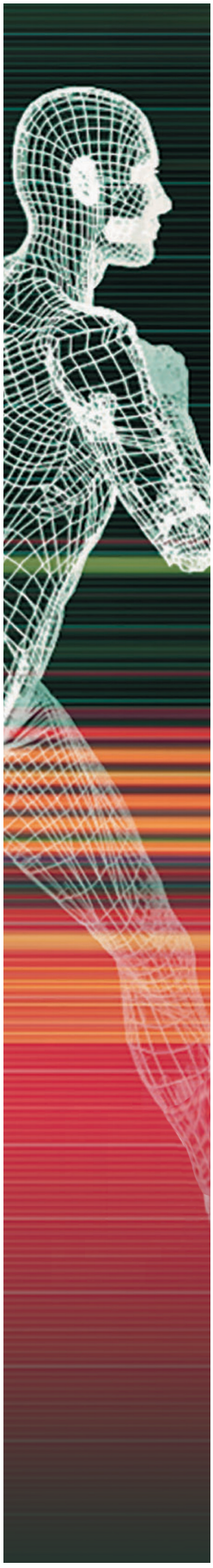
LAUREL BRIDGE
Orchestrating Medical Imaging Workflow

A White Paper



Anonymization of Patient Information in Medical Imaging; State-Of-The-Art and Challenges, Part 1

Written by: Herman Oosterwijk

Overview

Anonymization or de-identification of medical images has been a challenge for the past 30-40 years, ever since digital medical imaging was introduced. When medical imaging was film based, it was easy. To de-identify the image one would cut off the corner of the film where the patient information was recorded, or, when making a duplicate picture one would cover the patient information or use a black marker to obscure it.

With digital images, we need to anonymize the Personal Health Information (PHI) in the "meta-data," which is typically an encoded set of attributes known as the DICOM header. There can also be information embedded in the pixel data as well, as "burned-in" text, which creates its own set of issues.

One might wonder why we are still struggling with anonymization after these many years of experience, and the answer is that it is complicated. In part 1 of this paper, we will address the need for anonymization of images encoded in a DICOM format, how it works, standards guidance and conclusion. The next part addresses the regulatory guidance, top ten challenges, typical use cases and a final conclusion.

Why Anonymization?

1. To create teaching files to be shared between collegiate physicians and students, made available during conferences and in publications. The main purpose for anonymization is to maintain patient privacy. Most identifiers are replaced with random or dummy values and there is no traceability back to the original patient. Information that is clinically significant such as age, gender, body part, modality and procedure description have to be preserved. One needs to make sure that the new identifying information is identical to other studies for this patient so that one can see the progression and compare between different modalities and/or related information for that patient such as a lab results, genomics data, etc.
2. To generate data for clinical trials where certain patients are given a new or experimental drug or treatment and needs to be objectively evaluated. In this case the anonymization rules require one to be able to go back and check where the data came from, i.e. it needs to be traceable, therefore, random anonymization such as used for teaching files is not feasible.

Anonymization is needed for:

- Sharing teaching file images
- Clinical trial data
- AI training, test and validation datasets
- Image preparation for public AI cloud algorithms

Often, the DICOM metadata needs to be extended to include clinical trial identification information. Depending on the type of trial, more or less information can be eliminated, for example, sometimes patient sex is important, sometimes it is irrelevant as the trial is for one sex anyway (think prostate and/or ovarian cancer). As these studies must be traced back by authorized personnel, one uses artificial, traceable identifiers. This type of anonymization is typically referred to as de-identification in contrast to the randomized anonymization used for teaching files.

3. To provide training and test data-sets for AI algorithms that use machine and/or deep learning. Training data-sets are used during the development and for fine tuning of an AI algorithm, and test data-sets are needed for verification and validation of the algorithm. Especially for new diseases such as COVID-19, there is a big need for such data-sets to re-program existing AI algorithms to be used for detection and potential diagnosis.

There are two models with regard to accessing these data-sets, the first is a centralized model, in which one keeps the data in a central location that makes it publicly available. A good example of that is The Cancer Imaging Archive (TCIA)[1], funded by the National Cancer Institute (NCI). Another example is the Medical Image Resource Center (MIRC)[2] of the Radiological Society of North America (RSNA), which has both an anonymizer i.e. the Clinical Trial Processor (CTP) and an image repository.

The second model, which is the federated model, keeps the data on-site under the local jurisdiction and data protection rules. There are significant differences between regulatory requirements in different regions that can make public sharing of information non-compliant, in which case the federated model is preferrable.

4. To prepare imaging studies for processing by a "public" algorithm such as being provided in the cloud. Companies such as Amazon[3] and Google[4] provide public de-identification services but one might not be able or willing to use their services. Not only is the connection to the cloud possibly unsecured, but the party or application that submits the study to be processed might not have control or a HIPAA compliant Business Associate (BA) agreement with the algorithm provider to comply with the required patient privacy and security rules. This use case includes the transmission of studies over an unsecured channel such as a public network for a second opinion or over read.

How Does Anonymization Work And What Does It Include?

The information to be anonymized can consist two parts: 1) Burned-in identifiable text found in the pixel data. This is common with ultrasound, some X-ray devices like C-arms, and modalities based on visible light capture. This type of pixel data is often found when a screenshot of the display is taken and converted into a DICOM file. 2) The metadata as encoded in the DICOM header.

Regarding the DICOM header metadata, there are three categories of identifiable information. The first one includes the standard defined attributes, which includes information that can be used to trace back to the patient such as the patient demographics, i.e., name, DOB, sex, etc., and details about the equipment, operator, physicians and facility, all of which is relatively straightforward to identify and anonymize.

The second category is the private information added to the metadata by vendors. One could check the publicly available DICOM conformance statements to see which are used for patient identification, which is the approach that the TCIA folks took as the basis for their De-identification Knowledge Base[5], available on the TCIA website's wiki. However, this could be a challenge as these attributes are often not publicly documented; therefore a "selective" approach might be needed based on what is published and what is not.

PHI anonymization includes:

- Pixel data
- DICOM metadata (header)
- Standard and private tags as well as comment fields

The third category is the information that is "hidden" in comment fields, in study and series descriptions, or any field that can be entered or modified by a user. Especially if this information is combined with useful clinical information, it is a challenge as one cannot simply delete or replace the attribute. An example would be "Smith-CT/Head w/o Contrast" in a series description which combines both patient identifiable information and useful clinical data. There is probably a very good reason that this information is in this field as it supports the workflow for this institution that created the study, but these types of uses almost always require human inspection and intervention to eliminate or replace.

What About Re-identifying?

Re-identification of the data might be needed. For clinical trials it is a requirement to make sure that the data can be traced back to a "real" patient and is not faked to influence the trial outcome[6]. When submitting a data-set to a "public" processor, the processor might send back results either with the image or referring to the image. If, for example it includes an overlay or marker indicating a detection or finding, we want to make sure we can apply that back the to submitter who needs to see the original context and patient information.

Standards Guidance

There are two technical standards that give anonymization guidance to potential implementers and users, i.e. part 15 of the DICOM standard, which covers security profiles, and the definition in the Integrating the Healthcare Enterprise (IHE) profile[7] of the Teaching File and Clinical Trial Export (TCE).

Part 15 is very helpful as it identifies in a table all of the DICOM header attributes that potentially could or should be de-identified and lists several options for groups of attributes that should either be removed or retained based on the particular use case. A device can support one or more of the options listed in Part 15 in its DICOM conformance statement[8].

The IHE profile definition describes the teaching file and clinical trial use case defining an export manager which takes the selected images, de-identifies them, and forwards them to a destination accompanied with a so-called manifest that has a link to the de-identified images.

Conclusions

- The anonymization process cannot (yet) be fully automated. Processing an image with current anonymization software with confidence that it will be properly de-identified is not feasible unless one eliminates all potential PHI, which greatly reduces the clinical usefulness. This is despite its claim from several vendors, because it is difficult to identify all of the potential PHI in the different places in the metadata and in addition the image data where it can be hidden and/or mixed with useful clinical data.

- The process of anonymization is difficult. One needs a lot of flexibility as there are different interpretations, local rules, and a wide range of requirements depending on the use case.

- There is a need for a configurable and scalable anonymization processor that can handle different scenarios and complex workflows. When you consider a commercial solution, make sure to ask the right questions and look for a partner with a proven track record that can provide you with the flexibility needed to solve this complex problem.

In part 2 of this series, we discuss the regulatory guidance, top ten challenges, typical use cases and final conclusion.

About the author: Herman Oosterwijk is an experienced Health Care Imaging and IT professional who has been involved with PACS and related standards for over 30 years through participating in standards committees, publishing, training, and consulting.

1 https://www.cancerimagingarchive.net/
2 https://mircwiki.rsna.org/index.php?title=MIRC_Overview_-_CTP_and_TFS
3 https://aws.amazon.com/blogs/machine-learning/de-identify-medical-images-with-the-help-of-amazon-comprehend-medical-and-amazon-rekognition/
4 https://cloud.google.com/solutions/de-identification-of-medical-images-through-the-cloud-healthcare-api
5 https://pubmed.ncbi.nlm.nih.gov/25969931/
6 https://www.sciencedirect.com/science/article/pii/S1532046417302721
7 http://dicom.nema.org/medical/dicom/current/output/html/part15.html
8 https://www.laurelbridge.com/pdf/Compass-Dicom-Anonymization-Conformance-Statement.pdf