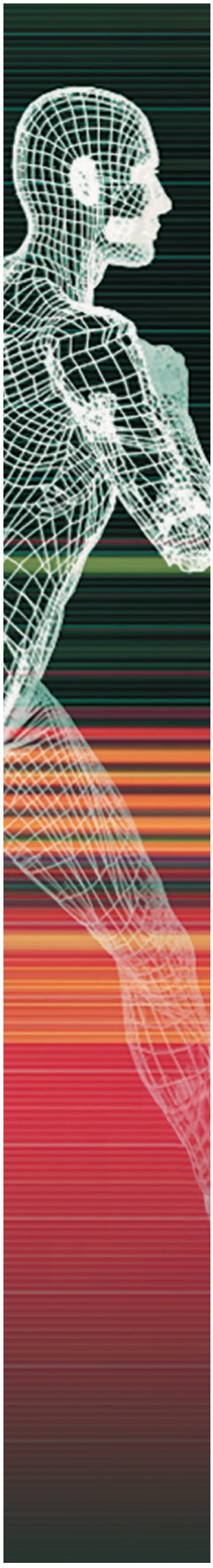
LAUREL BRIDGE
Orchestrating Medical Imaging Workflow

A White Paper



Anonymization of Patient Information in Medical Imaging; State-Of-The-Art and Challenges, Part 2

Written by: Herman Oosterwijk

Overview

In part 1 of this white paper series we discussed the need for anonymization of images encoded in a DICOM format, how it works, and standards guidance. In this part we'll cover the Top 10 anonymization challenges, typical use cases, and last but not least, regulatory guidance for anonymization.

Top ten anonymization challenges

There are several challenges that are related to the identification and de-identification process both with the implementation, and with the workflow related to the process and managing the potential results of the AI or clinical trial:

1. **There are different interpretations of regulatory requirements –** Interpretations of HIPAA/GDPR requirements range from very conservative by eliminating pretty much everything that could even remotely lead to the patient information, to more liberal approaches that take a risk that a study potentially could be traced back to a certain person although with considerable effort.

2. **It is hard to determine the clinical need in advance –** One cannot always predict how a data-set will be used in the future. The patient age, sex, weight and BMI might be irrelevant for a particular study, but for another study it could be a requirement. Once the data is anonymized, it is not possible to get access to the information that might be needed for another clinical use case. This creates a demand for a "dynamic de-identifier," which has access to the fully identified data, and creates a de-identified subset based on the specific requirements of the study instead of having a static data-set. This could be manageable for a particular university or enterprise but is hard to manage if one would like to provide open, public access to the data-set.

3. **Finding all of the PHI in the DICOM metadata tags –** There could be information in the following categories of data:

   - *Standard Tags as defined by the DICOM data dictionary* – De-identifying and/or anonymizing those would be straightforward.
   - *Standard tags that are misused* – An example is a physician name in the Accession Number or tags that violate encoding rules such as identifiable text in numeric fields. Such tag misuse is very hard to discover and often requires additional inspection.
   - *Private tags* – In many cases private tags are not documented in the vendor documentation or there could be undocumented tags that are added by an institution. These private tags often "hide" PHI but also can provide essential information to process the images. Therefore, simply eliminating all of the private tags is not acceptable. The most common approach is to identify those private attributes that are documented and validated to have no PHI and have those available as part of a "safe" data dictionary and remove anything that is not part of this data dictionary.
   - *Free text fields* – These fields could even contain a combination of clinical and PHI data. These attributes are also hard to de-identify.

Typical Anonymization Use Cases:

- Preparing studies for offsite, 3rd party research.
- Offsite AI algorithm processing.
- Preparing studies for use in teaching files or clinical trials.

4. **Finding all of the PHI in the "other-than-metadata" –** PHI can be "hidden" as:
   - Burned-in pixels such as screen saves or secondary capture files, for example from a bone densitometry (DEXA) scanner, ultrasound and C-arms often contain the information that is displayed on the screen including the PHI. It takes image processing such as Optical
   - Character Recognition (OCR) to identify any text and to identify what part of the text is PHI and what is useful clinical data. These pixels need to be replaced with dummy and/or black pixels.
   - Overlays are bitmaps or text fields that are either part of the metadata or encoded as a separate file as a Presentation State. The de-identification processor could flag suspicious overlays and make them available for a human for visual inspection.
   - Encapsulated PDF's and scanned documents are hard to de-identify and often require manual processing. Ophthalmology often uses pdf's for topographic maps and EKG's are often encoded as a pdf as well.
   - File names for an image might be stored on an exchange media or in a local file system with a file name that includes the unique identifier (UID), patient ID, name, or combination of those. Changing the names of the files is relatively straight forward but should not be forgotten.

5. **Determining if there are visual characteristics in an image –** This is especially important for dental and facial characteristics as there are many facial recognition applications even in the public domain that can be used for personal identification, such as to identify a person in a Facebook post, or to use for checking in at an airport, etc. One could try to "blur" the image but if the object of the image is to serve e.g. as input for plastic or reconstructive facial surgery, this would not be an option. The only possibility would be to get a patient consent; however, such consent might not cover any potential future use of these images.

6. **Determining visual characteristics in 3-D recons of image data-sets –** A head/neck cross-sectional image set such as a CT or MRI could be used to provide a 3-D surface reconstruction to identify the patient. Depending on the usage, for example if the data-set is used for neuro-imaging where the subject of interest is just the brain itself, one could "de-face" the image by distorting or removing the facial characteristics which has proven to be relatively effective[1]. But if the interest is for cranio-facial or facial and/or eye cancers, this is not possible. Facial recognition is currently widely available, unlike recognition of other body parts, but it has been proven that the signature or picture of a rib cage is as good as a fingerprint, therefore one might need to worry about other future identifiable body characteristics as well.

7. **Part numbers of prosthetics or implants –** A pacemaker has a unique serial number, as do other implantable body parts such as breast implants as they need to be tracked in case of potential manufacturing defects. These serial numbers could be visible through imaging and need to be de-identified as they could be used to trace back to a patient.

8. **Different real-time workflow requirements depending on how the de-identified data is used –** The usage of de-identified data e.g. by AI algorithms differs widely. Some algorithms need prior comparisons, some studies need to be staged so that the results are available before a physician looks at the actual image, some require prioritization if the AI determines triage, etc. Deploying AI and corresponding de-identification requires a flexible programmable workflow engine that can orchestrate the workflow as needed by the AI type and method.

9. **Different batch processing workflow requirements needing narrow data-set selection criteria –** A clinical trial or training data-set could include the selection of very narrow sets, e.g. fetch 3000 cases of COVID-19 diagnosis for male between 40 and 60 years old that have a CT 90 days prior and also an ultrasound. Again, one needs a very flexible configurable workflow engine to be able to do this.

10. **Scalability and efficiency –** Fast throughput to process many data-sets is important, maybe not so much if it is used in batch mode, but when an algorithm is to be applied real-time. If the algorithm to process the data-set is hosted in the cloud with potentially a slow public network connection, or when using an inefficient router that manages the de-identification, upload, and re-identification, the result might be too late to get back in time for the clinical decision process.

Examples of typical use cases/workflows:

- Exporting retinopathy images for research purposes. The Atlanta VA Health Care System (VAHCS) has extracted 100k+ diabetic retinopathy surveillance images to be used for validation of AI algorithms. A JSON object containing the applicable clinical information was added to the DICOM header in a text field (0040, A160) and the header data was de-identified using standard part 15 recommendations. The JSON object data was de-identified and dates for each patient were shifted by a random set amount. Birth dates were replaced with patient age.

- A cloud-based AI processing company might receive images that went through a local router that de-identified them, applies the algorithm, and responds by creating a new series that is sent back to a router, which in turn, re-identifies them and adds them to the study just in time before being interpreted by a physician. In some cases, the router will fetch applicable priors as well, which will follow the same path as the current study.

- Teaching file use cases, local and "public." When a physician selects one or more images, series or studies and labels them for "teaching file" use, additional information about the files could be added either in the metadata itself, e.g. as a comment field in the DICOM header or out-of-bound as additional information.

- Clinical trials workflow is similar to the teaching file case, i.e. images are labeled to be used for a clinical trial and exported to the clinical trial center. Instead of adding comments in the metadata (DICOM header) for the diagnosis, additional fields might be needed to indicate the clinical trial identification information.

Regulatory Guidance

The most important regulatory requirements are covered by the US federal HIPAA (Health Insurance Portability and Accountability Act) regulations and the European Community GDPR (General Data Protection Regulation) rules.

The US department of Health and Human Services (HHS) has issued a guidance document[2] to comply with the HIPAA Privacy Rule. Even though this guidance document is not

specific to imaging, there is no question that its compliance is required by anyone operating under the US jurisdiction. The HIPAA rule defines several categories of data elements which are a subset of the attributes defined in DICOM part 14 and therefore it appears that HIPAA compliance is achieved as long as Part 15 is followed. Interestingly enough, HIPAA also allows compliance as long as there is an "expert" who can make a documented argument that the information that is identified meet the privacy requirements.

Both clinical trials and AI development require access to data on a global scale. For example, COVID-19 drug trials are important to be performed based on data from different countries and continents. The same applies for collecting the data-sets for AI verification and validation data. As an example, the UK has established a repository of chest radiographs, CT and MR images and clinical data to support research and development of AI technology for the National COVID-19 Chest Imaging Database (NCCID[3]). Each data-set submitted needs to be de-identified meeting local regulations, which in Europe is the GDPR and in the US federal HIPAA requirements.

The GDPR ruling defines what they call pseudonymization[4] of PHI. Bryan Cave[5] published an article that defines the four levels of data and differentiates between the anonymization, de-identification and pseudonymization:

Pseudonymization involves replacing identifying fields by one or more pseudonyms, i.e. fictional identifiers. Pseudonymization is part of the data protection and privacy laws of the EU. Key to this technique is that it is reversible and that it allows linking of multiple data-sets, e.g. all the studies for a particular patient are identified with the same pseudonym and fictional ID. The reversal of pseudonymized data-sets is only allowed by authorized personnel otherwise it will be considered a privacy breach. It appears that pseudonymization

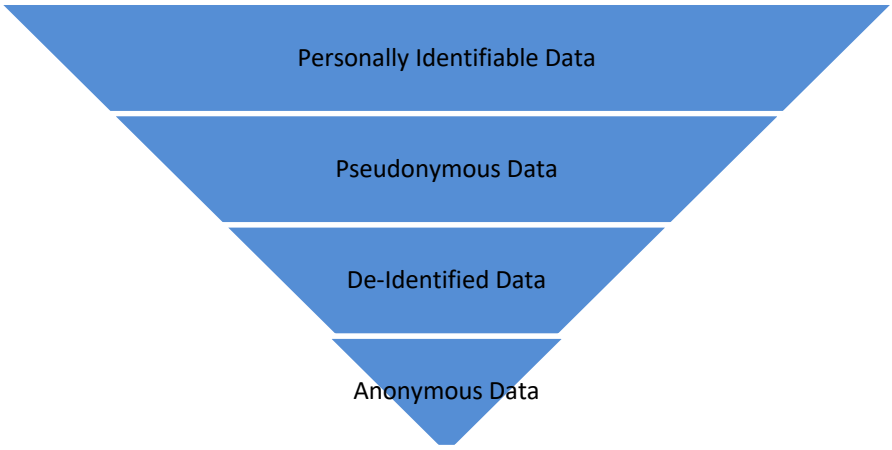


Fig 1: Four levels of anonymization per Bryan Cave

is a subset of de-identification for a specific use case, e.g. different identifiers are being replaced for example for a pharmaceutical trial than for exchanging the information with an outside laboratory.

In addition to the federal HIPAA and GDPR requirements, there could be other regulatory requirements from different countries such as Canada, China, India, and wherever there are large data-sets that are to be shared and accessed. To get around the different requirements, the federated model for testing AI algorithms is a solution, i.e. by passing the algorithm around to the different data-sets that stay local and guarded under the local jurisdiction. The local data management provider would pass the AI back with any results. This model opens up access to millions if not more cases to train AI with data-sets from different continents. Note the difference between the federated and distributed model, sometimes these models are mixed up, but the distributed model is merely used to distribute the processing power while the federated model uses distributed heterogeneous data-sets.

Final Conclusions

- There is no single one-fits-all solution for anonymization, de-identification or pseudonymization as each use case has its own set of requirements. Dynamic anonymization is often the only solution allowing narrow data-sets with specific requirements about case mix, priors, etc. to be created for specific use cases.
- Instead of using centralized file repositories for AI training data-sets, a better approach to meet local regulatory requirements is a federated approach by submitting the algorithm to the local repositories and receiving back the results.
- As mentioned in Part 1, the anonymization process is non-trivial. Fully automated processing is not feasible unless one decides to be ultra-conservative and remove and replace everything that remotely could lead to the patient identity, which makes the data-set potentially unusable for many clinical use cases. Many cases ultimately require a human to oversee the processing.
- And to reiterate the statement in Part 1, If you are in the market for an "anonymizer," make sure to ask the right questions and look for a partner who can show you that they have done this before and who understands the intricacies of this complex problem and who can provide you with the flexibility needed to provide a configurable and proven solution.

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://ieeexplore.ieee.org/document/8417227
2 https://www.hhs.gov/hipaa/for-professionals/privacy/special-topics/de-identification/index.html
3 https://www.nhsx.nhs.uk/covid-19-response/data-and-covid-19/national-covid-19-chest-imaging-database-nccid/
4 https://www.i-scoop.eu/gdpr/pseudonymization/
5 https://www.bclplaw.com/en-US/insights/at-a-glance-de-identification-anonymization-and-pseudonymization-1.html