J Digit Imaging (2012) 25:14–24
DOI 10.1007/s10278-011-9422-x

Image Data Sharing for Biomedical Research—Meeting HIPAA Requirements for De-identification

John B. Freymann · Justin S. Kirby · John H. Perry ·
David A. Clunie · C. Carl Jaffe

Published online: 29 October 2011

Abstract Data sharing is increasingly recognized as critical to cross-disciplinary research and to assuring scientific validity. Despite National Institutes of Health and National Science Foundation policies encouraging data sharing by grantees, little data sharing of clinical data has in fact occurred. A principal reason often given is the potential of inadvertent violation of the Health Insurance Portability and Accountability Act privacy regulations. While regulations specify the components of private health information that should be protected, there are no commonly accepted methods to de-identify clinical data objects such as images. This leads institutions to take conservative risk-averse positions on data sharing. In imaging trials, where images are coded according to the Digital Imaging and Communications in Medicine (DICOM) standard, the complexity of the data objects and the flexibility of the DICOM standard have made it especially difficult to meet privacy protection objectives. The recent release of DICOM Supplement 142 on image de-identification has removed much of this impediment. This article describes the development of an open-source software suite that implements DICOM Supplement 142 as part of the National Biomedical Imaging Archive (NBIA). It also describes the lessons learned by the authors as NBIA has acquired more than 20 image collections encompassing over 30 million images.

Keywords Data sharing · De-identification · Anonymization · Cross-disciplinary research · Open access · Open source · DICOM · Supplement 142 · Image archive · HIPAA · PHI · Common rule

This project has been funded in whole or in part with federal funds from the National Cancer Institute, National Institutes of Health, under Contract No. HHSN261200800001E. The content of this publication does not necessarily reflect the views or policies of the Department of Health and Human Services, nor does mention of trade names, commercial products, or organizations imply endorsement by the U.S. Government.

J. B. Freymann
SAIC-Frederick, Inc.,
EPN, Room 3006, 6130 Executive Blvd,
Rockville, MD 20892, USA

J. S. Kirby
SAIC-Frederick, Inc.,
EPN, Suite 317, 6130 Executive Blvd,
Rockville, MD 20892, USA
e-mail: kirbyju@mail.nih.gov

J. H. Perry
Radiological Society of North America,
820 Jorie Blvd,
Oak Brook, IL 60523, USA
e-mail: johnperry@dls.net

D. A. Clunie
CoreLab Partners, Inc.,
100 Overlook Center,
Princeton, NJ 08540, USA
e-mail: dclunie@dclunie.com

C. C. Jaffe
Boston University School of Medicine,
FGH Building 3rd Floor, 820 Harrison Ave.,
Boston, MA 02118, USA
e-mail: carl.jaffe@bmc.org

J. B. Freymann (✉)
SAIC-Frederick, Inc.,
EPN, Room 3006, 6130 Executive Blvd,
Bethesda, MD 20892-7412, USA
e-mail: freymannj@mail.nih.gov

Background

Advancing imaging research to serve as a critical element in clinical therapeutic trials requires that imaging methods be developed, optimized, and validated using commercial clinical imaging instruments. This applies particularly to quantitative imaging as a bio-marker for drug development or measurement of drug response. For example, there is a critical need to harmonize data collection and analysis across the different commercial platforms used in clinical practice to ensure robust correlation of image-derived parameters with clinical outcome. In addition, data integration with other laboratory-based molecular bio-markers requires a fundamental understanding of the physical and biological measurement uncertainty in order to convert data to knowledge or support a medical intervention. The National Cancer Institute (NCI) Cancer Imaging Program has supported research initiatives to improve the performance and reproducibility of imaging methods, including development of imaging technology, software tools for clinical decision making, and development of molecular probes to incorporate the molecular basis for clinical decision making. Central to these efforts is a fundamental need for a widely adoptable, image-focused informatics infrastructure along with data archives that provide a common framework for data exchange and shareable methods to validate current and emerging imaging agents and methods.

Public funding agencies have long recognized the importance of data sharing in cross-disciplinary research. National Institutes of Health (NIH), for example, has had a final statement for grantees on sharing research data since 2003 and a published guidance for grant recipients since 2006 [1]. Nevertheless, little data sharing has occurred outside the framework of prearranged links between research groups. One reason for the unwillingness of institutions to share clinical research data is the variety of local interpretations of Health Insurance Portability and Accountability Act (HIPAA) regulations enforced by HHS Office of Civil Rights. In this environment, the most comfortable stance for institutional IT departments has been to adopt risk-averse postures [2].

In the science community, mainstream stakeholders like NIH, FDA researchers, PhRMA, and the device industry continue to emphasize the importance of data and image sharing in policy statements. New societal attitudes toward funding science have focused renewed attention to data sharing as a way to break down silos, accelerate progress, and reduce research redundancy [3]. Besides access to a greater universe of data available for research purposes and assuring the validity of scientific claims, data sharing provides other advantages to individual researchers by producing more citations [4, 5]. Biomedical research containing clinical data in particular motivates new justification for encouraging data sharing since the bedrock of disease-based clinical genetics and cellular discovery rests on data derived from human subjects. Moreover, genetic research must rely on large population sample sizes, making conclusions derived from such data too costly to replicate by other investigators. The data from each individual is obtained at great cost and effort. If such data were sequestered in small isolated collections and cannot be cross-queried, the research community suffers. Investments in large-scale national and international bio-specimen genetic projects are underway by the NIH, including The Cancer Genome Atlas [6] and the Cancer Human Biobank [7]. To be adequately studied and analyzed, such tissue-specimen genetic data must be accompanied by the individual's clinical data, a key component of which could include non-invasive imaging obtained for diagnostic purposes. Sharing such images requires informed consent by the patient and robust removal of protected health information (PHI) from the images.

At a technical level, the field of diagnostic imaging has benefited from a long historical investment in the Digital Imaging and Communications in Medicine (DICOM) standard by equipment manufacturers and devoted personnel in the professional radiological societies [8]. In the context of image sharing, DICOM Working Group 18 has recently developed Supplement 142 (ftp://medical.nema.org/medical/dicom/final/sup142_ft.pdf, accessed 28 February 2011) that provides important guidance for de-identification of images and related data objects.

This manuscript describes the challenges faced and lessons learned during development and production implementation of an open-source suite of software that implements Supplement 142 for de-identification in the context of an NCI-sponsored public biomedical image archive, National Biomedical Imaging Archive (NBIA). These tools have matured through extensive field use over the past several years and offer a method sufficiently tested to assure de-identification, transfer, management, and distribution of DICOM images and XML objects. While this software suite is freely available for download and use [9], the focus of this paper is not to advocate for these specific implementations but rather to provide guidance for evaluating tools appropriate to a given context.

Technical Issues in Multi-center Data Sharing

Clinical trials and other research-driven image collection activities often produce a combination of image and non-image data objects. Preserving the interrelationships between these objects while de-identifying their PHI is challenging. Images are typically encapsulated in DICOM datasets that contain identifiers for a trial, a patient, a study, a series (of images), etc. Increasingly, non-image data objects are encapsulated in XML files. All data objects in a given research set must share common identifiers if the correspondences among them are to be preserved. Since the original identifiers inserted into the data objects when they were created can be PHI, they are almost always replaced by pseudonymous values (PHI encrypted by an appropriate authority) that maintain the relationships among the data objects but break the connection to the specific human trial participant [10]. When multiple data object types are present in a trial, the de-identification mechanism must support all the data types such that the identifying links between them are maintained.

It is possible to discern subtle differences in the meanings of the words “de-identification” and “anonymization,” but in this paper, they will be used as synonyms, with the former being preferred. In a multi-center clinical image collection project, images are generally received by a data system via the DICOM protocol, usually from a PACS workstation or modality. Non-image data objects are generally transferred to the clinical trial system via HTTP. Once the data objects have been received, they are de-identified and then transmitted to a principal investigator site, contract research organization, or a centralized archive, usually in another location, via the Internet.

Data Transmission

Although clinical image data are de-identified at the originating institution before transmission, many trials require that the data be transmitted using Secure Sockets Layer to provide encryption. Some trials use Transport Layer Security (TLS) to provide both data encryption and client/server authentication.

Most clinical image data transfer on the Internet requires the penetration of at least one firewall. Most projects employ software that makes outbound connections from the secure network at the image acquisition sites to the principal investigator site. This relieves the image acquisition sites from having to open a port to the Internet, but it requires one port to be open to the Internet at the principal investigator site—a requirement that some IT departments are unwilling to support (see Fig. 1). Some clinical trial transfer packages allow two programs to run together at the principal investigator site to pull data into the secure network from the DMZ without having to open a port to the secure network. A DMZ (demilitarized zone) is an interface sub-network that exposes an organization’s external services to a larger untrusted network. It provides an additional layer of security to an organization’s local network. Others use virtual private network technology to allow image acquisition sites to access the secure network at the principal investigator site. Most clinical trial data transfer packages support all those options.

Once in the secure network at the principal investigator site, data objects must be validated (checked that they belong to a specific trial), curated (assure that data file structure allows it to be viewable as an image), organized, and stored. This process, which varies from project to project, requires software that is flexible enough to allow human intervention in the process. In all projects, access to the stored data must be controlled. In large image archive acquisition projects, multiple layers of storage in staging servers may be involved prior to data being made available more generally.

De-identification

The objective of de-identification is to ensure that data objects cannot be connected to a specific human subject [11]. The HIPAA Privacy Rule [12] defines two approaches to removal of PHI: one that leaves the decision as to what constitutes PHI to a nominal expert and the other that pre-defines 18 categories of identifiers to specifically remove or conceal, i.e.,

> The following identifiers of the individual or of relatives, employers, or household members of the individual must be removed: (1) Names; (2) all geographic subdivisions smaller than a state, except for the initial three digits of the ZIP code if the geographic unit formed by combining all ZIP codes with the same three initial digits contains more than 20,000 people; (3) all elements of dates except year, and all ages over 89 or elements indicative of such age; (4) telephone numbers; (5) fax numbers; (6) email addresses; (7) social security numbers; (8) medical record numbers; (9) health plan beneficiary numbers; (10) account numbers; (11) certificate or license numbers; (12) vehicle identifiers and license plate numbers; (13) device identifiers and serial numbers; (14) URLs; (15) IP addresses; (16) biometric identifiers; (17) full-face photographs and any comparable images; (18) any other unique, identifying characteristic or code, except as permitted for re-identification in the Privacy Rule.

Note the ambiguity of item 18. The Federal Register in 2006 presents the rule [13], and NIH guidance is provided under the title “Research Repositories, Databases, and the HIPAA Privacy Rule” [14]. In research data, such information is typically replaced with pseudonymous values that allow trial subjects, studies, and data objects to be related to one another but not connected to a specific human being.

To fully de-identify a DICOM image, PHI must be removed from both the metadata elements and the pixels of the image itself. De-identifying metadata is complicated by the fact that manufacturers and even end users of medical imaging equipment often use DICOM elements in a way that legitimately extends or does not conform to the standard, resulting in PHI sometimes being found where not normally expected. In addition, manufacturers sometimes place PHI in private elements, the contents of which are unspecified in the

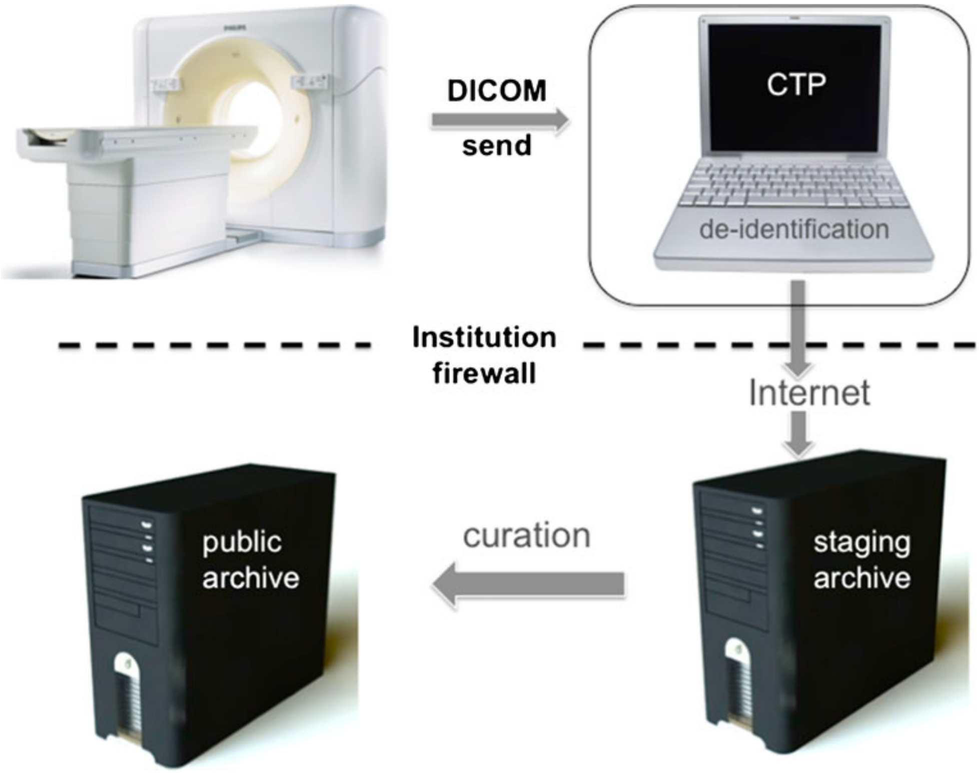

Fig. 1 CTP software performs custom scriptable de-identification behind the institution's firewall. The files are then securely transferred through the Internet to the host NBIA where they are re-inspected for DICOM validity and thorough de-identification before they are made publically accessible

DICOM standard, and not reliably clarified in conformance statements. These complications require a de-identification system to be flexible enough to be configured to handle special circumstances as they arise [15].

The removal of PHI burned into the pixels of diagnostic images is even more difficult. This can be performed completely manually (http://www.dclunie.com/pixelmed/software/webstart/DicomCleanerUsage.html—blackout accessed 28 February 2011), but several groups have developed approaches for discovering text information burned into the pixels of an image. In most of these efforts, image processors use optical character recognition to flag possible PHI. As yet, none seems provably robust enough to be acceptable for automatic processing without a human observer in the loop. The DICOM standard provides an element used to indicate that an image contains PHI, but the element is not universally supported, and in any case, it does not indicate where in the image the PHI is located. The best approach appears to be using the DICOM metadata elements to identify those images particularly at risk of containing burned-in PHI, such as specific modalities including ultrasound, or those images with elements suggesting that they are screen captures (e.g., of 3D reconstructions or other post-processed images). In some cases, specific templates for the locations of burned in text can be applied based on the device manufacturer and model. Care needs to be taken to address PHI present in the high (unused) bits of the pixel data that may be used as overlays.

DICOM Supplement 142

The DICOM standard provides important guidance for de-identification. In DICOM PS 3.15, Annex E, "Attribute Confidentiality Profiles," the standard defines the Basic Application Level Confidentiality Profile, which specifies requirements for applications that de-identify and/or re-identify dataset attributes and (in Table E.1-1) lists a set of attributes that are subject to the profile (ftp://medical.nema.org/medical/dicom/2009/09_15pu.pdf, accessed 28 February 2011). This profile was added in Supplement 55 in 2002 (ftp://medical.nema.org/medical/dicom/final/sup55_ft.pdf, accessed 28 February 2011), but it has proven to be insufficient for robust de-identification. During the development of the IHE Teaching File and Clinical Trial Export (TCE) profile (http://www.ihe.net/Technical_Framework/index.cfm#radiology, accessed 28 February 2011), additional standard material was added to elaborate on the issues of de-identification and pseudonymization, but it too does not define a comprehensive and detailed approach.

Accordingly, Supplement 142 (ftp://medical.nema.org/medical/dicom/supps/sup142_pc.pdf, accessed 28 February 2011) was developed, to provide more detailed guidance for de-identification of data objects for various purposes. The supplement is built on a Basic Profile that takes a very conservative approach to removing or replacing any information about the identity of the patient, their family members, any personnel involved in the procedure, the organizations involved in ordering or performing the procedure, and additional information that might be combined to associate the object with the patient.

Supplement 142 also provides several options appropriate to special situations. Two classes of options are defined, those that require significant and burdensome effort to remove additional information (and which may not be justified in low risk scenarios) and those that define retention of information that would otherwise be removed, but without which a particular type of research would be impossible. Common examples of the latter include the need to retain date information in therapeutic oncology trials, without which dates of progression or response cannot be determined, the need to retain patient characteristics related to body size for whole body PET studies, without which standardized uptake values cannot be computed, and the need to retain image and device (but not patient) unique identifiers that may be required for the audit trail. In such cases, the additional information that is needed for the conduct of the trial may not be permitted by regulation, and therefore, additional permission is required either from the subject or from the institutional review board (IRB) or ethics committee. The options defined in Supplement 142 are intended to provide a small and tractable set of standard definitions with accompanying justification, such that each IRB and consent form can reference the standard categories, rather than debating the merits of individual DICOM data elements.

The options defined in the supplement are:

- Clean Pixel Data Option: removal or distortion of the actual pixel data where there is identification information burned in as annotation text
- Clean Recognizable Visual Features Option: removal or distortion of the actual pixel data where there is possibility of visually identifying the individual in the images
- Clean Graphics Option: removal of identification information encoded as graphics, text annotations, or overlays (excluding Structured Report SOP classes)
- Clean Structured Content Option: removal of identification information in Structured Report SOP classes
- Clean Descriptors Option: removal of identification information from descriptive tags which contain unstructured plain text values over which an operator has control
- Retain Longitudinal Temporal Information Options: retention or modification of tags that contain dates or times
- Retain Patient Characteristics Option: retention of physical characteristics of the patient that are descriptive rather than identifying information (e.g., metabolic measures, body weight, etc.)
- Retain Device Identity Option: retention of information about the identity of the device used to perform the acquisition
- Retain UIDs Option: retention of the unique identifiers for studies, series, instances, and other entities in the DICOM model
- Retain Safe Private Option: retention of private attributes known to be safe

Supplement 142 was drafted by leading industry experts in DICOM Working Group 18. In particular, those involved in international pharmaceutical clinical trials for regulatory submissions were broadly consulted, and indeed, the work effort was initiated as a consequence of discussion during a Drug Information Association Medical Imaging Stakeholders Call for Action in 2007. Global regulations were considered, including not just the HIPAA Privacy Rule but also the European Privacy Directive. Supplement 142 provides a platform for consistent de-identification that meets global regulatory requirements and is thus a substantial contribution to medical research.

Methods

The NBIA [16] is an open-source software suite developed under the aegis of the caBIG program of the NCI's Center for Bioinformatics [17] and Information Technology [18]. The software has been installed at numerous institutions for use in sharing image collections. This section introduces the software and describes its use in the acquisition, management, and distribution of image collections by the NCI's Cancer Imaging Program and other institutions running the software.

National Biomedical Imaging Archive Project

NBIA [19] is a highly scalable, DICOM-based image archive that provides full submission-to-retrieval functionality optimized for the requirements of the *in vivo* medical imaging clinical and research communities. It combines image acquisition and processing capabilities with submission reporting and quality control tools to facilitate inter-institution data sharing. NBIA provides query access to more than 90 DICOM tag elements. These can be queried through three levels of search interfaces as well as an API. It integrates cine-view, thumbnails, and full DICOM element previews. A saved-query feature provides a unique reference keyword for direct linkage to data sets from publications, etc. Data download is supported through a Java download manager for larger collections. Non-DICOM metadata can be contained in XML or Zip files and linked at the image series level when appropriate. Images can be grouped within collections for specific research purposes, and the NBIA supports pop-up menus that can provide short summaries of these collections or link to external information sites such as Wikis or other web sites.

The NBIA web application allows users to search for, manage, and retrieve DICOM images. The web application is written in Java and relies on the JSF presentation framework. It is deployed on a JBoss application server. The image metadata indexed by the web application is stored in a MySQL or Oracle database. The DICOM images themselves are stored in a file system of the administrator's choice. NBIA provides a collection- and submission site-based authorization model that is implemented using NCI's Common Security Module. This allows an administrator to create public access and restricted access data sets as needed. Additionally, the NBIA system includes a caGrid data service based upon the caBIG NCIA_MODEL version 3 [20]. The grid service provides the ability to retrieve DICOM images using the caGrid Transfer service, allowing for multiple installations of NBIA to seamlessly communicate and share images in a federated manner.

NBIA integrates a separate software package, RSNA's Clinical Trial Processor (CTP), to manage the transfer of images into the NBIA system. In a project employing NBIA, CTP is installed at both the data acquisition sites and an NBIA site. These sites are often called client and server sites, respectively. CTP is configured to de-identify data objects at the client sites to ensure that PHI never leaves the originating institutions. At the client site, images are both de-identified and tagged with provenance information in private elements for use in indexing the images. The CTP at the client site then transmits the data objects to the CTP at the NBIA server site, which stores the images in a file system and extracts information from the DICOM elements for storage in the NBIA relational database.

NCI's Cancer Imaging Program has used NBIA to create more than 20 research image collections. These collections and more that will follow are intended to make medical imaging case studies available to a wide cross-disciplinary research community. NBIA has also been used to establish a nationwide infrastructure for sharing images, supporting stratification of patients in adaptive clinical trials, cross-disciplinary research on response measurement fundamentals, and increasing the research community's awareness of image reliability analysis.

NBIA's archive and open-source tools provide:

- Multiple research image data collections, encouraging development of reliable quantitative measurement of change over time by supplying longitudinal clinical response imaging case studies to a wide research community
- Real-time, multi-institutional image access, supporting protocol stratification strategies in adaptive trials
- Support for cross-disciplinary research on response measurement fundamentals and analysis of quantitative reproducibility studies

For clinical trial data residing in non-public-access archives, these same software tools implement role-based security to permit selected PHI to remain in place. In this situation, access to such images requires formal permission granted by the signing of a limited dataset agreement [21].

Clinical Trial Processor

CTP is a tool developed by the Radiological Society of North America (RSNA) for autonomously processing data objects in clinical trials. It is written entirely in Java and runs on Unix, Linux, Solaris, Mac OS, and Windows. It runs either standalone or as a Windows service on XP, Vista, and Windows 7. The program's interface is provided by an integrated web server with several servlets that provide access to status and configuration information. Complete documentation on CTP is located on the RSNA MIRC Wiki [22].

Processing in CTP is organized into pipelines [19], each consisting of a sequence of stages, where each pipeline stage is designed to perform a specific function. CTP is highly configurable, allowing administrators to construct pipelines to meet a wide variety of requirements. CTP currently provides 25 standard pipeline stages in four categories:

- Import Services receive data objects from external sources and queue them for subsequent processing.
- Processors receive a data object as it flows down the pipeline, take some action, and pass on the object to the next stage. Actions can range from simply logging the passage of the object to modification of the object. Processors are synchronous stages, not passing on the object until processing is complete.
- Storage Services receive a data object as it flows down the pipeline, store a copy of the object in some kind of storage system, and then pass the object on to the next stage. Storage Services are synchronous.
- Export Services receive a data object as it flows down the pipeline, queue a copy of the object for subsequent transmission to an external system, and then pass the object to the next stage. The queuing process is synchronous; the subsequent transmission occurs asynchronously.

CTP is designed to be easily extended by the addition of new pipeline stages and database adapters.

To be useful, a clinical data object must contain identifiers that relate it to other data objects. CTP supports four types of data objects, three of which provide standardized access to the identifiers and data they contain:

- FileObjects are data objects of indeterminate contents. This is the superclass of the other three types, but on its own, it is not useful because it does not provide access to the required identifiers.

- DicomObjects are DICOM datasets. This type provides all the necessary identifiers as defined in the DICOM standard.
- XmlObjects are XML documents. XML provides for the encapsulation of text-based data. Many XML schemas are in use in clinical trials today, and there is no standard definition of how the required identifiers are encoded. The CTP XmlObject attempts to find identifiers by looking in a sequence of commonly used schema locations.
- ZipObjects are zip files containing one or more data files plus a file called manifest.xml which contains the required identifiers. The manifest.xml file is located in the root of the zip file's directory tree, and it obeys a standard schema. The ZipObject provides a way to encapsulate collections of related data objects in any format while still carrying the identifiers which allow them to be related to other objects in the trial.

Since data objects in clinical image collections are generally produced by clinical systems, they almost always contain PHI. Among the most important standard pipeline stages in CTP are ones for de-identifying data objects. CTP provides four standard pipeline stages for modifying data objects to remove PHI and replace it with pseudonymous values:

- The DicomAnonymizer modifies DicomObjects in accordance with a script. The script is written in a simple language that provides many functions for handling specific types of DICOM elements. Both CTP and the independent clinical trial management software written by the American College Research Imaging Network use this language. CTP provides a special servlet to simplify the process of defining a DicomAnonymizer script. This servlet allows the administrator to define the rules for de-identification of each individual DICOM element. Since de-identification is a complex technical field, the DICOM committee has released Supplement 142 to the standard, specifying de-identification profiles and options for various purposes. One of the authors (JK) has written script implementations of all the Supplement 142 profiles and options, and these are built into CTP. The CTP DICOM Anonymizer Configurator also supports user-defined profiles. The default de-identification script is the most stringent one defined in Supplement 142 (the Basic Profile). This provides access to a de-identification mechanism that is in common use and has been vetted to meet regulatory requirements for protecting patient privacy. The configurator servlet allows the administrator to select a profile as a starting point and modify it to meet any special needs of the trial.
- The DicomPixelAnonymizer modifies DicomObjects by blanking regions of the pixels in a DicomObject in accordance with a script. The script consists of a sequence of signatures and region sets. A signature is a boolean calculation based on the contents of the DicomObject's elements. Each signature is accompanied by a list of rectangular regions to blank in images that match the signature. When processing a DicomObject, the DicomPixelAnonymizer computes each signature value in turn, chooses the first one that matches, and then blanks the regions associated with it.
- The XmlAnonymizer modifies XmlObjects in accordance with a script written in a language that is inspired by, but is much simpler than, XPath. CTP provides a special servlet to simplify the process of defining an XmlAnonymizer script.
- The ZipAnonymizer modifies the manifest.xml file in a ZipObject in accordance with a script written in a language that is identical to that used by the XmlAnonymizer. When de-identifying ZipObjects in a clinical trial, one must remember that since the ZipObject can contain files of any format, PHI may be contained in places that the ZipAnonymizer does not modify. For this reason, ZipObjects are most useful for encapsulating the analytic results of programs that operate on prior de-identified objects.

Import and export pipeline stages provide for the reception and transmission of data objects. CTP includes five standard import stages and five standard export stages that support the common protocols (HTTP(S), DICOM, and FTP) as well as manual import from directories and archives:

- HttpImportService receives data objects via the HTTP and HTTPS protocols.
- PollingHttpImportService makes an outbound connection to a PolledHttpExportService and receives data objects in the input stream of the connection, thus avoiding the necessity of opening a port for inbound connections.
- DicomImportService implements a DICOM Storage SCP for the receipt of DICOM data objects.
- DirectoryImportService imports (and removes) data objects that appear in a directory.
- ArchiveImportService copies data objects from a directory tree and processes them, leaving the objects in the original location unmodified.
- HttpExportService transmits data objects via the HTTP and HTTPS protocols.
- PolledHttpExportService serves data objects in response to received connections.
- DicomExportService implements a DICOM Storage SCU for the transmission of DICOM data objects.
- FtpExportService transmits data objects via the FTP protocol, organizing them on the destination server by study identifier.

- DatabaseExportService provides a queued interface to an external database.

In situations where a port to the internet cannot be opened on the secure network at a principal investigator site, two instances of CTP can be run, one on the secure network and one in the DMZ, using the polled HTTP stages to allow data objects in from the Internet without opening a port.

The DatabaseExportService interfaces with an external database through an extension of the standard CTP DatabaseAdapter class. In the NBIA project, the NCI wrote a DatabaseAdapter (NCIADatabase [sic]) that receives parsed data objects from the DatabaseExportService and extracts information for storage in an external SQL database (MySQL or Oracle). This mechanism provides a flexible way to build complex databases without having to manage the transfer, or even the parsing, of the data objects themselves.

A Survey of Image Collections and Tools

Several research alliances are actively developing both publicly accessible biomedical image databases and software tools to support them. In some cases, the tools themselves are accessible for download, allowing new research groups to utilize them in posting their own datasets. In other situations, the software is a customized solution with more limited scalability to other use cases.

To gain a better understanding of the characteristics of the various approaches, a search for biomedical imaging tools and archives was carried out. Since any such survey would be rapidly out of date, the information gathered is posted on a Wiki [23]. The goal of this resource is to allow members of the research imaging community to find image collections and tools for creating new collections, to participate in the review, and to ensure that posted information remains as accurate and up to date as possible. A well-maintained site that catalogues mostly open-source image software analytic tools is also available on another Wiki. [24]

Discussion

The process of building the collections housed at the NIH NCI NBIA produced a number of lessons learned with regard to effectively managing the process of collection, de-identification, and distribution of DICOM images for research. They are presented here as points to consider not only to users of the NBIA and CTP software suite but also to anyone developing or assessing similar tools. This section presents the key lessons learned.

Support Multiple Means for Submitting Data

Data have been submitted to the NCI NBIA archive from many sources via several communication protocols. Among the most common ways that DICOM objects have been imported into CTP at the client site are:

- Transmission via HTTP(S) on the Internet, usually from a tool such as RSNA's FileSender
- Transmission via the DICOM protocol from a PACS or workstation at the submission site
- Physical delivery on CD/Hard Disk via mail, some in the format of DICOM CDs, others simply image files

Any software suite must be able to import data from all these media. Although the transport protocol varies, DICOM is the dominant format for the image data itself. Occasionally, images have been received in a non-standard format, but we have found that converting such images to DICOM expands their utility.

Use DICOM Supplement 142 Profile Templates

Institutions and vendors vary widely in the ways they create and de-identify images. The de-identification rules for a collection depend on the intended use of the collection as well as the initial state of the images as they are acquired. For example, patient studies containing PHI must be de-identified fully, but previously de-identified studies obtained from another collection may require little or no additional modification. The de-identification process must therefore be very flexible.

Before the publication of Supplement 142, developing de-identification scripts for a variety of use cases required a thorough understanding of DICOM, and the scripts themselves took substantial time to write and test. Having implementations of the Supplement 142 profiles available in the CTP de-identification stages greatly simplifies the task and improves the confidence of the submitters and curators that regulatory requirements are being met. It also allows the de-identification rules to be changed quickly for specific submissions when necessary. Proper use of the Supplement 142 profiles also provides a historical record within each DICOM object detailing the previous profiles applied to de-identify the images. This practice, discussed in further detail below, allows consumers of the data to clearly evaluate how the image was de-identified and also clarifies what additional steps may need to be taken if the data are being repurposed for a new audience.

Do Not Overdo De-identification

Image collections generally contain data from many patients, each often having multiple studies and series. To maximize the benefits of such collections, the identifiers in the data objects must retain the ability to distinguish among patients, studies of a single patient, etc. Any implementation of the Supplement 142 profiles must be careful to provide pseudonyms for such identifiers rather than fixed values. For example, if every patient ID were named to the same value, then most DICOM software would treat your entire dataset as though it consisted of only a single patient.

Dates require special attention. Maintaining the temporal relationships among studies of the same patient adds significantly to the utility of an image collection, but original calendar dates themselves are PHI and must therefore be modified. This is addressed in the Supplement 142 "Retain Longitudinal Temporal Information Options." The simplest implementation is to offset dates by an interval that is the same for all images in the collection. Prior to the creation of Supplement 142, we had found it convenient to use intervals large enough that users of the collection do not question whether the dates have been modified. However, it was later discovered that offsetting the dates by large increments can cause problems in some DICOM software if the resulting dates are prior to the 1980s. Supplement 142 specifies that the Attribute Longitudinal Temporal Information Modified (0028,0303) should be populated with a value of "MODIFIED" to make it clear that dates have indeed been altered. This is a simpler and more effective solution.

Do Not Rely on DICOM to Indicate Burned-in PHI

PHI burned into the pixels of images poses a serious problem for public research archives. Technologists or PACS administrators are sometimes unaware that these types of images exist in their local systems. A wide variety of such images have been received for the NCI collections, including not only clinical images containing patient names in their pixels but also digitized billing records in DICOM wrappers. Of significant concern is the recent practice of scanning the patient exam request document into the DICOM study series to record the clinical need for the exam and validate billing. That scanned image, usually a final series in the study, is often full of PHI both in the DICOM tags and within the image. Most commercial software intended for de-identification fail to address the special content of that series.

Strategies for dealing with this issue are provided by the Supplement 142 "Clean Pixel Data" and "Clean Graphics" options, but the identification of the images themselves can be a problem. Some DICOM elements that can be useful are:

- (0008,0016) SOP Class UID: value indicating Secondary Capture and Ultrasound SOP Classes
- (0008,0008) Image Type: The values SECONDARY and SCREEN SAVE indicate a suspect image, but they are not definitive
- (0028,0301) Burned-in Annotation: The value YES is definitive, but this element is often not supplied in DICOM images, since it is optional for most objects and a relatively recent addition to the standard
- (0018,1016) Secondary Capture Device Manufacturer: The value of this element can be used to discriminate against certain image types that may contain PHI
- (0018,1018) Secondary Capture Device Manufacturer's Model name: The value of this element can be used to discriminate against certain image types that may contain PHI

Image collection tools must have a means for scanning such elements and segregating images for special attention based on defined criteria. CTP provides filter stages driven by a script language that allows testing the values of all DICOM elements and automatically quarantining objects that fail the test.

Keep an Audit Trail of De-identification History

It is often necessary to know the de-identification history of an image. DICOM Supplement 142 meets this need by defining standard profiles, the codes for which can be used as an audit trail. For example, if in the process of de-identification one used the Basic Application Confidentiality Profile with the option to Retain Longitudinal With Modified Dates, one would also populate the De-identification Method Code Sequence (0012,0064) with the corresponding Coding Scheme Designators for those changes. If the biomedical image community were to adopt this standard, it would be much easier to understand the history of how an image was de-identified and to make decisions on whether further changes are needed as images are repurposed for consumption by new audiences.

A separate audit trail of exactly what values have been replaced may also be maintained, but must be protected since by definition it may contain PHI. If this is done within the DICOM image file itself, it must be encrypted, and data elements are provided for that purpose; their use is deprecated, however, since any encryption scheme becomes vulnerable over time and such images may be archived indefinitely. Supplement 142 warns about this, and accordingly if any such audit trail is required, it should probably be maintained separately from the images and both logically and physically protected.

Enable Local Mapping Between Anonymized Identifiers and PHI

When questions arise about the integrity of the submitted data, it is often necessary for an administrator at the submitting site to examine the original data to determine whether the problems are within the original data or if they were created during the process of de-identification or transmission. To do so, the anonymized identifiers obtained from the collection curator must be translated back to the original PHI. The CTP IDMap stage can be used to provide this translation of identifiers. To have access to this function, a user must be authenticated and have administrator privileges. This functionality may also be necessary in situations where image data are to be correlated with additional data types that have not or will not be de-identified.

Provide End-to-End Transport Verification

In many clinical trials, each submission is accompanied by a case report form or an IHE TCE manifest. In most submissions to research image collections, however, no manifest is available to identify the individual images, series, etc. that have been transmitted. NBIA and CTP therefore include special tools to verify that the submitted data has been received and successfully processed.

Once such tool, the CTP Database Verifier, can be used at the submitting site to ensure that all transmitted data made it all the way into the NBIA database. This tool tracks the de-identified UIDs of every object that is sent to the archive and then periodically queries the NBIA server via its relational database to confirm the object was received and stored. This has saved both submitters and curators substantial effort. The NBIA View Submission Report function is also useful for comparing totals of data objects received by the system against counts of the original submissions, although this tool is more often used for general reporting and auditing of what has been archived.

Provide Multiple Levels of Data Verification

We have used CTP's filtering stages to verify that the metadata of images matches the protocol of a study and to quarantine images that fail before they are added to the collection. We have also used the QC Tool in the NBIA software to verify the content of the data manually. The tool is designed to allow a curator to see both the images and corresponding DICOM elements in a single view. Because PHI can occur both within the image pixels and the metadata elements, we have found that having the ability to view both simultaneously substantially decreases the level of effort involved in managing submitted data.

We have also found that having a built-in method for deleting images has been necessary more often than expected. This allows curators to easily remove data that have eluded detection for not matching the protocol or for containing PHI in unexpected places.

Carefully Estimate Resources Required

It is easy to underestimate the time and effort involved in collecting and managing images for image collections. While the maturity of CTP and NBIA has grown significantly over the past few years, it still requires between 1 and 4 h for an expert CTP user to provide training to a new site manager on how the submission process works and to do preliminary setup. In a large project, this justifies setting up a help-desk function. Preliminary setup is typically followed by small-scale submission tests to ensure the data arrives as expected (modality, number of images per study, de-identification completeness, etc.). Again, the use of CTP implementation of the Supplement 142 profiles has greatly reduced the amount of setup time required. It does not, however, completely remove the need for careful checking by a small test submission of the implementation before large-scale acquisition is started.

Although the combination of CTP and NBIA can be run autonomously, it is important to provide human oversight, not only to ensure that privacy regulations continue to be met as data from new acquisition sites are received but also to ensure that the data added to the collection are consistent with the collection's intended use. Tools such as the ones described here reduce the workload of the collection's human curator, but they do not eliminate it. Thus, anyone considering hosting a truly open biomedical image archive should also allocate staff resources for the collections' curators.

Conclusion

Publicly shared archives of image data are an increasingly critical element of cross-disciplinary research, especially for clinical biomedical research where diagnostic images of the spectrum of human disease and its response to therapy are a scarce commodity. As genetic biomedical understanding develops, one of the significant contributions of clinical imaging will be to produce very large collections that can be subjected to statistical tests of validity. Without a greater confidence in the image de-identification process, open-access DICOM archives that can be queried to correlate with genetics will never achieve their potential. Some international efforts besides those described in this paper are ongoing with the intent to achieve similar ends [25]. In the absence of community consensus on image de-identification and user-friendly tools and SOPs, researchers have been understandably reluctant to create publicly accessible image archives.

This paper suggests that developments in standards and technology have removed key stumbling blocks to the creation of these valuable archives. The DICOM Committee, through Supplement 142, now offers a robust framework for de-identification meeting the privacy regulations. The incorporation of these guidelines into easy-to-use image acquisition and management tools, coupled with the increasing availability of open archive solutions, should facilitate the creation of the image archives needed for the next generation of biomedical research.

Acknowledgments The authors wish to acknowledge the support of the Radiological Society of North America in developing and promoting the deployment of CTP. We would also like to recognize the extensive contributions of members of DICOM Working Group 18 in the development of Supplement 142.

References

1. National Institutes of Health. http://grants.nih.gov/grants/policy/data_sharing/. Accessed 28 February 2011
2. Ohm, Paul, Broken Promises of Privacy: Responding to the Surprising Failure of Anonymization (August 13, 2009). University of Colorado Law Legal Studies Research Paper No. 09–12. Available at SSRN: http://ssrn.com/abstract=1450006. Accessed 28 February 2011
3. Nelson B: Empty archives. Nature 461–10:160–163, 2009
4. Vickers AJ: Whose data set is it anyway? Sharing raw data from randomized trials, Trials 2006. BioMed Central 7:15, 2006
5. Piwowar HA, Day RS, Fridsma DB: Sharing detailed research data is associated with increased citation rate. PLoS One. 2(3):e308, 2007
6. National Institutes of Health. http://cancergenome.nih.gov/. Accessed 28 February 2011
7. National Institutes of Health. http://biospecimens.cancer.gov/archive/cahub/default.asp. Accessed 28 February 2011
8. Branstetter 4th, BF, Uttecht SD, Lionetti DM, Chang PJ: SimpleDICOM suite: personal productivity tools for managing DICOM objects. Radiographics. 27(5):1523–1530, 2007
9. National Cancer Institute, Cancer Imaging Program. https://wiki.nci.nih.gov/display/CIP/Incorporation+of+DICOM+WG18+Supplement+142+into+CTP. Accessed 28 February 2011
10. Noumeir R, Lemay A, Lina JM: Pseudonymization of radiology data for research purposes. J Digit Imaging. 20(3):284–295, 2007
11. Hrynaszkiewicz I, Norton M, Vickers AJ, Altman DG: Preparing raw clinical data for publication: guidance for journal editors, authors, and peer reviewers. Trials 11:9, 2010
12. Health and Human Services. http://www.hhs.gov/ocr/privacy/hipaa/understanding/summary/. Accessed 28 February 2011
13. National Institutes of Health. http://privacyruleandresearch.nih.gov/pdf/FinalEnforcementRule06.pdf. Accessed 28 February 2011
14. National Institutes of Health. http://privacyruleandresearch.nih.gov/research_repositories.asp. Accessed 28 February 2011
15. González DR, Carpenter T, van Hemert JI, Wardlaw J: An open source toolkit for medical imaging de-identification. Eur Radiol. 20(8):1896–1904, 2010
16. National Institutes of Health. https://imaging.nci.nih.gov/ncia/login.jsf. Accessed 28 February 2011
17. National Institutes of Health. http://ncicb.nci.nih.gov/. Accessed 28 February 2011
18. National Institutes of Health. https://wiki.nci.nih.gov/dashboard.action. Accessed 28 February 2011
19. National Institutes of Health. https://wiki.nci.nih.gov/display/CIP/NBIA+at+CBIIT+Image+Collections. Accessed 28 February 2011
20. National Institutes of Health. https://cabig.nci.nih.gov/tools/sharable/cagrid_overview?pid=primary.2006-07-07.4911641845&sid=caGrid&status=True. Accessed 28 February 2011
21. National Institutes of Health. https://wiki.nci.nih.gov/display/Imaging/Limited+dataset+—+user+agreement. Accessed 28 February 2011
22. National Institutes of Health. http://mircwiki.rsna.org/index.php?title=CTP_Articles
23. National Institutes of Health. https://wiki.nci.nih.gov/display/CIP/CIP+Survey+of+Biomedical+Imaging+Archives. Accessed 28 February 2011
24. National Institutes of Health. http://www.idoimaging.com/index.shtml. Accessed 28 February 2011
25. Lien C-Y, Onken M, Eichelberg M, Kao T, Hein A: “Open source tools for standardized privacy protection of medical images”, Proc. SPIE 7967, 79670M, 2011. doi:10.1117/12.877989