Published in final edited form as:

Acad Radiol. 2007 June ; 14(6): 757–764.

A Web-based Interface for Communication of Data between the Clinical and Research Environments without Revealing Identifying Information

Peyton H. Bland, Ph.D., Gary E. Laderach, B.S., and Charles R. Meyer, Ph.D.

Department of Radiology, University of Michigan Health System, BSRB, Room AC51, 109 Zina Pitcher Place, Ann Arbor, MI 48109-2200

Abstract

Recent health care policies and regulations have impacted the manner in which patient data – especially protected health information (PHI) – are handled in both the clinical and research settings. Specifically, it is now more challenging to obtain de-identified PHI from the clinic for use in research while adhering to the requirements of this new environment. To meet this challenge, we have devised a novel web-based interface that facilitates the communication of data (for example, biopsy results) between the clinic and research environments without revealing PHI to the research team or associated research identifiers to the clinical collaborators. At the heart of the scheme is a web application that coordinates message passing between the researchers (in general, the requesters of de-identified PHI) and clinical collaborators (who have access to PHI) by use of a protocol that protects confidentiality. We describe the design requirements of this communication scheme and present implementation details of the web application and its associated database. We conclude that this scheme provides a useful communication mechanism that facilitates clinical research while maintaining confidentiality of patient data.

Indexing words

PHI; protected health information; HIPAA; DICOM; anonymization; de-identification

Background and Problem Description

Recent years have seen significant changes in the management of patient data in the medical setting. These changes have been heavily influenced and, in some cases, dictated by regulations such as the rather broad mandates of the Department of Health and Human Services. Specifically, certain provisions of HIPAA (Health Insurance Portability and Accountability Act as detailed in DHHS 45 CFR Parts 160 to 164 [1]), notably, the Privacy Rule and the Security Rule [2], have greatly affected the manner in which medical staff and medical centers operate with respect to both clinical practice and research. Further, privacy practices in some hospitals and medical centers are influenced by policies of The Joint Commission (formerly the Joint Commission on Accreditation of Healthcare Organizations). On a local level, clinical research facilities are influenced by local institutional review boards (IRB). All of these are

Corresponding author: Peyton Bland, BSRB, Room A502, 109 Zina Pitcher Place, Ann Arbor, MI 48109-2200, voice: 734-647-0849, FAX: 734-615-1471, bland@umich.edu

Publisher's Disclaimer: This is a PDF file of an unedited manuscript that has been accepted for publication. As a service to our customers we are providing this early version of the manuscript. The manuscript will undergo copyediting, typesetting, and review of the resulting proof before it is published in its final citable form. Please note that during the production process errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

concerned with the proper handling of protected health information (PHI) which is any information about health and health care that can be linked to an individual. Confidentiality of PHI must be maintained for both legal and ethical reasons, and the adherence to appropriate standards is mandatory for facilities that receive federal funding. Clearly, a new set of procedures is needed to facilitate research while protecting PHI in this new environment.

As an example of PHI, consider the results obtained from a radiologic examination. Images from these exams are commonly made available in DICOM format (Digital Imaging and Communications in Medicine [3]). This widely used set of standards specifies not only the encoding and storage of image data but also specifies how meta-data relating to the patient, the imaging equipment, the imaging exam, etc. are to be stored along with the image data. The meta-data (the DICOM attributes, often referred to collectively as the “DICOM header”) are integral with the image data and thus cannot be lost or separated from it, a clear advantage to researchers. However, even while the DICOM standard has been a boon to researchers, it comes with a price: In the current environment of confidentiality practices, the information stored in the meta-data – much of it PHI – must be dealt with properly.

Continuing with the example above, consider the transfer of DICOM data files into the research environment. One way of complying with confidentiality regulations is to de-identify the DICOM attributes that contain PHI. (This process is sometimes referred to as “anonymization”.) In our lab, de-identification is performed automatically by use of a custom-designed DICOM receiver that processes DICOM image files as they come into the lab via the network. Some DICOM attributes are removed entirely by this process (or have their values set to null), while others are given new values. In the latter case, the original values are generally stored in an encrypted, protected database table along with the new values in such a way that the original values can be re-associated with the data when needed and when properly authorized. This process is described in more detail below.

Concerns with PHI, however, go beyond DICOM image files. In fact, the primary focus of this paper is not the transfer of de-identified DICOM images; instead, the primary focus is the communication of PHI that, in general, originates or exists independently of images (and the PHI that may be contained in them), such as clinical data from the patient interview, physical examination, reports from pathology and surgery, etc. In order to obtain local IRB approval, researchers must observe proper practices in handling and using these other types of PHI. Since much radiologic research involves the use of this information along with image data, the question arises how this information can be accessed and handled while maintaining confidentiality. At the current time, access to PHI typically results from direct communication between the researcher and a clinical collaborator – often via *ad hoc* or informal mechanisms. These conventional mechanisms do not guarantee protection of PHI nor can they reliably conceal the identity of patients from the research team.

One side effect of de-identifying DICOM images as described above is the increased difficulty in matching de-identified clinical PHI with research results that are derived from or related to the de-identified clinical data. At the same time, however, this separation works in the favor of increasing the integrity of the research since it prevents the researcher from being unduly influenced by knowledge of patient information, clinical data, diagnostic outcomes, etc. For example, knowledge of a patient's diagnosis could influence how research analysis results might be interpreted or used. However, if de-identified PHI and research results are matched in a blinded fashion, integrity can be maintained.

In summary, successful conduct of research often requires the transfer of de-identified patient-related information from the clinical environment to the researchers while maintaining confidentiality and integrity of the information. To fulfill this requirement, we have designed

a novel scheme built around a web application that we call “Web-CAP”, Web-based Communications Application for PHI. Web-CAP is designed as a tool to assist in the process of communicating de-identified PHI between the hospital and research environments by implementing a secure, well-defined, controlled protocol that improves upon ad hoc procedures sometimes used to manage PHI. Patients whose PHI is being communicated through the web interface have either given informed consent, or an IRB waiver has been obtained by the project's investigators. Web-CAP is not intended to circumvent any mechanisms or procedures that are in place for the protection of PHI; final authority for the manner in which PHI is managed in a medical center resides with the local IRB¹. Instead, proper use of Web-CAP and the systems that surround it facilitate the adherence to federal and local established regulations and thus potentially assist in obtaining IRB approvals by demonstrating to an IRB that proper safeguards for handling PHI are in place.

De-identifying DICOM Images

Since DICOM images are at the center of our research, as well as being obvious carriers of PHI, we designed Web-CAP to coordinate with and be consistent with the de-identification process that we routinely perform on DICOM images. Each DICOM image coming into our lab is modified by a custom-designed, de-identifying DICOM receiver that removes PHI or substitutes new, de-identified values for PHI contained in the value fields of the image's attributes. This receiver has been in routine use in our lab for approximately four years. In addition, it has been used by two other medical centers participating in the Lung Image Database Consortium [4] for approximately three years.

Our data typically originate in the PACS (Picture Archiving and Communication System) maintained by the Department of Radiology in the hospital. Images can be transferred to our lab by being “pushed” from workstations in the hospital. The process of receiving and de-identifying DICOM images is illustrated by the flowchart in Figure 1 and is described by the following steps that are also indicated in the figure.

Step 1: When a DICOM image arrives in our lab over the network, it is received and processed by the de-identifying DICOM receiver located in our lab. (We implemented this receiver by modifying the `simple_storage` application furnished as a part of the DICOM CTN software suite developed at the Mallinckrodt Institute of Radiology [5].)

Step 2: The modified receiver communicates with a database maintained in our lab (implemented using MySQL [6]) that coordinates various aspects of the de-identification process. Most importantly for the purpose of this discussion, it stores the correspondence between various pieces of PHI removed from the incoming DICOM files and the clear-text values substituted in their place by the de-identifying receiver. This is indicated by the correspondence table shown in the figure. Note that the PHI (in this case, the patient's hospital ID) stored in the database is stored in encrypted form.

Step 3: The de-identified DICOM file is written to a disk array on the fileserver in our lab.

The receiver and the de-identification process that it implements have a number of important characteristics. The de-identification process is performed automatically on all DICOM images coming into the lab. Processing is done “on-the-fly” as files arrive over the network so that they are not written to the disk array until *after* the de-identification process is complete. This minimizes the possibility that PHI from a DICOM file could be seen by unauthorized persons. To further minimize such a breach in confidentiality, image files are transferred over a

¹It is assumed that the local IRB has authorized the clinical collaborators (for example, clinical data technicians) using Web-CAP to have access to PHI and that it has authorized the research personnel using Web-CAP to have access to de-identified PHI. In this manuscript, the terms “clinical collaborator” and “researcher” imply persons who have been properly authorized by the local IRB.

restricted/private network maintained by our medical center's information technology group; traffic on this subnet is not visible outside of our medical center. Finally, the de-identification process takes place entirely inside our laboratory and as such does not interfere with clinical activities or require special software to be installed on any hospital systems.

The de-identification process performs one of three actions (removal, setting to null, assigning a new value) on each DICOM attribute that contains PHI. If the PHI stored in a particular attribute is not needed, the attribute is removed if it is a Type 3 attribute², or its value is set to null if it is a Type 2 attribute. The handling of Type 2 and 3 attributes is controlled by a configuration scheme implemented by a pair of database tables. If an attribute is needed for identification or other purposes, it is given a new de-identified value. A new value is also assigned to an attribute if it is Type 1 even if it is not needed for identification, etc.; this maintains compliance with the DICOM standard. In most cases where new values are assigned, the association between the original and new values is stored in an encrypted, protected database table.

Of particular interest in terms of de-identification is the patient's hospital ID stored in the "Patient ID" DICOM attribute as illustrated in Figure 1. The value of this attribute is replaced with a "research ID", a unique identifier that is automatically assigned as a part of the de-identification process when an image is transferred to our lab³. (Note that even though the original *patient* identifier has been replaced with a *research* identifier, the new identifier value is stored in the "Patient ID" DICOM attribute for compatibility.) The original value of the Patient ID attribute is encrypted as denoted by "encrypt(7654321)" in Figure 1 and is stored with its corresponding research ID in the correspondence table of the database. Study and series identifiers are similarly processed. Dates (such as the study date) are modified by creating a new date relative to some reference date and storing the new date in the de-identified DICOM file; this is particularly useful in serial studies on the same patient.

In the text that follows, we consider PHI to exist in one of two forms. When PHI exists in its original form in the hospital, it is typically a piece of information associated with a patient identifier: for example, "The patient having a hospital ID of 7654321 has positive biopsy results." However, in order to comply with confidentiality requirements, this information could only exist in the research setting as *de-identified* PHI: for example, "The patient having a research ID of 1234 has positive biopsy results." This conveys the needed information (the biopsy results), but it is no longer coupled with a traceable patient identifier. Moreover, since the research ID is unique, it is appropriate for associating the clinical outcomes data with research outcomes.

Web-CAP System Description

Based on this de-identification process, we can now design a mechanism to coordinate the communication of non-image data between the hospital and the research lab. Consider there to be a wall that separates the hospital and the research lab; on the hospital side, the patient ID identifies a patient, and on the laboratory side, the research ID identifies a patient. Web-CAP acts as an interface between the two sides, translating between the patient ID and the research ID using the encrypted correspondence table in the database. These IDs are used in the messages passed between users on the clinical and research sides: The researcher sees and has knowledge of only the research ID, and the clinical collaborator sees and has knowledge of

²Briefly stated, a Type 3 attribute is optional, a Type 2 attribute is required but its value is optional and can be null, and a Type 1 attribute is required and must have a non-null value. See the DICOM standard for details [7] (including information on conditional types which is not included in this brief description).

³Even though Patient ID is a Type 2 (value optional) DICOM attribute [8] when it appears in the Patient Module of an image, we have found that its value is always present in DICOM images at our institution and is thus a reliable key for tracking the patients in our database.

only the patient ID. With these identifiers, users of the system interact via Web-CAP over a secure, controlled protocol rather than via conventional methods such as face-to-face conversations, telephone, or email which cannot guarantee confidential handling of PHI.

Each set of messages passed between a researcher and a clinical collaborator in the process of transferring information is collectively referred to as a *transaction*. The content of each message along with status and tracking information is stored in the transaction table of the database, one record for each transaction. A transaction's database record is built-up incrementally as the transaction progresses. Thus, the state of a transaction can be constructed at any time by a query to the transaction table.

Figure 2 illustrates the process. We continue with the example in Figure 1 for the patient identified by research ID “1234”. (Recall that the images have already been received and de-identified by the research lab as described previously.) Suppose that this patient's images are being used in a research project, and the researcher needs to obtain the results of a biopsy that has been performed on this patient. The transaction begins when the researcher authenticates to Web-CAP and creates a request message (Figure 2a). For clarity, only a simplified version of the request message is shown. In practice, other information such as the sender's and recipient's names, time and date of the request, the name of the project, and study date are included. The steps for creating a request message are as follows:

- Step 1: The user enters information (the patient's research ID and the request) into the request form and submits it.
- Step 2: The information contained in the form is stored in the transaction table of the database.

Note that throughout this communication process, message content originating from the users is passed interactively into Web-CAP via web forms.

At some later time, a clinical collaborator for the research project authenticates to Web-CAP and sees the message sent by the researcher as an entry in his/her *worklist* – a table generated by Web-CAP (Figure 2b) containing all outstanding requests addressed to a particular clinical collaborator. A worklist paradigm is used to facilitate the management of requests originating from multiple senders. For simplicity, the worklist shown in this example contains only one request. The steps for responding to a request are as follows:

- Step 3: Web-CAP creates the worklist for the clinical collaborator. Note that creating the worklist involves accessing not only the database's transaction table to retrieve request message content but also the encrypted correspondence table to translate the research ID to the patient ID.
- Step 4: Web-CAP displays the worklist to the clinical collaborator.
- Step 5: The clinical collaborator selects an entry from the worklist.
- Step 6: Web-CAP creates a response form and loads the selected worklist entry into the patient ID and request fields.
- Step 7: To fulfill the request, the collaborator obtains the requested information – most likely, PHI – by accessing the appropriate clinical information system and adds the information (in this case, the requested biopsy results) to the form.

The response process continues as shown in Figure 2c:

- Step 8: The clinical collaborator submits the response form.
- Step 9: Submitting the response triggers Web-CAP to translate the patient ID to the research ID and to store the translated message in the transaction table of the database.

At some later time, the researcher authenticates to Web-CAP to check the status of any requests that s/he has made. Refer to Figure 2c:

Step 10: The researcher requests that his/her status information be displayed (retrieved from the transaction table).

Step 11: The researcher views the response message (with the ID in de-identified form) in the status table display, thus completing the transaction.

Note that the bi-directional translation between the patient ID and the research ID is automatic and is hidden from both the researcher and the clinical collaborator throughout the process. Further, the identifier field name and content change depending on the location of the message: “Patient ID” is used on the clinical side and “Research ID” is used on the research side so that each user sees only the identifier that is appropriate for his/her role in the process without the ability to associate the relationship between the two.

As described up to this point, communication using Web-CAP is initiated by the researcher. However, there are situations where the clinical collaborator would initiate a message. For example, as a part of a research protocol, the clinical collaborator sends the research team all CT exams of the lung. In addition to having the DICOM images of these exams, the researcher needs to know the smoking history of each subject in the study. Thus, after pushing a set CT images for a particular patient to the researcher's lab, the clinical collaborator would send a message via Web-CAP. This message would contain the patient ID (which would be translated to the research ID by Web-CAP by the time it is received by the researcher) along with the smoking history. Thus, the clinical collaborator's message serves three purposes: the researcher (1) knows that a case has been pushed, (2) knows its research ID (which can be used to locate the de-identified images on the lab's disk array), and (3) knows the smoking history of the subject.

Another variation on the use of Web-CAP is as follows. Even though the most common use of Web-CAP is in conjunction with DICOM images, it is possible to use Web-CAP to communicate PHI in cases where DICOM images are not present. To use Web-CAP in this way, it is only necessary to insert a record into the correspondence table that defines the mapping between the patient's hospital ID (encrypted, of course) and the research ID. (Ordinarily, this entry is inserted automatically as a part of the de-identification of incoming DICOM files.) As an example, consider the situation where a patient needs to be enrolled into a study that does not involve DICOM images (or where images will be considered later). In this case, the insertion of the record into the transaction table would be triggered by a “patient enrollment” message sent by the clinical collaborator to the researcher. When the message reaches the researcher, the researcher (1) knows that a new patient has been enrolled into the study and (2) knows the research ID that has been assigned to this patient. Then Web-CAP is used as described above to communicate PHI involved in the study.

Implementation and Use

As explained earlier, the database is used for two purposes and thus consists of two independent sets of tables. The first set is used by the de-identifying DICOM receiver (step 2 in Figure 1), as well as by Web-CAP (in read-only mode) for translating between the hospital and research IDs in the message-passing scheme (step 3 in Figure 2b and step 9 in Figure 2c). The second set of tables (used only by Web-CAP) holds authentication and configuration information, as well as the transactions themselves (step 2 in Figure 2a and step 10 in Figure 2c). Each transaction stored as a record in the transaction table is updated as the transaction progresses (that is, as the messages comprising the transaction are created and viewed). Each transaction record contains the content of the various messages along with message status and other “bookkeeping” information. These tables furnish the information from which the worklist and

status tables are constructed. Note that PHI is *not* stored in the second set of database tables; instead, for example, the transaction table stores the research ID rather than the patient ID.

Reports for monitoring and logging purposes can be easily generated from the transaction table. Any PHI in these reports would, of course, be presented in de-identified form. In addition, a separate database table (not shown in the figures) stores each interaction with Web-CAP as a separate, time-stamped record. These interactions include user authentication, data entry into the forms, interactions with the worklist, and interactions with the status table. Audit reports can be created from the information in this table. Reports of virtually any kind are easily generated by designing the appropriate database queries based on local requirements.

Inasmuch as the DICOM file de-identifier is a component used in the overall process of communicating PHI, any one of a number of de-identification applications could be used in this process; it is not mandatory to use the particular de-identifier that we developed from the Mallinckrodt DICOM software suite. There are, however, two requirements: First, the de-identifier must perform within the stipulations of the local IRB with regard to how it processes PHI that is present in the incoming DICOM images. Second, the de-identification scheme must include some sort of correspondence table (or similar construct or mechanism) that allows bi-lateral mapping between the patient's hospital ID (stored in encrypted form as described in the manuscript) and the research ID (step 2 in Figure 1, step 3 in Figure 2b, and step 9 in Figure 2c).

User interface considerations include simple, easy-to-use web forms for message creation and for interacting with the worklist and status tables. This type of user interface is employed by virtually all applications that are routinely encountered in everyday use of the web. The data entry steps employ automatic fill-in of existing, known information in as many of the fields as possible. Overall, it is anticipated that using Web-CAP to communicate de-identified PHI will incur little or no additional overhead compared with conventional methods that might involve composing email messages or filling-out hardcopy forms. A further advantage of Web-CAP is that all interactions with the system are stored in the database so that no additional record keeping on the part of the users is needed to track the status of a request for data.

A comment is in order concerning the text input widgets used to form the messages. A simple type-in box is appropriate for the research ID as shown in Figure 2a. However, the request and the response type-in boxes would, in general, need to accommodate richer data. Perhaps the most obvious solution is simply to furnish multi-line type-in widgets for free-form text entry by the user. This has the advantage of flexibility in terms of information content. However, there is a disadvantage: For example, due to user error, the text might contain PHI or other information that is not appropriate for transmission in clear text between the clinical and research sides. (This type of user error is, of course, also possible with conventional methods of PHI transmittal such as email or hardcopy forms.) It may be possible to design input validation schemes to detect certain types of erroneous or inappropriate data using an approach similar to that described in [9]. As an alternative to free-form text input for the request and the response, forms could be designed on a project-specific basis to contain more specific and restrictive input widgets such as checkboxes, pull-down menus, and radio buttons pre-programmed with acceptable choices.

In addition to the errors described above, other error detection and error handling capabilities should be included. The following situations could be anticipated: non-existent research IDs or patient IDs, missing or malformed data entered in the forms, and messages originating from the clinical side before images are received by the lab (and therefore no research ID corresponding to the patient ID), as well as network, database, or web server problems.

The software described herein (the de-identifying DICOM receiver, the database management system, and the Web-Cap application) run on a typical server. No part of this data communication scheme places undue demands on the server, and it is anticipated that a reasonably active set of research projects could be easily handled by a single server.

Summary and Conclusion

Regulatory requirements that define the handling of patient data and the concomitant changes in clinical and research practices mandate a new way of carrying-out research that uses clinical data. We have described Web-CAP, a web-based system that facilitates adherence to these regulations by providing a communication protocol to be used by researchers and their clinical collaborators. The protocol allows patient-related data to be communicated without revealing the identity of the patient to researchers and without revealing research identifiers to the clinical collaborators, thus ensuring the confidentiality of patient data, as well as enhancing the integrity of the research. Web-CAP is easy to use and should impose little or no additional overhead compared with existing practices. It replaces *ad hoc* methods of exchanging data with a secure, well-defined, auditable data communication protocol. It is implemented using standard web-based programming paradigms and conventional database technology, both of which are network-ready. It is easily adaptable, and can be customized to meet institution-specific requirements. We conclude that this scheme provides a useful communication mechanism that facilitates clinical research while maintaining confidentiality of patient data.

Acknowledgements

Supported in part by NIH/NCI 1P01 CA87634, Automated 3D Registration for Enhanced Cancer Management and NIH/NCI 1U01 CA91099, Lung Image Database Consortium

References

1. United States. Department of Health and Human Services. 45 CFR Parts 160 through 164. Standards for Privacy of Individually Identifiable Health Information. Aug 14. 2002 Retrieved October 4, 2006, from <<http://www.dhhs.gov/ocr/hipaa/guidelines/Introduction.rtf>>
2. United States. Department of Health and Human Services. National Institutes of Health. HIPAA Privacy Rule and Its Impacts on Research. Aug 4. 2004 Retrieved October 5, 2006, from <<http://privacyruleandresearch.nih.gov/healthservicesprivacy.asp>>
3. National Electrical Manufacturers Association. Digital Imaging and Communications in Medicine. DICOM Homepage. Retrieved October 6, 2006, from <<http://medical.nema.org/dicom/>>
4. United States. Department of Health and Human Services. National Institutes of Health. National Cancer Institute. Lung Image Database Consortium. Retrieved January 9, 2007, from <<http://imaging.cancer.gov/programsandresources/InformationSystems/LIDC>>
5. Mallinckrodt Institute of Radiology. MIR DICOM Central Test Node Software. Retrieved October 11, 2006, from <<http://wuerlim.wustl.edu/DICOM/ctn.html>>
6. MySQL AB. Retrieved October 11, 2006, from <<http://www.mysql.com/>>
7. National Electrical Manufacturers Association. Digital Imaging and Communications in Medicine (DICOM), Part 5: Data Structures and Encoding. 2006. p. 38
8. National Electrical Manufacturers Association. Digital Imaging and Communications in Medicine (DICOM), Part 3: Information Object Definitions. 2006. p. 255
9. Gupta D, Saul M, Gilbertson J. Evaluation of a Deidentification (De-Id) Software Engine to Share Pathology Reports and Clinical Documents for Research. Am J Clin Pathol 2004;121:176–186. [PubMed: 14983930]

Figure 1. De-identifying DICOM images

This flowchart illustrates how the “Patient ID” attribute is processed when a DICOM image is received by our lab. This attribute holds the patient's hospital ID when the image file leaves the PACS in the hospital. However, by the time the image file is written to the disk array in our lab, this value of this attribute has been changed to the so-called research ID. Note that the database table that stores the correspondence between the hospital and research IDs stores the patient ID in encrypted form.

Figure 2.

Flowchart of Web-CAP showing the life cycle of a request for information as initiated by the researcher. Type-in boxes indicate information that is entered by the users. For simplicity, the messages are shown in abbreviated form; in practice, the messages include additional identifying information. Part (a) depicts the creation of the request message by the researcher and its storage in the database. Part (b) shows the retrieval of the request by the clinical collaborator and the creation of a response message. Part (c) shows the storage of the response message in the database and the subsequent retrieval and viewing of it by the researcher. The Web-CAP application and the database are central to the process, acting in tandem as an interface between the clinical and research sides. Note how the identifier field name and content change depending on the location of the message: “Patient ID” on the clinical side and “Research ID” on the research side.