using simulations with random samples of users and patients.

   (b) Using simulations with permutations of users and patients.

   (c) Using a theoretical formulation.

3. Compare the observed frequency of high-risk accesses to the expected frequency of high-risk accesses, and determine the significance of the deviation between the observed and expected frequency.

4. Use the explanation-based method to identify the observed high-risk accesses that can be explained with clinical reasons, and determine the minimum rate that observed high-risk accesses can be explained.

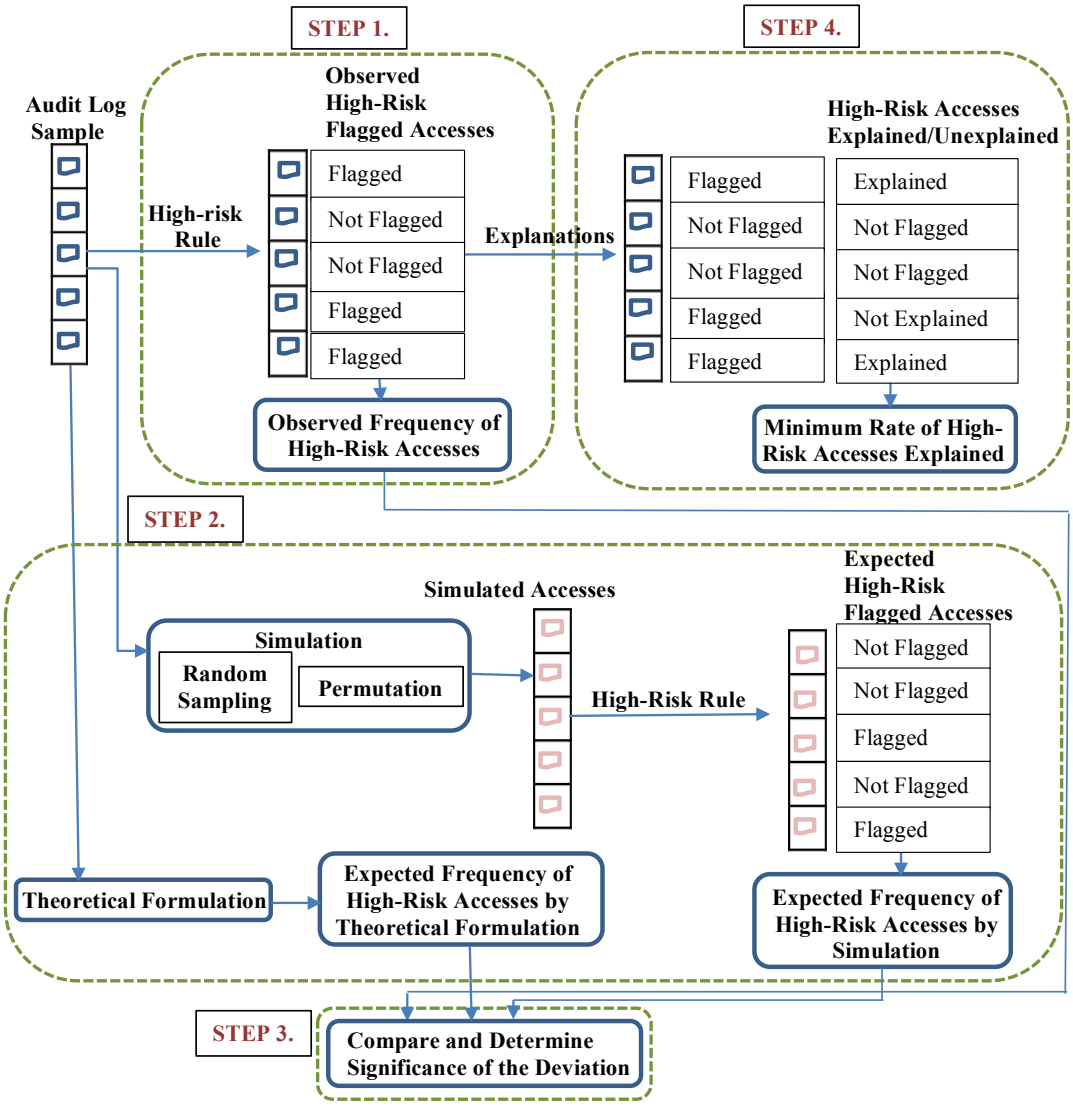


Figure 3.1: The steps to compute the observed and expected frequencies of high-risk accesses, and the minimum rates that high-risk accesses can be explained.

3.1.4 Observed and expected frequencies of high-risk accesses

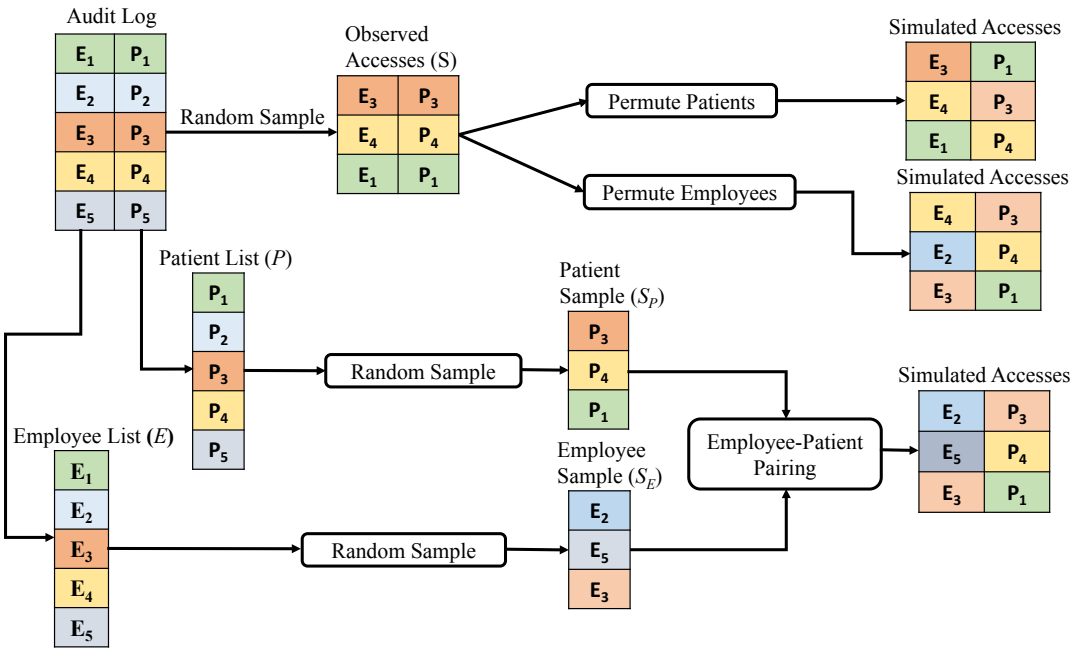


Figure 3.2: An overview of the process for sampling the observed accesses and simulation of the expected accesses.

Figure 3.2 depicts the method to obtain the observed accesses by sampling the audit log and simulation of the expected accesses using permutation and random sampling methods. We explain these methods in detail in the following sections.

Observed. We obtain the observed frequency of high-risk accesses empirically from the set of unique employee-patient pairs L_{EP}, where E is the list of employees (or users) and P is the list of patients in the employee-patient access pairs L_{EP}. These pairs are obtained from the 710,000 unique accesses in the audit log and are devoid of any self-access. We select a random sample S of 100,000 pairs from L_{EP}, each of which is assessed for the high-risk criteria. We count the occurrence of the high-risk accesses across the sample and calculate the frequency of the high-risk accesses in the sample as: $|high\text{-}risk\ accesses|\ /\ |S|$.

Expected. To simulate accesses and obtain an expected frequency distribution of high-risk behaviors we apply both permutation and random sampling methods. We use two distinct methods to confirm these simulation methods do not result in selection bias and

that the sample selected by our methods are representative of the population. We compare the results of the simulations to verify if the results lead to the same conclusion.

Expected: Permutation. In this approach, we construct simulated accesses by shuffling the data points in S. We use two types of permutation methods to simulate accesses and verify that results of both the methods lead to the same conclusion. 1) Permute Patients: This method shuffles the list of patients while holding the list of employees in sample S constant. 2) Permute Employees: This method shuffles the list of employees while fixing the list of patients in sample S.

Expected: Random Sampling. We obtain the employee list E and the patient list P from the set of employee-patient pairs L_{EP}. Next, we select a random sample of 100,000 employees S_E and 100,000 patients S_P (without replacement) from E and P, respectively. We then construct simulated accesses by randomly matching the records in S_E and S_P.

For each simulation, we calculate the frequency of high-risk accesses in sample S.

Expected: Theoretical Formulation. The expected frequency of high-risk accesses is computed empirically, using the probabilities of high-risk accesses occurring among the employees and patients in sample S. We determine the expected frequency for five of the six rules presented above.

The expected frequency of the high-risk accesses using probabilities is computed as $\sum_{i=1}^{|x|} P_{\mathrm{E}i} P_{\mathrm{Pt}i}$, where, $P_{\mathrm{E}i}$ = |*Employee with attribute value x*| / $|S_E|$, $P_{\mathrm{Pt}i}$ = |*Patient with attribute value x*| / $|S_P|$, *attributes*: [last name, first name, residential street name, work department name] and S_E and S_P are the lists of employees and patients in S, respectively.

Experimental Evaluation. We run 10 experiments each for the randomization and permutation methods to compute the observed and expected frequencies of high-risk accesses. We compute the ratio of the mean observed frequency to mean expected frequency for each of the high-risk rules. We also compute the percentage of observed and expected high-risk accesses for each high-risk type to determine the rate of observed and expected high-risk accesses.

Minimum rate of high-risk accesses explained. While there are many operational and clinical reasons that can explain the reason for accesses in an EHR, we select primary treatment, payment and healthcare operations (TPO) [20] to ascertain the extent to which high-risk accesses can be explained. We specifically focus on explanations in the form of 1) scheduled appointments, 2) ordered lab results, 3) ordered medications, 4) admission, discharge, and transfer events, and 5) clinical documentation. A high-risk access can have multiple explanations (e.g., patient had a scheduled appointment with the accessing employee, and patient also had a lab order with the accessing employee). We use the explanation-based approach to prioritize the observed high-risk accesses for further investigation by administrative officers, with unexplained accesses considered as high priority for the investigation. The explained accesses can be ranked using the type and number of explanations available for the access.

Since we do not exhaust the list of possible reasons, we compute the minimum rate at which high-risk accesses can be explained for each high-risk rule. Additional plausible explanations for the access exist (e.g., user performed surgery on the patient) and could be invoked to raise the rate.

Goodness of fit chi-square test. We apply a χ^2 test to determine the goodness of fit between the observed (empirical) and expected (simulated) number of occurrences of high-risk accesses. This test is designed to ascertain if there is a significant difference, such that these deviations are likely not the result of chance alone. The measure of goodness of fit is:

$$\chi^2 = \sum_{i=1}^{n} \frac{(O_i - E_i)^2}{E_i},$$

where Oi and Ei are the observed and expected high-risk event frequencies of type i, respectively.

We test this value against a χ^2 distribution with 1 degree of freedom. This is because there are two categories: 1) High-risk accesses, 2) Non-high-risk accesses. We perform

this test at the 0.01 significance level (i.e., we accept the alternative hypothesis when the value result is below this level).

3.2 Results

In this section, we summarize the deviation of the observed from the expected high-risk access rates obtained by four methods 1) Permute Patients, 2) Permute Employees, 3) Random Sampling and 4) Theoretical formulation, for each high-risk access rule. We begin by presenting the rate of observed and expected high-risk accesses. Next, we summarize the minimum explanation rate for the high-risk accesses. Finally, we report the statistical significance of the deviation between the observed and expected high-risk access rates.

3.2.1 Observed versus expected frequencies of high-risk access.

Table 3.2 summarizes the observed to expected frequency ratios for the various high-risk access rules. It was found that the ratio of observed to expected frequencies varies from 0.99 to 4.33 for the high-risk behavior rules. The observed frequency of the high-risk accesses is higher than the expected frequency for all the high-risk rules except for the HCO Co-Worker rule, which is approximately 1 (at 0.99).

Table 3.2: Observed to expected frequency ratio for the high-risk access rules.

	Observed / Expected			
High-Risk Rule	**Permute Patients**	**Permute Employees**	**Random Sampling**	**Theoretical**
HCO Co-Worker	1	1	0.99	1
First Name	1.17	1.17	1.15	1.12
Last Name	1.53	1.5	1.54	1.72
Geographic Proximity	2.51	2.34	2.54	Not computed
Residential Street	4.04	4.22	4.33	3.8
Department Co-Worker	3.22	3.14	3.25	2.41

As expected, the ratio of observed to expected frequencies for the First Name high-risk class ranges from 1.12 to 1.17 for the four methods, suggesting there is no significant deviation between the observed and expected frequencies for this rule.

While the Geographic Proximity rule identifies if the patient and employee live within a fixed distance (0.25 miles), the Residential Street rule identifies if the patient and user live on the same street. Limiting the high-risk criteria to street name results in the higher ratio of observed to expected for the Residential Street rule than the ratio of observed to expected for Geographic Proximity rule.

Table 3.3: The observed versus expected percentage for the high-risk access rules.

High-Risk Rule	**Observed**	**Expected**			
		Permute Patients	**Permute Employees**	**Random Sampling**	**Theoretical**
HCO Co-Worker	3.84%	3.84%	3.84%	3.86%	3.84%
First Name	0.25%	0.21%	0.21%	0.21%	0.22%
Last Name	0.14%	0.09%	0.09%	0.09%	0.08%
Geographic Proximity	0.16%	0.07%	0.07%	0.07%	Not computed
Residential Street	0.12%	0.03%	0.03%	0.03%	0.03%
Department Co-Worker	0.04%	0.01%	0.01%	0.01%	0.02%

Table 3.3 shows the percentage of observed and expected high-risk accesses for each high-risk type in a sample of 100,000 accesses. The average percentage of observed high-risk accesses ranged from 0.03% to 3.8%. Though the percentage of high-risk accesses for HCO Co-Worker is higher than other types (by more than 3%), the observed frequency of high-risk accesses does not deviate from the expected (see Table 2). This suggests that these accesses can be assigned the lowest priority for investigation. The percentage of observed and expected high-risk accesses for the rest of the high-risk rules is less than 1%, but given that millions of accesses are committed per week, this small percentage yields non-trivial numbers of high-risk accesses.

3.2.2 Minimum rate of high-risk access explained away.

Table 3.4 summarizes the average rate (over 10 experiments) at which the observed high-risk accesses can be explained with clinical reasons. Notably, the selected set of explanations accounted for less than 50% of the accesses.

Table 3.5 summarizes the distribution of the explanations per high-risk rule. The highest number of high-risk accesses is explained with the Clinical documentation explanation

Table 3.4: The rate at which high-risk alerts would be explained away.

High-Risk Rule	Observed Accesses Explained Away	Standard deviation
HCO Co-Worker	38.78%	0.67
First Name	35.59%	2.54
Last Name	21.43%	2.88
Geographic Proximity	24.79%	3.13
Residential Street	16.11%	4.88
Department Co-Worker	43.90%	9.32

for all high-risk rules, with the percentage of accesses explained in the range of 15% to 43%. Scheduled Appointment explains 2% to 8% of the high-risk accesses. The other four explanations explain less than 5% of the high-risk accesses for all high-risk rules. The Clinical Documentation explanation shows high standard deviation (9.26) for the high-risk rule Department Co-Worker because of two out-lier experiments with the highest and lowest number of explained accesses, respectively.

Table 3.5: Distribution of explanations per high-risk rule (STD DEV = Standard deviation).

	% Observed Accesses Explained Away									
	Scheduled Appointment		Ordered Lab		Ordered Medications		(Admission, Discharge and Transfer)		Clinical Documentation	
High-Risk Rule	%	STD DEV	%	STD DEV	%	STD DEV	%	STD DEV	%	STD DEV
HCO Co-Worker	7.5	0.3	1.28	0.22	0.006	0.01	0.43	0.05	37.94	0.67
First Name	8.76	1.29	1.59	0.67	0	0	0.85	0.54	34.6	2.51
Last Name	2.23	1.01	0.43	0.37	0.22	0.35	0.22	0.35	20.69	3.15
Geographic Proximity	6.33	1.47	1.07	0.65	0	0	0.35	0.49	24.25	3.38
Residential Street	2.99	1.5	0.69	0.37	0	0	0.15	0.33	15.68	5.06
Department Co-Worker	8.26	4	1.4	2.02	0	0	1.1	1.43	43.6	9.26

3.2.3 Hypothesis test.

Table 3.6 shows the χ^2 result for a sample S of 100,000 unique accesses that are devoid of self-accesses. The expected number of accesses for this experiment was simulated through the permutation method (i.e., shuffling the list of patients and keeping list of employee fixed in the observed accesses). It should be noted that we did not include the Co-Worker rule in the χ^2 test because the results showed that there was no difference between the observed and expected accesses for this high-risk class.

The result of the χ^2 for high-risk rules Last Name, Geographic Proximity, Residential Street and Department Co-Worker indicated a probability < 0.0001. This is below the 0.01 significance level, such that we accept the alternative hypothesis for these high-risk rules

(i.e., the difference between the observed and expected frequency of high-risk accesses for these rules is statistically significant). The result of the χ^2 for First Name indicated a probability of 0.0113, which is above the 0.01 significance level, such that we reject the alternative hypothesis (i.e., there is no significant difference between observed and expected frequencies of high-risk accesses). This is notable because it suggests that our control rule is functioning correctly.

Table 3.6: Results of the χ^2 test for goodness of fit between the observed and expected with one degree of freedom per experiment. * denotes: Significance at 0.01 level.

High-Risk Rule	**Observed**	**Expected**	**Chi-Square**	**Probability**
First Name	245	208	6.42	0.0113
Non-High-Risk	99755	99792		
Last Name	140	91	25.87	<0.0001*
Non-High-Risk	99860	99909		
Geographic Proximity	166	66	150.1	<0.0001*
Non-High-Risk	99834	99934		
Residential Street	115	28	267.29	<0.0001*
Non-High-Risk	99885	99972		
Department Co-Worker	36	11	54.58	<0.0001*
Non-High-Risk	99964	99989		

3.3 Discussion

This study examined the extent to which high-risk EHR access rules are plausible in practice. Our empirical investigation illustrates that the observed rate at which high-risk rules are triggered is higher, at a statistically significant level, than what one would expect at random for several typical classes of high-risk behavior. This significant deviation suggests that there may be systematic EHR user behavior that requires further investigation, implying those rules may hold merit. Still, not all rules deviate to the same degree. In this respect, we further believe that the magnitude of the deviation of the observed frequency of

high-risk accesses from their expected frequency obtained from each high-risk class may be a plausible measure to assist in the prioritization of auditing rules in emerging game theoretic frameworks [21].

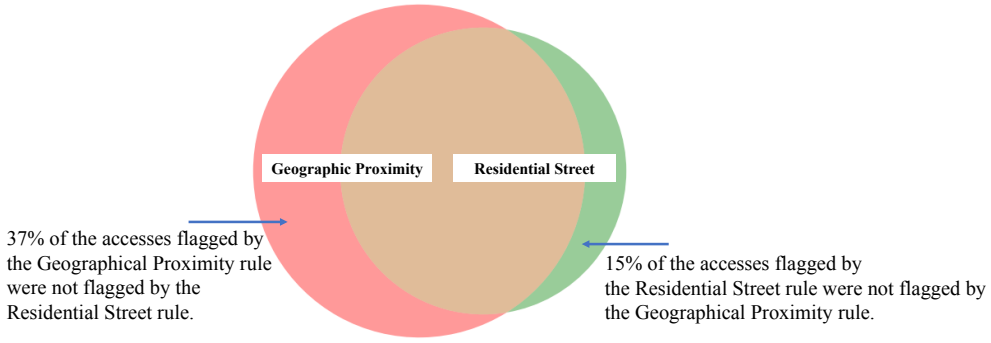


Figure 3.3: Geographic Proximity and Residential Street rules yield different results.

Geographical Proximity and Residential Street rules are designed to capture the same high-risk behavior (i.e. a user accessing records of a patient living in close geographic vicinity of the user). However, these two rules yield different results in terms of number of accesses flagged and the deviation of observed frequency from expected frequency of flagged accesses. Notably, the user and patient in 15% of the accesses flagged by the Residential Street rule do not live within 0.25 miles of each other, and the user and patient in 37% of the accesses flagged by the Geographical Proximity rule do not live on the same residential street, as depicted in Figure 3.3. Also, the length of the streets in the city varies from 0.4 miles to over 10 miles leading to a non-uniform application of the geographic vicinity criteria. This result indicates that rule definitions play an important role in effectively capturing high-risk behavior.

Despite their potential, high-risk access rules often have a high false positive rate. This makes them prohibitively expensive for HCOs to systematically investigate, which is a concern given the limited budgets available to privacy officers. However, we show that high-risk access rules can be complemented through an explanation-based model, such

that many accesses can be explained away by valid TPO reasons (16% to 44% depending on the high-risk rule at minimum). We believe this is notable because it suggests that high-risk rules and explanations are not correlated. Nonetheless, we believe that, in this setting, the explanations can be used to prioritize high-risk accesses for manual investigation. The unexplained accesses can be considered high priority for investigation, while the explained accesses can be ranked using the type and number of explanations available for the access.

There are, however, several limitations of this study that we wish to highlight for future investigations. First, an explanation-based system relies solely on the data stored in the database to generate explanation for an access. Missing information (or non-documented relationships) may result in few unexplained appropriate accesses. For example, EHR systems maintain records of patient appointments with doctors, but they do not explicitly record the relationship between the doctor and the nurse working together at the appointment. Thus, the system cannot readily explain the access of patient's record by the nurse working with the doctor, though the access in this case is appropriate. Other research has posited enhancing explanations with additional data learned from diagnosis information [19]. Second, in this study we only consider simple high-risk rules. In a future investigation, we plan to study more complex and nested high-risk rules. Fourth, this study suggests that different high-risk rules yield different results, but does not investigate the reasons for the differences.

3.4 Conclusions

In this chapter, we examined the rate of high-risk access rules in the electronic health record of a large healthcare organization. Specifically, we compared the observed and expected rates to ascertain the extent to which such rules are potentially useful in practice. The primary finding of this investigation was that such rules appear to detect behaviors that are statistically significantly different than what would transpire under random activities. There are many reasons why such deviation might transpire, but our investigation shows

that such rules should not be dismissed.

Chapter 4

Flagging and Ranking Suspicious Accesses in Electronic Health Record Systems Using Machine Learning Based Prediction Methods

Hospitals are facing steep challenges to protect patient data in EHR from insider threat. As per HIPAA Breach Notification Rule [22], hospitals are required to maintain an audit log of accesses to EHR and report the breaches within a specified time frame. To detect the breaches, the audit log is manually reviewed and investigated by the administrative officers. However, given the high volume of accesses per day in large hospitals [8], it may not be feasible to detect the breaches within the specified time frame. Hence, to enable prompt detection of insider misuse, hospitals need automated inappropriate access detection mechanisms.

During the course of care, a patient can have multiple clinical encounters, such as Appointment, Labs, etc. EHR system records information for each of these patient encounters including the encounter date. When a hospital employee accesses a patient's records, that information is recorded in the audit log as an access along with the access date. An access by an employee can occur on/before/after a patient's encounter. For example, a surgeon accesses a patient's records prior to the surgery encounter whereas a lab technician accesses a patient's records after the labs have been ordered by the physician. We observe that, the patient encounter and audit log information can be utilized to identify suspicious accesses to patient EHR. More specifically, encounter and access dates can be utilized to predict the probability of access on a given date for the observed encounters. If the probability of access is below a certain threshold and the access occurs, then the access can be automatically flagged as a suspicious access.

We leverage supervised machine learning technique to build a prediction model that utilizes a patient's encounter and diagnosis information to predict the probability of access

to patient's EHR. We observe that, access patterns may vary significantly across different hospital departments. Hence, we construct three different types of models for prediction:

1. Unique Model for Each Department: This model predicts the probability of access of a patient's EHR by employee of the given department on a given date,

2. Unique Model for Each Cluster of Departments: This model predicts the probability of access of a patient's EHR on a given date by employee of any department from given cluster of departments and

3. Single Model for All Departments: This model predicts the probability of access of a patient's EHR on a given date by an employee who can belong to any department.

4.1 A Motivating Scenario

We hypothesize that the access pattern of a department is determined by patient's specific encounter types and diagnoses. To test our hypothesis, we analyze the audit log of two months from VUMC. We define a metric using access date and encounter date as (access date - encounter date), to quantify how far apart an access occurs from an encounter. We use the prior access and encounter information stored in the audit log to derive a distribution for our metric. A large hospital like VUMC can have multiple departments. Each department can exhibit a different pattern for patient accesses. Hence, we derive distributions for our metric in context of each department. Figure 4.1. shows the distribution of our metric for four different departments. To further understand the statistical properties of these distributions we harness skew and kurtosis.

Table 4.1. lists the statistics that describe the access distributions presented in Figure 4.1. Skewness is a measure of the asymmetry of a probability distribution of a variable. Zero skew value indicates that the distribution is symmetric about the mean whereas negative skew value indicates that the distribution is left tailed i.e. left tail of the distribution is heavier than the right tail and a positive skew value means that the distribution is right

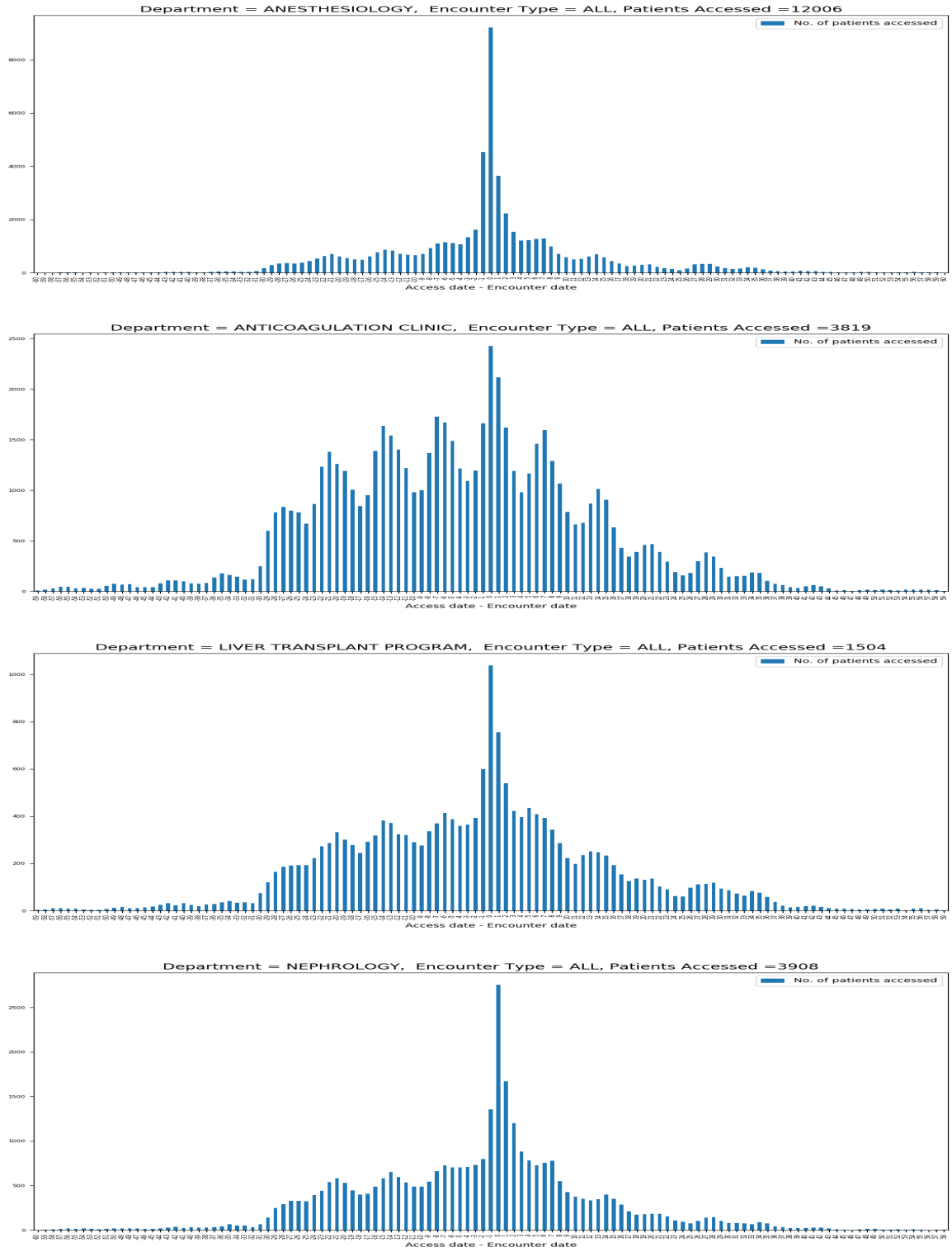


Figure 4.1: Distribution of Metric (Access Date - Encounter Date) for Four Different Departments, Encounter Type = ANY.

tailed [23]. We performed this statistical analysis for all the 1737 departments in our data and we observe that the Skew values range from 0 to 9.89. It should be noted that all the departments had distribution with positive skew value i.e. there were more accesses after a clinical encounter compared to before the encounter for all departments.

Kurtosis is a measure of tailedness / peakedness of a distribution. A high value of kurtosis means the distribution has heavy tails compared to a normal distribution and a low value of kurtosis means that the distribution has light tails compared to a normal distribution. We observe that the kurtosis values for the departments in our dataset range from -3 to 99.02 which suggests that the tailedness varies for all department distributions. Out of the four departments in the table 4.1, the Anticoagulation Clinic has a negative kurtosis value which means it is lightly tailed and has flat peak, and Anesthesiology department has high value of kurtosis which means it is heavily tailed and has a sharp peak.

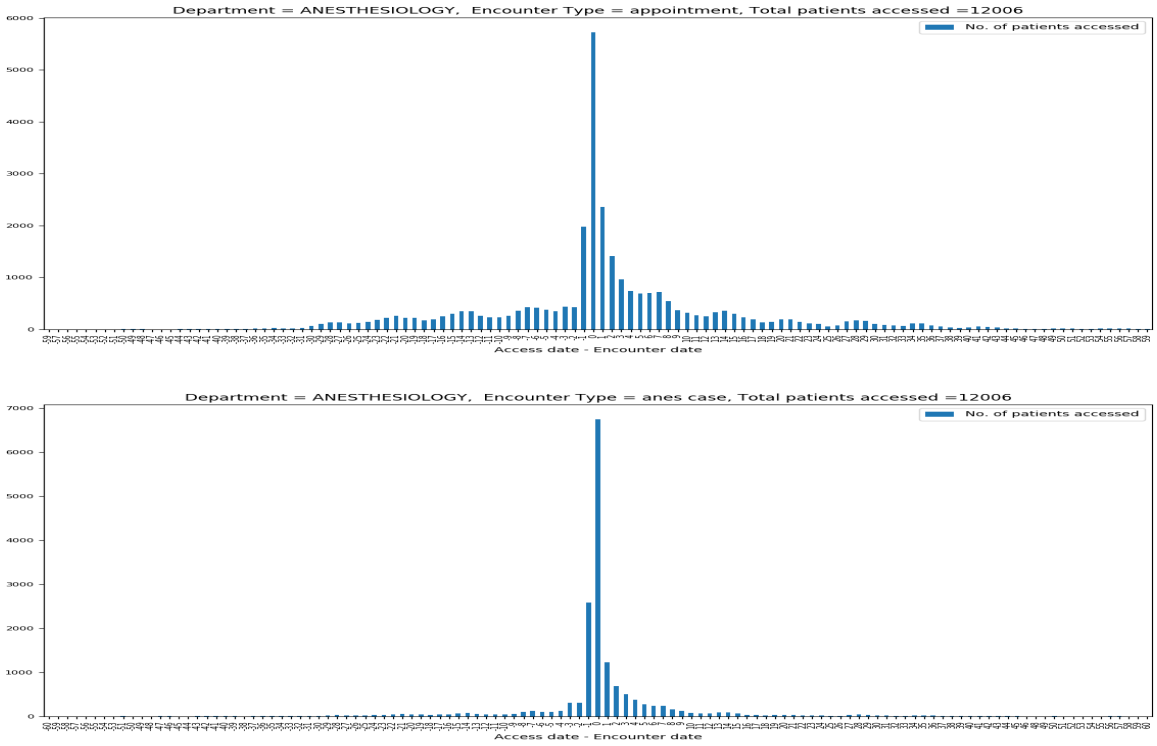
Table 4.1: Statistics of Distribution of Metric (Access Date - Encounter Date) for Four Different Departments, Encounter Type = ANY.

Department	**Minimum**	**Maximum**	**Mean**	**Variance**	**Skew**	**Kurtosis**
Anesthesiology	1	9212	480.03	1057367.45	5.93	43.62
Anticoagulation Clinic	5	2425	526.78	339468.04	0.98	-0.07
Liver Transplant Program	2	1038	153.12	31418.53	1.76	4.62
Nephrology	1	2752	273.5	151441.54	2.99	13.64

We perform similar analysis for all VUMC departments and the results enable us to conclude that access patterns vary significantly across different departments.

A clinical encounter is further classified into different types including appointment, labs etc. To assess whether the distribution of our metric for a department changes for different encounter types, we evaluate the distribution of our metric for each encounter type. Figure 4.2 shows that the distribution for two different encounter types (anesthesiology-case and appointment) of the Anesthesiology department are different. These results inform us that the prediction model needs to be aware of encounter types.

Figure 4.2: Distribution of Metric (Access Date - Encounter Date) for Anesthesiology Department, Encounter Type: 1) Appointment, 2) Anesthesiology Case



4.2 Methodology

4.2.1 Data Overview

We perform experiments on the data from VUMC. Table 4.2. summarizes the statistics of the data used in this study.

Table 4.2: Summary statistics of the data used in this study.

Patients	Employee Departments	Encounter Types	ICD Chapters	Access Dates	Encounter Dates
433254	1737	54	21	2017/10/01-2017/11/30	2017/10/01-2017/11/30

Figure 4.3 depicts the relationships in our dataset. For each access, the Audit log records, ID of the patient whose records are accessed, ID of the employee who made the access and the time of access. The EHR stores information for each clinical encounter of a patient by recording the encounter type and encounter time. EHR system also stores pa-

tient diagnosis in terms of International Classification of Diseases, Ninth Revision (ICD-9) and Tenth Revision (ICD-10) codes [24]. The ICD codes are mapped into 21 ICD chapters depending on the subject of the ICD codes. To reduce the size of the feature matrix for prediction models we use ICD chapters instead of ICD codes. In addition, the VUMC dataset includes the list of departments each employee is affiliated with.

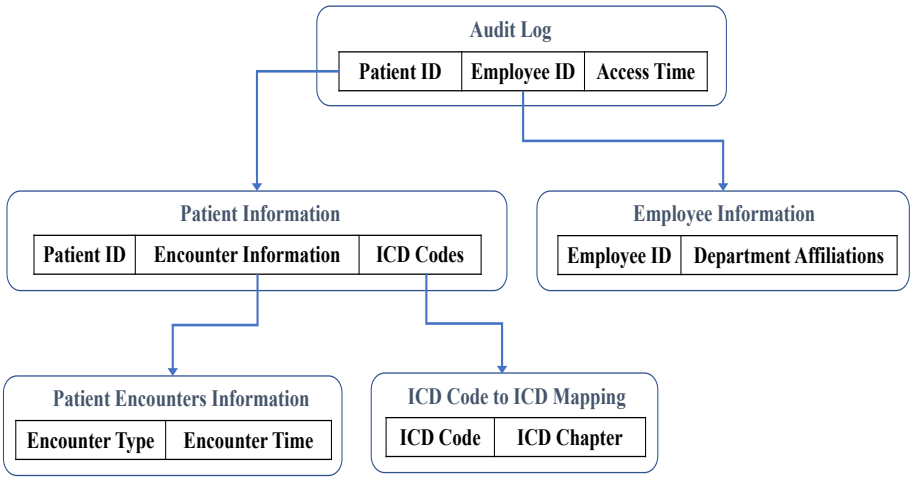


Figure 4.3: Input Data Integration.

4.2.2 Feature Extraction

We constructed features using our metric (access date - encounter date) for each encounter type and diagnosis information (ICD chapters). Table 4.3. depicts patient data example.

Table 4.3: Patient Encounters and ICDs

Patient	Appointment	Labs	ICD Chapters
P1	2017/10/04	NO	1
P3	2017/10/07	NO	NO
P4	2017/10/07	2017/10/06	21

Features: Encounter Information. Our dataset includes 54 types of clinical encounters e.g. appointment, lab order (labs), medication order, anesthesiology case etc. For each

patient in our data, we obtain the encounter-date i.e. when each patient had these encounters. We then obtain the absolute difference between access date and encounter date for each of these encounters of the patient. If the patient had multiple occurrences of same type of encounter we consider only the encounter date that is closest to the access date. For the encounters that did not occur for a patient we default the value to 100.

Features: Diagnosis Information. The EHR system stores ICD codes for each patient. We obtain the ICD chapters for each patient by mapping these ICD-codes to ICD chapters. If a patient has a ICD chapter assigned in EHR, the value is considered as 1 for corresponding chapter in the feature matrix, else it is default to 100.

Feature Matrix. We construct feature matrix using the audit log (sample shown in Table 4.4), the metric data (access date - encounter date) for 54 encounter types and diagnosis (ICD chapter), for 433254 patients over two-month period as shown in Table 4.5.

This feature matrix covers following cases:

- Patients had clinical encounters and their EHR was accessed by hospital employees,
- Patients did not have any clinical encounters but their EHR was accessed by hospital employees and
- Patients had clinical encounters but their EHR was not accessed by any hospital employee.

Table 4.4: Audit Log Sample

Patient	Access Date	Department of Employee in the Access
P1	2017/10/04	D1
P1	2017/10/05	D2
P3	2017/10/07	D3
P4	2017/10/04	D4

Table 4.5: Feature matrix Sample

Patient	Access Date	Department of Employee in the Access	Appointment	Labs	-	ICD Chapter 1	-	ICD Chapter 21
P1	2017/10/04	D1	0	100	-	1	-	100
P1	2017/10/05	D2	1	100	-	100	-	100
P3	2017/10/07	D3	0	100	-	100	-	100
P4	2017/10/04	D4	3	1	-	100	-	1

4.2.3 Classifier construction

We evaluate four classifiers including Random Forest Classifier (RFC), Stochastic Gradient Descent, Gaussian Naive Bayes and Logistic Regression. The results show that RFC performs better than the other models hence, we primarily include results of RFC in this study (see section 4.3.1). Random forest is an ensemble learning approach for building a predictor with a set of decision trees which grow in random subset of data [15]. Random forest fits the set of decision tree classifiers on the various subsets and utilizes averaging to achieve optimal predictive accuracy and controls over-fitting of data.

We build RFC using scikit-learn which is a python machine leaning library [25]. We train and test the classifier on five-fold cross-validation. In a n-fold cross-validation the dataset D is randomly split into n mutually exclusive subsets D1, D2, ..., Dn of same size to ensure every sample in the dataset D has equal probability of appearing in the training and test set [26]. The classification model is trained and tested n times with the training data as D/Dt and tested with Dt, where t ε 1, 2, ..., n.

Model Types. We trained three separate models:

1. Unique Model for Each Department,

2. Unique Model for Each Cluster of Departments and

3. Single Model for All Departments.

The feature matrix remains same for all three models. The models differ in the output matrix, the 'Unique Model for Each Department' predicts the probability of access of a patient's EHR by employee of the given department on a given date, 'Unique Model for

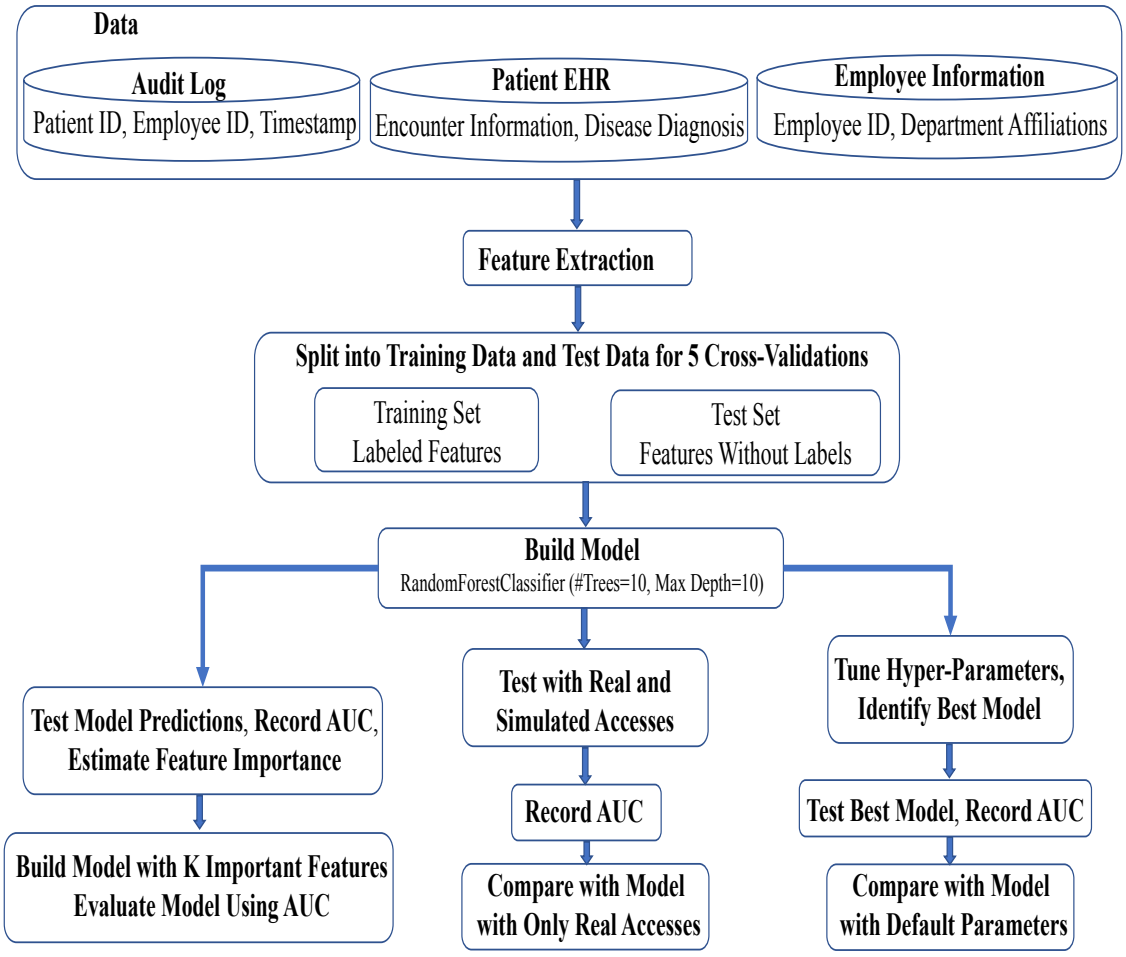


Figure 4.4: Steps to Build and Evaluate Prediction Model

Each Cluster of Departments' predicts the probability of access of a patient's EHR on a given date by employee of any department from given cluster of departments and 'Single Model for All Departments' predicts the probability of access of a patient's EHR on a given date by an employee who can belong to any department. In this study, we used a prediction time window of one day. In future, we would like to experiment with different time window sizes. Table 4.6 depicts the output matrix for these models.

Table 4.6: Model Outcome Matrix

		Unique Model for Each Department	**Unique Model for Each Cluster of Departments**	**Single Model for All Departments**
Patient	Access Date	Access by Department 'Dn'	Access by Any Department in Cluster 'Cn'	Access by Any Department
P1	2017/10/04	1	0	1
P1	2017/10/05	0	0	0
P3	2017/10/07	1	0	1
P4	2017/10/04	0	1	1

Department Clustering. We use the Louvain method for community detection in large networks to obtain department clusters [27]. Louvain method is a heuristic method based on greedy network modularity optimization. Modularity of a network is measure of strength of division of a network into modules/clusters/communities [28]. A network with dense connections between the nodes within modules but sparse connections between nodes in different modules has high modularity value. The data used for detection of department communities is the list of patients accessed by each department.

4.2.4 Model Evaluation and Optimization

Area Under ROC Curve (AUC ROC). We use the average area under the receiver operating characteristics curve (AUC ROC) of all cross-validations to evaluate the performance of the classifier [16].

Feature Importance Evaluation, Build Model with Important Features. Importance of a feature in the input feature matrix with respect to the predictability of the target variable is computed using the relative depth of the feature used as decision node in a tree. Features that are used at the top of a decision tree contribute to the decision of a larger fraction of input sample. The relative importance of each feature is measured using the expected fraction of input samples they contribute to [25]. We estimate the importance of each feature in the feature matrix of the predictor. We rank the features according to their importance level then, we identify the importance level of 10th feature from the top and select this importance value as a threshold importance level. We build a new model with only the features having importance value equal or above this threshold importance level. We compute the AUC ROC of the new predictor with these k important features, with five-fold cross-validation. We then compare the AUC ROC of predictor with all features and predictor with only k important features.

Model with Simulated Accesses. To determine if an access is an appropriate access or an inappropriate access, needs manual investigation by privacy and administrative officers.

Our dataset does not include this information about the accesses in the audit log and we do not know if our dataset includes any true inappropriate accesses. Hence, to evaluate the performance of our predictive models on audit log containing inappropriate accesses, we mix real accesses with simulated accesses and test our models. We generate the simulated accesses by randomly pairing patient and department together. We assume that the randomly generated accesses would not follow the access patterns expected by various hospital departments. We train and test the predictor with mix of real and simulated accesses and compare the performance of this model with performance of model built with only real data.

Model Optimization. We perform hyper-parameter tuning on RFC using Hyperopt-sklearn [29]. We obtain the best model and compute the AUC ROC using best model.

Figure 4.4 depicts the steps in our methodology.

4.3 Results and Discussion

4.3.1 AUC ROC

Table. 4.7 shows the AUC ROC values using different Machine Learning Algorithms including RFC, Logistic Regression, Stochastic Gradient Descent and Gaussian Naive Bayes. Results of this test show that, RFC performs better than all tested Machine Learning Algorithms hence, we select RFC to build our prediction framework.

Table 4.7: AUC ROC Using RFC VS Other Machine Learning Algorithms

	Single Model for All Departments	**Unique Model for Each Department**	
		INTERNAL MEDICINE	ANESTHESIOLOGY
Random Forest Classifier	0.81	0.79	0.87
Logistic Regression	0.52	0.67	0.83
Stochastic Gradient Descent	0.51	0.66	0.81
Gaussian Naive Bayes	0.51	0.64	0.75

Table 4.8. shows the average AUC ROC for the three different model types. Results indicate that there is no clear winner. We observe that 'Unique Model for Each Department' has a wide range for AUC ROC. Further investigation shows that, in general the

departments that have very few accesses, have low model AUC ROC. 'Unique Model for Each Cluster of Departments' provides the AUC ROC for an entire cluster of departments. However, it is not clear what is the contribution of each department to overall AUC ROC score. Similar concerns can be raised for 'Single Model for All Departments' which is a specific case where all departments belong to a single cluster.

Table 4.8: AUC ROC for Each Type of Model

Model Type	**AUC ROC**	**AVERAGE AUC ROC**
Single Model for All Departments	0.78	0.78
Unique Model for Each Department	0.6 - 0.9 (No. of departments = 1737)	0.84
Unique Model for Each Cluster of Departments	0.77 - 0.87 (No. of clusters = 15)	0.8

4.3.2 Feature importance and AUC ROC with important features

In section 4.2.4, we describe our method to identify important features and reduce the feature set size. Table 4.9 shows the top 5 important features of the example department and cluster models. From the table, we see that the important features for each department specific model are different. In addition, the important features (Encounter Type and Disease codes) of the predictor are directly related to the department. Similar trend is observed in cluster specific models. This confirms our hypothesis that the access pattern of a department is determined by patient's specific encounter types and diagnosis.

Table 4.9: Top Five Important Features Detected for Few Models

Model Type		**Top Five Important Features**
Unique Model for Each Department	ANESTHESIOLOGY	Anesthesiology Case, ADT, Labs, Documents History And Physical, Appointment
	INTERNAL MEDICINE	Labs, Appointment, ADT, Documents Clinical Communication, Documents Medication Administration
Unique Model for Each Cluster of Departments	UROLOGY Cluster	Appointment, ICD Chapter 14: 'Diseases of the Genitourinary System', ICD Chapter 11: 'Diseases of the Digestive System, Complications Of Pregnancy, Childbirth, And The Puerperium', Anesthesiology Case, Labs
	OPHTHALMOLOGY, EYE CLINIC Cluster	ICD Chapter 7: 'Diseases of the eye and adnexa, Diseases Of The Circulatory System', Appointment, ICD Chapter 8: 'Diseases of the ear and mastoid process, Diseases Of The Respiratory System', Labs, Anesthesiology Case
	ENT, AUDIOLOGY, ALLERGY Cluster	Appointment, Labs, ICD Chapter 10: 'Diseases of the respiratory system, Diseases Of The Genitourinary,System', Anesthesiology Case, ADT

We compute the AUC ROC of the new model with only important features with five-

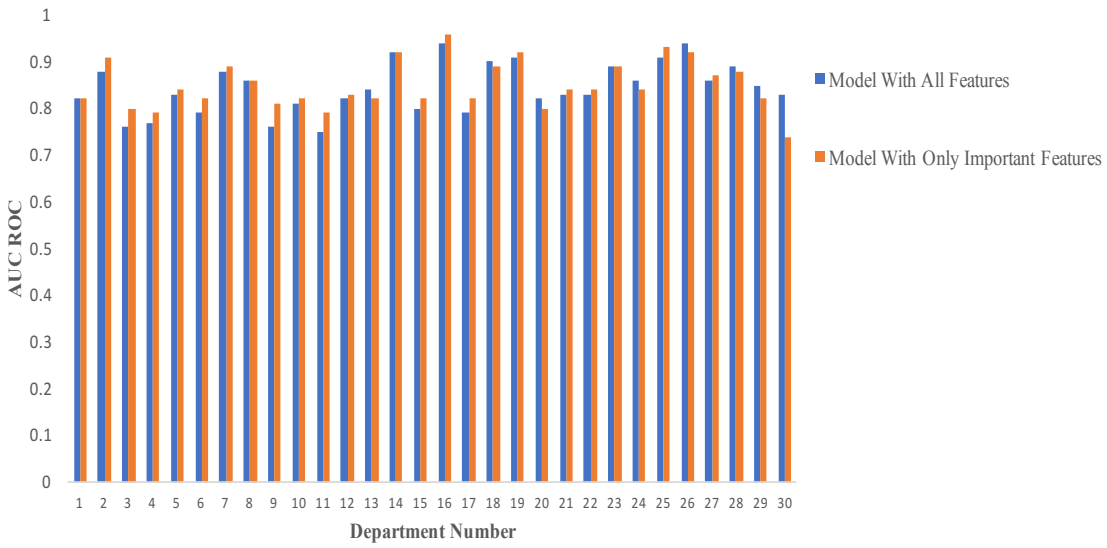


Figure 4.5: AUC with All Features VS AUC with K Important Features

fold cross-validation. We observe that AUC ROC values of model built with only important features exhibit minimal difference as compared to the AUC ROC values of model built with all features as shown in Figure 4.5. We believe, in case of very large dataset model with only important features can be used to reduce feature vector size while achieving similar model accuracy as model with all features.

4.3.3 Real and Simulated Accesses

If we introduce x number of simulated accesses in each department model then the performance of the model depends on the number of real accesses by that department as seen in the table 4.10.

Table 4.10: AUC with Real and Simulated Accesses

Department	Number of Real Accesses by Department	AUC with Real Accesses Only	AUC with Real + Simulated Accesses
INTERNAL MEDICINE	86816	0.67	0.67
ANESTHESIOLOGY	27573	0.78	0.76
DENTAL-VAV CLINIC	259	0.77	0.74

4.3.4 Optimized Model Performance

We obtain the best parameters for the RFC using hyper-parameter tuning method. We compute the AUC ROC using the RFC with optimized hyper-parameters. Table 4.11. shows the results of the model optimization experiment. The results suggest that, in general hyper-parameter tuning improves the performance of the prediction models.

Table 4.11: AUC ROC with Optimized Parameters VS AUC ROC with Default Parameters

Model	**Department**	**No. of Accesses**	**With Optimized RFC**	**Without Optimization** (RFC (n_estimators=10, max_depth=10))
Unique Model for Each Department	INTERNAL MEDICINE	86816	0.81	0.67
	ANESTHESIOLOGY	27573	0.85	0.78
	NEPHROLOGY	17813	0.84	0.84
Single Model for All Departments	ALL	2108607	0.78	0.77

4.4 Conclusions

In this chapter, we propose a supervised machine learning framework to detect suspicious accesses to sensitive patient health information. This framework uses patient clinical encounter information and diagnosis information to predict which hospital department employee will access a patient's EHR and when this access will occur. We empirically evaluate our prediction models on two months of audit logs from VUMC EHR system. The average AUC over different hospital departments in our dataset is 0.84. Results of our experiments indicate; this automated identification of suspicious accesses can be utilized to significantly reduce the manual effort in EHR auditing.

Chapter 5

Conclusions and Future Work

Hospitals are facing steep challenges to protect the privacy of patient data in EHR from insider threats. To achieve fast detection of insider misuse and reduce further harm, large hospitals need automated suspicious access detection mechanisms. This work presents fundamental results towards designing efficient automated suspicions access detection systems.

HCOs primarily use rule-based auditing to identify suspicious insider behavior. However, rule-based methods have several limitations. First, rule-based auditing systems have not been evaluated empirically. Second, rule-based auditing systems rely on predefined rules and are oblivious to the statistical properties of the EHR data. To this end, we propose a principled approach to evaluate the effectiveness of rule-based auditing methods in identifying suspicious behavior. Then, we propose an auditing method based on supervised machine learning techniques which utilizes clinical context of the accesses in the EHR data to identify suspicious behavior. Our detailed contributions in this work are as follows:

- We examined the rate of high-risk access patterns and minimum rate of high-risk accesses that can be explained with appropriate clinical reasons in a large EHR system. An analysis of 8M accesses from one-week of data from the VUMC shows that specific high-risk flags occur more frequently than theoretically expected and the rate at which accesses can be explained away with five simple clinical reasons is 16 - 43%.

- We build a machine learning model to predict the probability of access of a patient's EHR by the specified department on the specified date based on the clinical encounter and diagnosis information. To empirically evaluate our prediction models, we perform an analysis with two months of audit logs from VUMC EHR system. The average AUC over different hospital departments in our dataset is 0.84. Results of our

analysis indicate; this automated identification of suspicious accesses can be utilized to significantly reduce the manual effort in EHR auditing.

There are several limitations in our study which provide directions for future investigations. First, this study does not test if a flagged suspicious access is in fact an inappropriate access. A flagged suspicious access needs to be investigated manually by a privacy officer to determine if it is a true inappropriate access. However, this manual investigation is beyond the scope of this study. Second, this investigation focused on data from only a limited time period (one week and two months) from a single medical center. As such, it will be necessary to validate these findings with data from a broader time period and over other healthcare organizations.

BIBLIOGRAPHY

[1] Menachemi N, Brooks RG. Reviewing the benefits and costs of electronic health records and associated patient safety technologies. J Med Syst. 2006; 30(3): 159-68.

[2] U.S. Department of Health and Human Services. HITECH Act Enforcement Interim Final Rule. 2009 https://www.hhs.gov/hipaa/for-professionals/special-topics/HITECH-act-enforcement-interim-final-rule/. Accessed February 26, 2017.

[3] Blumenthal D, Tavenner M. The meaningful use regulation for electronic health records. N Engl J Med. 2010;363(6):501-504.

[4] Blobel B. Authorisation and access control for electronic health systems. Int J Med Inform. 2004; 73: 251-7.

[5] Røstad L, Edsberg O. A study of access control requirements for healthcare systems based on audit trails from access logs. Proc Annual Computer Security Applications Conference 2006: 175-86.

[6] Ferreira A, Cruz-Correia R, Antunes L, Chadwick DW. Access control: How can it improve patients' healthcare? Stud Health Technol Inform. 2007; 127: 65-76.

[7] U.S. Department of Health and Human Services. Modifications to the HIPAA Privacy, Security, Enforcement, and Breach Notification Rules under the Health Information Technology for Economic and Clinical Health Act and the Genetic Information Nondiscrimination Act. March 26, 2013. https://www.federalregister.gov/documents/2013/01/25/2013-01073/modifications-to-the-hipaa-privacy-security-enforcement-and-breach-notification-rules-under-the

[8] Malin B, Nyemba S, Paulett P. Learning relational policies from electronic health record access logs. J Biomed Inform. 2011; 44(2): 333-42.

[9] Boxwala AA, Kim J, Grillo JM, Ohno-Machado L. Using statistical and machine learning to help institutions detect suspicious access to electronic health records. J Am Med Inform Assoc. 2011; 18: 498-505.

[10] Fabbri D, LeFevre K. Explanation-based auditing. Proceedings of the Very Large Data Bases Endowment. 2011; 5: 1-12.

[11] Asaro PV, Herting RL Jr, Roth AC, Barnes MR. Effective audit trails−−a taxonomy for determination of information requirements. Proc AMIA Symp. 1999:663-5.

[12] Herting RL, Jr, Asaro PV, Roth AC, et al. Using external data sources to improve audit trail analysis. Proc AMIA Symp. 1999:795-9.

[13] Asaro PV, Ries JE. Data mining in medical record access logs. Proc AMIA Symp. 2001:855.

[14] S. B. Kotsiantis. Supervised Machine Learning: A Review of Classification Techniques. In Proceedings of the 2007 conference on Emerging Artificial Intelligence Applications in Computer Engineering: Real Word AI Systems with Applications in eHealth, HCI, Information Retrieval and Pervasive Technologies. 2007. 978-1-58603-780-2. 3-24.IOS Press.

[15] Leo Breiman. Random Forests. Mach. Learn. 45, 1 (October 2001), 5-32. DOI: https://doi.org/10.1023/A:1010933404324.

[16] Tom Fawcett. An introduction to ROC analysis. Pattern Recogn. Lett. 27, 8 (June 2006), 861-874. DOI=http://dx.doi.org/10.1016/j.patrec.2005.10.010

[17] Li X, Xue Y, Malin B. Detecting anomalous user behaviors in workflow-driven web applications. Proc 31st IEEE Symposium on Reliable Distributed Systems. 2012: 1-10.

[18] Zhang H, Mehotra S, Liebovitz D, Gunter CA, Malin B. Mining deviations from

patient care pathways via electronic medical record system audits. ACM Transactions on Management Information Systems. 2013; 4(4): 17.

[19] Fabbri D, LeFevre K. Explaining accesses to electronic medical records using diagnosis information. J Am Med Inform Assoc. 2012; 20(1): 52-60.

[20] U.S. Department of Health & Human Services. 45 CFR 164.501. Uses and Disclosures for Treatment, Payment, and Health Care Operations. https://www.hhs.gov/hipaa/for-professionals/privacy/guidance/disclosures-treatment-payment-health-care-operations/index.html. Accessed July 6, 2017.

[21] Laszka A, Vorobeychik Y, Fabbri D, Yan C, Malin B. A game-theoretic approach for alert prioritization. Proc AAAI Workshop on Artificial Intelligence for Cyber Security. 2017.

[22] Health Information Privacy, Submitting Notice of a Breach to the Secretary, https://www.hhs.gov/hipaa/for-professionals/breach-notification/breach-reporting/index.html, Last Accessed Jan 28, 2018.

[23] Groeneveld, Richard A., and Glen Meeden. "Measuring Skewness and Kurtosis." Journal of the Royal Statistical Society. Series D (The Statistician) 33, no. 4 (1984): 391-99. doi:10.2307/2987742.

[24] CDC/National Center for Health Statistics. International Classification of Diseases, Tenth Revision, Clinical Modification (ICD-10-CM). https://www.cdc.gov/nchs/icd/index.htm. Accessed Jan 28, 2018.

[25] Pedregosa et al. Scikit-learn: Machine Learning in Python. JMLR 12, pp. 2825-2830, 2011.

[26] Ron Kohavi. A study of cross-validation and bootstrap for accuracy estimation and model selection. In Proceedings of the 14th international joint conference on Artificial

intelligence - Volume 2 (IJCAI'95). Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 1137-1143.

[27] Vincent D Blondel and Jean-Loup Guillaume and Renaud Lambiotte and Etienne Lefebvre. Fast unfolding of communities in large networks. Journal of Statistical Mechanics: Theory and Experiment. 10. P10008. http://stacks.iop.org/1742-5468/2008/i=10/a=P10008. 2008.

[28] M. E. J. Newman. Modularity and community structure in networks. Proceedings of the National Academy of Sciences Jun 2006, 103 (23) 8577-8582; DOI: 10.1073/pnas.0601602103.

[29] Brent Komer, James Bergstra, Chris Eliasmith. Hyperopt-Sklearn: Automatic Hyperparameter Configuration for Scikit-Learn. PROC. OF THE 13th PYTHON IN SCIENCE CONF. (SCIPY 2014).