Health Information Identification and De-Identification Toolkit

Isaac S. Kohane,*MD, PhD Hongmei Dong,*MS, Peter Szolovits,† PhD

*Children's Hospital Informatics Program, Children's Hospital and Harvard Medical School;
†Clinical Decision-Making Group, MIT Laboratory for Computer Science

Health identifiers are required for health information systems ranging in scope from the national to the smallest clinical study. Identification systems will differ in the tradeoffs of privacy and control that they represent. The Health Information Identification and De-Identification Toolkit (HIIDIT) is a generator of health identification systems that allows a system architect to specify the set of tradeoffs that are desired for any particular health information system.

Introduction

Patient identification is one of the higher and more controversial priorities for the implementation of health information systems[1,2]. There is a broadly shared goal to better understand the long-term health status of patients when addressing their immediate needs, to study the effectiveness of different patterns of care, to investigate the long-term outcomes of proposed interventions through clinical research studies, and to optimize the system of healthcare delivery. These all create the need for coherent, comprehensive longitudinal records about the care of individual patients. Because most data are, however, collected in disparate, unintegrated ways, it is vitally important to be able to identify the same individual's data though coming from different institutions and collected by different means and at different times. The most convenient way to address this problem would be to associate with each individual a unique, permanent identification number that would be used universally in every database that collected information about that individual. Under such a scheme, every database in the country could, at least in principle, be *joined* on this common key to produce a complete database about everyone. This ability is, of course, both the advantage and the defect in this simple scheme. Although it makes collation of data relatively easy, our national traditions of privacy and patients' expectations of confidentiality of their health data are too easily violated.

We describe in this paper our preliminary work to develop a set of tools that will allow the creation of a broad range of patient identification systems. This work is part of a larger project to develop the Health Information Identification and De-Identification Toolkit (HIIDIT, pronounced "Hide It"), funded by the National Library of Medicine. In addition to its Health Identification Toolkit (HIT) component, HIIDIT also has a de-identification component designed along the lines of Latanya Sweeney's SCRUB program[3] which will not be discussed in this paper. Heretofore, in this paper, when we refer to HIIDIT, we only are specifying its HIT component.

HIIDIT allows the designer of health information systems to select different judgements or trade-offs between competing desiderata for an identification system, including dimensions such as who controls the creation and dissemination of identifiers, the extent to which the same identifier can be used for multiple purposes, the source of trust who certifies the identity of a patient or institution, the degree to which the identifier itself is kept secret, and the complexity of the resulting system of identification. That is, HIIDIT is not of itself a health identification system, but rather a generator of health identification systems. HIIDIT's cryptographic infrastructure is based upon a recently proposed "Simple Distributed Security Infrastructure" (SDSI[4]) that provides a small, powerful set of security capabilities in terms of the underlying cryptographic techniques.

In this paper, we first review earlier work that lead to the conception of the HIIDIT project. We then sketch a brief outline of the framework for identification systems that we are capturing in HIIDIT. Subsequently we review SDSI functionality. Finally we provide an example of an identification system for a multi-center genomics study that represents one set of tradeoffs among a broad continuum of tradeoffs implementable with HIIDIT. In the discussion section, we review the challenges for broad implementation.

Background

In response to the mandate for national health identifiers in the Kennedy-Kassebaum bill[5], several proposal have been made to adopt or adapt one of several existing identification systems. Among the most cited proposals is that for the use of the Social Security Number (SSN) or a simple modification of it (to include a check digit) as the only practical approach that can make possible the short-term adoption of large-scale information sharing systems.

The proponents of the use of the SSN describe its advantages and cost-efficacy as overwhelming. Several national studies[6], by contrast, have cautioned against the adoption of the SSN, and privacy advocates almost universally oppose the idea. Nevertheless, some responsible and influential organizations endorse it and are working actively to assure its adoption [1].

We have argued[7] that before paying the cost of lost confidentiality, we must explore technical alternatives that enable convenient access to data but still protect privacy. In the earlier publication, we described one instance of the use of public key cryptography to create institution-specific identifiers that could be nonetheless be used to link patient records, with proper authorization, across multiple institutions. We suggested that similar mechanisms could be used to implement a wide range of identification systems responding to different societal goals. HIIDIT represents our attempt to move in this direction.

The HIIDIT Identifier Framework

As mentioned above, HIIDIT is not an identification system but a generator of identification systems. It enables the health information systems designer to choose between a large range of implementations of different social/security policies regarding identification of patients and their records. HIIDIT makes no commitment to a particular social policy but it does define the major dimensions of the properties of any identification system and provides mechanisms for implementing various identification systems located in different loci within these dimensions. We list below, just a few of the dimensions that are encompassed by HIIDIT.

1. Directory locus.

With one or more patient identifiers for patients seen at one or more institutions, there can be one or more directories to link patients to these institutions. These directories are necessary if longitudinal medical records are to be generated across institutions and if multi-center studies are to be effective. Just where the directories will be located and who controls them will determine the degree of patient consent in information access, the performance of access and update procedures as well the inconvenience of access for providers and researchers. Examples of who may hold and control the directory include: 1) Patients. 2) Provider. 3) Provider organization. 4) Trusted escrow third party. 5) Governmental authority.

2. Scope of Identification

For some purposes and societal goals, a single identifier for each patient may be desirable. With different premises about ownership of patient information and the role of the patient in enabling or allowing collation and distribution or their data, multiple identifiers may be desirable. These schemes, with appropriate directory loci (see above), all permit the maintenance of longitudinal medical records across multiple institutions and national outcomes studies. Scope of identification in fact represents two orthogonal dimensions: the geographical or organizational scope of the identification (e.g. national or a single institution study) and the nature of the data linked to a particular identifier (e.g. the entire patient record, billing information, sexual history, or just the address).

3. Certifying Authority

Patient identifiers must be issued by a Certifying Authority (CA) that certifies, with varying degrees of authority and credibility, that the identifier does indeed correspond to a particular patient. Note that the CA can provide the directory locus for an identifier but also may not. For example, a governmental agency could be the CA but the directory loci could be the institutions at which the patient had received care or the patients themselves. Examples of some of the CA's that HIIDIT has to support include: 1) Patient. 2) Non-provider patient-driven local or regional authority. 3) Healthcare provider (institution). 4) Insurer. 5) Government

4. Scope of Identifier Secrecy

Assumptions about how important it is to keep a patient identifier confidential and to whom such identifier(s) should be disclosed vary considerably in this country and others[8,9]. HIIDIT permits the specification of the scope of secrecy of identifiers and provides the cryptographic tools to enforce this specification. The scope of secrecy can be, and often will be distinct from the directory locus. This can be illustrated by considering an information system for a multi-center study. In one such study, each institution at which the patient was studied or cared for has its own ID for the patient. To get the patient's identify from the institutional ID at each institution, one needs the institutional secret key; but in order to find out at which institution the patient was seen, one needs the study center's directory. That is, the scope of secrecy would be with the patient's healthcare institution(s) but the directory locus would be at the study center. Examples of various scopes of secrecy include: 1) Just the patient.

2) Patient & family, friends or guardians. 3) Provider. 4) Class of Providers. 5) All providers. 6) Healthcare institution. 7) Insurer. 8) Government. 9) Combinations of the above

The above dimensions are only a few of the dimensions of identifier generation that HIIDIT must cover. We wish to emphasize that HIIDIT does not cover the related but broader task domain of the use and maintenance of master patient indices (MPI), which is being addressed by larger efforts such as the MPI Workshop out of Los Alamos National Laboratories. The particular system of identifiers used by an MPI is only a subset of the MPI's functionality.

SDSI—HIIDIT's Cryptographic Infrastructure

HIIDIT uses the Simple Distributed Security Infrastructure proposal[4] (SDSI) to provide its underlying cryptographic services. SDSI is intended to allow implementation of a large variety of cryptographic systems based on a small set of common principles. The most important for our work are:

Principals (the individuals and institutions that partake in information exchange) are represented by their *public keys* in an RSA public-key cryptosystem. These keys may be used both to encrypt secrets to be sent to the principal and to verify the digital signature on messages received from the principal.

Each principal can create and share with others *local name spaces* with which she can refer to other principals. HIIDIT relies on these abilities to refer to the set of patients at a site and to allow references at one institution to others where the patient may have been seen.

SDSI certificates may include information on membership of an individual in a group. HIIDIT uses this ability, for example, to represent a patient's membership in a hospital plan within the patient identifier.

An Example

We take as our example application the design of a regional genomic database, which we use to illustrate HIIDIT's capabilities and limitations. We do not necessarily advocate the particular application design, though it is defensible. At this early stage in its development, HIIDIT has not yet been deployed for any clinical systems in production. The example databases were modeled after existing databases but were populated by pseudo-randomly generated data.

Genomic databases are growing in popularity as multiple disease-specific regional studies (e.g. regional oncology groups) ally their efforts with national genomics centers (e.g. the recent cancer genetics network initiative funded by the National Cancer Institute). These bioinformatics databases have begun to generate a lot of confidentiality concerns. Such data may accurately predict the risks that a patient and others in her family may have or develop grave (and costly) diseases, and these predictions may occur without any consent by the patient (or her relatives) for donation and analysis of their DNA.

For the purpose of this example, we will define multiple *source sites*, the local institutions (e.g. hospitals or practices) where the patients are usually seen for their care. It is at these source sites that patients provide blood samples for analysis and consent to inclusion in the genomic database. The source sites have *source* databases which contain identified (e.g. name, address) information about the patient. The processing of the blood samples is done by a third party that specializes in mass production. The results of sequencing a particular cancer-predisposing gene are sent to a *central study site* which has its own *study* database, but the blood sample itself is returned to the source site from which it originated. That is, none of the patient's blood is stored in the study database and therefore the information available to the central study site is only the information conveyed by the source sites and the sequence information from the third party. Additional sequencing of the patient's data would therefore require obtaining a sample from the source sites.

With this context, we define a set of common goals that would have to be agreed upon by the source and central study sites. For purposes of illustration we choose the following set of goals: 1) Only data that are duly authorized for release from the source site are entered into the study database. 2) The identity of the patient should not be trivially obtainable just by looking at that patient's study database record in isolation. That is, the study database is *anonymous* and the central study site should operate without knowing the patient's identity. 3) It should be practically impossible to read the sequence data in the study database without approval by the institutional review board of the central study site. 4) It should be possible to reliably add new information obtained from the source sites to a patient's record in the study database without requiring that the patient be identified to the central study site. 5) If the central study site's ombudsman agrees to it, she will be able to decode the identity of the source site from which a patient came from, but not the local (source site) identifier for that patient. This will allow the source sites, with consent of their local IRB, to identify the patient (e.g. to ask them additional clinical questions). That is, it becomes possible but non-trivial to find the patient's identity and then ask them for more information or for tissue or blood samples.

Before we describe the HIIDIT scheme designed to meet the above goals, it is important to underline the limitations of this approach to protecting privacy. Any cryptographic scheme, no matter how much protection it provides against attempts to "crack" encrypted or securely signed messages, remains vulnerable to subversion by non-cryptographic methods. For example, if the private key of an ombudsman is distributed widely throughout the central study site, then any data encrypted with the ombudsman's public key immediately can be decrypted. As noted in the medical privacy literature[10], "insider" access remains the most prevalent conduit for breach of medical record privacy. Cryptographic schemes such as the one described below will only inhibit, not prevent breaches by insiders.

The viability of truly *anonymous databases* is also questionable. A truly private, untraceable patient identifier lacks "face" identification. That is, just by examining or manipulating the ID or looking it up in other databases, one cannot obtain more information about the patient (as opposed to a picture ID that might easily identify the individual). However, with enough patient data, even an anonymous database (e.g. without a unique key, without name or date of birth) can be re-identified by combining its partial information with other databases. This has been demonstrated in detail by Latanya Sweeney in her work on re-identification[11]. Nonetheless, de-identification of databases makes it far more difficult and costly to look up details about a patient, and reduces the likelihood of accidental or non-malicious investigations. To protect against re-identification or to protect particularly sensitive material (as in the genetic sequence strings of the example), encryption of the content itself (in this case the genetic code) may be necessary.

Use of HIIDIT in the Example

In the sense of the SDSI definitions above, the HIIDIT user first specifies who the principals are (each represented by a public key). These are the patient(s), the source site(s), the providers at the source site(s), the source sites' IRB's, the central study site, the central study site IRB, and the central study site ombudsman.

HIIDIT then needs to be told the locations of the source site and central study site patient directories and who controls them and their scope of secrecy. In this instance, there are patient directories at each source site controlled by the care providers at that site, and a patient directory at the central study site controlled by the study ombudsman. The scope of secrecy is restricted to those who control their respective patient directories. Additionally, HIIDIT allows the specification of sensitive data items that should be encrypted within the scope of secrecy. In this case these are the strings that represent the DNA sequence information stored in the study database.

Also, HIIDIT must be told who are the certificate authorities (CA). Because each source site is in the best position to verify the identity of the patient whose data are entered into the source site database, each source site serves as its own CA for patient identifiers as well as the identity of the providers. Similarly, the central study site is well placed to verify the identity of the various legitimate source sites for the study as well as the central study site IRB and the ombudsman. The central study site therefore serves as the CA for these principals. Because of SDSI's ability to export name spaces, using the Internet, HIIDIT's key distribution task is considerably simplified.

To be most helpful in automating the identification system, HIIDIT should be informed of the likely data flow between the principals. In this example, these would include one or more transfers of patient data from the source site to the study site, decryption and retrieval of encrypted DNA sequence data (upon IRB approval), requests to the source site from the central study site for re-identification of the patient. These are given to HIIDIT as template function calls using its limited set of reserved words. For example, the template in this example for exporting patient data from a source site to the study site is:

TRANSFER(<patientID>,<source site>,
<study site>,<patient data>)
Given the above definitions of the principals, the patient directories, the CA's, the scope of secrecy and the required data flow, HIIDIT generates a patient identification system as follows:

HIIDIT generates the patient directory definitions and their corresponding data structures. In this case, the data structures are implemented as relational tables. For each new patient added to a source database, HIIDIT provides the function call to the designated CA for that directory which provides the patient's keys. In our notation, we designate a principal's public key as $\text{Principal}^{\text{public}}$ and the corresponding private key as $\text{Principal}^{\text{private}}$. We denote encryption of a message using one of these keys as Key(message). So, for example, $\text{Principal}^{\text{public}}$ ($\text{Principal}^{\text{private}}$(message)) describes the application of first the principal's private key and then her public key (which returns the original message).

HIIDIT generates the patient identifiers for the source site by encrypting the patient's public key with the source site's private key. That is $\text{Source}^{\text{private}}$ ($\text{Patient}^{\text{public}}$). We'll call this the *Source ID.* Note that the Source ID is provably specific to the source site, because only that site's public key will, when applied to the Source ID, recover the patient's public key.

The identifier for the central database is generated by encrypting the Source ID with the public key of the source site's IRB, appending the source site public key and then encrypting the result with the public key of the central study site ombudsman. This ID is called the Study ID That is:

Study ID = $\text{Ombudsman}^{\text{public}}$(
Source $\text{IRB}^{\text{public}}$(Source ID),$\text{Source}^{\text{public}}$)
This allows the ombudsman to use her private key to decrypt the Study ID and determine which was the source institution for that patient (from the source site's public key, $\text{source}^{\text{public}).}$ However, the Ombudsman cannot determine what the Source ID is. To obtain the identity of a patient, the ombudsman would have to contact the source site and have the IRB of that source site apply its private key to obtain the source ID. Only then could the public key of the source site be used to decrypt the source ID, revealing the patient's public key.

All designated sensitive data (in this case, the DNA sequences) are encrypted with the public key of the principals responsible for the scope of secrecy of data (in this case the Study Site IRB). Consequently, only these same principals, using their private key, could decrypt the sensitive data.

In addition to the functions specific to generating identifiers, HIIDIT uses the given specifications to generate a number of support procedures to help maintain the identification system. These include:

a) Data transfer functions (e.g. generating the Study ID from the Source ID at the source site, appending the patient data and sending the package to the Study Site as an encrypted message over the Internet).

b) Data encryption functions (e.g. automating the encryption of patient data as directed by the scope of secrecy statements).

c) Authenticated encryption and decryption procedures.

d) Generation of cryptographically authenticated transaction logs for all HIIDIT-derived procedures.

Summary and Discussion

We have outlined here one half of HIIDIT's function: the generation of health identification systems. This function rests on four attributes: 1) The adequacy of our definitions of the different dimensions of identification systems. 2) The expressivity of HIIDIT's specification language required to describe a particular set of choices along these dimensions. 3) The efficient compilation of these choices into a set of data structures and support procedures. 4) The use of a capable security infrastructure (e.g. SDSI). Although HIIDIT can meet the tasks set by the example presented in this paper, it has not been used for any "production" clinical application. Consequently, the adequacy of the current set of attributes remains unknown.

Although our initial design of HIIDIT was motivated by the national debate on universal health identifiers, the ability of HIIDIT to be configured to meet a variety of policy goals suggests broader applicability. For example, HIIDIT can be used to configure identification systems for highly sensitive databases whether they include genomic data or social history. Similarly, for sharing data between health care institutions that are competing in the market, HIIDIT can generate identification systems that allow sharing limited data on well-circumscribed populations without otherwise compromising what the institutions believe to be their intellectual property and the private data of their patients.

Acknowledgments

This research was supported in part by the National Library of Medicine (R01 LM06587-01) and an equipment grant from Hewlett-Packard.

References

1. Board of Directors of the American Medical. Informatics Association. Standards for medical identifiers, codes, and messages needed to create an efficient computer-stored medical record. J Am Med Informatics Assoc 1994;1:1-7.

2. Fitzmaurice JM, Murphy G, Wear P, Worpman R, Weber G. Patient identifiers: stumbling blocks or cornerstones for CPRs (computer-based

medical records))? Healthcare Informatics 1993;10(5):38-40.

3. Sweeney L. Replacing personally-identifying information in medical records, the SCRUB system. In: Cimino J, ed. Proceedings of the AMIA Fall Symposium. Washington, DC: Hanley & Belfus, 1996:333-337.

4. Rivest R, Lampson B. SDSI A Simple Distributed Security Infrastructure. Cambridge, MA: http://theory.lcs.mit.edu/~cis/sdsi.html, 1996:

5. Health Insurance Portability and Accountability Act of 1996

6. Institute of Medicine. Health Data in the Information Age. 1994. (Donaldson MS, Lohr KN, eds.

7. Szolovits P, Kohane I. Against simple universal health identifiers. Journal of the American Medical Informatics Association 1994;1(4):316-319.

8. Baitty RL, Jain RB, Hager C, Pope W, Goosby EP, Bowen GS. Protecting confidentiality in a national reporting system for HIV services. Int Conf AIDS, 1993:PO-D36-4384.

9. Gardner RM. Integrated Computerized Records Provide Improved Quality of Care with Little Loss of Privacy. J. Am Med Informatics Assoc 1994;1(4):320-322.

10. Clayton PD, Boebert WE, Defriese GH, et al. For the Record: Protecting Electronic Health Information. Washington, DC: National Academy Press, 1997.

11. Sweeney L. Guaranteeing Anonymity When Sharing Medical Data: The Datafly System. In: Masys DR, ed. AMIA Fall Symposium. Nashville, TN: Hanley & Belfus, 1997:51-55.