

Securing Electronic Health Records without Impeding the Flow of Information

Rakesh Agrawal, Christopher Johnson
IBM Almaden Research Center
650 Harry Road, San Jose, CA 95120

Abstract

The European Union Directive on Data Protection requires member states to enact laws that impose strict limitations on the processing of personal data. Recent laws in the United States, Canada, Australia, and Japan also protect the privacy and security of personal data. These laws significantly impact the management, sharing, and analysis of electronic health records. We present an integrated set of technologies, known collectively as the Hippocratic Database, that enable compliance with privacy and security requirements of data protection laws without impeding the legitimate flow of information. These technologies include (1) active enforcement of fine-grained data disclosure policies, (2) efficient auditing of past database access to verify compliance with policies, (3) privacy-preserving data mining, (4) de-identification of personal data using an optimal method of k -anonymization, and (5) secure information sharing among autonomous data sources. We describe the functionality of each component, offer example scenarios to demonstrate their usefulness, and identify remaining research challenges in securing electronic health records.

1 Introduction

The 1995 European Union Directive on Data Protection (“Directive”) [1] set forth stringent cross-industry standards regarding privacy and security of personal data. Pursuant to these standards, EU member states adopted data protection laws that obligate controllers of health data to provide all data subjects with: (1) notice of the purposes for which they collect and use personal data; (2) choice regarding whether their data may be disclosed to third parties or used for a different purpose than it was originally collected or subsequently authorized; (3) reasonable assurance that the data will be secured and its integrity maintained; (4) access to the data and the opportunity to correct inaccuracies; and (5) legal recourse to ensure compliance with data protection requirements. States may allow processing of health data without owner consent for purposes of preventative medicine, diagnosis, treatment, management of medical services, or otherwise under professional confidentiality obligations, only if suitable safeguards are provided.

Similar laws in the United States [2], Canada [3], Australia [4], and Japan [5] require healthcare institutions

to protect the privacy and security of personal health data. Advisory reports commissioned by the United States government [6] [7] stress the importance of developing secure, interoperable electronic health records systems that preserve patient privacy. As countries around the world transition from paper-based to electronic health records infrastructures, compliance with data protection laws will require sophisticated information management technologies. Healthcare organizations must implement privacy and security protections such that they do not unduly constrain proper use and dissemination of health data or impede scientific discovery. Technical and policy challenges concerning the widespread adoption of electronic health records have been discussed, for example, in [8] and [9].

The Hippocratic Database (“HDB”) [10] is a set of technologies that manages disclosure of electronic health records in compliance with data protection laws without impeding the legitimate flow of information. HDB’s active enforcement component limits disclosure of personal health information at a fine-grained level in strict accordance with enterprise policies, legal regulations, and individual patient choices. Its compliance auditing component efficiently tracks past disclosures to verify compliance with these policies. Finally, its data mining, de-identification, and information sharing components enable organizations derive maximum value from sensitive data without compromising privacy or security.

The remainder of this paper is organized as follows. Sections 2 and 3 describe HDB active enforcement and compliance auditing. Sections 4, 5, and 6 discuss privacy-preserving data mining, optimal k -anonymization, and sovereign information integration. In each section, we include example scenarios demonstrating practical applications of these technologies. In Section 7, we suggest a number of opportunities for further research in securely managing electronic health records. We conclude in Section 8.

2 Active Enforcement

HDB active enforcement (“AE”) [11] is a disclosure management component that is transparent to enterprise applications and agnostic to database systems. It resides in a layer above the database, rewriting user queries to conform to the organization’s data disclosure policies and individual patient choices. AE enforces disclosure

policies down to the cell-level in the database, allowing health organizations to comply with detailed requirements of data protection laws without recoding their applications. HDB policy controls are more fine-grained than conventional role-based access controls [12], as they account for the purpose of access, the intended recipient of the information, and patient consent rights, in addition to the user's access privileges. The complete AE solution is comprised of three stages – policy creation, preference negotiation, and application data retrieval. (See Figure 1.)

In the *policy creation* stage, the healthcare organization specifies a data disclosure policy through the HDB control center. The policy governs the access privileges for each role within the organization according to the category of information sought, the purpose of the request, and the intended recipient of the results. It may also provide individual patients with the opportunity to express opt-in or opt-out choices regarding the disclosure of their personal information, also according to category, purpose, and intended recipient. For example, a patient may opt into sharing his medical information with universities for research purposes, but opt not to share his contact information with pharmaceutical companies for marketing purposes. Policies are expressed in a language such as P3P [13] and installed in the database in a form amenable to symbolic manipulation. The organization may update or replace policies through a one-step installation process in the control center. The database stores multiple policies and versions of policies.

In the *preference negotiation* stage, the patient is notified of the health organization's policies concerning data use and disclosure, advised of any conflicts with his own privacy and security preferences, and allowed to express personal opt-in or opt-out choices. This fully automated process is completed before the patient

provides any personal data to the organization. The patient first uses the HDB preference interface to express his preferences concerning the use and disclosure of his personal data. This information is then specified in a preference language [14] and matched with the health organization's privacy and security policies to identify any conflicts. The patient is advised of these conflicts and given an opportunity to resolve them or terminate the process. Lastly, the patient is provided opt-in or opt-out choices regarding whether his data may be disclosed to third parties or used for a different purpose than it was collected. These choices are recorded in the database and factored in at the time of query processing. A successful preference negotiation confirms agreement between the patient and health organization concerning processing of his personal data.

In the *application data retrieval* stage, all queries to be executed on the data source are programmatically modified so that the application only retrieves results that are compliant with disclosure policies, including legal requirements and patient opt-in and opt-out choices. The query rewrite process transparently enforces cell-level access controls based upon the user's role, purpose, and intended recipient. This ensures that queries from any application return all responsive data that the particular user is entitled to access, but none that he is not. HDB active enforcement can also be configured to support policy rules granting read-only or write access, on a cell-by-cell basis, depending on the context of the query.

AE is integrated into existing environments through a database interface such as ODBC or JDBC. Its fine-grained enforcement capability implements cell-level restrictions, such as opt-in and opt-out choices, without requiring any changes to enterprise applications. Further, AE actually improves query processing speed in

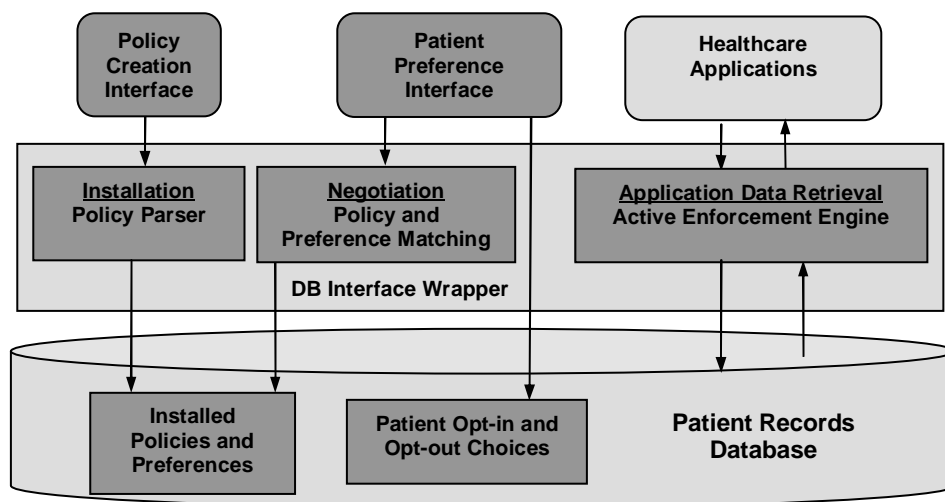


Figure 1: HDB Active Enforcement Architecture

the typical instance because rewritten queries benefit from the optimizations and performance enhancements of the database engine [2]. AE can also be combined with techniques that allow queries over encrypted numeric data without significantly degrading performance [15].

2.1 Active Enforcement Scenario

Richard is a young professional who has recently moved to a new city and would like to select a local healthcare provider and schedule his annual physical. He is considering Continental Hospital, a well-regarded healthcare organization that is part of a large network with many patients and locations. Prior to scheduling an appointment, Richard must register for membership on Continental's website.

Policy Creation: As part of its transition to an electronic records infrastructure, Continental has recently installed HDB active enforcement software. The first step in the enforcement process is for Continental to create a data disclosure policy. Hospital management starts by reviewing Continental's existing policy to ensure it is consistent with current data protection laws and its own business objectives. Amanda, its Chief Privacy Officer, then specifies the policy in the chosen privacy language and installs it through the HDB administrative console.

Preference Negotiation: After entering the registration page on Continental's website, Richard is notified of the hospital's data disclosure policy prior to submitting any personal information. He submits a list of preferences regarding the use of his personal data through an extension of his web browser. Among other preferences, Richard indicates that he does not want to share his medical information with government agencies for any purpose and does not want to share his telephone number with third parties for marketing purposes.

HDB automatically compares Richard's preferences with Continental's data protection policy and uncovers one potential conflict. Continental's policy is to release medical information to relevant government agencies if necessary to verify an employee disability claim or comply with a court order. After being notified of this conflict, Richard decides to waive his preference regarding disclosure to government agencies and completes Continental's registration form.

Prior to submitting his completed registration, Richard is provided with two opt-in choices. These choices are intended to allow the patient and healthcare provider to reach an agreement concerning the provider's discretionary use of his personal information, in compliance with national data protection laws. For the first choice, Richard consents to share his medical information with third parties for research purposes. For the second choice, however, he does not opt to share his

medical information with affiliates of Continental for marketing purposes.

Application Data Retrieval: Several months after his annual physical, Richard injures his ankle playing basketball. His doctor sends him to RadioTech Labs, a Continental affiliate, to have a series of X-rays performed. The lab technician types Richard's name into a computer terminal and requests access to his medical records. In the absence of HDB controls, the technician would see all of Richard's personal health records stored in Continental database. However, with HDB active enforcement in place, the application returns only Richard's contact information and the records of his latest hospital visit, but no other health records. This complies with Continental's data disclosure policy and Richard's privacy preferences.

2.2 Information Sharing with Active Enforcement

A second scenario demonstrates how AE can be used to facilitate policy-compliance information sharing among multiple organizations.

Joan is a professor at Northern University Medical School with access to Continental's patient database under a joint research agreement. She is currently working on a project to evaluate whether various environmental and genetic factors contribute to high cholesterol levels. To begin her research, Joan logs into Northern's web portal and submits the following SQL query to the Continental Hospital database:

```
Select * from patients where total cholesterol ≥ 200
```

Without HDB controls, Joan would be given total access to the records of all patients with total cholesterol levels of 200 and above. This is a violation of Continental privacy policy and EU data protection laws, because not all patients have consented to reveal their health information to third parties for research purposes. With HDB in place, the AE engine rewrites Joan's query to comply with Continental's data disclosure policy and patient opt-in and opt-out choices. Thus, AE filters out the personal data that patients did not opt to share with third parties for drug research purposes and returns the remaining data that is responsive to the query.

3 Compliance Auditing

Pursuant to the EU Directive and member state laws adopted thereunder, health organizations must be accountable to patients for all processing of their personal data. Upon request, patients are entitled to a description of the data disclosed, the recipients of the data, and the purposes of the processing. Further, member states must provide all persons with a remedy for any breach of their rights under national data protection laws. In the United States, the Health Insurance Portability and Accountability Act ("HIPAA") [2] requires healthcare

organizations to account for certain disclosures of patient health data upon request and provides penalties for unlawful disclosures. Accordingly, there is a critical need for auditing systems that track past disclosures of information and verify whether they complied with applicable laws and policies.

HDB compliance auditing [16] enables organizations to investigate past disclosures without the performance and overhead burdens of other auditing systems. Using an audit application over existing database infrastructure, HDB allows auditors to track the identities of users who have accessed any cell in the database, the date and time of access, the purpose of the access, the recipient of the information, and the exact information disclosed. Thus, HDB auditing provides reliable and efficient means for health organizations to account for its processing of personal information.

The HDB compliance auditing system consists of two parts – a logical logging system and an audit application. The logical logging system records all queries and contextual information (i.e., identity, time, purpose, recipient) in query logs. It also stores all data updates, insertions, and deletions in backlog tables. These backlog tables are populated using database replication logs, triggers, or point-in-time query features.

The audit application provides a simple user interface that allows an auditor to formulate an audit query specifying the data she wants to audit. Upon receiving the audit query, the application generates a list of suspicious queries. Using the query logs and backlog

tables, the application then produces an audit report that identifies the user, time, purpose, recipient, and exact information disclosed for each suspicious query.

HDB is superior to auditing systems that log the actual results of database queries, because it does not incur a cost for read queries or otherwise log redundant data. By logging only the queries and changes to the database, HDB operates much more efficiently and requires far less overhead than result logging systems. HDB also has a security advantage in that it captures information revealed by a query that may not be reflected in the output. For instance, the query “Select value 1 if patient ‘Jane’ has diagnosis ‘diabetes’” would not be tracked by auditing systems that log the output of queries. The same is true for queries that aggregate values from the records accessed. In contrast, an HDB audit would show the precise data revealed in these situations.

In the following scenario, a healthcare organization uses HDB compliance auditing to investigate a claim that it unlawfully disclosed a patient’s personal health data.

3.1 Compliance Auditing Scenario

Palmer is a candidate for political office and a patient of Continental Hospital. Shortly before the election, a local newspaper story discloses portions of Palmer’s personal health records indicating that he has been treated for depression. He believes that Continental is responsible for this unlawful disclosure and threatens to sue the hospital under national data protection laws.

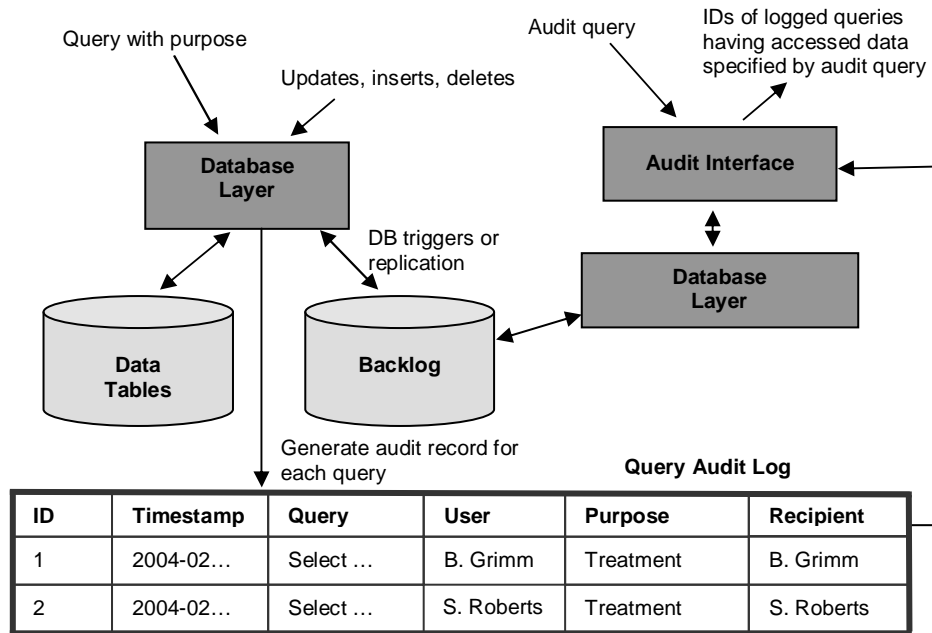


Figure 2: HDB Compliance Auditing

Continental's president is very concerned about this high-profile accusation and requests that Amanda, the Chief Privacy Officer, immediately provide him with an accounting of all who have accessed Palmer's personal health data. He also demands that Amanda conduct a more specific investigation to determine who, if anyone, was actually responsible for the disclosure.

HDB Audit Specification: Amanda logs into the HDB audit interface to begin the investigation. The system is preset with a number of common tasks that a hospital auditor might want to perform. Examples of such tasks include "Accounting of Access and Disclosure", "Who Accessed Medical Information", "Which Third Parties Accessed Information", and "Frequent Access of Information." Alternatively, auditors can declaratively specify statements in a SQL-like syntax to execute custom audits. Auditors can also indicate the exact timeframe of the disclosures they would like to audit.

Investigation of Suspicious Access: Amanda would first like to know the identities of all persons who have accessed Palmer's medical information in the past year. To accomplish this, Amanda selects the "Accounting of Access and Disclosure" task and restricts her search to only Palmer's medical information, rather than all of his personal records (e.g., address, telephone number, payment information), and defines the audit timeframe as the past twelve months. The audit application identifies suspicious queries that accessed Palmer's medical records during last year and returns a list of users who accessed them, the time and purpose of each access, and the exact data returned in response to each query. Amanda quickly provides a printed report of this accounting to the hospital president and proceeds with her investigation.

Amanda notices that the results show a large number of queries accessing Palmer's medical records, but not all of those queries revealed the diagnosis of depression or his prescription for anti-depression medication. Thus, she adds a custom column to the audit based on *diagnosis* to further sort the information, so that she can isolate those queries which accessed information about Palmer's diagnosis of depression. She repeats the same process for the *prescription* column. These views show many queries that returned information about Palmer's past diagnoses and treatment for influenza and strep throat, but nothing about depression. These queries can be disregarded, as they could not have resulted in the wrongful disclosure.

Among the queries that accessed Palmer's depression diagnosis or treatment, Amanda sorts the results by user. Comparing the user identities with her record of Palmer's treating physicians, she notes that his primary physicians and nurses frequently accessed medical data relating to his depression. However, another physician, Dr. Roberts, who is not listed as one of Palmer's treating physicians,

also accessed this data several times over a short period, purportedly for treatment purposes.

Amanda is suspicious of this access pattern, so she specifies another audit to determine the precise records that Dr. Roberts has accessed. She notices that Dr. Roberts has made only a few queries in the system, but has accessed a large number of patient's records, all with diagnoses related to depression. Wondering whether this type of search is a common occurrence, Amanda proceeds to specify another task, this time to isolate physicians that accessed over 200 depression patient records at a time. Still, Dr. Roberts is the only physician that has conducted such a query. Amanda heads off to interview Dr. Roberts to continue her investigation.

In this scenario, a manual audit would have required countless hours of searching through files and notes and interviewing various hospital employees, with little hope of locating the actual source of the leak, if any occurred. In contrast, HDB Compliance Auditing allows a hospital auditor to conduct a series of audits, in a matter of minutes, to reliably isolate potential sources of the leak. In fact, Amanda could have reduced the steps above by formulating a more precise initial audit query.

An audit may either reveal the actions of a malicious employee or serve as evidence that the hospital is not responsible for the disclosure. Moving forward, Amanda can initiate proactive audits to investigate the effectiveness of the hospital's disclosure controls. Further, if employees are aware that auditors have the ability to track past disclosures, HDB compliance auditing will provide a significant deterrent to unlawful access and disclosure in the future.

4 Privacy-Preserving Data Mining

HDB's Privacy-Preserving Data Mining ("PPDM") [17] allows mining of aggregate data without revealing precise information in individual records. Thus, it enables analysis of large data sets for epidemiological studies and other medical research without violating patient privacy.

PPDM uses a randomizing function to perturb sensitive values in a patient's record such that they cannot be estimated with reasonable precision. From the randomized data, it reconstructs the original data distribution to allow data mining at the aggregate level, without revealing individual values. Algorithms for building classification models and discovering association rules on top of privacy-preserved data can be used on the randomized data with only a small loss of accuracy [18].

4.1 Privacy-Preserving Data Mining Scenario

Continental recently began a home health monitoring program in which patients measure their vital statistics at

home on a daily basis. Scales, blood pressure monitors, cholesterol monitors, thermometers, and other pervasive devices wirelessly feed data into a web application on the patient's home computer that transmits this data to the hospital. The data is fed into patient medical records and used to monitor and diagnose various health conditions. Hospital management recognizes that these large sets of patient health data would also be valuable for a variety of data mining purposes. They would like to be able to share this information with third party researchers, on an ongoing basis, without revealing private information.

Continental decides to solve this problem using PPDM. As patient data is received from the home monitoring system, one copy is sent to the standard patient database and another copy is sent to a PPDM randomizer. Upon receiving each data item, the randomizer perturbs the data and sends it onto a research database consisting of only privacy-preserved data. Continental can provide access to the research database to third party researchers, who can run PPDM algorithms to reconstruct the original data distributions. Researchers can then construct data mining models on the reconstructed data, with an insignificant loss of accuracy.

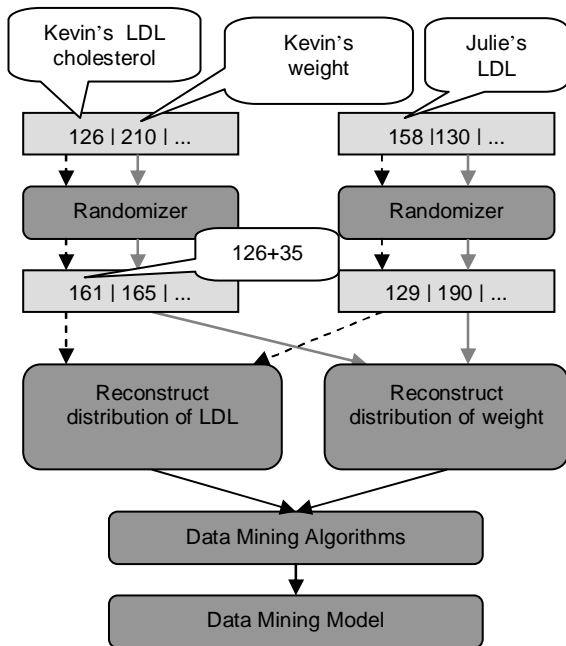


Figure 3: HDB Privacy-Preserving Data Mining

5 Optimal k -anonymization

The EU Directive generally prohibits the processing of personal health data without patient consent, unless required in connection with the provision of medical care. However, member states are allowed to lay down exemptions to this prohibition if data are processed under an obligation of professional secrecy or for reasons of

substantial public interest, subject to suitable safeguards. In the US, HIPAA allows healthcare organizations to process personal data without patient consent if they use statistically acceptable methods to de-identify the data.

HDB's optimal k -anonymization [19] component provides for an optimal method of de-identifying sensitive data sets that protects the privacy of data subjects, but maintains the value of the data for research purposes. In a k -anonymized data set, each record is indistinguishable from at least $k - 1$ other records [20]. The process of k -anonymization involves data suppression (deleting cell values or entire tuples) and cell-value generalization (replacing specific values with more general ones). The larger the value of k , the greater the privacy protection.

The k -anonymization method is superior to naïve de-identification approaches that simply remove certain identifiers. These naïve approaches are prone to data linkage attacks that combine the subject data with other publicly available information to re-identify represented individuals. The k -anonymization method was designed to avoid such linkage attacks, while preserving the integrity of the released data. Unlike other disclosure protection techniques that involve condensation, data scrambling and swapping, or adding noise, the records that remain in a k -anonymized data set are truthful [20].


Unfortunately, even simple restrictions of optimized k -anonymity are NP-hard [21], leading to significant computational challenges. Our new approach [19] to exploring the array of possible anonymizations tames the combinatorics of the problem. Our experiments on a real data set show that the resulting algorithm can find optimal k -anonymizations under two representative cost measures and a wide range of k . Our algorithm can also produce useful anonymizations in circumstances where the input data or input parameters preclude finding an optimal solution in a reasonable amount of time.

5.1 Optimal k -anonymization Scenario

Continental Hospital would like to share de-identified data sets with Northern University for medical research purposes. However, removing personal identifiers such as name, street address, telephone number, is insufficient because it leaves the data set prone to data linkage attacks. While no records in the de-identified data set contain a single identifying value, many of them may contain unique value combinations. An individual who is the only Caucasian male born in 1925 living in a sparsely populated area could have his age, race, gender, and zip code joined with a voter registry from the area to obtain his name and mailing address. This would reveal all of the individual's private medical information. However, removing all information that could possibly be used for data linkage attacks would render the data useless for research purposes.

Optimal k -anonymization strikes a balance between protecting the individual privacy and maintaining useful data for analysis. Rather, than categorically removing or revealing columns of information, k -anonymization removes certain cells of data and generalizes others so that every record is indistinguishable from $k - 1$ records. In Figure 4 below, the records in the top table are de-identified such that $k = 2$ with respect to name, address, city, and age. Accordingly, names are suppressed and addresses and ages are generalized to the extent that each record is indistinguishable from at least one other record. The remaining data is truthful and valuable for research.

Name	Address	City	Age	Diagnosis
Eric	7, rue du Mont Dore	Paris	26	Influenza
Paul	13, rue des Canettes	Paris	42	Hypertension
Marc	48, rue du Four	Paris	47	Diabetes
Henri	21, rue du Mont Dore	Paris	28	Asthma


 k -anonymization
 ($k=2$, on name,
 address, city, age)

Name	Address	City	Age	Diagnosis
*	17 th Arrondissement	Paris	20-29	Influenza
*	6 th Arrondissement	Paris	40-49	Hypertension
*	6 th Arrondissement	Paris	40-49	Diabetes
*	17 th Arrondissement	Paris	20-29	Asthma

Figure 4: HDB Optimal k -anonymization

6 Sovereign Information Integration

HDB's Sovereign Information Integration ("SII") [22] component enables two or more autonomous entities to run queries across their databases in such a way that the results of the query are revealed, but no other data is exposed among the databases. SII uses a web services infrastructure to apply a set of commutative encryption functions to uniquely identifiable data in different orders and at different locations. The multiply encrypted values are then compared, and the query results provided, without compromising the security of either data set.

Unlike other data integration approaches, such as centralized data warehouses and mediator-based data federations, which reveal all data among the databases, SII only reveals results of the query. This allows collaborating parties to perform a variety of joins and other operations across their databases without revealing unnecessary information. SII is a scalable software solution that can be integrated seamlessly into existing data environments without the need for a trusted third party or any anonymization of the original data.

In the following scenario (depicted in Figure 5), SII presents an ideal solution to a research problem requiring secure sharing of information among autonomous entities.

6.1 SII Clinical Genomics Scenario

Walter is a medical researcher at Northern University who would like to test hypotheses concerning correlations between certain genetic expressions and efficacy of a new diabetes drug, Glucotin. Specifically, Walter believes that Glucotin is ineffective in patients with a specific DNA sequence and highly effective in patients with another specific DNA sequence.

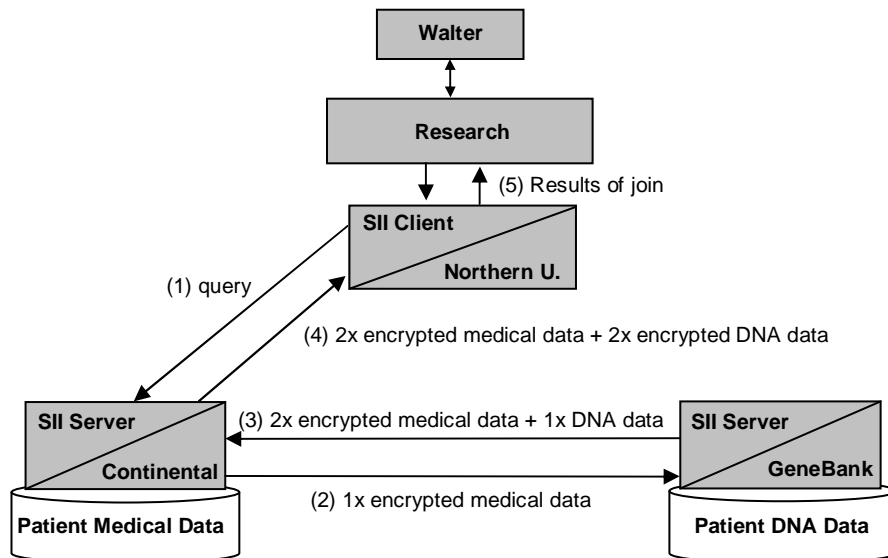


Figure 5: HDB Sovereign Information Integration

To test these hypotheses, Walter must have access to the medical records of patients who are taking Glucotin as well as genetic information about these same patients. Walter is aware that Continental and GeneBank have a number of common patients, many of whom have been prescribed Glucotin. However, national data protection laws prohibit Continental and GeneBank from revealing personally identifiable information without patient consent. Thus, Walter would like to investigate the correlation between the two specific DNA sequences and the efficacy Glucotin, without revealing any other information among the three organizations.

Continental, GeneBank, and Northern have installed SII to facilitate secure, privacy-preserving information sharing. Figure 5 illustrates the process of Walter's join operation in the following five steps. (1) To determine whether the first DNA sequence correlates with ineffective Glucotin treatment, Walter sends an intersection query to the Continental SII service via Northern's client application. (2) Continental then encrypts the patient table with its own key and sends the table to GeneBank's SII service. (3) Next, GeneBank encrypts Continental's singly encrypted patient table and its own DNA table with its own key and sends both tables back to Continental SII service. (4) Continental then encrypts GeneBank's singly encrypted table so that both data sets are now doubly encrypted. (5) Finally, SII joins both doubly encrypted tables and sends the number of matching results to Northern's application.

7 Research Challenges

7.1 Policy Specification

Effective HDB active enforcement controls rely on the ability of policies to capture the intent of the policy maker accurately. At the same time, the policy specification should be clear enough that the patient can easily understand the policy and the implications of his choices. While privacy policy specification languages such as P3P offer vast improvement over long legal texts of privacy policies and make policies amenable to symbolic manipulation, they fall short on readability and understandability. Thus, there is a major challenge in designing a policy language that reconciles the goals of understandability and efficient computation.

7.2 Sticky Policies

As healthcare organizations share personal health data with multiple entities, they should be assured that the original policy controls will be enforced over that data as a condition of transfer. When a patient agrees to provide personal data to a healthcare organization under a set of policies and preferences, he is entering into a contract regarding the handling of his data. If the policies allow the data to be transferred to another entity, the

patient should be assured that the same disclosure rules will apply to the data after transfer. Thus, it is necessary to have "sticky policies" that transfer with the data and remain with it after consolidation. The transferee should be capable of applying the source disclosure policies to any information in its database. Currently, such data sharing is governed by contracts that require transferees to apply appropriate privacy and security controls to data they receive. Assuming interoperable enforcement systems, sticky policies would be much more effective in ensuring that personal data is always processed in accordance with the patient's expectations.

7.3 Data Pointillism

As electronic health records become more prevalent, patients are likely to have personal health data stored in a variety of distributed data sources. Physicians with assorted specialties may be located in different areas, and patients may change healthcare providers when they relocate, change jobs, or switch insurance companies. To provide physicians with a complete health history for each patient, there is an important need for technologies that unambiguously identify patients and link their information from multiple sources. Such consolidation greatly assists physicians in diagnosis and treatment decisions and reduces the cost of duplicative and unnecessary procedures. Healthcare providers should be able to integrate patient information coherently by combining small, continuously arriving "points" of data. Several techniques exist for this type of data integration [23] [24], but further research is needed to accommodate and correct errors in the data, incorporate different data types, and limit false positives. Mechanisms are also needed to enable patients to check the accuracy of their data and make corrections in case errors are found.

7.4 Management of Massively Distributed Data

There are many questions raised by the growing amounts of personal health data stored on inexpensive personal devices such as memory keys, portable disks, and smart cards. In addition, pervasive devices such as wireless monitoring devices are becoming increasingly important for modern healthcare. Accordingly, new technologies are needed to protect the security of the information on these devices, enable selective sharing of this information, and create back-up mechanisms to prevent data loss.

7.5 User Authentication and Authorization

Secure access to health information requires mechanisms for accurately identifying those accessing and modifying patient records and ensuring that they have proper authorization. Currently, there are not defined standards for electronic authentication of users and transmitting instant authorizations. To enable information

sharing among a network of unaffiliated healthcare organizations, research should define extensible trust hierarchies and authentication standards [6]. Adequate data protection can be assured only if there are accepted and reliable methods for verifying the identities of users accessing sensitive data.

7.6 Data Lifecycle Management

As electronic health records are stored in databases, technologies that facilitate data life cycle management will become crucial. Data controllers should be able to define retention periods for data based upon legal requirements and patient specifications. At the end of the retention period, storage systems should have methods to remove expired data and forget any persistent data that would allow recreation. Because healthcare organizations require superior availability and reliability of data, storage systems must be secure from data contamination, loss, and leakage and provide methods for establishing the truthfulness of data.

7.7 Interoperability

Another technical challenge facing the healthcare industry is interoperability. Effective sharing of health information requires the ability to communicate among sovereign systems, using standard data formats and clinical vocabularies. While there has been progress toward developing messaging standards such as HL-7, standard vocabularies such as SNOMED-CT, and document standards such as CDA and CCR, much further work remains to be done to ensure that patient health records are complete and healthcare organizations have access to all information necessary for diagnostics, treatment, and medical research [25]. An intriguing research direction worth exploring is the use of mass collaboration [26] to define clinical vocabularies and taxonomies.

8 Conclusion

We have shown how Hippocratic Database technologies protect the security of personal health records without sacrificing the value of information for diagnosis, treatment, or research purposes. Our example scenarios demonstrate how each of these technologies enables efficient management, sharing, and processing of sensitive data in compliance with the principles of the EU Directive and other data protection laws. We have also identified a number of significant technical challenges that remain in this area. We hope that the technologies outlined herein serve as a foundation for modern health records infrastructures and inspire productive research in secure information management.

References

- [1] European Union Directive on Data Protection, *Official Journal of the European Communities*, 23 November 1995 No L. 281 p. 31.
- [2] Health Insurance Portability and Accountability Act of 1996, United States Public Law 104-191.
- [3] Personal Information Protection and Electronic Documents Act, Second Session, Thirty-sixth Parliament, 48-49 Elizabeth II, 1999-2000, Statutes of Canada 2000.
- [4] Privacy Act of 1988, Commonwealth of Australia, Act No. 119 of 1988 as amended.
- [5] Law on the Protection of Personal Information, promulgated by the Diet of Japan on May 30, 2003.
- [6] President's Information Technology Advisory Committee, "Revolutionizing Health Care Through Information Technology". Report to the President of the United States, June 2004.
- [7] Commission on Systemic Interoperability, "Ending the Document Game: Connecting and Transforming Your Healthcare through Information Technology". United States Government Printing Office, October 2005.
- [8] B. Humphreys, "Electronic Health Record Meets Digital Library," *Journal of the American Medical Informatics Assoc.*, Vol. 7(5) Sep-Oct 2000, pp. 444-52.
- [9] I. Iakovidis, "Towards Personal Health Record: Current Situation, Obstacles and Trends in Implementation of Electronic Healthcare Record in Europe," *International Journal of Medical Informatics*, Vol. 52, No. 1, October 1998, pp. 105-115.
- [10] R. Agrawal, J. Kiernan, R. Srikant, Y. Xu, "Hippocratic Databases". *Proc. of the 28th Int'l Conf. on Very Large Databases*, Hong Kong, China, August 2002.
- [11] K. Lefevre, R. Agrawal, V. Ercegovac, R. Ramakrishnan, Y. Xu, D. DeWitt. "Limiting Disclosure in Hippocratic Databases". *Proc. of the 30th Int'l Conf. on Very Large Databases*, Toronto, Canada, August 2004.
- [12] R. Sandhu, E. Coyne, H. Feinstein, C. Youman, "Role-Based Access Control Models" *IEEE Computer*, Vol. 29, No. 2, February 1996, pp. 38-47.
- [13] L. Cranor, M. Langheinrich, M. Manchiori, M. Presler-Marshall, J. Reagle, "Platform for Privacy Preferences 1.0 (P3P1.0) Specification". *W3C Recommendation*, April 2002.
- [14] R. Agrawal, J. Kiernan, R. Srikant, Y. Xu, "An XPath-based Preference Language for P3P". *Proc. of the 12th Int'l World Wide Web Conference*, Budapest, Hungary, May 2003.
- [15] R. Agrawal, J. Kiernan, R. Srikant, Y. Xu, "Order-Preserving Encryption for Numeric Data". *Proc. of the ACM SIGMOD Conference on Management of Data*, Paris, France, June 2004.
- [16] R. Agrawal, R. Bayardo, C. Faloutsos, J. Kiernan, R. Rantau, R. Srikant, "Auditing Compliance with a

-
- Hippocratic Database". *Proc. of the 30th Int'l Conf. on Very Large Databases*, Toronto, Canada, August 2004.
- [17] R. Agrawal, R. Srikant, "Privacy-Preserving Data Mining". *Proc. Of the ACM SIGMOD Conference on Management of Data*, Dallas, Texas, USA, May 2000.
- [18] A. Evfimievski, "Randomization in Privacy-Preserving Data Mining". *SIGKDD Explorations: Newsletter of the ACM Special Interest Group on Knowledge Discovery and Data Mining*, 4(2), December 2002, pp. 43-48.
- [19] R. Bayardo, R. Agrawal, "Data Privacy Through Optimal k -Anonymization". *Proc. of the 21st Int'l Conf. on Data Engineering*, Tokyo, Japan, April 2005.
- [20] P. Samarati, L. Sweeney, "Generalizing Data to Provide Anonymity when Disclosing Information". *Proc. of the 17th ACM SIGMOD-SIGACT-SIGART Symposium on the Principles of Database Systems*, 188, 1998.
- [21] H. Lewis, C. Papadimitriou, *Elements of the Theory of Computation* (2d Ed.), Prentice Hall, 1998, pp. 293-98.
- [22] R. Agrawal, A. Evfimievski, R. Srikant, "Information Sharing across Private Databases". *Proc. of the ACM SIGMOD Conference on Management of Data*, San Diego, California, June 2003.
- [23] O. Benjelloun, H. Garcia-Molina, J. Jonas, Q. Su, J. Widom, "Swoosh: A Generic Approach to Entity Resolution". Stanford University Technical Report, March 2005.
- [24] S. Ellard, "System and Method for Indexing Information about Entities from Different Information Sources". United States Patent No. 5,991,758, Issued November 23, 1999.
- [25] California Healthcare Foundation, "Clinical Data Standards Explained," November 2004.
- [26] M. Richardson, R. Agrawal, P. Domingos, "Trust Management for the Semantic Web". *2nd Int'l Semantic Web Conf.*, Sanibel Island, Florida, October 2003.