Journal of Digital Imaging

Pseudonymization of Radiology Data for Research Purposes

Rita Noumeir, Alain Lemay, and Jean-Marc Lina

Medical image processing methods and algorithms, developed by researchers, need to be validated and tested. Test data would ideally be real clinical data especially that clinical data is varied and exists in large volumes. Nowadays, clinical data is accessible electronically and has important value for researchers. However, the usage of clinical data for research purposes should respect data confidentiality, patient right to privacy, and patient consent. In fact, clinical data is nominative given that it contains information about the patient such as name, age, and identification number. Evidently, clinical data needs to be de-identified to be exported to research databases. However, the same patient is usually followed during a long period of time. The disease progression and the diagnostic evolution represent extremely valuable information for researchers as well. Our objective is to build a research database from de-identified clinical data while enabling the data set to be easily incremented by exporting new pseudonymous data, acquired over a long period of time. Pseudonymization is data de-identification, such that data belonging to an individual in the clinical environment still belong to the same individual in the de-identified research version. In this paper, we explore various software architectures to enable the implementation of an imaging research database that can be incremented in time. We also evaluate their security and discuss their security pitfalls. As most imaging data accessible electronically is available with the digital imaging and communication in medicine (DICOM) standard, we propose a de-identification scheme that closely follows DICOM recommendations. Our work can be used to enable electronic health record (EHR) secondary usage such as public surveillance and research, while maintaining patient confidentiality.

KEY WORDS: Research database, confidentiality, security, privacy, de-identification, pseudonymization, nominative health care data, radiology, medical imaging

INTRODUCTION

Medical image processing consists of defining methods and algorithms to enhance the image quality or to detect specific characteristics and aid in the diagnostic. These methods, developed by researchers, need to be validated and tested with real clinical data. Furthermore, clinical data includes the diagnostic data such as radiology reports. Nowadays, clinical data is accessible electronically and has important value for researchers. However, the usage of clinical data for research purposes must respect data confidentiality and patient right to privacy. In fact, clinical data is nominative as it contains information about the patient such as name, age, and identification number. Evidently, clinical data should be de-identified to be exported to research databases. De-identified data does not allow the researcher to trace back its origin to the patient to whom it really belongs. But, how much information should and can be kept?

A patient being usually examined and followed during a long period of time, the disease progression and the diagnostic evolution represent extremely valuable information for researchers as well. Time series of medical data is thus quite common, but still, a few inference based on temporal evolution are performed automatically. Signal processing mainly focuses on spatial data such as radiography, computed tomography (CT), magnetic resonance imaging (MRI), or positron

From the École de Technologie Supérieure, 1100 Notre-Dame West, Montreal, QC, Canada H3C 1K3.

Correspondence to: Rita Noumeir, École de Technologie Supérieure, 1100 Notre-Dame West, Montreal, QC, Canada H3C 1K3. email: noumeir@ele.etsmtl.ca

Online publication 28 December 2006
doi: 10.1007/s10278-006-1051-4

emission tomography (PET) scans; however, functional and physiological information require analysis based on the temporal evolution of data. Functional brain imaging, either for cognitive experiments or mental disorder diagnosis, is about the most elaborated temporal medical data. Time scale may vary from a few milliseconds during the acquisition to a few days or weeks to observe the functional reorganization of brain activities (plasticity). Diseases such as Alzheimer and strokes also require temporal monitoring.

Another image processing field of interest is breast imaging where spatial signal processing is used to segment and to characterize textural regions associated with masses and densities. Detection of abnormalities is not limited to anatomical features extraction. The temporal evolution of physiological information (from near infrared spectroscopy imaging for instance) is also of interest. Research algorithms are needed to measure shape evolution of the breast anatomy, for example. Finally, time series are of great importance for drugs impact evaluation and for the monitoring of medical treatments.

Our objective is to build a research database from de-identified clinical data while enabling the database to be easily incremented over time. That is, exporting to the research database new clinical data, acquired over a long time period, for the same patient, such that data that belongs to the same identified patient in the clinical database, belongs to the same de-identified patient in the research database as well. This is known as pseudonymization. Figure 1 depicts the pseudonymization problem. A patient has many exams over time. Each exam is composed of many related information objects. Patient set continuously increases. Solid lines represent links between information objects that should be preserved. Dashed lines represent nonreversible links to "real world" patients.

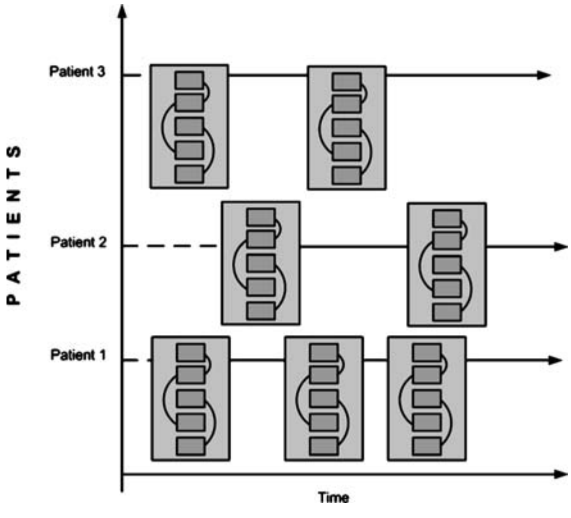

Fig 1. Pseudonymous data for various patients over time.

In this paper, we explore several architectures to enable the implementation of a research database that can be incremented in time. Our main contribution consists in a thorough discussion and analysis of various architectural solutions that allow de-identified imaging clinical data to be exported to a research system. The exported data is identified with a patient's pseudonym enabling the research data set to be incremented in time. As most imaging data accessible electronically is available with the digital imaging and communication in medicine (DICOM) standard, we have proposed a de-identification scheme that closely follows DICOM recommendations. Our work can be used to enable electronic health record (EHR) secondary usage such as public surveillance and research, while maintaining patient confidentiality.

The paper is organized as follows: In the Definitions section, we recall the definitions of de-identification, anonymization, and pseudonymization; in the Architectural solution section, we provide and discuss methods to achieve de-identification, anonymization, and pseudonymization of radiology data; finally, we provide a summary and propose how this work can be used for teaching files and clinical trials.

DEFINITIONS

De-identification

De-identification consists in stripping patient identifiers such as name, address, hospital identification number, from the image headers or substituting a false value for the real identifier.

Anonymization

It is wrong to believe that sensitive medical information is kept confidential simply by de-identification. De-identified information is not likely to be anonymous, as a combination of attributes may enable one to almost uniquely identify a patient by linking this information with a publicly available external source of information. For example, a combination of ethnicity, zip code, and date of birth may be linked with publicly available census data to identify a patient. The Webster's dictionary defines anonymous as:

(1) not named or identified;
(2) of unknown authorship or origin;
(3) lacking individuality, distinction, or recognizability.

De-identification is compatible with the first definition of anonymous. However, to fulfill the third definition, one should ensure that within the de-identified data set, any given combination of feature values matches more than one individual. Therefore, anonymization methods rely on ambiguation of the de-identified data; ambiguation ensures that each patient in the de-identified set cannot be distinguished from at least one other patient with respect to information that may be used for linking. Ambiguation is based on the generalization of some data field by merging bins, that is, by grouping several consecutive numerical values to form intervals. The date of birth field, for example, contains month/day/year information; it can be generalized by dropping the day information. Similarly, people can be grouped by weights of within 5 kg.

Many anonymization systems have been proposed in the literature. They enable the user to query a database either to link its content with other sources or to access it with structured query language (SQL) queries. Their architectural models consist in filtering the query response by generalizing values within fields, suppressing cells,[1,2] and removing extreme outlier information. The ambiguation decisions are based on some confidentiality measures such as bin sizes[3,4], and nonuniqueness[1]. Most methods can also be used for database anonymization before it is released for research use.

The fact that a user can deduce classified information from unclassified information is not a new issue. This inference problem has been studied in the context of statistical databases. A statistical database enables the user to retrieve statistics for a subset of database entries such as the mean, count, and standard deviation. Even though the goal of recent anonymization systems is to publish the data itself rather than its statistics, techniques developed to deal with inference in statistical databases have been very useful for anonymization. Anonymization for privacy protection still faces several challenges[5]. Metrics that measure the degree of relative anonymity and the degree of protection for sensitive information still need to be developed.

In our case, we are interested in exporting images for research purposes incrementally over time. With time, new images will be exported for patients that already exist in the research set. Also, images for new patients will be added to the research database. Therefore, the patient set will increase over time; it is incomplete when the research database is initially created and the first images are exported. All anonymization methods deal with a fixed set of individuals and known characteristic distributions; therefore, anonymization of evolving data sets is still to be studied.

Moreover, we expect the number of patients that will have the same image acquisition history to be extremely small if not null. The image acquisition history for a specific patient is a list composed with the date and type of acquisition. The acquisition history is almost unique per patient, and therefore, constitutes an identifier for the patient when linked back with the initial clinical database source. This does not represent a risk for patient confidentiality as access to the original clinical database is controlled and granted only for authorized users. However, we will use this characteristic to enforce data integrity.

Pseudonymization

Pseudonymization means that a true identifier such as name or patient identification number is replaced by a pseudonym that is unique to the individual but bears no relation to the person "in the real world". Pseudonym cannot therefore be used as a means of identification. This is because in pseudonymization, the information that reveals who the pseudonym relates to will be held securely, and separately, from the data being processed.

There are 2 types of pseudonymization. When the relationship between the pseudonym and the patient identifier is kept securely, usually by a trusted pseudonymization party, the re-identification of the patient is possible, and the pseudonymization is reversible. However, when it is not possible to re-identify the patient, even for the entity that performed the pseudonymization, because the relationship between the pseudonym and the patient identifier is not kept or the pseudonym generation process is not reversible, then the pseudonymization is irreversible. Pseudonymization is necessary to identify information of a given patient consistently over time. A specific patient pseudonym is attached to all de-identified information related to that patient regardless of the information de-identification time.

Pseudonymization is, thus, attaching a unique identifier to de-identified and ideally anonymous information that relate to a specific patient. In practice, the pseudonymization process would be carried out fully automatically. Reversibility may be useful if in accordance with ethical policies and patient consent, some extra information is needed from the patient. For instance, if the patient agrees, he may be notified about the research findings. However, in many cases, the patient is not interested in the research results, and patient re-identification is not necessary for research purposes. For example, research about cerebral plasticity during a postsurgery recovery of the brain functionality may be accomplished with a nonreversible pseudonymization.

Pseudonymization is very common in market and social research, where the personal reference of the collected data is removed and cannot be restored by the recipient. It is usually done by storing separately the respondent address from the response data, with code numbers that are common to both. Address and response data may be recombined temporarily for the purposes of carrying out quality controls on the data collected or to be able to link the response data from different follow-up surveys.

ARCHITECTURAL SOLUTION

De-identification

Our main objective is to de-identify clinical data by creating a new instance that is exported and used in a different database environment. We focus on radiology clinical data and use the DICOM standard. DICOM has recently published a new supplement entitled Supplement 55: Attribute Level Confidentiality (including de-identification).[6] This supplement lists the attributes that need to be protected to provide a minimal level of confidentiality from identification. It proposes to protect the attributes by removing them from the instance and replacing them by a pseudonym or a dummy value. It also provides a mean to encrypt the original attributes and store them inside the de-identified instance along with the key needed for decryption. Of course, the enclosed key alone is not sufficient to recover the identity of the patient. Decryption requires access to the private recipient key that is kept secret with the recipient.

This supplement provides a mean to exchange de-identified DICOM objects in a safe and standard way. We use it to encode the de-identified object. Moreover, this supplement provides a list of attributes that typically need to be protected to provide a minimal level of confidentiality from identification; but to be able to use the information objects in image processing research projects, the value of some attributes are needed, and therefore, will not be protected. Table 1 lists various attributes with their de-identification actions. The de-identification of these attributes is discussed hereinafter:

- Private attributes are removed as their values are meant to be used only by the equipment that have created them, and therefore, are not generally useful in research projects.
- Attributes related to the institution, physicians, and operator are cleared because their values are not normally relevant for research on image processing or aided diagnosis algorithms. However, equipment attributes are kept.
- Study description, series description and protocol information such as the protocol name are not de-identified because their values are important for image processing algorithms. Requested procedure description, scheduled procedure step description, performed procedure step description are not required to be de-identified by DICOM, and we do not de-identify their values for the same reason.
- Additional patient's history and patient comments are free text attributes. Their values are rarely checked by clinical applications. Even though they may contain valuable information, we believe that their value should be cleared in the image information objects.

Table 1. De-identification of DICOM Attributes

Attribute Name	Tag	Action	Comments
Station Name	(0008,1010)	Removed	Their values are only relevant to the equipment
Device Serial Number	(0018,1000)		
Institution Name	(0008,0080)	Removed	Their values are not normally relevant for research on image processing or aided diagnosis algorithms
Institution Address	(0008,0081)		
Referring Physician's Name	(0008,0090)		
Referring Physician's Address	(0008,0092)		
Referring Physician's Telephone Numbers	(0008,0094)		
Institutional Department Name	(0008,1040)		
Physician(s) of Record	(0008,1048)		
Performing Physicians' Name	(0008,1050)		
Name of Physician(s) Reading Study	(0008,1060)		
Operators' Name	(0008,1070)		
Admitting Diagnoses Description	(0008,1080)		
Derivation Description	(0008,2111)		
Additional Patient's History	(0010,21B0)	Removed	Free text attributes that are considered not relevant for research purposes
Patient Comments	(0010,4000)		
Image Comments	(0020,4000)		
Other Patient Ids	(0010,1000)	Removed	Attributes that can be used to identify a patient but are not relevant for research
Other Patient Names	(0010,1001)		
Medical Record Locator	(0010,1090)		
Ethnic Group	(0010,2160)		
Occupation	(0010,2180)		
Patient's Birth Date	(0010,0030)	Ambiguated	The day is set to 01
Instance Creator UID	(0008,0014)	Replaced	Generated UIDs should preserve the object hierarchy and their respective relationships
SOP Instance UID	(0008,0018)		
Referenced SOP Instance UID	(0008,1155)		
Study Instance UID	(0020,000D)		
Series Instance UID	(0020,000E)		
Frame of Reference UID	(0020,0052)		
Synchronization Frame of Reference UID	(0020,0200)		
UID	(0040,A124)		
Storage Media File-set UID	(0088,0140)		
Referenced Frame of Reference UID	(3006,0024)		
Related Frame of Reference UID	(3006,00C2)		
Patient's Name	(0010,0010)	Blanked	Attributes that can be used to identify a patient but are not relevant for research
Patient's Age	(0010,1010)		
Patient's Birth Time	(0010,0032)		
Accession Number	(0008,0050)	Changed or Blanked	Attributes that are not relevant for research
Study ID	(0020,0010)		
Requested Procedure ID	(0040,1001)		
Scheduled Procedure Step ID	(0040,0009)		
Performed Procedure Step ID	(0040,0253)		
Request Attributes Sequence	(0040,0275)		
Patient ID	(0010,0020)	Changed	Replaced by a pseudonym
Study Description	(0008,1030)	Unchanged	Their values are important for image processing algorithms
Series Description	(0008,103E)		
Protocol Name	(0018,1030)		
Patient's Sex	(0010,0040)	Unchanged	Attributes that may be relevant for research algorithms
Patient's Size	(0010,1020)		
Patient's Weight	(0010,1030)		
Requested Procedure Description	(0032,1060)	Unchanged	Their values are important for image processing algorithms
Scheduled Procedure Step Description	(0040,0007)		
Performed Procedure Step Description	(0040,0254)		

- All unique object identifiers, the UIDs, are replaced with new unique identifiers. The format of the generated UID is well defined and wholly constrained by DICOM. It is globally unique, and its creator is the de-identification entity. Moreover, the generated UIDs should preserve the object hierarchy and their respective relationships. This latter constraint presents great implementation challenges and will be further discussed later.
- Accession number and study ID are required to be de-identified by DICOM. We replace their values with new unique identifiers. The new identifiers are not globally unique as with UIDs, but they should be unique in the exported data domain. Therefore, the same challenge as for preserving the object hierarchy and their respective relationships should be addressed. Similarly, although not required by DICOM, we replace the requested procedure ID, the scheduled procedure step ID, and the performed procedure step ID.
- patient's name, patient's birth time, other patient Ids, other patient names, medical record locator, ethnic group, and occupation are cleared. An empty value field contains a specific value "EMPTY VALUE" to prevent some database engines to fail. On the other hand, the patient's birth date is ambiguated by keeping the month and year and by changing the day to 01. This is equivalent to encoding the patient's age. Patient's sex, patient's size, and patient's weight are kept as these attributes may be relevant for research algorithms. Patient ID is replaced with a pseudonym.
- De-identification of structured reports content has not been studied. Therefore, content sequence (0040,A730) is removed.

Object identifiers, globally unique such as UIDs or application unique such as IDs, should preserve the objects' hierarchy and relationships. This is a difficult challenge. If, for example, a series is exported, its image instances will contain the study instance UID of the study to which the series belongs. If, a new series that belongs to the same study is exported later in time, the de-identified object instances of the newly exported Series should contain the same study instance UID. The challenge arises from the fact that the de-identification entity should remember the mapping between the study instance UID of the clinical identified study and the study instance UID of the research de-identified study. This mapping could be used to reidentify a patient; it is thus a threat for patient confidentiality and should be adequately protected. This issue is similar to pseudonymization. Solutions that enable a single study export over time are similar to those discussed later for pseudonymization. However, as the DICOM UID has a specific format, it cannot be generated by applying a reversible or a one-way function unless the function is applied to generate the last UID subcomponent only. However, saving the mapping between the old UID and the new one provides a relatively easy way to implement the solution. Whether UIDs need to be changed or whether UID mapping needs to be protected depends on the desired confidentiality level. UIDs are globally unique, and one can identify the patient from an image UID. However, this requires having access to the clinical database. In any case, the methods described in this paper to create a patient pseudonym can be applied to create new UIDs.

A simple solution is achievable if we constrain the de-identification application to export a complete study. The challenge is transformed. In fact, it is almost impossible to know when a study was completely received, especially if the de-identification application receives its information over the network with the DICOM protocol. Moreover, the clinical application can add images or other information objects to the study at any time. To avoid this problem with this simple solution, a study should be exported only once.

The main disadvantage of this solution concerns references to objects outside the study that are broken. We believe that this limitation has relatively small impacts because practically, references are within the same study except for prior reports. Nevertheless, future researcher input will certainly help in assessing the impact of this design decision.

To export a study only once, the clinical system needs to keep track of what information objects have been exported. Similarly, the receiving system needs to know that all expected objects had been received. Using a manifest can fulfill both needs. A manifest is a concept recently introduced in DICOM. It is based on the Key object selection with specific semantics. Basically, it is a DICOM object that only contains references to other objects such as images, reports, or presentation states. The purpose of a manifest is to create a list of objects that are transferred together. Therefore, the clinical system can create a manifest for every exported study. Furthermore, it can archive that manifest as a DICOM object. By getting a manifest, the receiving system can identify what information objects to expect and determine whether they have been completely received or not. There are some constraints that apply on a manifest that are not discussed here. However, the most important one is that a single manifest can reference objects that belong to a single patient.

On the other hand, information such as region of interest, text, or graphical layers can be associated with specific images. This type of information is encoded in separate DICOM objects, such as presentation state or key object selection that usually belongs to the same study as the image and reference the image using its UID. These DICOM objects are subject to the same de-identification, anonymization and pseudonymization steps as DICOM images. Moreover, changing their UID as described earlier ensures that the relationship between the image and objects referencing it is preserved. Obviously, text may contain information that could lead to patient identification, but we have not investigated, in more detail, this identification hazard and how to mitigate it.

Finally, to de-identify images from modalities such as ultrasound which burn patient and institution information into the pixel data, special attention is required. The image region, where the information is located, needs to be removed by setting its pixels to the background color. We have not addressed this problem, but if the region location and size are fixed per equipment, then the information removal would be achieved relatively easily; however, if the region size changes, for example, optical character recognition (OCR) methods or user input would be necessary.

Anonymization

The anonymization level achieved by de-identifying specific patient attributes and ambiguating patient birth date is hard to evaluate as existing anonymization techniques operate on a fixed patients set, while our patients set may be continuously increasing over time. Effective data ambiguation requires prior knowledge of the data distribution. In this paper, we have not studied anonymization of data sets that increase over time. Even though this seems to be an important step in the general pseudonymization process, more investigations are needed to achieve and quantify anonymization of data accumulated during a long time period.

Pseudonymization

We need to ensure that we always use the same pseudoidentification number in the de-identified information object for the same real world patient. This new patient identification number, the pseudonym, cannot be used to identify the real world patient.

There are two kinds of pseudonyms:

- one-way pseudonyms that cannot be reversed;
- reversible pseudonyms that enable patient re-identification.

Each pseudonym type requires a different solution. The choice of the pseudonym type depends on the research project requirements: it is mainly dictated by the patient re-identification need. Pseudonymization is applied to change the patient ID into a new one, a pseudonym. Therefore, patient ID is not changed until the pseudonymization step is performed. Moreover, pseudonym based on hash function can be used as hash function input such as patient ID, study description, and study date are not changed during the de-identification or anonymization steps.

Reversible Pseudonym

Reversible pseudonym can be implemented either by saving the mapping between the patient identification number and the pseudonym or by saving the parameters of a reversible mapping function that can be applied, either on the patient identification number to generate the pseudonym, or on the pseudonym to generate the patient identification number.

As reverse mapping enables patient re-identification, it should be protected and made possible only after an explicit consent by the patient. Therefore, if mapping between the patient identification number and the pseudonym is saved, this persistent mapping can be encrypted with a key that is kept secret. Similarly, if a reversible function is used to generate the pseudonym, the parameters of that function can be kept secret. Figure 2 shows a pseudonymization system that is responsible for mapping the patient identification number to a pseudonym and vice versa. It also shows a de-identification system that is responsible for blanking out specific fields or replacing others while ensuring consistency between multiple information objects of the same study. We have preferred to group the de-identification system with the anonymization system that is responsible for data ambiguation, as a minimum level of anonymization is reached in our case by blanking many of the identifying fields. The research system receives pseudonymous information objects. Received information can be indexed in its own database in which case the stored data cannot be linked back to the real patient.

Data flow for generating the pseudonym is depicted with plain arrows while data flow to reverse the pseudonym is depicted with dashed arrows. The pseudonymization system is shown as a separate system from the de-identification and anonymization system. The pseudonymization service can therefore be outsourced. But in most situations, de-identification, anonymization, and pseudonymization will be implemented within the same system. The mapping function is easier to manage and maintain because less information needs to be saved. It is discussed in more detail hereinafter. As the mapping function should be hard to reverse by an unauthorized person, a secret key algorithm provides an appropriate solution. Therefore as a result, the pseudonym is an encrypted patient identification number, and to reverse it, a secret key is necessary.

We propose to use the DES encryption algorithm to generate the pseudonym from the patient identification number.[7] DES is a symmetric algorithm that takes 64 bits for input and generates 64 bits for output. The same key is used for encryption and decryption. The algorithm security is granted by the key secrecy. The output length of DES makes it suitable for encoding a DICOM Patient ID field as this field is constrained to be of maximum, 64 characters.

In case where information from multiple institutions need to be exported and used in a single research data pool, the issuer of patient ID can be used in conjunction with patient ID to ensure the uniqueness of the identification. The issuer of

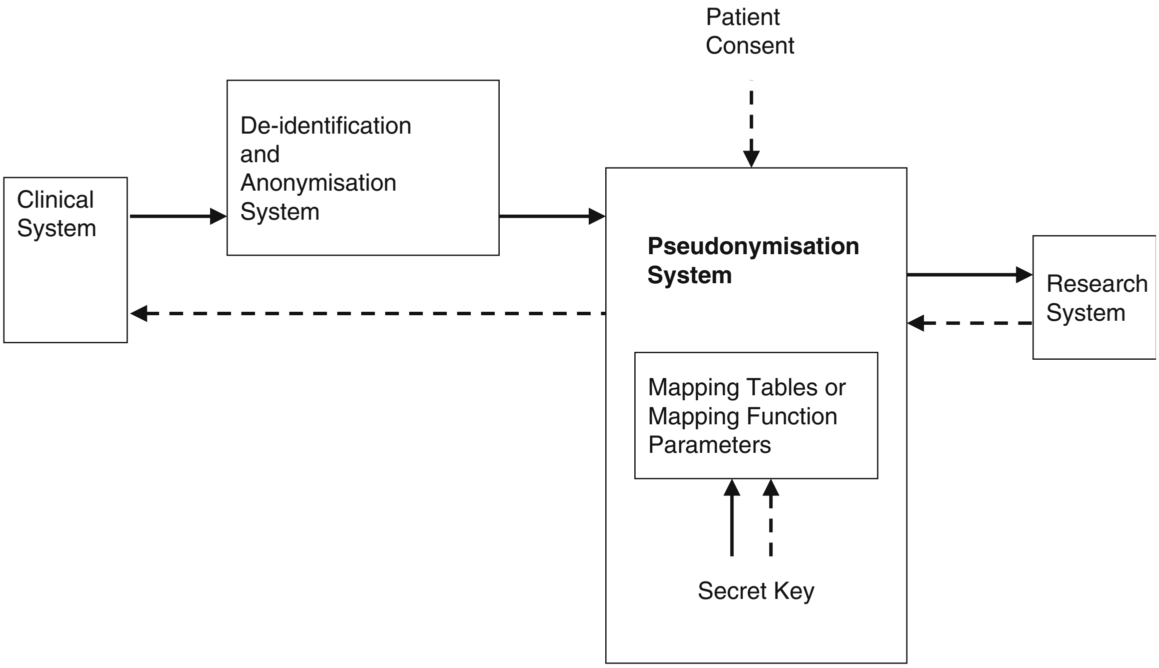
Patient Consent
De-identification and Anonymisation System
Clinical System
Pseudonymisation System
Research System
Mapping Tables or Mapping Function Parameters
Secret Key

Fig 2. Reversible pseudonymization system and its interaction with other systems.

patient ID, which is also constrained to a maximum of 64 characters in DICOM, can be encrypted with the same method used for patient ID to generate an issuer pseudonym.

Re-identification should be restricted. Rigorous procedures implemented under the patient control are essential. A patient may authorize a pseudonymous submission of his clinical information for research purposes. He may preauthorize reidentification, restrict it to some conditions or disallow it regardless of any research outcome. Although we have not studied patient consent management, we think that a consent manager is needed. It is responsible for managing the patient consent and authorizing the pseudonym re-identification. Consent management is a very recent requirement in electronic health record (EHR) architectures; it still needs a thorough investigation.

DICOM protocol is used to exchange images and other information objects between systems. As de-identification is to be performed on all instances from a specific study at the same time to ensure UID consistency, two paradigms are possible. A query/retrieve from the de-identification system to the clinical data source, a pull model, can be implemented with any existing DICOM archive. It places the responsibility of issuing the export action, which should be constrained to a specific time period to ensure that images are exported only once, on the de-identification system. The second paradigm is based on a push model where the clinical data source issues a simple DICOM store to export instances. The push model has been chosen for implementing the Integrating the Healthcare Enterprise (IHE) teaching file and clinical trial profile[8]. When a manifest that references all exported information objects is also exported, as prescribed by IHE, it helps keep track of exported objects and ensures that all objects of the same study are de-identified and pseudonymized at the same time.

After the data is de-identified, it is submitted for ambiguation and then submitted for pseudonymization. The pseudonymization service can be performed inside the institution within a single secured network, or it can be performed by a trusted third party outside the institution. Audits and nodes authentication are essential for security. They can be achieved by implementing the Audit Trail and Node Authentication Profile (ATNA) defined by IHE[9]. Moreover, encryption of the clinical information would increase the confidentiality level as described in Pommerening and Reng[10].

To our best knowledge, articles on pseudonymization are extremely rare. Only Pommerening and Reng[10] and its referenced articles have addressed this problem. High-level possible model architectures are proposed for a secondary use of EHR information in Pommerening and Reng[10]. Five models are proposed in all. They differ in the number of data sources involved, the type of secondary use that is one-time secondary use or long-term data accumulation, and the need for patient re-identification. Pseudonymization is performed by a trusted third party service that is separated from the data source and the secondary use data pool. Other third party services are involved in some of the models. A patient identifier cross-reference service is used to map identifiers from multiple sources and to provide an identifier to be used in the pseudonymization. The major characteristic of models in Pommerening and Reng[10] is the confidentiality level that is achieved by requiring the data source to use asymmetric encryption to encrypt the clinical data with the key of the final receiver before sending the non-encrypted identifier with the encrypted data to the pseudonymization service. The pseudonymization service cannot read the clinical data. It uses the identifier to generate the pseudonym. The receiver receives a pseudonym with the encrypted data. It decrypts the data but cannot reverse the pseudonym.

One-Way Pseudonym

One-way pseudonym can be implemented with the help of a hash function. A hash function $H(M)$ maps a string M with an arbitrary length onto a fixed-length hash value $h=H(M)$, where h is of length m. Hash functions are widely used in database indexing, in cryptography, and digital signature. With a one-way hash function, it is extremely hard to reconstruct the input string from its hash value in a reasonable amount of time. One-way hash functions have the following characteristics:

given M, it is easy to compute h;
given h, it is hard to compute M;
given M, it is hard to find another message M' such that $H(M)=H(M')$

One-way hash functions are used in cryptography because they provide a unique fingerprint of a message, and at the same time, enforce the property that the knowledge of the fingerprint does not enable the attacker to guess the original message.

Undesirably, a hash value is prone to collisions. In fact, two input messages may map to the same output hash value. Therefore, in practice, collision-resistance hash functions that minimize the probability of collisions are needed. The collision-resistance requirement is satisfied when it is hard to find two random input messages that map to the same hash value. There are many one-way hash functions in use,[7] such as SHA-1, Snefru, MD4, and MD5. Figure 3 shows the information flow between various systems involved in a one-way pseudonymization.

SHA is largely used; it generates a 160-bit-long hash value. MD5 generates a hash value of 128 bits long. SHA is more secure than MD5 because it generates a longer hash value. We propose to apply SHA on the patient identification number to generate a pseudonym that will be encoded as a DICOM Patient ID.

However, as we propose to use the pseudonym as the new patient identification number, collisions introduce inconsistency by assigning the same pseudonym to different patients. Therefore, it is important to use a hash function in a way that minimizes the upper bound on the probability of collisions and/or to detect collisions, preventing data from different patients from being grouped with an identical pseudonym.

The work in Peyraviana et al[11] studied the relationship between: r the length in bits of an input message, $Q_{\max}$ the maximum permissible value of the collision probability expressed as 2^{-t} for some t, and k, the length in bits of the hash function values. It is found that this relationship satisfies $k>r+t$. Therefore, one can choose the length of the input message to achieve a collision probability that is kept below a given value.

In addition, we propose to use the patient history to detect hash collisions. Patient history in a radiology context is a list of radiology procedures for which images were acquired. But, as we are working with DICOM data at the persistency level, we consider the patient history as a list of studies where each study is identified by a study date and a study description. Patient history is practically unique per patient, as the probability of having two patients with the same history is almost zero. Consequently, to detect hash collisions, we propose to perform the following process while exporting the data.

- For a specific patient, query for all studies that are exclusively older than the study that is being exported.
- Calculate the history hash and attach it to the images in a secondary field such as other patient ID (see Figure 4).

Upon the reception of de-identified and pseudonymous images, the receiving application would calculate the history hash value based on its own internal patient history and would reject the study if the calculated history hash is found different from the encoded history hash value.

Then again, hash values are vulnerable to the dictionary attack where a malevolent user can use software to guess the patient ID by applying the hash function on all possible values. A common surprisingly successful example of such attack consists in finding the password of a user's UNIX account[12]. Pseudonym generation that is based on hash functions is particularly vulnerable to such

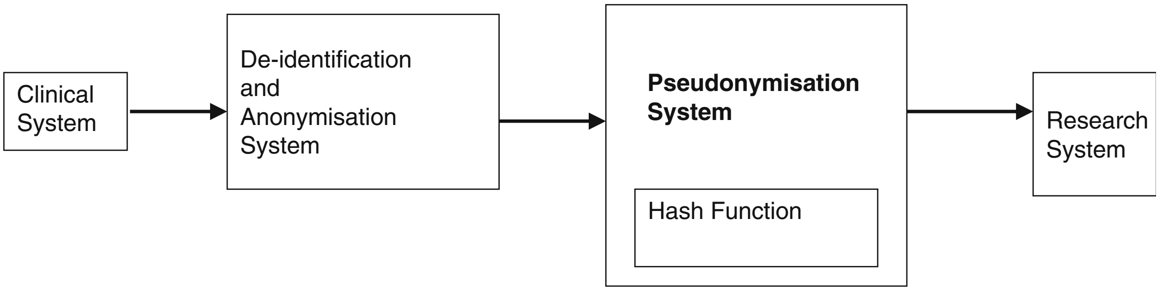

Fig 3. One-way pseudonymization system and its interaction with other systems.

attacks because of the data used as input to the hash function. Unlike a password file whose data entropy is relatively high,[13] the patient ID has usually a well-known format. As an example, the patient ID could be 12 digits having the following format:

three positions representing the first three letters of the family name ("AAA" to "ZZZ");
one position for the first letter of the surname ("A" to "Z");
six positions for the birth date in the format ("YYMMDD");
two last positions for an arbitrary sequential number.

A hash value, generated from such data, is prone to the known-plaintext attack. A malicious person could easily generate a file of patient ID from data publicly available like a phone directory. To make this kind of attack harder, the hash function can be applied on a combination of several patient information such as patient ID and patient name. However, this implies that some patient information would be available at the pseudonymization system and that patient information is consistent throughout the clinical database. Consistency is not guaranteed in real world. Patient name, for instance, may be recorded inconsistently.

Other methods to overcome dictionary attack are described in the literature. One technique consists in concatenating a random value, called salt, to the input data before processing it by the one-way hash function. This method is not infallible. However, it makes it harder to an attacker by forcing him to try encryption on each salt value.[10] Another technique consists in using a message authentication code (MAC). MAC-based hash function requires a secret key.[10]

CONCLUSION

The use of clinical data for research purposes is a vital need shared by all medical researchers. Availability of this extremely valuable data to sustain scientific progress is in conflict with the patient right to privacy. Approaches that offer reasonable protection of patient clinical information confidentiality while enabling researchers to access the large and rich medical information pool are essential.

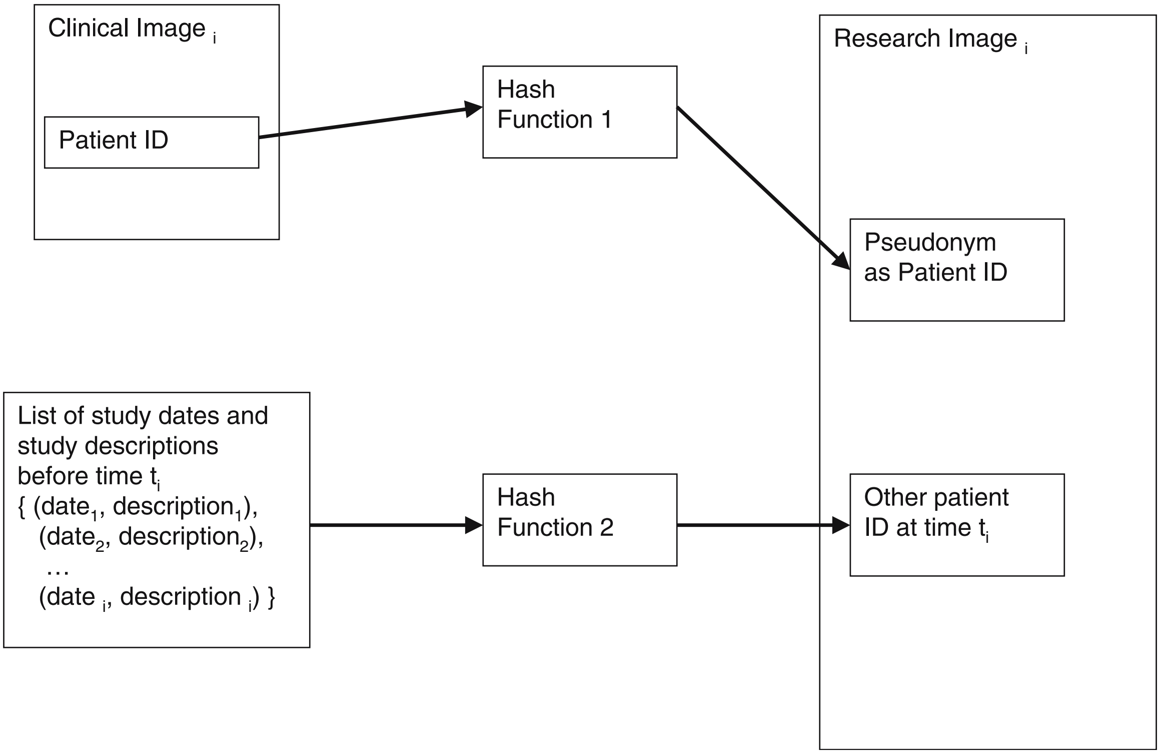
Clinical Image $_i$
Patient ID
Hash Function 1
Research Image $_i$
Pseudonym as Patient ID
List of study dates and study descriptions before time t_i
{ (date$_1$, description$_1$), (date$_2$, description$_2$), … (date $_i$, description $_i$) }
Hash Function 2
Other patient ID at time t_i

Fig 4. Using study history to detect hash collisions.

We have presented and discussed de-identification of radiology clinical data. We have also proposed two architectural models that achieve radiology clinical data pseudonymization to collect data for research over a long period of time. These models are based on the implementation of reversible and one-way pseudonyms. Pseudonyms are used as the patient identification number in DICOM images and other information objects. Reversible pseudonyms are based on a DES symmetric encryption algorithm, whereas one-way pseudonyms are based on the SHA hash function.

Pseudonymization of radiology data is useful for teaching files and clinical trials. Teaching file is very important for medical education. It requires a radiologist to access de-identified images and other radiology information that have been previously identified as a pertinent teaching case for authoring and documenting that case. Therefore, images and related information need to be de-identified before their transfer to an authoring system. Our solutions can be used for the de-identification of teaching file information. While changing the patient ID is essential to ensure de-identification before the data is exported, assigning the same pseudonym to the same patient over time, may be useful for teaching files but is not crucial for most cases.

On the other hand, clinical trials have similar needs to teaching files, in that images and related information need to be de-identified and exported to another system that is outside of the institution security network. Clinical trials may involve multiple organizations; it requires, in addition to pseudonymization, the insertion of trial-specific attributes such as information about the trial protocol. Our solutions can be used for the de-identification and pseudonymization of radiology clinical trial data. Trial-specific data insertion can be achieved by a separate system.

IHE teaching file and clinical trial export (TCE) profile defines methods to flag relevant images and other information objects on radiology systems such as acquisition modality or work station, to record any additional information and to transfer them to an exporter/de-identifier system and then to a teaching file authoring or clinical trial receiver system. While IHE TCE profile describes the pseudonymization process, and the cases where it should be applied, no information is provided on how it can be achieved. The methods presented in this paper help implementers choose the pseudonymization architecture that best fulfills their requirements. Although teaching files and clinical trials have similar needs in terms of de-identification and pseudonymization, gathering radiology data for the same patient during a long period of time, his lifetime probably, is unique to research applications.

We finally conclude by pinpointing a significant challenge that faces anonymization of medical images. In fact, volumetric images can be segmented and three-dimensionally displayed to represent the patient's skin. Even though the data could have been completely de-identified, the reconstruction of the patient's face from the sole pixels, even partially, leads inevitably to the patient identification.

ACKNOWLEDGMENTS

This work was supported by the Natural Sciences and Engineering Research Council of Canada.

REFERENCES

1. Chiang YC, Hsu TS, Kuo S, Liau CJ, Wang DW: Preserving confidentiality when sharing medical database with the Cellsecu system. Int J Med Inf 71(1):17–23, 2003
2. Ohrn A, Ohno-Machado L: Using Boolean reasoning to anonymize databases. Artif Intell Med 15(3):235–254, 1999
3. Hundepool A, Willenborg L: μ- and τ-ARGUS: software for statistical disclosure control. Proc. Third International Seminar on Statistical Confidentiality. Bled, 1996
4. Sweeney L: Datafly: a system for providing anonymity in medical data. In: Lin TY, Qian S (eds). Database Security XI: Status and Prospects. New York: Chapman and Hall, 1998
5. Ohno-Machado L, Silveira PSP, Vinterbo S: Protecting patient privacy by quantifiable control of disclosures in disseminated databases. Int J Med Inf 73(7-8):599–606, 2004
6. Digital Imaging and Communications in Medicine (DICOM), Supplement 55: Attribute Level Confidentiality (including De-identification), National Electrical Manufacturers Association, 2002
7. Schneier B: Applied Cryptography Second Edition: protocols, algorithms, and source code in C, John Wiley, 1996
8. Radiology, IHE Technical Framework, http://www.ihe.net
9. IT Infrastructure, IHE Technical Framework, http://www.ihe.net
10. Pommerening K, Reng M: Secondary use of the electronic health record via pseudonymisation. In: Bos L, Laxminarayan S, Marsh A (eds). Medical Care Compunetics 1. Amsterdam: IOS Press, 2004, pp 441–446
11. Peyraviana M, Roginskya A, Kshemkalyanib A: On probabilities of hash value matches. Comput Secur 17(2):171–174, 1998
12. Klein DV: 'Foiling the Cracker': A Survey of, and Implications to, Password Security. Proceedings of the USENIX UNIX Security Workshop, 1990, pp 5–14
13. Feldmeier DC, Karn, PR: UNIX Password Security-Ten Years Later. Advances in cryptology-CRYPTO '89 Proceedings, Springer, Berlin Heidelberg New York 1990, pp 44–63