Redactable and Auditable Data Access for Bioinformatics Research

Jordan Brown[*]*; Mustaque Ahamad*[†], PhD*; Musheer Ahmed*[†]*; Douglas M. Blough*[*], PhD*; Tahsin Kurc*[††], PhD*; Andrew Post*[††], MD, PhD; and *Joel Saltz*[††], MD, PhD

[*] School of Electrical and Computer Engineering, Georgia Institute of Technology, Atlanta, GA
[†] School of Computer Science, Georgia Institute of Technology, Atlanta, GA
[††] Department of Biomedical Informatics, Emory University, Atlanta, GA

Abstract

Presently, the process of extraction and dissemination of data subsets for research from clinical data warehouses is cumbersome and error prone. Furthermore, large-scale research projects often involve multiple users of the same data extract; each of these users may be authorized to access different data elements and specific subsets of the data extract. Once initial data extraction has been done for a research project, capability to transform the data for individual users and track which data are being accessed by which users in a secure environment is lacking in existing systems. This paper describes several methods that the authors are integrating into a system designed to provide secure, flexible, and auditable support for supplying users with data subsets. The methods implement secure, redactable, and auditable mechanisms for data extraction and dissemination. This paper describes the architecture along with an initial proof-of-concept implementation. Preliminary performance measurements show that the approach manages clinical data in redactable and auditable form with reasonable overheads.

Introduction

As the adoption rate of electronic health records grows, enormous amounts of health information are being stored in these systems. It is also becoming increasingly common for these data to be used for clinical research. Due to the data's sensitive nature, security and privacy are of the utmost importance. In particular, there is a tremendous need for mechanisms to determine and control which information can be released to which research projects/users according to data access policies, and to monitor data use. Even for de-identified datasets, there are regulations regarding what data subsets can be accessed, when the data subsets can be accessed, and who can access specific data subsets. It is generally not feasible, or possible from a regulatory perspective, for a researcher to directly access a clinical data warehouse or access an entirety of a dataset even if it is de-identified.

Presently, the process of data subset extraction and dissemination is cumbersome and error prone. Data managers often receive high-level and verbal descriptions of the data subset to be extracted. They then review authorizations for data access and convert the descriptions to database queries while complying with stated restrictions. Another complication is that a large-scale research project often involves multiple users of the same data extract; each of these users may be authorized to access different data elements and specific subsets of the data extract. Simply sharing the entire data extract with all members of the research team is not a satisfactory solution. Members might not have the same data access permissions. Furthermore, best data management practices require disclosure of *minimum necessary* data subsets to each individual. In this way, the harm caused by unintended data disclosures, due for example to a researcher's laptop being lost or stolen, is limited as much as possible.

Consider a large clinical project that studies patients with hypertension. Project data could consist of patient demographic data, medication data, lab data, disease diagnosis information, blood pressure history, procedures, and genomic data such as microarray data and sequence data. The project encompasses investigators, biostatisticians, bioinformaticians, and postdoctoral researchers from multiple sites. Project datasets are managed by a data coordination team. Depending on the IRB protocol, the project leader, who is also a physician, might be authorized to view data with partial protected health information. However, the other investigators, biostatisticians, and bioinformaticians might only be allowed to see de-identified data. A biostatistician may see a subset of demographic data, medication data, lab data, and procedures for statistical analyses; a bioinformatician may see genomic data and disease diagnosis information for bioinformatics analyses. Moreover, depending on which institution an investigator is from, she may only be allowed to see certain data elements and data subsets. In this setting, the data coordination team receives data extracts from institutional clinical data warehouses. This forms the first set of data releases; the data warehouse teams have to create the appropriate data extracts and track them. Once the data are in the project databases, the data coordination team becomes authorized to create restricted and transformed data subsets (e.g., a transformed dataset may contain aggregation of attributes to comply with data access policies) for the researchers in the project. This creates a chain of data releases (from data extracts from data warehouses through data subsets

created by the data coordination team) that need to be managed. Furthermore, any data leaked from a bioinformatician's data subset should not cause other data (e.g., medication and lab data) to be revealed.

Some of the desirable properties of a data subset extraction and dissemination environment are as follows. Researchers should be provided assurances that the data they are accessing or have received are authentic and have not been tampered with. Any legitimate alterations to the data should be tracked in a secure way. Auditing has to be implemented for each data subset created. Records should clearly show which researchers were given access to which data elements, no matter how many levels of data subsetting and transformation are done and by which systems/users. A trail of accountability has to be implemented and maintained so that if an unauthorized access or update to data occurs or a data breach occurs, appropriate actions can be taken.

In this paper, we present an architecture and methods that provide secure and auditable data extraction and dissemination. Ours is a policy-driven approach, wherein different entities, e.g. an IRB, a data warehouse administrator, or an honest broker, digitally sign data to authorize its release and to indicate that the appropriate policies have been verified. A key component of our approach is the use of *redactable signatures* on data [3]. Redactable signatures allow arbitrary subsets of data elements to be removed from signed data while maintaining desirable properties. Integrity of the remaining data subset can still be verified through the signature and data alterations can be detected. The data subset maintains consistency with the original data set, i.e. its data items are linked to the corresponding items in the source database. Finally, all signatures demonstrating authorization of the extracted data set are maintained. Auditability is provided by our approach through the use of *designated verifier signatures* [7]. Designated verifiers permit data releases to be logged even when releases are done in a highly decentralized fashion, which facilitates auditing of data use policies and monitoring to detect anomalous accesses.

Related Work

A variety of platforms have been developed to support research projects involving heterogeneous clinical datasets generated by multiple institutions [2], [3], [9-11]. These platforms all provide support for decentralized data management, data sharing, and database federation. The systems presently have limited support for signature based data redaction, source integrity verification, and data flow tracking. Our approach complements the security mechanisms implemented in these systems, i.e. it could be leveraged by these systems. Our goals include the design of support to add a verification aspect to the dynamic view of these patient records, view tracking, and auditing capabilities to the standard scenario. The need for secure, auditable subset extraction and redaction is evident in many large-scale research projects that involve medical information sharing, e.g., comparative effectiveness research [4]. In the work of Bouhaddou, et al. [1], a group of collaborating facilities perform patient matching through patient discovery on the Nationwide Health Information Network. This interesting effort was made more challenging by the lack of a capability to redact specific information, which necessitated a manual process to obtain patient approval for release of protected information. This cumbersome process of retrieving approvals turned out to be both expensive and error prone.

Methods

Basic Concepts

In what follows, we provide descriptions of basic concepts upon which our approach is built.

Merkle Hash Tree: To provide the capability to redact signed information, we use redactable signatures based on Merkle hash trees (MHTs) [5, 6]. MHTs are implemented using a collision free hash function, which maps arbitrarily sized data to a fixed length record, while preventing an adversary from finding two inputs that map to the same output. The MHT structure is constructed as a binary tree. The tree contains a set of data items within its leaf nodes. Every granule of the data to be signed is passed into a collision free hash. The hash value is stored in the accompanying node. By recursively using pairs of hash values to form a new hash value, each node concatenates and hashes the hash values found in its children. This process is repeated until one root hash value is generated. The root hash represents the entire data set. With the assumption of a collision free hash, the root value of the tree, in fact the root value of any sub-tree, is only reproducible through the same combinations of the data items. By signing the root value of the tree, the data source confirms its approval of this specific set of data. By using signed hashes from the roots of trees, a grouping of signed data sets may be signed by another party in a similar fashion. Verification of a signed data set requires recreation of its root hash. Using the same combination order as the initial structure, hashes are generated and compared and finally the signature on the root hash is verified. To redact information, the data associated with a hash value is removed from the transaction. The hash value provided for that data is assumed to be correct if, when paired correctly, it is capable of producing the signed root hash value.

In an earlier work [5], we demonstrated the applicability of this approach in the medical context by building and evaluating software to sign and redact Continuity of Care Documents (CCDs). The MHT approach is extremely efficient, because it relies primarily on inexpensive hash calculations and minimizes the number of expensive public-key cryptography operations. Refer to [5] for a more in depth analysis and graphic depiction of the approach.

Monitoring Agent: A trusted monitoring agent or a reference monitor provides mediation for sensitive data accesses to support auditability and accountability. In a decentralized setting, cryptographic primitives can be used to implement such a monitoring agent by requiring its involvement in the creation, update and use of sensitive health data. Whenever data is created or updated by an entity, it is signed using a special type of signature known as a designated verifier signature [7] and encrypted using the monitoring agent's public key requiring decryption by the monitoring agent upon verification. Then, whenever a researcher wants to use this data, she must also contact the monitoring agent, which may designate the researcher as the verifier for the signature. This allows the monitoring agent to record all uses of the data. Thus, the trusted monitor can support full auditability and accountability. In an earlier work [8], we demonstrated the applicability of this approach in the medical context by building and evaluating a monitoring agent for health care transactions and integrating it with NHIN Direct.

View Tracking: To assist with tracking the information seen by clients who receive copies of the records from secondary sources, such as a research collaborator, section signatures are needed. In contrast to signing once over all elements it provides, a data source divides the data into sections and provides a list of sub-trees, each with its own signed hash. This approach provides more information for auditing purposes as to who has accessed which data.

Approach

Figure 1 depicts the major components and the data flow for our system. The major components consist of a transactions manager, which is responsible for managing the data flow of the system, a record handler, and the monitoring agent. Patient health records, or simply records, are stored here in two parts. A relational database maintains the records in such a way as to facilitate simple queries on the system. Another database maintains the Merkle hash trees (MHTs) corresponding to these relational representations of the records. Access policies and their related documents, such as patient consent forms or IRB protocols, are also maintained within the manager to control data release. The monitoring agent provides the mechanism by which we audit data flow by tracking verification of the data. This agent could also be linked with the access policies to help enforce proper data flow by refusing to verify data and flagging access upon a policy breach. The record handler, at some stages, could be seen to act similarly to the transactions manager. Often, in research, the data used is not the full set available within the main repository but is still a large set. Therefore, each user in the system would likely require storage similar to the transactions manager for maintaining the MHT structures as well as the relational database versions of the data.

The data flow in our system is as follows. Using a designated verifier signature in conjunction with the MHT structure, the research lead or an outside source can provide the transactions manager with patient records. These records are maintained in a secure database upon verification of the signature with the monitoring agent. The research lead, or someone with proper administrative privileges, can also provide the transactions manager with release policies and documentation supporting these release decisions. These policies can either be global or local release policies, which specify applicable users, time frame for access approval, viewable data, and signatures required for release. Global policies would define release operations for users who require common access to specific sets of data (e.g. medications). Local policies would allow setting release standards for a given patient record (e.g. hiding mental illness).

When, for example, the lead of a research project queries the transactions manager, the enforcement of the policies relevant to that query redacts any information that has not been approved for release to the project from the query result. The redacted data is then provided after confirming that it matches the original query. During this transaction, the system logs all data that has been accessed. Statistical queries are able to provide users a general view of the available data. These general queries, e.g. number of patients taking a specific medication, would not result in receiving MHT structures as the data provided is not linked to specific patient records.

The research lead verifies the data at this point by contacting the monitoring agent to confirm that the data is valid and trusted. The monitoring agent also logs the data being accessed by this researcher. After this, the researcher has full access to consume and further release the data. The research lead may also need to update the data within the database. This change could be made by providing a newly signed hash tree to the transactions manager and specifying the old version of the records which should be replaced. If the data is needed by another user, for example a biostatistician, the project lead uses the record handler to redact information not needed by the

biostatistician before passing the data set along. The biostatistician then contacts the monitoring agent to verify this subset of the data and uses it internally as well.

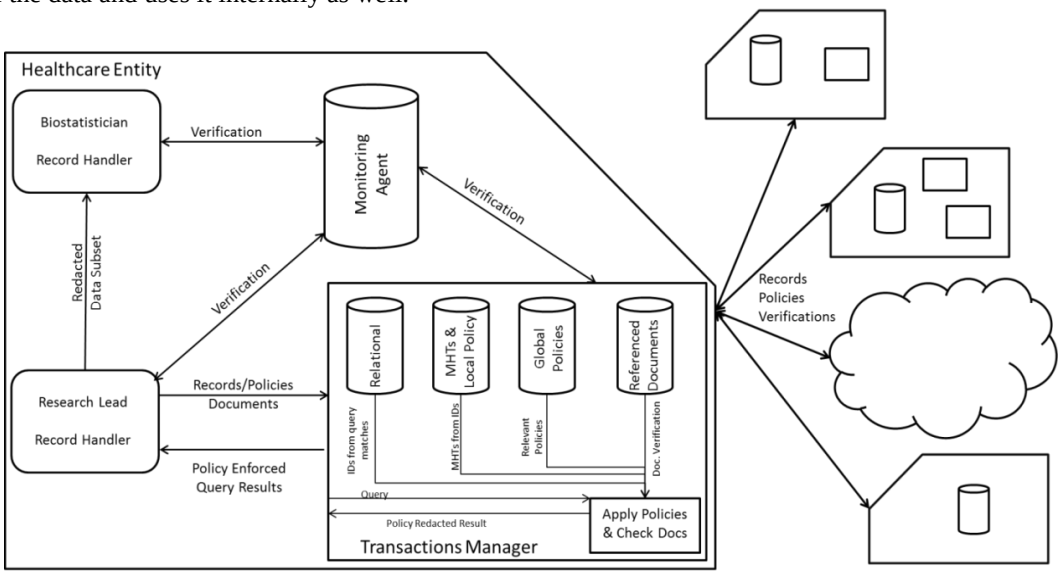

Figure 1: System Architecture and Data Flow

Data transfers are not necessarily limited to a single organization. In addition to sharing the information locally, a research team might need to collaborate with other researchers from a different institution. The required records for the collaboration could be transferred to a research project representative at the other institution. The verification of this data would still be logged, noting that the new researcher has accessed elements of the shared records. In some cases, patient records may be built from various environments and similarly stored in each repository. Verification of this data would communicate with the respective monitoring agents and be logged for record keeping.

Implementation

Using Continuity of Care Documents (CCDs) as a form of patient records, progress has been made in developing the system. A version of the transactions manager has been developed, which is capable of handling CCDs, which are an XML standard for storing health record information. These documents consist of two primary portions. They contain visual representations of the record as well as machine readable elements to express the same data. Another part of our implementation combines a basic record handler with functionality to verify documents at a monitoring agent. That implementation demonstrates the capability of integrating the MHT structures with a monitoring agent. However, the current handler and monitoring agent primarily operate on small sets of records and more work is needed to scale them to work with the data set sizes typical of research scenarios.

The transactions manager works with the MHT structures also. However, the current version does not use a designated verifier but rather a standard electronic signature. The manager generates a set of MHT structures and unique identifiers for the patients represented by the records. Using these data, the XML elements stored in the structure are used to populate versions of the documents in an XML database. This allows for the database to be queried using XPath to find desired information. Information regarding the signatures on the data is also inserted within the documents to assist in finding data signed by specific entities. Each unique identifier maintains a single XML document, which may be used to identify the corresponding MHT, stored in a relational database.

Current redaction operations on the data allow for specific CCD sections to be redacted or data with specific signatures to be kept or redacted. Using the certificates within the data structure and their names, redaction policies may be created that only allow data to be released that has been authenticated by the proper authorities. Currently, policies may be defined to give users or groups access to specific sections of data. A GUI has been developed that allows a user to permit or deny access to a set of data provided the required release signatures are present on that data. These policies are saved in a database accessible by the transactions manager and used at the time of request to determine the appropriate data to return. To err on the side of caution, any case not covered by explicitly defined policies results in a denial of access. The sets of data, controlled by these access policies, may be seen as the sections within the CCD. However, the system allows mapping of data from a generic document to user defined sections. These generic mappings allow importing and exporting of data to and from the MHT structure in a variety of data formats, of which CCD is only one example.

Using an early implementation with the CCD as a primary record source, we were able to estimate the overhead of our implementation. A standard XML database approach was used to store 1000 de-identified CCDs, which required 0.58 seconds to insert. Our approach, which stored the XML document as well as the MHT structure for each entry, required 0.64 seconds for insertion (an 11% overhead). Using these same documents, a query was run that extracted the results section for records of patients born before the year 1928 (317 out of the 1000 patients matched this query). The average time to extract the required sections, using the standard XML approach, for these patients was 0.49 seconds. The additional time required by our approach (to collect all patient identifiers, retrieve corresponding MHTs, and redact all information other than the results section) was 1.26 seconds. We recall that our approach targets a one-time data extraction so this extraction cost would be infrequent. More detailed overheads of MHT structures and the monitoring agent may be found in earlier works [5, 8].

Discussion

The field of clinical research requires efficient methods by which data may be dynamically distributed under different views while maintaining source integrity and the ability to track data accesses. This paper discusses a set of initial implementations as well as an achievable final goal for providing these characteristics with low overhead. Integration and reconfiguration of the existing components, development of proper query components, and exploration of data encryption as an alternate way to enforce access logging are major components of the work still to be completed. However, the reported implementations provide a basic proof of concept for such a system.

Acknowledgments. This research was funded, in part, by grants PHS UL1TR000454 from the Clinical and Translational Science Award Program, National Institutes of Health, National Center for Advancing Translational Sciences, RC4MD005964 from the National Center on Minority Health and Health Disparities, R24HL085343 from the National Heart Lung and Blood Institute, R01HSS019828 by the National Institutes of Health, and R01LM009239 and R01LM011119 from the National Library of Medicine.

References

1. Bouhaddou, O. and Bennett, J. and Cromwell, T. and Nixon, G. and Teal, J. and Davis, M. and Smith, R. and Fischetti, L. and Parker, D. and Gillen, Z. et al. The Department of Veterans Affairs, Department of Defense, and Kaiser Permanente Nationwide Health Information Network Exchange in San Diego: Patient Selection, Consent, and Identity Matching. *AMIA Annual Symposium Proceedings*. 2011; 135 – 143.
2. PoPMedNet, Distributed Research Network Technologies for Population Medicine, http://www.popmednet.org
3. Lin, C.P. and Black, R.A. and LaPlante, J. and Keppel, G.A. and Tuzzio, L. and Berg, A.O. and Whitener, R.J. and Buchwald, D.S. and Baldwin, L.M. and Fishman, P.A. et al. Facilitating Health Data Sharing Across Diverse Practices and Communities. *AMIA Summits on Translational Science Proceedings*. 2010; 16-20.
4. Toh, S. and Platt, R. and Steiner, JF and Brown, JS. Comparative-Effectiveness Research in Distributed Health Data Networks. *Clinical Pharmacology & Therapeutics*. 2011; 90(6): 883-887.
5. Brown, J. and Blough, D. Verifiable and Redactable Medical Documents. *AMIA Annual Symposium Proceedings.* 2012; 1148 - 1157.
6. Merkle R. A certified digital signature. In: *Advances in Cryptology-CRYPTO'89*. Springer; 1990. p. 218-238.
7. Steinfeld, R., Bull, L., Wang, H., and Pieprzyk, J. Universal designated-verifier signatures. *Advances in Cryptology-Asiacrypt 2003 Proceedings*, pp. 523–542, 2003.
8. Daisuke Mashima and Mustaque Ahamad, "Enabling Robust Information Accountability in E-healthcare Systems." *Proc. 3rd USENIX Workshop on Health Security and Privacy*, 2012, to appear.
9. Winslow, R. L., Saltz, J., Foster, I., et al. The CardioVascular Research Grid (CVRG) Project. *Proceedings of the AMIA Summit on Translational Bioinformatics*. 2011; 77-81.
10. Behrman, R. E., J. S. Benner, J. S. Brown, M. McClellan, J. Woodcock and R. Platt (2011). "Developing the Sentinel System—a national resource for evidence development." New England Journal of Medicine 364(6): 498-499.
11. Sittig, D. F., B. L. Hazlehurst, J. Brown, S. Murphy, M. Rosenman, P. Tarczy-Hornoch and A. B. Wilcox (2012). "A Survey of Informatics Platforms That Enable Distributed Comparative Effectiveness Research Using Multi-institutional Heterogenous Clinical Data." Medical Care 50: S49-S59.