Ref. Ares(2022)2414739 - 31/03/2022

A European Cancer Image Platform Linked to Biological and Health Data for
Next-Generation Artificial Intelligence and Precision Medicine in Oncology

Deliverable D4.1:
GDPR compliant anonymisation POSDA Suite

Reference	D4.1_ EuCanImage_institute_UAMS
Lead Beneficiary	University of Arkansas for Medical Sciences (UAMS)
Author(s)	Prior, Fred (UAMS) Smith, Kirk (UAMS) Nordell, Anders (CM) Harms, Alexander (EMC)
Dissemination level	Public
Type	Report
Official Delivery Date	March 31, 2022
Date of validation of the WP leader	March 31, 2022
Date of validation by the Project Coordinator	March 31, 2022
Project Coordinator Signature	

EuCanImage is funded by the European Union’s H2020 Framework
Under Grant Agreement No 952103

Version log

Issue Date	Version	Involved	Comments
29/03/2022	V1.0	Prior, Fred	Draft version for review
30/03/2022	V2.0	Prior, Fred	Revised and corrected final version.
31/03/2022	V3.0	Alexander Harms, Stefan Klein	Feedback WP3.
31/03/2022	Final	Karim Lekadir, Isabell Tributsch	Submitted version.

Executive Summary

Deliverable 4.1 aims to define for EuCanImage appropriate tools and procedures for acquiring anonymised or pseudonymised data in a GDPR compliant manner to meet Objective 4.1 "Create a comprehensive suite of open source tools and procedures for data anonymization that meet the legal requirements of all EU partners." Drawing on the decade of experience provided by TCIA (UAMS), Task 4.1 was undertaken to determine what modifications might be needed to adapt TCIA tools and procedures to accomplish this objective.

The document is structured in three main parts: **(1) Comparison of Anonymization Tools**. As an initial task we developed a synthetic data set and a procedure to compare the performance of anonymization tools and used these components to compare Posda tools with those provided by CMRAD as a commercial reference. (2) **Image Pseudonymization and Annotation Project.** The goal of this part was to test the end-to-end image transfer and annotation process starting with the appropriate legal framework at each data controller institution, the installation and operation of pseudonymization software at each institution, transfer of properly pseudonymized data to the cloud based annotation platform, completion of expert annotation and upload of images and annotation objects to the Euro-BioImaging repository. (3) **Validation of Posda GDPR Compliance**. The Clinical Trail Processor (CTP) software is used both by TCIA and Euro-BioImaging for anonymization and secure transfer of anonymized imaging data. This component of Posda was validated by EMC for GDPR compliance using the procedure used for validation of CTP for use by Euro-BioImaging.

Table of Contents

Acronyms

Name	Abbreviation
Perl Open Source DICOM Archive	Posda
General Data Protection Regulation	GDPR
Extensible Neuroimaging Archive Toolkit	XNAT
Clinical Trial Processor	CTP
The Cancer Imaging Archive	TCIA
Digital Imaging and Communications in Medicine (ISO 12052:2017)	DICOM®
Generative Adversarial Network	GAN
Protected Health Information	PHI
Personally identifiable information	PII
Collective Minds Radiology	CMRAD
Data Transfer Agreement	DTA
Artificial Intelligence	AI
European Union	**EU**

1 Introduction

This report will contribute to the achievement of the final goal of EuCanImage, the creation of a GDPR compliant integrated platform for large-scale cancer imaging, and AI solutions.

Working with WP1 we identified three use cases:

- Full anonymization,
- Pseudonymization within a GDPR compliant legal framework,
- Distributed annotation and machine learning where data does not leave the data controller site.

Our initial task was to determine the relative capabilities of existing anonymisation tools available to the EuCanImage data submission sites (data controllers) relative to best practices defined by TCIA in the US. The Posda tool suite[1] combined with procedures developed by and for TCIA implement anonymisation and pseudonymisation in compliance with the Digital Imaging and Communications in Medicine (DICOM®) Standard (ISO 12052:2017)[2]. Section 2 of this report summarizes a process for comparing anonymisation tools and the results of executing this process to compare CMRAD's Data Anonymiser and Posda.

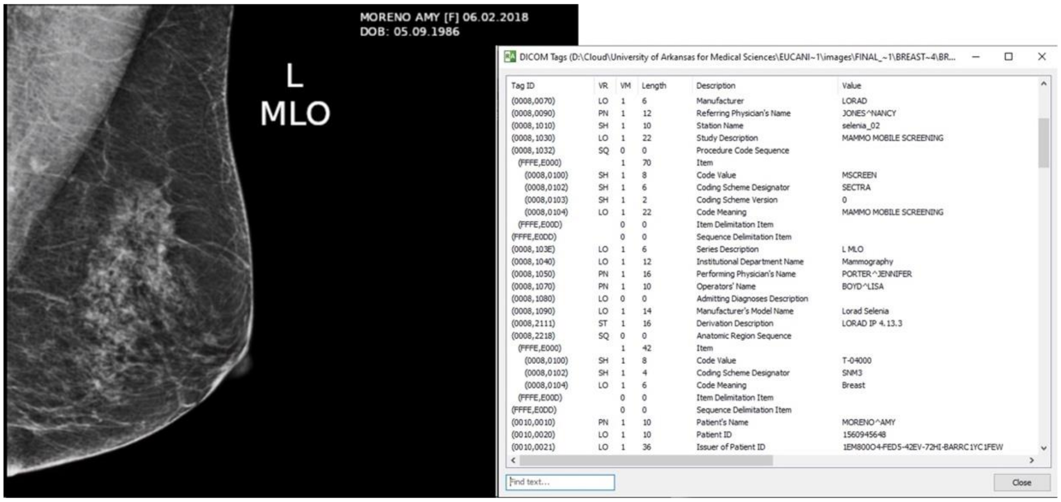
While the ultimate goal of EuCanImage is to employ full anonymization and make data available to other researchers outside of GDPR, the standards for claiming anonymization of data related to European persons are unclear. Section 4 of this report details our approach for dealing with this issue and highlights the need for clarity concerning the fate of the pseudonymization key in certifying GDPR compliant anonymization. For the bulk of work to be done within EuCanImage we will employ pseudonymization within a GDPR compliant legal framework and distributed machine learning approaches which do not require data to leave the data controller's site. Section 3 of this report in combination with D1.3 describe our initial implementation of this approach using CMRAD tools for pseudonymization.

2 Comparison of Anonymisation Tools

Synthetic data and an evaluation procedure were created and used in an experiment to compare Posda with similar tools provide by CMRAD. The goal of the experiment was to understand the differences resulting from the application of each approach and to review with each image submission site the significance of these differences relative to their site's national and institutional data sharing policies and regulations.

2.1 Anonymisation Tool Evaluation Dataset

The Figure 1 illustrates the basic structure of the synthetic data used for image anonymisation tool evaluation. Each DICOM data object (image) consists of a header containing metadata and a pixel matrix. Header data elements were created using a previously documented procedure[3] and the image pixels were generated using a GAN.



DICOM Tags (D:\Cloud\University of Arkansas for Medical Sciences\EUCANI~1\images\FINAL_~1\BREAST~4\BR...

Tag ID	VR	VM	Length	Description	Value
(0008,0070)	LO	1	6	Manufacturer	LORAD
(0008,0090)	PN	1	12	Referring Physician's Name	JONES^NANCY
(0008,1010)	SH	1	10	Station Name	selenia_02
(0008,1030)	LO	1	22	Study Description	MAMMO MOBILE SCREENING
(0008,1032)	SQ	0	0	Procedure Code Sequence	
(FFFE,E000)		1	70	Item	
(0008,0100)	SH	1	8	Code Value	MSCREEN
(0008,0102)	SH	1	6	Coding Scheme Designator	SECTRA
(0008,0103)	SH	1	2	Coding Scheme Version	0
(0008,0104)	LO	1	22	Code Meaning	MAMMO MOBILE SCREENING
(FFFE,E00D)		0	0	Item Delimitation Item	
(FFFE,E0DD)		0	0	Sequence Delimitation Item	
(0008,103E)	LO	1	6	Series Description	L MLO
(0008,1040)	LO	1	12	Institutional Department Name	Mammography
(0008,1050)	PN	1	16	Performing Physician's Name	PORTER^JENNIFER
(0008,1070)	PN	1	10	Operators' Name	BOYD^LISA
(0008,1080)	LO	0	0	Admitting Diagnoses Description	
(0008,1090)	LO	1	14	Manufacturer's Model Name	Lorad Selenia
(0008,2111)	ST	1	16	Derivation Description	LORAD IP 4.13.3
(0008,2218)	SQ	0	0	Anatomic Region Sequence	
(FFFE,E000)		1	42	Item	
(0008,0100)	SH	1	8	Code Value	T-04000
(0008,0102)	SH	1	4	Coding Scheme Designator	SNM3
(0008,0104)	LO	1	6	Code Meaning	Breast
(FFFE,E00D)		0	0	Item Delimitation Item	
(FFFE,E0DD)		0	0	Sequence Delimitation Item	
(0010,0010)	PN	1	10	Patient's Name	MORENO^AMY
(0010,0020)	LO	1	10	Patient ID	1560945648
(0010,0021)	LO	1	36	Issuer of Patient ID	1EM80004-FED5-42EV-72HI-BARRC1YC1FEW

Figure 1: Illustration of synthetic data used to evaluate Anonymisation Tools

For the experiment described in section 2.2 a total of 50 mammography images representing one image for each of 50 patients as created. Figure 2 outlines the steps used to create the synthetic data set. The data included difficult cases drawn from TCIA experience and DICOM syntax errors.

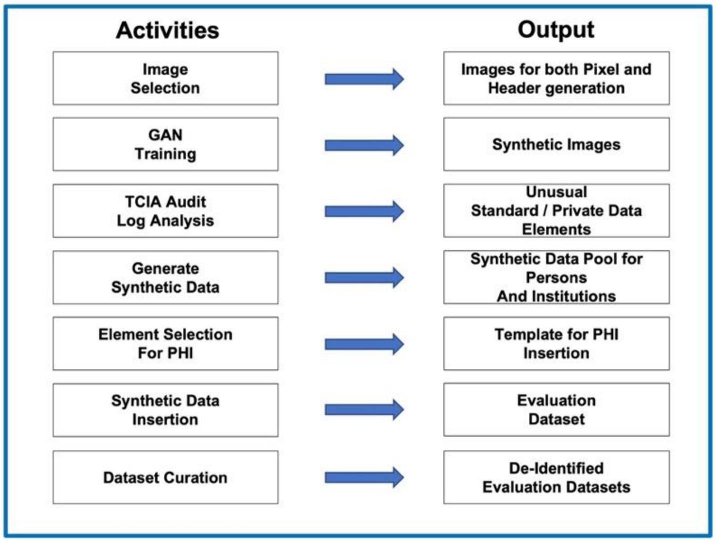

Activities
Output
Image Selection
Images for both Pixel and Header generation
GAN Training
Synthetic Images
TCIA Audit Log Analysis
Unusual Standard / Private Data Elements
Generate Synthetic Data
Synthetic Data Pool for Persons And Institutions
Element Selection For PHI
Template for PHI Insertion
Synthetic Data Insertion
Evaluation Dataset
Dataset Curation
De-Identified Evaluation Datasets

Figure 2: Illustration of Synthetic Data Creation Procedure

2.2 Anonymization Tool Evaluation Procedure

The DICOM standard defines profiles that detail what data elements contained in a DICOM information object (e.g., image, structured report, segmentation object) need to be modified and in what manner to achieve specified levels of anonymization and pseudonymization. Posda in combination with TCIA curation procedures de-identifies and "minimizes" data based on the DICOM Standard (PS3.15 2022a - Security and System Management Profiles) "**Basic**

Application Level Confidentiality Profile.” This profile is amended by inclusion of profile options:

- Clean Pixel Data Option,
- Clean Descriptors Option,
- Retain Longitudinal With Modified Dates Option,
- Retain Patient Characteristics Option,
- Retain Device Identity Option,
- Retain Safe Private Option.

The CMRAD Data Anonymizer also complies with DICOM Standard (PS3.15 2022a - Security and System Management Profiles) but retains and removes DICOM attributes to minimize data to meet European national and institutional regulations and requirements.

Figure 2 shows an overview of the Evaluation Experiment.

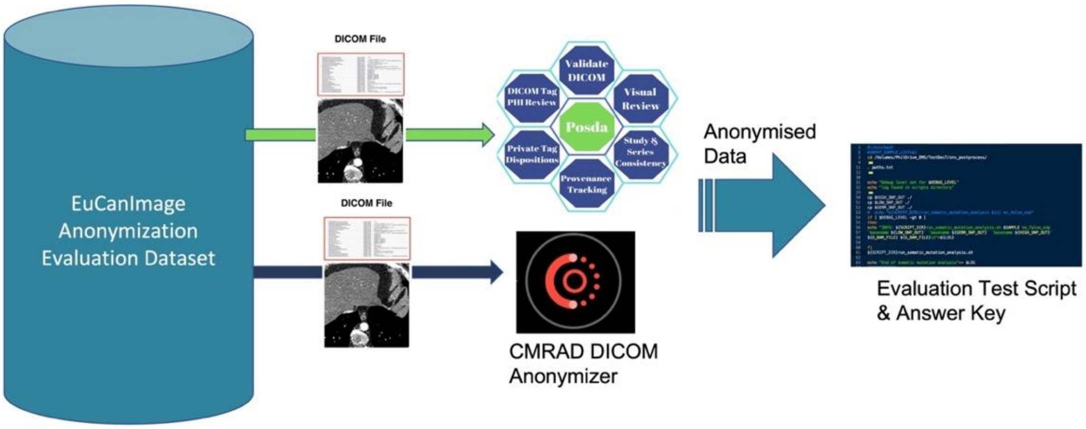


Figure 3: Illustration of Anonymisation Tool Evaluation Experiment

The Posda tools and TCIA procedures used in this experiment are described at: https://wiki.cancerimagingarchive.net/display/Public/Submission+and+De-identification+Overview. Similarly, the CMRAD tools and procedures are described at: https://www.cmrad.com/privacy.

Because the data used in the test was synthesized with PHI and PII placed in inappropriate locations (all locations based on actual TCIA experience), it was possible to create an Answer Key that identified situations that must be addressed by any anonymization procedure. Since the amount of data involved is too large for human visual inspection, a test script was developed that compared anonymized data objects that resulted from each arm of the experiment on a data element by data element basis and generated a text report of differences, as well and confirming that all unusual cases covered by the Answer Key were identified and addressed.

2.2.1 Experimental Results

Table 1 summarizes the results of the experiment comparing TCIA procedures and Posda tools with the corresponding procedures and tools used by CMRAD.

Type of Data	TCIA Action	CMRAD Action
PHI/PII in the image pixel data	Replaces pixels with black and retains the image	Deletes the image
Institution Name	Removed all instances	Removed some but not all instances
Patient size/weight	Retained	Minimized
Dates	Shifted so the actual date is unknown	Retained original date
Times	Retained	Removed
Private Data Elements	Retained scientifically valuable data	Removed
Images deleted	0%	22% (2,456/11,122)

Table 1: Summary of Experimental Results comparing Posda to CMRAD Data Anonymiser

2.2.2 Summary of Findings

The experiment identified key differences between the tools and procedures, most significantly the amount of image data that was lost in the CMRAD process. This was primarily the result of cases where PHI/PII was embedded in the image pixels. The CMRAD tools did not have the ability to remove all instances of data elements that were embedded in nested data structures (sequence data elements) in the DICOM image header. The other differences relate to accommodations made by CMRAD for European national and institutional requirements that are not taken into account by TCIA.

3 Image Pseudonymization and Annotation Project

The goal of the Project is to test the annotation process including the creation of DICOM annotation objects and the submission interface to the Euro-BioImaging Image Archive. To achieve this goal the project required implementation of the appropriate GDPR legal framework at each clinical site (data controller) and appropriate, site approved, pseudonymisation and data minimisation. The workflow at each clinical site includes:

- Use existing CMRAD pseudonymization/submission software available at each participating site,
- Submit data to a CMRAD data management system,
- Perform annotation using cloud based CMRAD tools,
- Create DICOM Structure Set annotation objects as a result of expert annotation,
- Create and utilize an API for submission to the Euro-BioImaging Archive (XNAT).

Deliverable D4.2 provides a complete description of the annotation project and its results. In this report we consider only the pseudonymisation and data submission process.

3.1 Legal Framework

WP1 and WP4 team members worked with each EuCanImage clinical site (data controller) to establish necessary DTAs and other legal agreements in accordance with the “Data processing and sharing based on data provider’s decisions and instructions” scenario defined in D1.1 for the centralised image analysis model. Data Processing and Sub Data Processing agreements were put in place between the data controllers and Collective Minds, and between Collective Minds and Euro-BioImaging as illustrated in Figure 3.

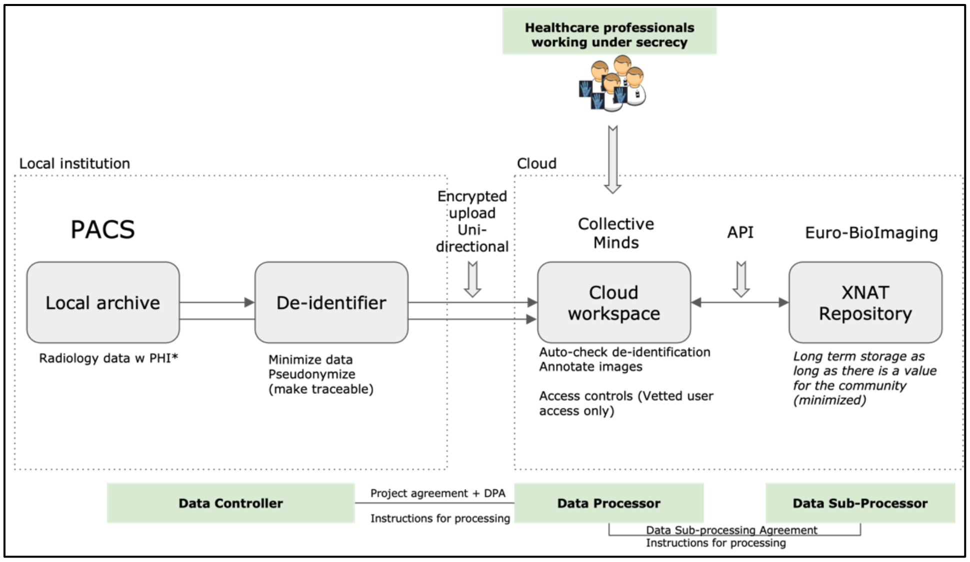

Healthcare professionals working under secrecy
Local institution
Cloud
PACS
Encrypted upload Uni-directional
Collective Minds
API
Euro-BioImaging
Local archive
De-identifier
Cloud workspace
XNAT Repository
Radiology data w PHI*
Minimize data
Pseudonymize
(make traceable)
Auto-check de-identification
Annotate images
Access controls (Vetted user access only)
Long term storage as long as there is a value for the community (minimized)
Data Controller
Project agreement + DPA
Instructions for processing
Data Processor
Data Sub-Processor
Data Sub-processing Agreement
Instructions for processing

Figure 4: Organizational and legal Framework according to GDPR with CM as Data Processor

3.2 Pseudonymization Process

Before data is transferred from the local institution it is pseudonymized using software and procedures where the identifiable characteristics are replaced with a hash key and a subject ID (pseudoID). This makes patient data traceable only for the data controller. Figure 4 illustrates this process.

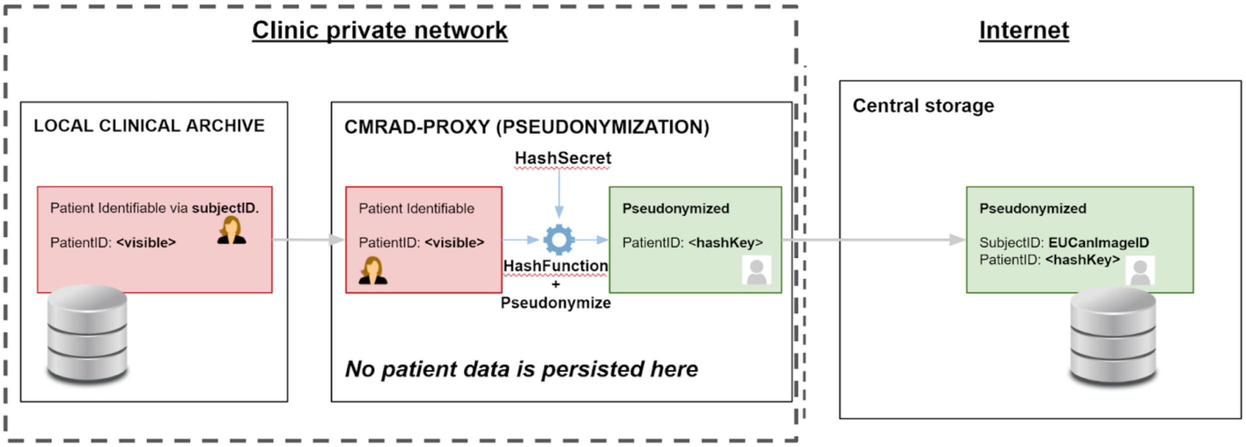


Figure 5: Pseudonymization process

Pseudonymization is achieved using the hashing algorithm SHA-2 512/256. The hash function inputs the patientID used by the data controller to identify a specific person and a hex encoded 64-character string (https://tools.ietf.org/html/rfc4648). The output from the hashing process is a unique non-traceable 265-bit hash Key which represents the subject ID when the data leaves the local institution (data controller).

For radiology images the above process of pseudonymization is automatic using a CE-marked tool (the CMRAD-PROXY) to ensure consistency. For other clinical data the same hashing system as for images will be applied to ensure data integrity and that any data parameter belonging to the data subject will be represented under the same subject ID. A hashing function is provided in an excel sheet or similar.

3.3 Results

We have now validated this framework, pseudonymization tools and procedures with all five clinical sites in EuCanImage (GUMED, UNIPI, UMAE, FCRB, KAUNO). Hence weare now proceeding with the data annotation process for all use cases and defined in D4.2.

4 Validation of Posda GDPR Compliance

CTP is a key component of Posda that is used to perform data pseudonymization and de-identification. EMC created a procedure for evaluating GDPR compliance of their image anonymisation and transfer pipelines that are also based on CTP; this procedure has been adopted by the Dutch national "Health-RI" infrastructure for ensuring GDPR compliance in multi-center medical imaging studies. The standard configurations of CTP used by TCIA and Health-RI were compared to determine the degree to which they differ.

As a starting point for the anonymization process, Health-RI uses a strict CTP anonymization script which conforms to the DICOM Basic Application Level Confidentiality Profile with no options, but with some modifications, e.g., it retains data elements such as series description which would be deleted under the Basic profile. Using this anonymization script all dates are set to the current date so all temporal information between image series is lost. Any sequence data element that is not specifically removed, will be retained and its contents (which may contain PHI) not modified. All private data elements are removed, even those containing critical scientific information. Key patient attributes (age, size, weight) and study specific attributes (e.g., whether contrast is used) are removed. Using this strict procedure, and following the EMC analysis procedure, CTP was found to fully anonymize data. From this point on Health-RI, together with the researchers of the particular study, adjust the CTP configuration on a per project basis in order to strike a balance between privacy and the scientific usefulness of the data set.

While this strict script differs greatly from TCIA's standard procedure, it can be used by TCIA and therefore by Posda to implement the initial phase of anonymization. It is worth noting that Posda supports a secondary stage of anonymisation incorporated into its curation procedures that is designed to detect any residual PHI/PII and remove it. While the evaluation of this stage is beyond the scope of this deliverable, it is logical to assume that it would only improve the completeness of anonymization provided by the Health-RI CTP script.

In summary, the data submission and primary anonymization component of Posda (CTP) has been evaluated by Health-RI and found to be GDPR compliant using a 'strict' anonymization pipeline. An analysis of CTP scripts used by Health-RI and by TCIA identifies critical issues related to the loss of scientifically valuable information when fully anonymizing DICOM data objects. Thus, while proving it is possible to fully anonymize data in a GDPR compliant manner using Posda, the result may not be optimal for EuCanImage AI analyses.

5 Final Considerations

Our analysis indicates that a strict anonymization of DICOM image data may render that data unsuitable for EuCanImage research due to loss of scientifically key information. We plan to work toward a compromise position that unites what we have learned from our experience in TCIA, in the image pseudonymization and annotation project (Section 3) and the analysis presented in Section 4, to develop more a optimal anonymization solution for EuCanImage over its duration. The pseudonymization tools and procedures currently in place and described in Section 3 provide EuCanImage an effective solution with which to move forward in this project, to enable the implementation of the research questions. Enhancements will be added to this basic framework as they are developed throughout the project. Complete anonymization will be deployed at a later stage to enable more open sharing of EuCanImage data.

6 References

1. Bennett W, Smith K, Jarosz Q, Nolan T, Bosch W. Reengineering Workflow for Curation of DICOM Datasets. Journal of digital imaging. 2018;31(6):783-91.
2. NEMA PS3 / ISO 12052, Digital Imaging and Communications in Medicine (DICOM) Standard Rosslyn, VA, USA: National Electrical Manufacturers Association; 2020. Available from: http://medical.nema.org/.
3. Rutherford M, Mun SK, Levine B, Bennett W, Smith K, Farmer P, Jarosz Q, Wagner U, Freyman J, Blake G. A DICOM dataset for evaluation of medical image de-identification. Scientific Data. 2021;8(1):1-8.