DICOM METADATA - A USEFUL RESOURCE FOR BIG DATA ANALYTICS

Francisco Maestre[1], Ana Paula Sorrell[2], Christian Mata[2,3] and Miguel Cabrer[1]
[1] *Idonia Medical Exchange, Palma de Mallorca, Spain*
[2] *Universitat Politècnica de Catalunya, Barcelona, Spain*
[3] *Fundació de Recerca Sant Joan de Déu, Barcelona, Spain*

ABSTRACT

This project provides an overview of new ways to represent data combining patient access and Medical Images (DICOM) information, advanced use of medical imaging metadata and the process to anonymize the personal health information contained to facilitate medical research.

KEYWORDS

DICOM Metadata, Big Data, Personal Health Information, Anonymization

1. INTRODUCTION AND OBJECTIVES

Data is the world's most valuable resource and it is possible to find data everywhere. In medical images, data covers not only gigapixel images, but also metadata and quantitative measurements (Aiello et al. 2021). DICOM (Digital Imaging and Communications in Medicine) is a clear source of medical data, since it is the current standard for storing and transmitting medical images (Aiello et al. 2021) and related information (Savaris et al. 2014); this means it contains raw data imaging and all metadata related to the procedures of image acquisition and curation (Aiello et al. 2021).

Idonia is a medical imaging exchange platform that facilitates the collection, storage, delivery and visualization of medical images for medical centers, professionals and patients. Over 200 million medical images (DICOM) have been processed and delivered up to today; with them, a necessity arose: this was studying the medical data contained in them in order to see if any conclusions could be drawn from such examination. This was possible thanks to a service called Magic Link, a tool to deliver medical images to patients that replaces the use of CD/USB, which simplified the process of accessing the data.

So the purpose and objectives of the internal research project were basically two:

- Analyze that the metadata contained into the DICOM images is valuable data with concreate patient related information.
- Demonstrate that DICOM metadata and medical reports can be processed and anonymized to facilitate clinical research and collaboration between organizations.

2. DATA LAKE FOR MEDICAL IMAGING ACTIVITY

DICOM is the current standard for storing and transmitting medical images (Aiello 2021) and it is defined as the international standard for medical images and related information. It was originally developed by the American College of Radiology (ACR) and the National Electrical Manufacturers Association (NEMA) (Savaris et al. 2014). Its first publication, in 1993, revolutionized the practice of radiology, allowing the replacement of x-ray film with a fully digital workflow (Aiello et al. 2021).

The DICOM standard comprises a set of specifications regarding structure, format, and exchange protocols for digital-based medical images. In other words, it defines the formats for medical images that can be exchanged with the data and quality necessary for clinical use (Aiello et al. 2021). This means it contains raw

data imaging and all metadata related to the procedures of image acquisition and curation (AAPM 2008).

DICOM images consist of textual metadata (Kathiravelu et al. 2021); in fact, a DICOM file contains both the image and a large variety of data in the header (Aiello et al. 2021). Physically, the content of a DICOM file can be seen as structured at the data element level (Savaris et al. 2021), which means that the information recorded on the file are the attributes; these shall be ordered by increasing data element tag number and shall occur at most once in a data set (Aiello et al. 2021).

After two years of research, we created the Data Lake under an R&D project supported by the CDTI (*Centro para el Desarrollo Tecnológico Industrial*) from the Ministry of Science and Innovation of Spain. The purpose of this tool was to analyze all the information provided by the clients –hospitals at most– to provide them with relevant and new information about their own data and also to facilitate a way to analyze, process, anonymize, extract and exchange data in a legal way to facilitate clinical research and collaboration.

So, the data lake metadata information was created with standard data warehouse techniques based on cloud computing (ETLs, Bigquery and Google Cloud). The datalake was containing the metadata information combined with the accessibility information, who has accessed the images obtained from Idonia Magic Link functionality.

The Data Lake obtains and analyses some relevant information from DICOM metadata:

Figure 1. DICOM Metadata analyzed by Data Lake

The first analysis was done around medical imaging activity combined with some information stored in DICOM images, such as the imaging technique, manufacturer of the medical device, the patient's body part examined and modality of study, etc. This information was shared with clients via the development of a command center functionality that went beyond traditional dashboards: it also contained access information from both professionals and patients.

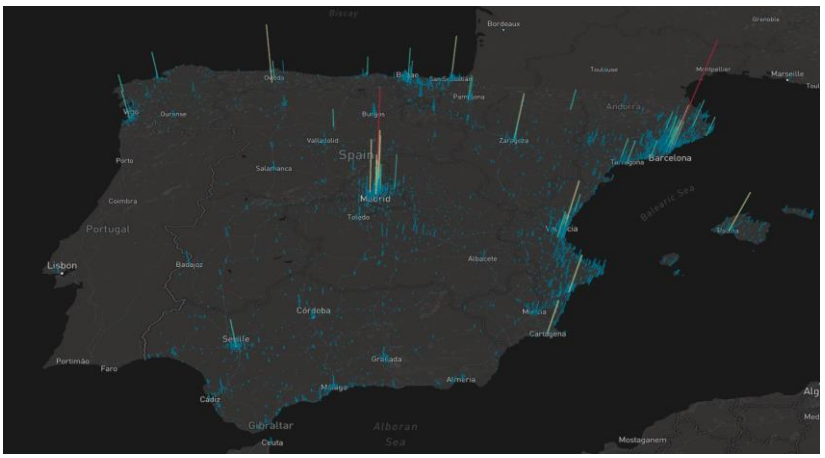

Figure 2. Example of information from patients accessing their medical studies

It was also possible to generate a *medical device activity map*, that provided some useful information about medical devices, generated information, their activity and remote access to their generated content.

Figure 3. Example of information from medical devices

2.1 Data Analytics Around Radiation

In the first objective of the research, we analyzed several DICOM Metadata Tags willing to encounter valuable concrete medical information that is relevant and is normally out of the scope of a traditional Electronic Medical Record. The radiation information contained in the DICOM Images was capturing our attention.

After years of studying medical related information, it was found that the most relevant data relied on radiation dose parameters, and from this, the idea of developing a digital tool to analyze radiation-related data originated.

In the current defined Directives and Regulations, no limits on the radiation dose are stipulated for patients undergoing diagnostic nor treatment procedures. There is proof that ionising radiation has direct implications in human health (ICRP 1990), which is why measures need to be taken as soon as possible.

These actions start with being able to quantify the radiation received by a patient in studies over time, which can be done by means of a DICOM dataset of metadata. With the available DICOM data, it is possible to quantify radiation with the effective dose, a parameter that serves as a dose descriptor (AAPM Report 2008).

The purpose of this research part of the project (Sorrell et al. 2022) was to extract and analyze the data and compare it with the SEPR (*Sociedad Española de Protección Radiológica*) and the ICRP (*International Commission on Radiological Protection*) recommendations. The result of this is differently structured/organized data sets, which makes data analysis a complex challenge. Since the information is valuable, different techniques can be applied in order to collect that information.

The imparted ionising radiation to a patient has always been the main concern in radiology, since there is substantial evidence of adverse effects due to radiation exposure. The process of ionization that occurs when imparting radiation to a patient changes atoms and molecules and may sometimes damage cells (ICRP 1990).

The European Council Directive 2013/59/EURATOM, of 5 December 2013, establishes uniform basic safety standards for the protection of the health of individuals subject to occupational, medical and public exposures against the dangers arising from ionizing radiation. It defines medical exposure as the exposure incurred by patients in order to be diagnosed or treated of any disease. In disparity with the dose limits for professional workers and the public in both occupational and public exposures, neither the Directive nor any of the Spanish Regulations establishes any limits on patient dose. In fact, in Article 6.1, it says that the radiological protection of the exposed patient will be optimized in order to keep the individual doses *as low as reasonably possible* (ALARP). Article 6.2 follows up with this and establishes that, in medical exposures, dose restrictions will only apply with respect to the protection of caregivers and volunteers involved in medical or biomedical research.

Let’s take Computed Tomography (CT) as an example. CT scans consist of a computerized x-ray imaging procedure in which a narrow beam of x-rays is aimed at a patient and quickly rotates around the body.

CT images are based on the different x-ray absorption rates of the various organs of the human body, which is why it provides both good soft tissue resolution (contrast) as well as high spatial resolution

(Zhanli et al. 2009). To ensure the best resolution, the dose imparted must be considerably high; in fact, the dose levels imparted in CT exceed those from conventional radiography and fluoroscopy and the use of CT continues to grow, often by 10% to 15% per year, which leads to a discussion of radiation risk versus medical benefit (AAPM Reports 2008). At the end of the day, what matters are the long-term repercussions of radiation exposure, which is why the American Association of Physicists in Medicine (AAPM) has defined several dose parameters to provide guidance on reasonable CT dose levels on routine examinations (AAPM 2008).

3. PROCESSING DICOM METADATA AND MEDICAL REPORTS FOR RESEARCH PURPOSES

The research organizations need real medical content in order to analyze the information and be able to cross data, build AI models or medical research. Idonia has been creating a relevant database of medical images (200 millions) combined with access information from patients and doctors. This allows to profile patients and medical devices and brings interesting information that can be analyze or combined with a medical data warehouse to complement and cross more information.

The Medical information is a high-level protected information by all country regulations (GDPR in Europe, HIPAA in the United States). So, in order to be able to facilitate a data lake that can be used for researchers or collaboration purposes between organizations, it has to be properly anonymized / de-identified.

Once the Data lake was created, the second step was to create a tool to search and navigate on the data lake and build the processes to anonymize data and process.

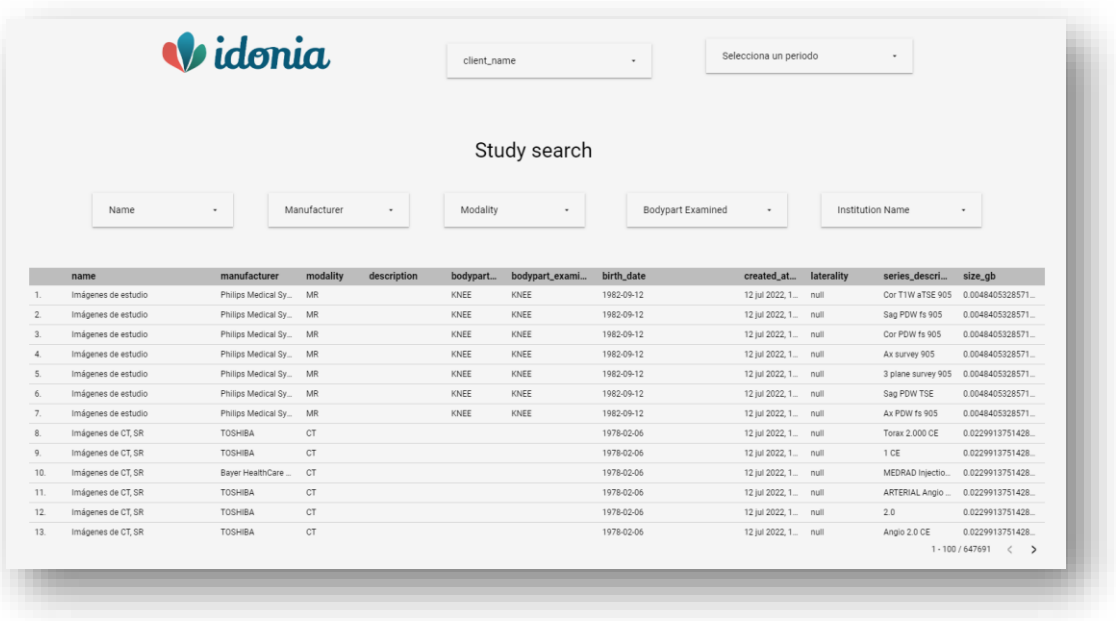

Figure 4. Tool to search for medical images on the DICOM Metadata Data Lake

The information associated to medical images is normally containing a DICOM structure (metadata and images) and medical reports (normally PDFs). The process of data manipulation the data was divided in two:

- DICOM Anonymization: through tag morphing techniques and ID Hash (tag removal, tag redaction, tag shifting) combined with PixelData De-identification in case image has printed data on it.
- Reports in SR, PDF, TXT format conversion, identification and removal of sensitive data, translation in multiple languages (to facilitate international research initiatives) exporting a de-identify and translated report.

The process took around 8 months to be implemented with double processed in some cases to ensure personal data is not involved. Techniques used for all the processes were OCR, Reports Translation, Tag removal, Hashing, data loss prevention, pixel data analysis.

4. CONCLUSION

The delivery of medical images in a secure and efficient way is a necessary functionality up to date, and specially patients are taking benefit from it. Accumulating those delivered images in a cloud infrastructure, analyze the data generated around them may bring great benefits for the medical centers to better understand their patient profiling and medical activity and promote research. An infinite loop of value is created.

DICOM files contain both the image and a large variety of metadata. This metadata provides valuable information for many different applications, including a radiation-related study. At the presented project, a Metadata Data Lake was created with deep analysis of valuable DICOM Tags (like radiation dose).

The data exploration executed was remarkable throughout and allowed a deep understanding on how DICOM files are constructed. The value of aggregating all the information from different sources in one common dataset (Data lake) opens new possibilities to analyze or enrich the data that can benefit all data providers.

Then the objective of demonstrating that that can be prepared, manipulated, combined and extracted without patient information and in a way that researchers can reuse it for their R&D projects was also achieved. It took time and the need of combining different techniques but finally a successful test was performer and validated by some institutions involved.

DICOM has some very interesting medical metadata available, thus can complement a Big Data and Analytics project in the medicine scope. The medical imaging delivery service based on cloud technology allows to combine disparate information like the remote accessibility from patients and professionals with the DICOM metadata. A proper data warehousing and personal data removal techniques offer the possibility to reuse the content for medical research purpose. This unique way to analyze the information is just a first step but the potential is huge. The aggregate information obtained from different sources through cloud infrastructure not only enriches the data source but provides more potential capabilities for deep learning data analytics.

REFERENCES

Aiello M, Esposito G, Pagliari G et al. (2021) *How does DICOM support big data management? Investigating its use in medical imaging community*. Insights into Imaging. 12(1):1-21.

Savaris A, Härder T, von Wangenheim A (2014) Evaluating a row-store data model for full-content DICOM management. Proceedings - IEEE Symposium on Computer- Based Medical Systems. 193–198.

Zhanli H, Hairong Z, Jianbao G, Ying Z (2009) *Real-time gray and coordinate statistics methods of medical CT image*. 3rd International Conference on Bioinformatics and Biomedical Engineering. 1-4.

Juszczyk J et al. (2021) *Automated size-specific dose estimates using deep learning image processing*. Medical Image Analysis. 68:101898.

Kathiravelu P, Sharma A, Sharma P (2021) *Understanding Scanner Utilization with Real-Time DICOM Metadata Extraction.* IEEE Access. 9:10621–10633.

ICRP (1990) ICRP Publication 60: 1990 *Recommendations of the International Commission on Radiological Protection*. Available at https://journals.sagepub.com/doi/pdf/10.1177/ANIB_21_1-3

American Association of Physicists in Medicine (2008). *The Measurement, Reporting, and Management of Radiation Dose in CT.* Available at https://www.aapm.org/pubs/reports/rpt_96.pdf

SEPR, "*SEPR - Requisitos básicos para los sistemas de registro y gestión de dosis en pacientes sometidos a exploraciones de diagnóstico por imagen*," 2020. Accessed: Jun. 14, 2022. [Online]. Available: https://revistadefisicamedica.es/index.php/rfm/article/view/328

"*Computed Tomography (CT).*" https://www.nibib.nih.gov/science-education/science-topics/computed-tomography-ct (accessed Jun. 14, 2022).

Ana Paula Sorrell Meriño, Christian Mata. *Analysis of Radiation Dose Using DICOM Metadata*. EEBE, Universitat Politècnica de Catalunya (UPC), Bachelor's Degree in Biomedical Engineering, defended 28/06/2022, Barcelona, Spain