Using External Data Sources to Improve Audit Trail Analysis

Robert L. Herting Jr, MD, Phillip V. Asaro, MD, Allan C. Roth, PhD, Mike R. Barnes, MD
Department of Health Management and Informatics,
University of Missouri-Columbia School of Medicine, Columbia, MO

Audit trail analysis is the primary means of detection of inappropriate use of the medical record. While audit logs contain large amounts of information, the information required to determine useful user-patient relationships is often not present. Adequate information isn't present because most audit trail analysis systems rely on the limited information available within the medical record system. We report a feature of the STAR (System for Text Archive and Retrieval) audit analysis system where information available in the medical record is augmented with external information sources such as: database sources, Lightweight Directory Access Protocol (LDAP) server sources, and World Wide Web (WWW) database sources. We discuss several issues that arise when combining the information from each of these disparate information sources. Furthermore, we explain how the enhanced person specific information obtained can be used to determine user-patient relationships that might signify a motive for inappropriately accessing a patient's medical record.

INTRODUCTION

The main tool for detection of inappropriate use of the medical record is audit trail analysis; moreover, user awareness of a complete and frequent audit trail analysis will discourage inappropriate use of the electronic medical record[1]. Recognizing this fact, the proposed Health Insurance Portability and Accountability Act (HIPAA) mandates regular audit trail analysis[2]. However, except for some early work on use-pattern deviation[3], audit trail analysis is usually performed exclusively by manual review of audit logs[4]. Thus, there is a perceived need for more effective and automated tools to maintain continuous surveillance of audit trail information in health care[3, 5-6].

In consideration of automating audit trail analysis, it is important to first try to specify as many as possible of the "indicators", or potential evidence, that a confidentiality breach has occurred. By considering scenarios of confidentiality breaches (imagined situations in which a breach might occur), indicators of breach can be developed. Such indicators might relate to: (1) user behavioral deviations, such as unexpected log-on characteristics (unexpected site of log-on, unexpected time of day, etc.); (2) characteristics of the patient or patient record, such as the presence of sensitive diagnoses, the presence of sensitive test results, or a patient who is a high profile public official; (3) a relationship between the user and the patient that would raise the level of suspicion of inappropriate record access (such as fellow employee, spouse, ex-spouse in a legal child custody dispute, etc.); and (4) the established provider role of the user in a user-patient relationship, such as the patient's primary care provider (Note: here the fourth indicator would be used as a "negative" indicator of confidentiality breach, offsetting the other three "positive" indicators just mentioned).

The problem is that although there may be considerable information in most medical system audit logs, they typically do not contain adequate information to support the type of analysis indicated above; or more specifically they don't contain the information needed to determine useful user-patient relationships that might signify a motive for inappropriately accessing a patient's medical record (see point number three above). This information is not present because most audit trail analysis systems rely on the limited information available within the medical record system itself[7, 8]. By expanding consideration beyond the medical record system, useful information may be found within the institution (such as employee information on an LDAP server), or even outside of the institution (for example, publicly available web databases). Information obtained from these sources, such as the fact that a patient is an employee at the same institution as the user, could reveal a potential motive raising the suspicion of an inappropriate access. Thus, if the user was not known to be actively involved in caring for the patient, such an inappropriate access could be flagged as highly suspicious (i.e. warranting manual review and inquiry) because an illegitimate motive is suggested.

OVERVIEW

We report on a feature of the audit analysis system of our medical results retrieval system, STAR. This feature of the audit analysis system augments the information already available in the medical record using software agents that retrieve data from external information sources. These information retrieval agents fit into three categories (See Figure 1 on the next page):

1091—8280/99/$5.00

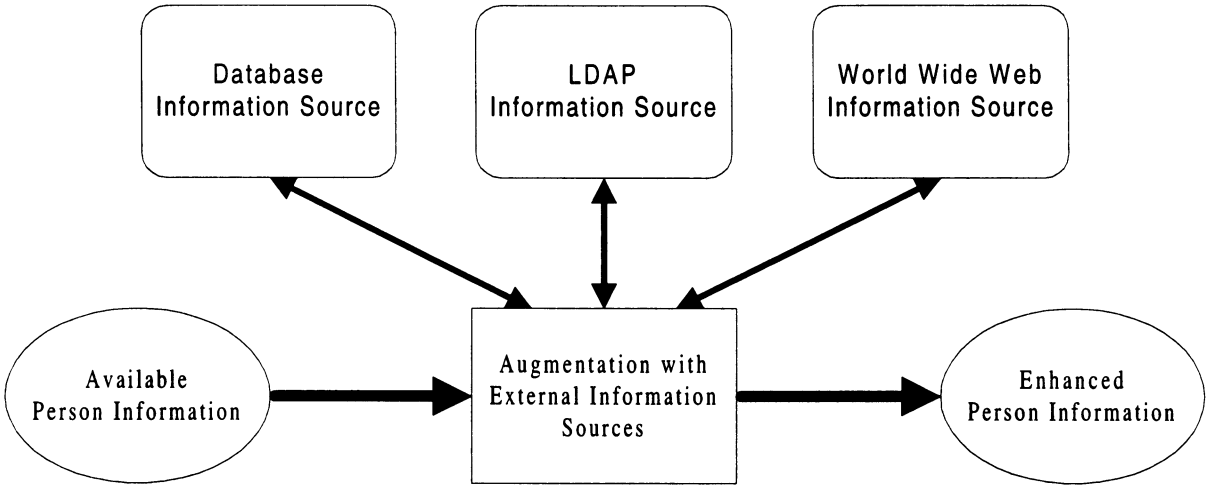

Figure 1 - Augmenting Available Person Information

(1) Database agents - These agents extract information from Structured Query Language (SQL) compliant databases. Accessing a database local to the institution, such as insurance claim information, is an example of this type of agent.

(2) LDAP agents - These agents query LDAP servers on a local network or on the Internet for additional demographic information and other specific information, such as answering the question: Is the patient an employee within the same department and/or institution as the user?

(3) WWW agents - These agents search web databases for additional person specific information. For example, an agent could look up web pages that map an address to latitude and longitude and tell the absolute distance between two addresses. Such information would be useful in establishing a neighbor relationship between a user and a patient based on the absolute distance between their home addresses. If a neighbor relationship were established, it could reveal a motive (a user looking up a neighbor) signifying that an inappropriate access might have occurred. As another example, an agent could query the American Medical Association (AMA) web site to find out whether a person went to medical school; and if so, it could retrieve the name of the medical school, the year of graduation, and the residency program attended. Such information could reveal a motive as to why an inappropriate access might have occurred (eg. a physician looking up a medical school colleague). Lastly, of special note are large legal databases available on the web that provide marriage license information, birth certificate information, lawsuit information, etc. Such information could also be helpful in identifying relationships as motivational indicators of inappropriate access.

Several points of discussion follow from combining the information from each of these disparate information sources: (1) there is a need for a standardized nomenclature; (2) each datum of personal information retrieved from a data source should have a reliability score attached to it; (3) information sources may need to be searched again as data they need to conduct their query becomes available; (4) there are sensitive data security issues surrounding the transference of data to and from the information source; and (5) there may be advantages to mirroring some information sources into a local database. Each of these topics will be discussed separately in the sections that follow.

The Need for a Standardized Nomenclature

The disparate information sources refer to the same data items by different terms necessitating the use of a standardized nomenclature to allow reuse or sharing of the retrieved data. Such a standardized nomenclature provides each agent with a common term to use when referring to any piece of data it needs for its search and retrieval function, or its storage function. For example, most LDAP services use the attribute "sn" to represent a person's surname (or last name), but STAR identifies this field by the term "LAST_NAME." In order for the STAR audit trail analysis system to integrate this data item, the labels "sn" and "LAST_NAME" need to be mapped to a standard nomenclature term, such as "lastName." In addition, this standardized nomenclature can be used by other programs that analyze the data obtained from these disparate sources (such as a rule-based system to determine if an inappropriate access occurred).

Factors that Determine Reliability Scores for Data Obtained from Information Sources	Lower Reliability Score	Higher Reliability Score
Information Source Reliability		
How frequently the information source is updated	Infrequently (eg. once a year)	Frequently (eg. once a month)
The quality control of the data entry process	Frequent data errors / Missing data	Few data errors / Complete data
Strength of the Match Between the External Information Source and the Internal Medical Record System Query Data		
The number of data items that match between the external information source data and the internal medical record system query data	Some items match, eg. Only last name AND first name match	More items match, eg. Last name AND first name AND middle name match
The degree of ambiguity of the data that matches	More ambiguous data matches, eg. Last name AND first name match	Less ambiguous data matches, eg. Last name AND social security number match*
* In our example, it is implied that the social security number in combination with the last name narrows the possible matching people down to one person, whereas a first name in combination with last name may match multiple people.		

Figure 2 - Factors that Determine Reliability Scores

It should be noted that in addition to mapping labels, the data might need to be morphed into a specific format. For example, any data labeled by the 'lastName' term must be converted to only upper case letters.

Reliability Scores

Because we are using outside data sources, the degree of reliability of those sources is not always certain. Thus, each piece of data that is retrieved from a data source should have a reliability score attached to it that is based on two factors: (1) the external data source's reliability, which is a function of how frequently the data is updated, and the quality control of the data input into the database; and (2) the strength of the match between the outside information source's data and the internal medical record system data (See Figure 2 on the next page). Hypothetically, if the American Academy of Family Physicians (AAFP) Web Site only updates its data once a year, whereas the AMA web site updates its data monthly, then the reliability score of data from the AAFP might be lower than that from the AMA. Similarly, if the AAFP site was known to have frequent data errors or missing data compared to the AMA site, then the reliability score of data from the AAFP would be lower. Additionally, the more matches between the query search data sent and the data that an information source retrieved, the higher the reliability score will be for the retrieved data. For example, if a search matched on first name, middle name and last name, then the data retrieved would have a higher reliability score than if a search matched on just the first and last name. Similarly, if the query search data and information source data matched on the social security number and last name, that would be a better match than just matching on first name and last name, so the former would have a higher reliability score.

Rationale to Search an Information Source Again

Data obtained from one information source may be useful as input data to a different, previously searched information source, since the latter may need the new information to retrieve results or make the data it previously retrieved more reliable. This implies two important points. First, if agents access

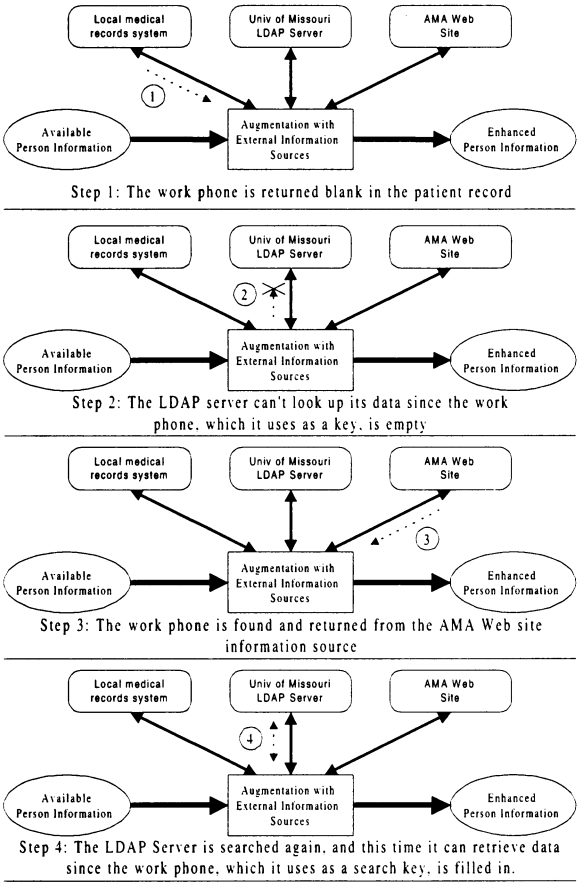

Figure 3 - Rationale to Search an Information Source Again

Person Information Record	
Information Available within the Medical Record System/Current Audit Log	
Information Source	*Personal Attributes Present*
STAR Patient Table	Patient ID, last name, first name, middle name, name prefix, name suffix, professional title, social security number, date of birth, gender, ethnic group, religion, marital status, address, city, state, county, zip code, country, home phone, and work phone
STAR User Table	User ID, password, provider ID, last name, first name, middle name, and work phone
Information Available from External Information Sources for Augmenting the Sources Above	
Information Source	*Personal Attributes Present*
AMA Physician Select Web Site	Last name, first name, middle initial, whether a member of the AMA, city, state, zip code, gender, medical school name, medical school city, medical school state, medical school zip code, year of graduation from medical school, residency training program name, primary practice specialty, and work phone
University of Missouri LDAP Server	Last name, first name, middle initial, work address, work city, work state, work zip code, work country, professional title, company, department, work phone, work fax, home phone, mobile phone, pager, and email address

Figure 4 – External Information Source Augmentation of Existing Person Specific Information

information sources and are unable to retrieve results or retrieve results with a low reliability score, then they should note the fields needed to improve their search so if this data is found in a different information source, the first source can be searched again with the new data. For example (see Figure 3), if the work phone of the patient record were blank, then an LDAP server information source, which uses the work phone as a key to look up its data, would be unable to do its search. However, if the AMA Web site finds that this patient is a physician and returns a work phone, then the LDAP information source should be searched again since the work phone is now present.

The second important point is that external information sources should be searched in an order so as to minimize the number of information sources that need to be queried again. Thus, information sources that provide the most data should be searched first. Moreover, if an information source depends on other data as input to its query, then whenever possible it should be searched after that data has been obtained.

Data Security Issues of Information Sources

A serious confidentiality breach could occur if the information sources logged the query data sent to their servers and subsequently built a list of names that could be interpreted as patients from a particular institution. In addition, since transmission to many of these sources is not encrypted, the query data sent can be intercepted. Thus, if audit trail analysis involves augmenting the person information as we have suggested, then patients may need to sign disclosures stating that they would like this augmenting service to take place. Patients could refuse this service with the knowledge that doing so may prevent effective detection of illegal accesses to their patient information. Ultimately, applicable laws and institutional policies will decide many of the security issues involved in retrieving information from external data sources.

Mirroring Heavily Used Information Sources

Gathering information from some external information sources may take a lengthy amount of time. Therefore, it may be prudent to mirror the information source into a local database. This has two major advantages: (1) queries are typically much faster; and (2) some of the privacy concerns mentioned above may be lessened since the entire data source is mirrored (ie. the server could no longer track the individual patient lookups, and query information containing patient identifying information would not be transmitted over non-encrypted network lines).

IMPLEMENTATION

Our experiences in developing an improved audit log for our medical results retrieval database, STAR (System for Text Archive and Retrieval), suggest the feasibility of using external sources to augment the existing audit log information from STAR. While we have not developed a fully functional automated audit trail system, we have begun developing STAR audit modules that access a variety of external sources.

The first module uses the Java Database Connectivity (JDBC) routines to retrieve patient and user attributes from tables within STAR (as shown in figure 4). This information provides a demographic base from which to search external data sources.

The second module uses the Java Naming and Directory Interface (JNDI) to connect through our local intranet to the publicly available University of Missouri LDAP server. The module runs patient work phone numbers from the STAR database through our JNDI-LDAP interface. The resulting matches provide additional personal attributes that enhance those established by the STAR patient and

user tables, namely whether the patient is an employee of the University (See Figure 4).

The third module uses Java's networking package to access World Wide Web information sources, for example the AMA Physician Select web site (http://www.ama-assn.org/aps/amahg.htm). Physician background details obtained from the AMA web site could be used to augment the personal attributes established by the STAR patient and/or user tables (See Figure 4).

DISCUSSION

The primary use for the proposed additional person information is to increase the detection of possible improper accesses by searching for motives as to why a user might have looked at a given patient's record. Motives may include looking up a family member, a neighbor, a fellow employee, a colleague from medical school or residency, etc. Such motives could not be elucidated from currently available data stored in medical record systems, but these motives are more likely to be found through the use of enhanced personal information obtained using external information sources.

OPPORTUNITIES FOR RESEARCH

It should be noted that our development efforts using external data sources for STAR audit trail analysis suggest several future research opportunities. One opportunity for research is to examine the user-patient relationship motives in combination with other "positive" indicators of confidentiality breach (such as evidence that the user was looking at sensitive patient data like HIV results). These additional "positive" indicators of breach would raise the suspicion of confidentiality breach further. In addition, these "positive" indicators of breach could also be considered in light of "negative" indicators. For example, if a user is the primary care provider, is this "negative" indicator enough to negate the suspicion of breach? Any one of these indicators used alone would likely raise excessive "false positives." However, when used together, such indicators could be very useful in an automated audit trail analysis system.

CONCLUSION

Current audit trail analysis is limited by the information available in the audit log, which is usually a subset of the information available in the medical record system. We report on audit trail modules that use external sources of information to augment the personal information available in the STAR medical record system. These modules use three types of external data sources to augment the information available in STAR. Future research will determine if this augmentation could facilitate access violation detections that were not possible before. Ironically, obtaining information from outside sources may in and of itself involve an improper disclosure of patient information. Therefore, careful attention as to how this information is gathered and used is paramount and must be weighed against the potential benefits of detecting inappropriate access violations. In some cases, the use of external data may be the only method available to obtain information essential to audit trail analysis. We believe that the careful construction of such systems will improve the detection of improper access of patient information that is not possible with existing systems.

Acknowledgements

Drs. Herting, Asaro, and Roth are supported by NLM Training Grant LM07089-07. Dr. Barnes is supported in part by MIAMS Grant LM05415.

References

1. Bowen JW, Klimczak JC, Ruiz M, Barnes M. Design of access control methods for protecting the confidentiality of patient information in networked systems. Proc AMIA Annu Fall Symp 1997:46-50.
2. Carrington C. Keeping data safe: New HIPAA regs hit hard. Telehealth 1998;4(6):30-33.
3. White GB, Fisch EA, Pooch UW. Auditing and intrusion detection. In: Computer system and network security. Boca Raton: CRC Press; 1996. p. 91-115.
4. Council NR. For the Record: Protecting Electronic Health Information. Washington, DC: National Academy Press; 1997.
5. Bauer DS, Eichelman FR, II, Herrera RM, Irgon AE. Intrusion detection: An application of expert systems to computer security. Proc 1989 Int Carnahan Conf Secure 1989:97-100.
6. Hayam A. Security Audit Center--a suggested model for effective audit strategies in health care informatics. Int J Biomed Comput 1994;35 (Suppl):115-27.
7. Murphy GF. Audit logs--a security tool for CPRs. Journal of Ahima 1996;67(6):40-3.
8. Safran C, Rind D, Citroen M, Bakker AR, Slack WV, Bleich HL. Protection of confidentiality in the computer-based patient record [see comments]. MD Comput 1995;12(3):187-92.