Safe (Compliant) De-identification of Medical Images With Maximum Retention of Scientific Research Value

Stephen Moore[1], Justin Kirby[2], Kenneth Clark[1], John Freymann[2], Bruce Vendt[1], Carl Jaffe[3], Kirk Smith[1], David Maffitt[1], Lawrence Tarbox[1], Fred Prior[1]

[1]*Mallinckrodt Institute of Radiology, Washington University School of Medicine, St. Louis, MO 63110*

[2]*Clinical Research Directorate / CMRP, SAIC-Frederick, Inc., NCI-Frederick, Frederick, MD 21702*

[3]*Boston University School of Medicine, Department of Radiology, Boston, MA 02114*

Abstract

- Patient data made publicly available, through repositories such as The Cancer Imaging Archive (TCIA[1]), may constitute a breach of personally identifiable information if not properly de-identified.
- Imaging data are especially at risk as some intricacies of the DICOM format are not widely understood by researchers.
- To properly de-identify image data, one must understand what protected health information (PHI) exists and where that PHI resides and must have the tools to remove PHI without compromising the scientific integrity of the data themselves.
- DICOM public elements are defined by the DICOM Standard.
- However, private elements are not standardized and a common de-identification practice is to delete all private elements, a practice that removes the scientifically useful data as well as PHI.
- This poster describes some of the issues with protecting PHI while making the imaging data useful to researchers and presents a thorough methodology and set of open source tools for image de-identification.

Experience in De-Identifying TCIA Images

The Cancer Imaging Archive is a project funded by the National Cancer Institute to publish images collected from clinical trials for secondary research. This section describes some of the issues we have encountered during that process; we believe these are not unique to this project.

- **DICOM standard elements with well defined semantics are abused at the time of collection.** Some elements are encoded as text strings and are taken from technologist input at the console. Rather than using the intended purpose for the field, local practice may allow the technologist to enter PHI. For example, a site may enter the patient's Social Security Number in a field for image comments.
- **Modality vendors use private elements to encode acquisition parameters not yet documented by the DICOM Standard**. They also use private elements to record study or demographic information to support legacy data structures. Locating the proper DICOM Conformance Statement that documents the private elements is critical for proper de-identification, but there are some challenges.
- **Data elements that are essential for scientific measurements (e.g., MR Diffusion imaging parameters) are stored in private elements rather than the standard elements.** These data must be preserved for the images to be reused for future research.
- **Image providers or others involved in the original image submission remove information from the images that identifies the vendor model and software version.** This information is required to locate the proper DICOM Conformance Statement for the acquisition modality.
- **Even when present in the images, the software version can not always be traced back to a DICOM Conformance Statement published by a manufacturer.** The string for the software version in the image may use a different scheme than that published by the manufacturer. More simply, the manufacturer might not have published a DICOM Conformance Statement or gone out of business.
- **We do not have the ability to contact the original submitters and ask questions about the acquisition equipment and software versions.** As the management of TCIA, we are far removed from the original imaging studies. Many of the collections were acquired five or ten years before they were transmitted for publication. We have little ability to interview the original staff and determine the exact equipment type and software version.
- **Manufacturers do not always document all private elements in their DICOM Conformance Statement.** We believe we have located the proper conformance statement, but not all private elements found in a set of images are documented.
- **Sites may include screen captures with PHI or billing documents with DICOM data submitted for a research trial.**
- **Additional requirements for meeting US HIPAA requirements are documented by Freymann, et. al[2].** DICOM elements that will help indicate the presence of PHI in places such as the pixels themselves are not yet universally supported by manufacturers.

DICOM Attribute Confidentiality Profiles

- Tables 1 and 2 are taken from DICOM PS 3.15, Appendix E[3] and represent a methodology used to de-identify standard elements. The Table 1 codes define actions to be performed on standard elements.
- Table 2 is an extract of Table E.1-1 from DICOM PS 3.15. The columns in the table refer to different profiles and options with different levels of confidentiality.
- The DICOM Standard does not describe how to select or combine profiles and options. The goal of TCIA is to retain as much scientifically useful information in the images as possible while removing all PHI. These requirements mean that we cannot take the most simple approach that would include:
  - Delete all private elements
  - Delete or clean all standard elements that could possibly have PHI without review
- For the TCIA publication process, we have chosen the Basic Application Level Confidentiality Profile with the following options:
  - Clean Pixel Data
  - Clean Graphics
  - Clean Descriptors
  - Retain Longitudinal Temporal Information with Modified Dates
  - Retain Patient Characteristics
  - Retain Device Identity
  - Retain Safe Private Tags
- Conversely, we have explicitly chosen to not implement these options:
  - Clean Recognizable Visual Features (we do not obscure facial features)
  - Clean Structured Content (we do not publish DICOM SR objects)
  - Retain Longitudinal Temporal Information with Full Dates

Solution

Our goal is to release images for public use that contain no embedded PHI (standard or private elements) but contain as much data as possible for future researchers. Figure 1 shows the process and applications we use to de-identify images. The Knowledge Base contains the action codes defined for DICOM standard elements in PS 3.15 and action codes we have defined for manufacturer private elements based on reading conformance statements.

- In Step 1, contributing sites use the RSNA Clinical Trial Processor (CTP[4]) and a common script that we provide to partially de-identify and submit images to our central collection system. These are the Contributed Images in Figure 1.
- In Step 2, our Extraction Tool is used to organize images by manufacturer, modality, model and software version. This sorting is used in the next step.
- In Step 3, a Tag Sniffer application records the values for each standard and private element. The application uses the action codes in the Knowledge Base combined with the Extraction Tool output to generate a report that identifies elements that might contain PHI. In the first report, the Tag Sniffer does not report elements that we know are going to be changed (e.g., Study Date)
- In Step 4, a senior analyst writes a script for the RSNA CTP application based on the Tag Sniffer report.
- Images are de-identified using the CTP script in Step 5.
- The Tag Sniffer is run a second time on the de-identified images in Step 6. A more verbose report is generated in this step. We are checking that values that should have been changed (e.g., Study Date) are changed.
- Trained data analysts review the verbose output generated by the Tag Sniffer. They look for any data that contains PHI. Should any data be found, the CTP de-identification script will be updated and applied to the images again.
  - Any text element, standard or private, that is text based and might contain PHI is carefully reviewed at the end of the process.

Table 1. DICOM Action Codes for Confidentiality

Action Code	Intended Action
D	replace with a non-zero length value that may be a dummy value and consistent with the VR
Z	replace with a zero length value, or a non-zero length value that may be a dummy value and consistent with the VR
X	Remove
K	keep (unchanged for non-sequence attributes, cleaned for sequences)
C	clean, that is replace with values of similar meaning known not to contain identifying information and consistent with the VR
U	replace with a non-zero length UID that is internally consistent within a set of Instances
Z/ D	Z unless D is required to maintain IOD conformance (Type 2 versus Type 1)
X/ Z	X unless Z is required to maintain IOD conformance (Type 3 versus Type 2)
X/ D	X unless D is required to maintain IOD conformance (Type 3 versus Type 1)
X/ Z/ D	X unless Z or D is required to maintain IOD conformance (Type 3 versus Type 2 versus Type 1)
X/ Z/ U*	X unless Z or replacement of contained instance UIDs (U) is required to maintain IOD conformance (Type 3 versus Type 2 versus Type 1 sequences containing UID references)

VR: Specifies the data type and format of the Value(s) contained in the Value Field of a Data Element.

Type 1 Data Elements … shall be included and are mandatory elements.

Type 2 Data Elements …shall be included and are mandatory Data Elements. However, it is permissible that if a Value for a Type 2 element is unknown it can be encoded with zero Value Length and no Value.

Type 3 Data Elements … are optional Data Elements.

Table 2. Extract From DICOM Application Level Confidentiality Profile Attributes

Attribute Name	Tag	Basic Profile	Retain UIDs Option	Retain Patient Chars. Option	Retain Long. Full Dates Option	Retain Long. Modif. Dates Option	Clean Desc. Option
Accession Number	(0008,0050)	Z					
Acquisition Comments	(0018,4000)	X					C
Acquisition Date	(0008,0022)	X/ Z			K	C	
Patient ID	(0010,0020)	Z					
Patient's Birth Date	(0010,0030)	Z					
Patient's Name	(0010,0010)	Z					
Patient's Sex	(0010,0040)	Z		K			
Study Instance UID	(0020,000D)	U	K				

Knowledge Base

- Over time, we are able to build a knowledge base of private elements by reading DICOM Conformance Statements (DCS). As we publish imaging data from more and more trials, we add to the knowledge base for each different acquisition modality we discover. Over time, we begin to see similar modalities at different sites and are able to reuse the existing data in the knowledge base.
- The Knowledge Base is available on the TCIA wiki (https://wiki.cancerimagingarchive.net) and as a searchable database
- Figure 2 is an extract of a PDF document available on our wiki. It lists a subset of the private elements defined by GE for Signa MR modailties.
  - Cell marked with '*' means the element is defined in the DCS
  - Cell marked with 'X1' means the elements is listed in the DCS as no longer supported by the manufacturer.
  - Multiple models and software versions are included in one summary.
- Figure 3 shows a prototype of a web based interface that is available to the public. Researchers who receive images might discover private elements that their software does not understand. The web based system will allow researchers to enter some information about the private element (hexadecimal tag, manufacturer, modality, private creator ID) and find all private elements in our database that match the criteria.
- The first search returns a list of all elements that match the query criteria. A user may select any individual element from that list, and the software will make a further search and show a set of documents that are relevant to the private element:
  - DICOM Conformance Statements
  - Our summary documents (Figure 2) on our wiki
  - Spreadsheets that contain the action codes we have defined for private elements
  - CTP de-identification scripts

Figure 1. Process Diagram for Image De-Identification

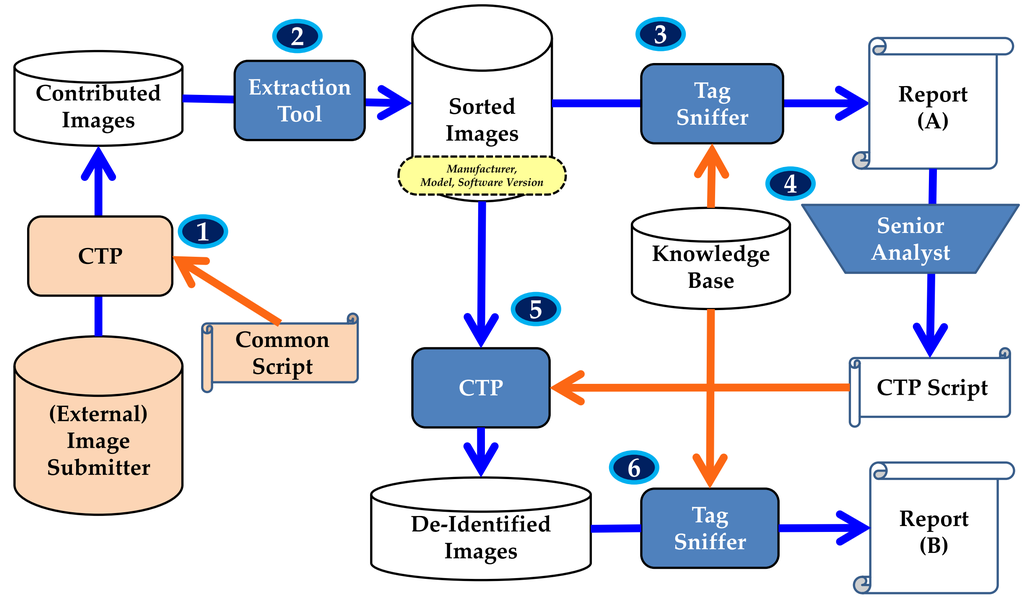

Figure 2. Example Summary of Private Elements

GEMS_ACQU_01

				EXCITE 3T/1.5T (11.0)	EXCITE 1.5T (11.1)	Signa EXCITE HD 0xarion(12.0)	Signa HDx 3.0T/1.5T (14.0)	Signa HDx 3.0T/1.5T (14.0) DOC0099380 Rev. 3
GEHC Private Creator ID	0x00190010	LO	1	*	*	*	*	*
Horiz. Frame of ref.	0x0019100f	DS	1	*	*	*	*	*
Series Contrast	0x00191011	SS	1	*	*	*	*	*
Last pseq	0x00191012	SS	1	*	X 1	*	X 1	X 1
Series plane	0x00191017	SS	1	*	X 1	*	X 1	X 1
First scan ras	0x00191018	LO	1	*	X 1	X 1	X 1	X 1
First scan location	0x00191019	DS	1	*	X 1	X 1	X 1	X 1

Figure 3. Prototype Private Element Query Interface

Enter group number (hex, 4 digits): 0019
Enter element number (hex, 4 digits): 105A
Modality: ---
Manufacturer: General Electric Medical Systems
Private Creator: ---
Lookup | Cancel

Private Element

Select	Group	Element	Modality	VR	Manufacturer	Private Creator	Description
Select	0019	105A	MR	FL	GEMS	GEMS_ACQU_01	Acquisition Duration

Modality Profiles

This section contains the set of modality profiles that reference the private element selected from the table above.

Profile Name	Profile Description	Documents
SIGNA MR	SIGNA MR	Signa Product Line DICOM Conformance Statement (Software Version 14.0) 2008-02-11 GE
		Signa Product Line DICOM Conformance Statement (Software Version 14.0) 2007-02-12 GE
		Signa Product Line DICOM Conformance Statement (Software Version 12.0) 2006-04-04 GE
		Signa Product Line DICOM Conformance Statement (Software Version 11.1) 2007-06-13 GE
		Signa Product Line DICOM Conformance Statement (Software Version 11.0) 2003-06-27 GE

Conclusion

- We have defined what we believe to be a rigorous system to de-identify public collections based on DICOM standard practices and manufacturer conformance statements.
- We have created open source tools and a database of private elements that will help researchers faced with similar tasks.

References

[1]

[2] Freymann JB, Kirby JS, Perry JH, Clunie DA, Jaffe CC., Image data sharing for biomedical research--meeting HIPAA requirements for De-identification. JDigit Imaging. 2012 Feb;25(1):14-24.

[3] DICOM Standard Part 15, Appendix E Attribute Confidentiality Profiles, pp. 60-92, PS 3.15-2011 (de-identification guidelines)

[4] CTP – The RSNA Clinical Trial Processor: Clinical_Trial_Processor

ements

: SAIC-F (PI: Prior) Image Archive

t with federal funds from the ealth, under Contract No.

y reflect the views or policies of the oes mention of trade names, orsement by the U.S. Government.

Contact: Steve Moore. Email: moores@mir.wustl.edu