EUCAIM
CANCER IMAGE EUROPE

Project title: European Federation for Cancer Images
Project acronym: EUCAIM
Grant Agreement: 101100633
Call identifier: DIGITAL-2022-CLOUD-AI-02

D5.9: The EUCAIM tools for Data Preprocessing and Data/Metadata Management & Interoperability

Responsible partner(s): MEDEX

Author(s): Laure Saint-Aubert (MEDEX), Celia Martin Vicario (QUIBIM), Alejandro Vergara (Quibim), Xavier Rafael Palou (Quibim), Ana Jiménez (Quibim), Thomas Trouillard (MEDEX), Carina Soler Pons (HULAFE), Pedro Miguel Martinez (HULAFE), Adrian Galiana (HULAFE), Aikaterini Vraka (HULAFE), Manuel Marfil (HULAFE), Leonor Ceda (HULAFE), Matias Fernandez (HULAFE), Alexandra Kosvyra (AUTH), Dimitris Fotopoulos (AUTH), Ioanna Chouvarda (AUTH), Feriel Ramdhane (CNR), Dario Longo (CNR), Diego Silveira (ITI), Silvia Ruiz (ITI), Dimitrios Zaridis (FORTH), Eleftherios Trivizakis (FORTH), Katerina Dovrou (FORTH), Katerina Nikiforaki (FORTH), Vasilis Pezoulas (FORTH), Nikolaos Tachos (FORTH), Stelios Sfakianakis (FORTH), Valia Kalokyri (FORTH), Daniel Capellan (UPM), Tobias Kussel (DKFZ), Enola Knezevic (DKFZ), Moanes Ben Amor (DKFZ), David Rodriguez Gonzalez (CSIC-IFCA), Ines Victoria Rodriguez (CSIC-IFCA), Miguel Chambel (F. Champalimaud), José Almeida (F. Champalimaud), Christian Salvatore (DeepTrace), María Jesús Ledesma Carbayo (UPM).

Reviewer (s): Maria Gonzalez Lopez (SAS), Carlos Luis Parra-Calderon (Institute of Biomedicine of Seville)

Date of delivery: 30 December 2025

Version: 1.0
Link to demonstrable:
https://www.youtube.com/playlist?list=PL3Q1XjQpjfg_GEmwPDrQeESh6nqCMnYyR

Table of contents

Disclaimer

Abbreviations

AI: Artificial Intelligence

API: Application Programming Interface

DH: Data Holder

DICOM: Digital Imaging and Communication In Medicine

DIPG: Diffuse Intrinsic Pontine Glioma

ETL: Extract, Transform, Load

EUCAIM: European Federation for Cancer Images

FAIR: Findable, Accessible, Interoperable, Reusable

FLAIR: Fluid-attenuated inversion recovery

GDPR: General Data Protection Regulation

HCC: Hepatocellular carcinoma

IDAT: identifying attributes

MR: Magnetic Resonance

OMOP: Observational Medical Outcomes Partnership

OCR: Optical character recognition

PIDs: Personal Identifiers

1. Introduction

Aim and scope of the deliverable

This report completes the information presented in the demonstration video and jointly contributes to Deliverable D5.9 The EUCAIM tools for Data Preprocessing and

Data/Metadata Management & Interoperability. It should be seen as an update to the reports and demonstrators delivered at month 24 of the project under D5.4 “Data Pre-processing Tools and Services” as well as under D5.5 “The EUCAIM tools for Data and Metadata Management and interoperability”. At the time of those deliverables, only 14 tools were ready to be presented in demos, among which several tools have gone through improvements and refactoring since. The data preparation workflow has also been revised and is now extended to clinical data in addition to imaging data. It will all be presented in this deliverable.

As a reminder, The European Federation for Cancer Imaging (EUCAIM) provides preprocessing tools and services that prepare data for GDPR compliance, annotation, harmonization, quality checks, and sharing within the EUCAIM infrastructure for future reuse. These tools also ensure adherence to FAIR principles (Findable, Accessible, Interoperable, Reusable). In a federated environment, data originates from diverse sources with varying formats, standards, and inconsistencies. Consistent preprocessing reduces these disparities, creating a reliable foundation for data-driven applications such as developing and validating Artificial Intelligence (AI) algorithms.

This document aims to add context to understand the demonstration videos. It is structured as follows: section B provides a short description of the data preprocessing tools on-premises by category (for de-identification, quality assessment, etc). Section C details the preprocessing of data in the reference nodes, including further quality assessments, annotation and harmonisation tools, but also FAIRness assessment. Finally, section D outlines future work regarding EUCAIM preprocessing tools.

The demonstrator from previous D5.4 showed videos of preprocessing scenarios with specific use cases, and focusing on imaging data. The present deliverable instead gives a comprehensive description of all the preprocessing tools currently registered to the tool catalogue of EUCAIM, with individual demo videos when available.

Preprocessing tools preparation and validation

Currently, **55** preprocessing tools are part of the EUCAIM catalogue, from which 13 are applicable for data holders to prepare their datasets on premises (Figure 1), and 32 are available for data users to run from the reference nodes. Yet, additional tools may later be incorporated into the tools catalogue. **Thirty-four tools** are already showcased using demonstrator videos accompanying this document. The link to access all demos is the following:
https://www.youtube.com/playlist?list=PL3Q1XjQpjfg_GEmwPDrQeESh6nqCMnYyR.

EUCAIM defines three levels or “Tiers” for data compliance, to facilitate the participation of new data holders (DHs) and ease the incorporation of their datasets into EUCAIM data collections. The Tiers (thoroughly described in D4.3) have different data preprocessing requirements, and DHs are encouraged to curate their data to reach higher Tiers. The minimum requirements for all Tiers’ datasets is as follow:

- Image and clinical data must be linked using the same patient ID.
- No entity can exist more than once within your dataset.
- Imaging data should be in DICOM format and associated annotations and segmentations, when available, should be in DICOM-SEG format. Specific cases will be considered if a DH has diagnostic images in different data formats.
- All images should include a minimum set of imaging metadata and comply to the DICOM standard
- Imaging data must be accompanied by a set of minimum clinical metadata.

Most existing preprocessing tools apply to all Tiers, unless stated otherwise in the description below.

To ensure that data processing respects data and infrastructure privacy, and data integrity, a validation process for tool incorporation has been defined together with WP6 and WP7, and updated since last description in D5.4. This validation is a three-steps process that each requires several actions:

1. Software on-boarding
    - Registration of the tool in the bio.tools catalogue
    - Completion of technical documentation
    - Completion of a self-risk assessment questionnaire
    - Provision of a licence and description of said licence in bio.tools
2. Software evaluation (done by peer-reviewers)
    - Technical check (done by peer developers)
    - Legal check (done by EUCAIM legal team)
3. Dockerization and registration to the EUCAIM harbor (software registry)
    - Upload of the docker image (or executable files) to the harbor
    - Creation of a demo video

This comprehensive process ensures having stable and well-validated tools provided in the EUCAIM catalogue, as no tool can be released to the end-user until the full validation framework is complete. It also requires important work from the tool developers. Some of the tools below are fully developed, validated and accessible in the EUCAIM harbor. When this is the case, a link to a demo video is given at the end of the tool description. The current status (at the time of submitting this deliverable, at M36) of each tool’s readiness is available in Annex 1.

2. Preprocessing tools on premises

Some of the EUCAIM preprocessing tools are dedicated to DHs to prepare their datasets to fulfil the mandatory requirements for EUCAIM data collections, before being shared in EUCAIM. They are all optional tools downloadable form the EUCAIM harbor, that may be used at the DH premises.

2.1. Access and download local tools from the harbor

Data holders and data users can get information about the data preparation tools (listed in the following subsections) in the bio.tools catalogue (https://bio.tools/t?domain=eucaim). The binaries of the tools can be downloaded from the EUCAIM Software artifacts registry (https://harbor.eucaim.cancerimage.eu/harbor/projects/3/repositories). Only data holders and project members can download the tools: the access to the registry requires a valid account and additional permissions that can be requested on the first access to the registry. Instructions on the downloading and usage of each tool are provided in the description of the tools in the bio.tools catalogue. For clarity, a step-by-step guide on how to access the Harbor repository and download the required tools is available here: Accessing the Harbor and Downloading Tools.

There are currently 12 tools in the EUCAIM catalogue for data preparation by DHs (Figure 1). To that is added one tool for data transfer to UPV reference nodes.

Figure 1: Data preparation tools available in the EUCAIM catalogue, to be used on clinical premises by data holders.

2.2. Data preparation workflow on premise

Data preparation should follow a step-wise workflow, as described in the Handbook section on data preparation (see Annex 2).

For each step, EUCAIM preprocessing tools may be used by DHs on their premise for dataset preparation. No EUCAIM tool is considered mandatory to use.

2.2.1. Tier 1

- Step 1: Imaging annotation (optional): DHs may want to enrich their datasets with imaging annotations, or convert existing annotations to a valid format. See section 2.3.1.

- Step 2: De-identification: DHs must ensure that no identifiable information (direct or indirect) is present in their dataset. See section 2.3.2.

- Step 3: Data quality check: DHs must check that their datasets comply with the EUCAIM data quality framework. See section 2.3.3.

- Step 4: Data transfer (optional): This only applies to datasets intended to be transferred to a EUCAIM reference node, and does not apply to federated nodes. See section 2.3.5.

2.2.2. Tiers 2 and 3

- Step 1: Clinical data structuring: In order to have interoperable data that can be queried and processed, DHs must provide additional information regarding their dataset structure. No EUCAIM tools apply to this step. See Annex 2.

- Step 2: Imaging correspondence with clinical data: In order to successfully link the imaging exams with the clinical information provided by DHs, especially the timepoints of each episode, the correspondence between each imaging study and each clinical episode must be retrieved. See Annex 2.

- Step 3: Imaging annotation (optional): DHs may want to enrich their datasets with imaging annotations, or convert existing annotations to a valid format. See section 2.3.1.

- Step 4: De-identification: DHs must ensure that no identifiable information (direct or indirect) is present in their dataset. See section 2.3.2.

- Step 5: Data quality check: DHs must check that their datasets comply with the EUCAIM data quality framework. See section 2.3.3.

- Step 6: Data conversion to EUCAIM CDM: Clinical and imaging datasets mapping to the EUCAIM CDM is recommended for Tier 2 and mandatory for Tier 3 datasets. See section 2.3.4

- Step 7: Data transfer (optional): This only applies to datasets intended to be transferred to a EUCAIM reference node, and does not apply to federated nodes. See section 2.3.5.

2.3. Description of the preprocessing tools on premises

2.3.1. Data annotation

DHs may want to annotate their imaging data to enrich their dataset. Annotations should be provided in DICOM-SEG format. EUCAIM annotation and converter tools will ensure this format is respected (Figure 2).

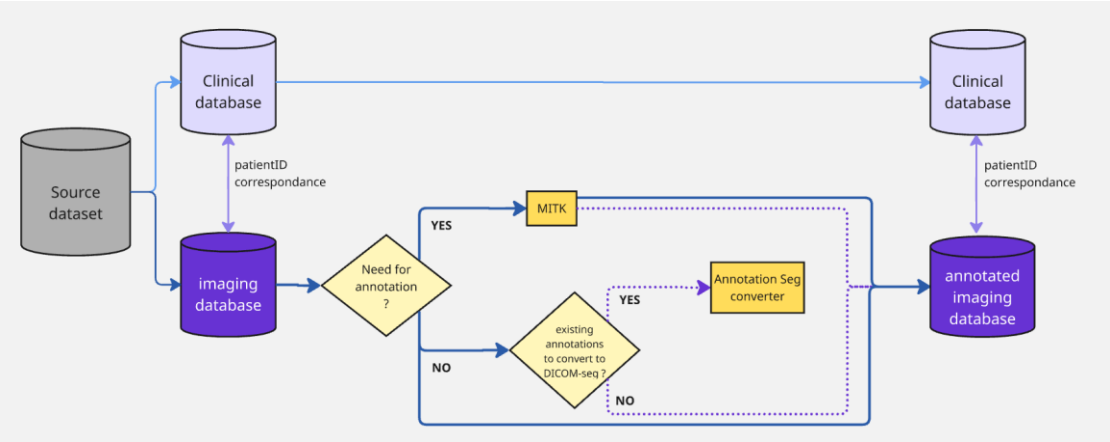

Figure 2: Annotation workflow using EUCAIM tools.

2.3.1.1. Annotation SEG Converter (SynLab)

This tool provides capabilities for segmentation files conversion to DICOM Segmentation (DICOM SEG) format, supporting conversion from both DICOM RTSTRUCT and NIfTI-based segmentations. The tool converts DICOM Radiotherapy Structure Set (RTSTRUCT) files, together with their referenced DICOM image series, into DICOM SEG objects. In addition, the tool supports bidirectional conversion between NIfTI and DICOM SEG, enabling comprehensive interoperability centered on the DICOM SEG standard. A demo video of the tool is available here: https://youtu.be/w4CVG1qrqak?si=M08RY0Wi0Tjx6_OF

2.3.1.2. Medical Imaging Interaction Toolkit (MITK) viewer (DKFZ)

The MITK is a free open-source software for the development of interactive medical image processing software. Based on MITK, there is the MITK Workbench, a powerful and free application to view, process, and segment medical images. In addition, MITK integrates EUCAIM vocabulary support.

2.3.2. De-identification

DHs must ensure that no identifiable information (direct or indirect) is present in the dataset they will share. Special attention should be given to **embedded text** in images, that may contain patient-identifiable information, as well as **skull and head images** that pose a risk of patient re-identification. DHs may need to apply additional de-identification techniques to mitigate this risk. The tools from the EUCAIM catalogue (Figure 3) provide ways to fully de-identify imaging and clinical datasets as per EUCAIM requirements, and to assess the risks for re-identification.

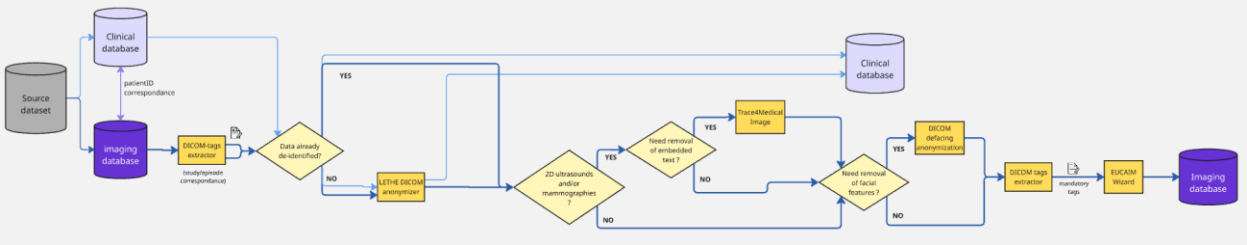

Figure 3: De-identification workflow using EUCAIM tools.

2.3.2.1. LETHE Anonymizer (FORTH)

A new de-identification tool has replaced the former one ("EUCAIM DICOM anonymizer") in the EUCAIM catalogue. This tool has new features that will further improve user experience and the de-identification process:

- The DICOM Anonymization Pipeline is now a **dockerized** version instead of standalone;
- The tool performs **OCR** on DICOM pixel data to remove sensitive information (burned-in information) - note that this step is optional, and by default no OCR will run; both Tesseract and PaddleOCR engines may be selected.
- Input and output directories can now be specified as parameters
- The tool works **both as an anonymization and pseudonymization** tool
- The tool can **restructure and rename** image folders ensuring no PHI is present
- The tool can **store anonymized Study IDs** in a dedicated csv file in order to associate clinical data and episodes/events to specific imaging studies in order to not lose association.

A demo video of the tool is available here: https://www.youtube.com/watch?v=yM9skIu_UfI&list=PL3Q1XjQpjfg_GEmwPDrQeESh6nqCMnYyR&index=34.

2.3.2.2. Mainzelliste (DKFZ)

Mainzelliste is a web-based first-level pseudonymization service. It allows for the creation of personal identifiers (PID) from identifying attributes (IDAT), and thanks to the record linkage functionality, this is even possible with poor quality identifying data. The functions are available through a RESTful web interface. Mainzelliste also includes built-in consent management and supports FHIR APIs

2.3.2.3. Trace4MedicalImageCleaning (Deep Trace)

This tool is aimed at automatically detecting and removing text in medical images, with a specific focus on 2D ultrasound and mammography studies. The tool applies to images in DICOM format.

2.3.2.4. DICOM Defacing Anonymization tool (UMEA)

This tool removes facial features ("defacing") from DICOM images. It may be used 1) only for facial removing, or 2) for a complete import and anonymization using DICOM transfer. Defaced images must be visually checked, which can be done through the tool's web interface.

2.3.2.5. The EUCAIM Wizard tool (FORTH)

The Wizard tool aims to perform an analysis of data re-identification risks for both imaging and clinical data that follow the EUCAIM CDM. It includes and uses an EUCAIM specific configuration of the ARX Data Anonymization Tool (biotools:arx), by supporting a wide variety of privacy and risk models as well as methods for analyzing the usefulness of output data. The tool receives as input the metadata in tabular format (CSV or XLSX), as they are provided for instance by the DICOM tag extraction tool. The results provided by this tool are for informational purposes only and should not be considered as professional advice. The output risk assessment must be reviewed by the DHs and assess whether corrections must be made to the dataset before submitting it to the EUCAIM technical team for certification of the dataset's compliance. A demo video of the tool is available here: https://www.youtube.com/watch?v=MDwKyuvw8q4&list=PL3Q1XjQpjfg_GEmwPDrQeESh6nqCMnYyR&index=2.

2.3.3. Data quality assessment

As per the EUCAIM data quality framework, datasets must be:

- **Complete**: all required data values are present
- **Unique**: no entity exists more than once within the dataset

- **Consistent**: dataset values of two sets of attributes within a record / within a data file / between data files / within a record at different points in time, comply with a rule
- **Accurate**: correspondence between dataset values to real values
- **Showing integrity**: absence of data value loss or corruption

Several tools from the EUCAIM catalogue can help DHs to assess the degree of compliance of their dataset to these principles (Figure 4). The tools below, except for the last one, were all already described in D5.4 deliverable. Demo videos are now made available for most of them.

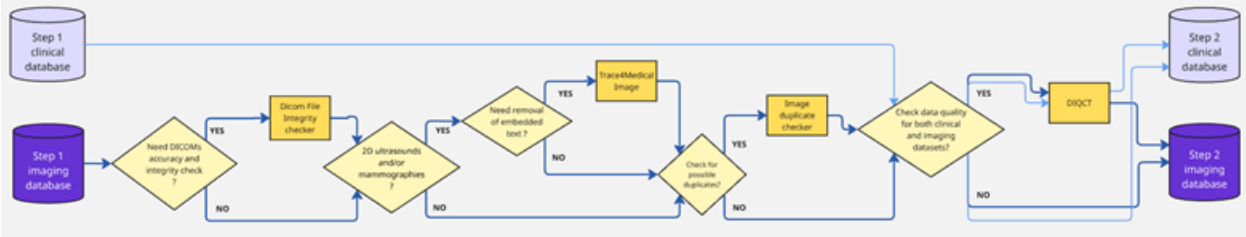

Figure 4: Data quality assessment workflow using EUCAIM tools. Note that the use of EUCAIM tools is not mandatory to achieve all the steps described. The DIQCT is only applicable to Tier 2 and 3 datasets, while all other tools mentioned can apply to any tier's imaging dataset.

2.3.3.1. DICOM file integrity checker (HULAFE)

The DICOM File Integrity Checker tool has already been described in its final version in D5.4, and a demo video was provided then. In short, the tool performs a DICOM quality check in terms of 1/ correct number of files per sequence, 2/ corrupted files, 3/ precise directory hierarchy, 4/ separated dynamic series merging them, 5/ interest series filtering/selection by specific series description lists, 6/ diffusion sequence identification by b-values. It applies the desired changes to the dataset and generates a report containing information about the selected sequences, corrupted files, missing files and merged file. In addition, the tool may be used to detect information in DICOM metadata to allow federated searches, allowing Tier 1 datasets' conversion to Tier 2 level for imaging data.
The demo video of the tool is available here: https://www.youtube.com/watch?v=oUebkjLYeSs&list=PL3Q1XjQpjfg_GEmwPDrQeESh6nqCMnYyR&index=3.

2.3.3.2. Data Integration Quality Check Tool (DIQCT) (AUTH)

The tool performs a dataset quality assessment based on a series of requirements that are set specifically for each use case. Specifically, the tool checks: the clinical metadata

quality (validity, completeness); the integrity between images and clinical metadata provided, as well as their accuracy; the de-identification protocol applied; the existence of annotation together with the consistency between the images and the annotation files. The tool informs the user of corrective actions that need to be taken prior to the data provision, to ensure that the data provided is of high quality, and reports a series of quality metrics characterizing the dataset. This tool only applies to Tier 2 or Tier 3 datasets. A demo video of the tool is available here: https://drive.google.com/file/d/1MxQstbCasmvQfvY3vwD7KJDukZooG5dz/view?usp=drive_link.

2.3.3.3. Tabular data curator (FORTH)

This tool has been integrated as a modular part of the EUCAIM ETL, although it can be used as a stand-alone tool for tabular data quality check and curation. This fully automated tool can be applied on any kind of tabular data (e.g. clinical) to automatically identify duplicated fields (lexically similar and/or highly correlated features), outliers, data inconsistencies. It can also deal with missing values through the application of data imputers. A demo video of the tool is available here: youtube.com/watch?v=eowjz6c8wf4&list=PL3Q1XjQpjfg_GEmwPDrQeESh6nqCMnYyR&index=7&pp=iAQB.

- Image duplicates checking tools

Two tools have been added to the EUCAIM catalogue for checking the presence of duplicate files in imaging datasets:

- Image duplicates checker (AUTH)

This tool detects duplicate or visually similar DICOM series combining metadata analysis, hash-based comparison, and pixel-level similarity metrics. It ideally serves for: 1/ High-level series comparison and classification, and 2/ Detecting duplicates even after post-processing (e.g., denoising, filtering). For strict, image-level duplicate search based on pixel content, fully agnostic to metadata changes, the tool below should be preferred. A demo video of the tool is available here: https://www.youtube.com/watch?v=14NAjHHxEK0&list=PL3Q1XjQpjfg_GEmwPDrQeESh6nqCMnYyR&index=20.

- Image duplicate check tool (F. Champalimaud)

This new tool, developed by F. Champalimaud partners, detects strictly identical DICOM images based on their pixel content, agnostically to metadata. It relies on a pixel data analysis, generating content-based hashes, and comparing them to identify identical

images. It ideally serves cases where metadata may differ, such as when a subject's data have been included in two different studies but with different de-identification pipelines. For duplicate detection after post-processing steps, where image pixels might no longer be identical, the tool above should be preferred.

2.3.4. Data conversion to EUCAIM CDM

Transformation of the clinical and imaging datasets in accordance with the EUCAIM CDM is recommended for Tier 2 nodes and mandatory for Tier 3 nodes.

The transformation step requires:
a) the mapping between the source metadata (clinical and imaging) and the EUCAIM CDM.
b) the actual transformation of all the clinical and imaging data to a format compliant with the EUCAIM CDM through the use of the EUCAIM ETL (Figure 5). Further details on all preliminary steps is provided in Annex 2.

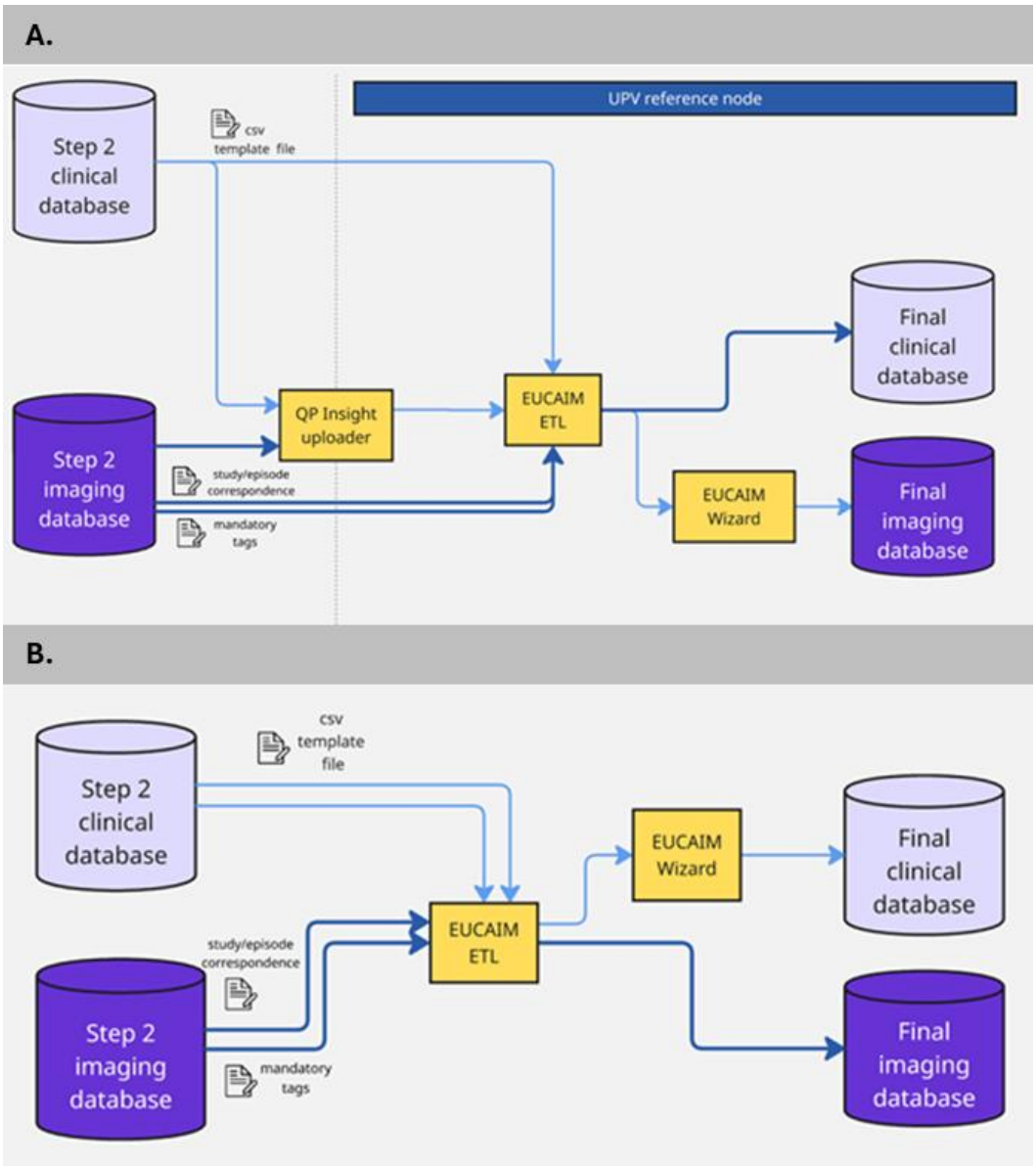

Figure 5: A. Data mapping and transfer workflow using EUCAIM tools. This workflow only applies to datasets aimed at being transferred to the UPV reference node. **B. Data mapping workflow for federated nodes using EUCAIM tools.**

2.3.4.1. EUCAIM ETL (BAHIA)

ETL stands for “Extract, Transform, and Load”. The ETL tool has already been described in its earlier version in D5.5: in short, it has been developed to ingest, transform and map both clinical data and imaging metadata to the EUCAIM common data model (CDM). It is designed to meet the specific requirements of the EUCAIM workflow, accommodating the expected wide variety (in terms of both format and standards) of datasets from DHs, and minimizing manual data preparation efforts. The architecture is designed to be flexible,

allowing for customization and integration with other EUCAIM components. It supports input datasets in CSV, JSON or XSLX format. This tool only applies to Tier 2 or Tier 3 datasets. Many refinements have been made to the EUCAIM ETL, based on interaction with DHs, and adjustments to the general data preparation workflow. The concepts, design, and architecture presented in D5.5 have proven to be sound. However, an additional output relational database was introduced. The ETL now operates with two EUCAIM CDM–compliant databases: one for data ingestion and processing, and a second one containing only fully transformed datapoints, enforcing a stricter and cleaner relational model. The output is generated in two formats: CSV files and the output relational database.

The newest version of the EUCAIM ETL now includes:

- Integration of the Tabular Data Curator tool (see section e on data quality) as a module invoked via its HTTP REST API from the pipeline for data curation.
- Imaging metadata processing. Imaging metadata and clinical data are linked using the imaging timepoints. At the level of the CDM, the instances of ImageStudy are linked with the corresponding instances of ImagingProcedure.
- A log output to report the EUCAIM hyperontology codes that were mapped for the codeable concept values in the original dataset. This log serves as a tool to adjust the ETL mapping and correct technical and conceptual problems in a first deployment and execution of the ETL for a new dataset.
- The ETL can be deployed using two alternative deployment mechanisms: Docker Compose for Local Nodes, and a Helm Chart for the environment deployment on the Reference Node.

A demo video of the tool is available here: https://www.youtube.com/watch?v=I1LSNLzmkFc&list=PL3Q1XjQpjfg_GEmwPDrQeESh6nqCMnYyR&index=30.

2.3.4.2. DICOM tag extraction tool (MEDEX)

The DICOM tags extractor has already been described in its earlier version in D5.5: in short, it scans the DICOM files of a defined directory at the series and study level, and produces as output one single file containing all the predefined DICOM tags, either at the series or at the study level. Several improvements have been made to this tool to adapt to the data preparation workflow and address gaps in-between steps. The tool now runs with different scripts to either 1/ extract from an imaging dataset all preselected DICOM tags (specified in an input file) contained in all detected *series*, or 2/ extract all patientIDs and the corresponding StudyUIDs at the *study* level. Both functions serve to provide input files useful for the EUCAIM ETL and the EUCAIM Wizard tool (see de-identification section 2.3.2). A demo video of the tool is available here:

https://www.youtube.com/watch?v=kAzKq0fQ2eo&list=PL3Q1XjQpjfg_GEmwPDrQeESh6nqCMnYyR&index=33.

2.3.5. Data transfer

Once preprocessing is complete, DHs who have chosen to transfer their datasets to a EUCAIM reference node can do so with the appropriate tool. Transfer to Health-RI reference node goes through XNAT, while transfer to UPV reference node goes through QP Insight uploader. This desktop application developed by Quibim partners enables users to upload DICOM data along with associated clinical information towards a data management platform called "QP insight", deployed at the UPV reference Node (Figure 5). A demo video of the tool is available here: https://www.youtube.com/watch?v=F_nK-V676n8&list=PL3Q1XjQpjfg_GEmwPDrQeESh6nqCMnYyR&index=14.

3. Preprocessing tools on reference nodes

3.1. Access to tools

In addition to the EUCAIM preprocessing tools dedicated to DHs to be used at the DH premises, other preprocessing tools are dedicated to data users, to be used from the reference nodes and applied to the EUCAIM datasets, either to verify/improve some aspects of the dataset's quality, or to enrich the dataset.

The preprocessing tools on reference nodes are also accessible in the bio.tools catalogue (https://bio.tools/t?domain=eucaim), with instructions on the usage of each tool in the description section. Only project members with valid account and permissions can get access to the tools from the registry. More information on how to access and use the tools from the reference nodes is available here: https://eucaim.gitbook.io/enduserguide/4-userguideforresearchers#id-4.8.-reference-node-at-upv. There are currently **32 tools** in the EUCAIM catalogue for data users.

3.2. Description of the preprocessing tools on reference nodes

3.2.1. Data annotation and segmentation

3.2.1.1. Breast dense tissue segmentation (ITI)

The tool takes a digital mammogram and performs an automatic segmentation of the breast area and the dense tissue. After the mammogram segmentation, the tool returns a DICOM-SEG image with both the dense tissue and the breast tissue mask combined.
A demo video of the tool is available here: https://www.youtube.com/watch?v=_X68b2wjrj0&list=PL3Q1XjQpjfg_GEmwPDrQeESh6nqCMnYyR&index=27.

3.2.1.2. HCC segmentation tool for contrast-enhanced abdominal CT (F.Champalimaud)

The tool features an nnU-Net based segmentation tool capable of performing hepatocellular carcinoma (HCC) segmentation for both arterial and portal venous phase in polyphasic contrast-enhanced abdominal CT. The tool receives DICOM CT images as input (NIFTI format is also supported) and can issue various output files (segmentations, maps, logs) in a variety of possible formats.

3.2.1.3. Prostate zone segmentation tool (F.Champalimaud)

This tool automatically segments the prostate into two zones: the central and transition zones (TZ+CZ) and the peripheral zone (PZ). It takes as input a T2-weighted image in DICOM (NIFTI format is also supported) and produces a segmentation in the same input format.
A demo video of the tool is available here: https://www.youtube.com/watch?v=WdjWF4w_ZD8&list=PL3Q1XjQpjfg_GEmwPDrQeESh6nqCMnYyR&index=19.

3.2.1.4. MR-based neuroblastoma tumour detection and segmentation (HULAFE)

HULAFE's team has provided this tool specifically designed and externally validated for automated detection and segmentation of neuroblastic tumours in T2-weighted magnetic resonance images (T2-MR) using deep learning. It processes DICOM or NIfTI input data and outputs in NIFTI or DICOM SEG.
A demo video of the tool is available here:

https://www.youtube.com/watch?v=c7XOEGRA9aQ&list=PL3Q1XjQpjfg_GEmwPDrQeESh6nqCMnYyR&index=12

3.2.1.5. CT-based neuroblastoma tumour detection and segmentation (HULAFE)

HULAFE's team has provided this tool specifically designed and validated for automated detection and segmentation of neuroblastic tumours in contrast-enhanced CT images (CE-CT) using deep learning. It processes DICOM or NIfTI input data and outputs in NIFTI or DICOM SEG. A demo video of the tool is available here: https://drive.google.com/file/d/1-mXbFLpAcXAudSeHrlq03dAu1zWPVt15/view?usp=drive_link.

3.2.1.6. MR-based DIPG tumour detection and segmentation (HULAFE)

HULAFE's team has provided this tool specifically designed and validated to perform an automatic segmentation of the possible DIPG (Diffuse Intrinsic Pontine Glioma, also called DMG - Diffuse Midline Glioma) tumours on MR images. T1W and T2-weighted or FLAIR magnetic resonance images are needed as input. The tool includes a complete workflow from DICOM images to DICOM seg tumoral masks. Also NIFTI is supported. A demo video of the tool is available here: https://drive.google.com/file/d/1zFiIOO-jjZK6e6lZX5kljV1OlBUdjn6v/view?usp=drive_link.

3.2.1.7. MR-based glioblastoma tumour detection and segmentation (HULAFE)

HULAFE's team has provided this tool specifically designed and externally validated to perform an automatic segmentation on MR images (T1W, T2W and FLAIR are needed) of the possible glioblastoma tumours and its subregions: necrosis (Intratumoral necrotic core), edema (Peritumoral vasogenic edema), enhancing (Contrast-enhancing tumor region), total (Total tumor including edema and necrosis by a single model) and total-fused (Total tumor fusioning of necrosis+edema+enhancing). It takes DICOM images as input and generates tumoral masks in DICOM SEG or NIFTI formats.
A demo video of the tool is available here:
https://www.youtube.com/watch?v=8_OJPTQUKAw&list=PL3Q1XjQpjfg_GEmwPDrQeESh6nqCMnYyR&index=21

3.2.1.8. Brain segmentation tools (UPM)

Partners from Universidad Politécnica de Madrid (UMP) and Children's National Hospital developed a set of software tools for the segmentation and analysis of specific cerebral

tumours in MRI scans. Built in Python, the tools enable precise quantitative analysis to support clinical decision-making in both diagnosis and prognosis.

Name of the tool	Cerebral cancer type	Demo	Modality supported
Pediatric Brain Tumor Segmenter	Pediatric Brain Tumor	video	MR
Brain Metastasis Segmenter	Brain Metastasis	video	
Intercranial Meningioma Segmenter	Intercranial Meningioma	video	
Brain Glioma Segmenter	Brain Glioma	video	
Sub-Saharan Africa Brain Glioma Segmenter	Sub-Saharan Africa Brain Glioma	video	

3.2.1.9. nnU-net segmentation tools (DKFZ)

Partners from DKFZ (Division of Medical Image Computing) and their collaborators from the Applied Computer Vision Lab (ACVL) of Helmholtz Imaging developed the “nnU-Net” self-configuring method for deep learning-based biomedical image segmentation. It is designed to automatically adapt to a given dataset, analyzing the provided training cases to configure a matching U-Net-based segmentation pipeline without requiring expertise from the user. The tool receives DICOM images as input, and provides pretrained models for various organs and tumours:

Name of the nnU-net tool	Organ	Cancer type	Modality supported
Pancreas and pancreas tumor segmentation	Pancreas	Pancreas tumor	MR
Colon cancer primaries segmentation		Colon cancer primaries	MR
Abdominal Organ Segmentation	Abdominal Organ		MR

Liver and liver tumor segmentation	Liver	Liver tumor	MR
Kidney and kidney tumor segmentation	Kidney	Kidney tumor	MR
Brain Tumor segmentation in PET		Brain tumor	PET
Hippocampus segmentation	Hippocampus		MR

A demo video of the tool is available here: https://www.youtube.com/watch?v=7-dBoUN45Bs&list=PL3Q1XjQpjfg_GEmwPDrQeESh6nqCMnYyR&index=31.

3.2.2. Data quality assessment

All the tools listed below were already presented in D5.4 deliverable. Links to demo videos have now been added for most of them.

3.2.2.1. Time coherence tool (HULAFE)

This tool assists visual validation of the chronological order and logical consistency of dates associated with a patient's medical history. It generates a timeline visualization for each patient from an Excel file and highlights rule violations. A demo video of the tool is available here: https://www.youtube.com/watch?v=j4J_UPQqYR4&list=PL3Q1XjQpjfg_GEmwPDrQeESh6nqCMnYyR&index=10.

3.2.2.2. Image Quality assessment metrics for XNAT platform (CNR)

The tool offers a suite of four Image Quality Assessment (IQA) metrics designed to evaluate the perceptual quality of preclinical and clinical image datasets within the XNAT platform at the scan and subject levels. The tool outputs Excel files containing the mean and median for each metric and scan, as well as PNG files with corresponding histogram plots.

3.2.2.3. NLmCED denoising Filter (CNR)

NLmCED stands for “Non Local mean Coherence Enhancing Diffusion”. The NLmCED denoising filter is a tool aimed at reducing Rician Noise in 3D MR images. It is a combination between two filters - the Non-Local mean filter and the Anisotropic Diffusion tensor method- with an estimator of Rician noise. The NLmCED filter can benefit from the noise reduction capabilities of the NLM filter while maintaining the edge-preserving characteristics of anisotropic diffusion.

3.2.2.4. Deep Learning Noise Reduction (DLNR) (FORTH)

The Deep Learning Noise Reduction (or DLNR) tool is meant to reduce the noise on already noisy T2w MR prostate images. Quality control is required prior to applying DLNR

as this module will compromise the image quality if high-quality examinations are used as input. A demo video of the tool is available here: https://www.youtube.com/watch?v=DsX42vSAKv0&list=PL3Q1XjQpjfg_GEmwPDrQeESh6nqCMnYyR&index=5.

3.2.2.5. N4 Bias Filter (FORTH)

This tool was already described in D5.4 and included in the demonstrator of said deliverable. In short, the tool aims to reduce the bias field effect, improving the quality of the image. Two main functionalities are offered: (1) Apply N4 filter to images, with either default parameters values or user-defined parameters values, (2) Find the optimal configuration of the N4 filter for specific images by measuring the Full Width at Half Maximum (FWHM) of the periprostatic fat distribution. The demo video of the tool is available here: https://www.youtube.com/watch?v=e9pO1lfsg1A&list=PL3Q1XjQpjfg_GEmwPDrQeESh6nqCMnYyR&index=14.

3.2.2.6. RACLAHE (FORTH)

This tool was already described in D5.4 deliverable. In short, RACLAHE stands for “Region-Adaptive Contrast Limited Adaptive Histogram Equalization”. The tool is an image enhancement method specifically designed for improving CNN-based segmentation of the prostate and prostatic zones in T2-Weighted MR images. A demo video of the tool is available here: https://www.youtube.com/watch?v=EQDzQ-99uV8&list=PL3Q1XjQpjfg_GEmwPDrQeESh6nqCMnYyR&index=6.

3.2.2.7. IQA DCE tool (FORTH)

This tool was already described in D5.4 deliverable under its former name “MR image quality tool”.The tool provides a complete framework for automatic detection of image quality for Breast DCE MR images. The assessment is performed on the first post-contrast dynamic phase in order to assess the most clinically relevant sequence among a number of identical acquisitions. Two categories are available for image classification, i.e. high and low quality, depending on the level of noise, degree of blurring and presence of artifacts. A demo video of the tool is available here: https://www.youtube.com/watch?v=z2sSu1eIXTg&list=PL3Q1XjQpjfg_GEmwPDrQeESh6nqCMnYyR&index=6.

3.2.2.8. ML model for MR series categorisation (HULAFE)

This AI-based tool performs a categorisation of MRI series by using standardized DICOM tags. The categorisation includes the type of sequence (e.g. spin echo, gradient echo), the weighting (e.g. T1W, T2W, DCE, ...), the presence of fat suppression and the

detection of non-relevant / junk series (e.g. localizers, calibrations, screenshots, etc.). The demo video of the tool is available here: https://www.youtube.com/watch?v=gudDCiuJIf8&list=PL3Q1XjQpjfg_GEmwPDrQeESh6nqCMnYyR&index=11.

3.2.2.9. Denoising-Inhomogeneity Correction Tool (HULAFE)

The tool is designed to reduce noise and inhomogeneity field effect in a customized fashion, thus improving image quality and reproducibility of radiomics features. It consists of two independent steps: 1/ denoising, using one of 5 integrated filters (Bilateral Filter, Anisotropic Diffusion Filter (ADF), Curvature Flow Filter (CFF), SUSAN and Non Local Means (NLM)), 2/ ANT's N4 bias correction filter. The parameter configuration of this tool has been optimised for TW1, T2W, DWI and DCE sequences in neuroblastoma (NB) and paediatric brain tumours, but it can also be configured with some of their parameters using a JSON parameter configuration file. The demo video of the tool is available here: https://www.youtube.com/watch?v=HkHqFGXGEbo&list=PL3Q1XjQpjfg_GEmwPDrQeESh6nqCMnYyR&index=9.

3.2.3. Data harmonization

This task is focusing on providing advanced statistical and machine learning techniques to harmonise datasets that potentially originated from different populations, acquisition protocols and/or devices. This ensures their comparability and enables more accurate downstream analysis. To achieve this, recent contributions and research conducted by the partners about data harmonisation were proposed, filtered, evaluated, and integrated into the EUCAIM project. The first four tools listed below were already presented in a previous deliverable (D5.4). Five additional tools have been integrated since.

3.2.3.1. Biologically motivated normalization techniques (FORTH)

This tool was already described in D5.4 and included in the demonstrator of said deliverable. In short, it is designed to perform normalization at the image-level, based on the intensity values of specific tissues, to reduce the variability in the intensity values of MR prostate images due to different scanners, acquisition protocols and conditions. The tool implements three biologically-motivated intensity normalization techniques: (1) The fat-based normalization method, (2) The muscle-based normalization method, and (3) The single tissue (fat or muscle) piece-wise normalization method.
The demo video of the tool is available here: https://www.youtube.com/watch?v=NtwamIRRJ6o&list=PL3Q1XjQpjfg_GEmwPDrQeESh6nqCMnYyR&index=17.

3.2.3.2. MRI image intensity harmonization (Quibim)

This tool was already described in D5.4 and included in the demonstrator of said deliverable. In short, this tool harmonizes the intensity dynamic ranges of MR T2W prostate images through a deep learning model that transforms the input image into an harmonized one closer to a curated reference set of high-quality contrast images.
The demo video of the tool is available here: https://www.youtube.com/watch?v=BQ8aYk2p99c&list=PL3Q1XjQpjfg_GEmwPDrQeESh6nqCMnYyR&index=16.

3.2.3.3. Trace4Harmonization (Deep Trace)

This tool was already described in D5.4 and included in the demonstrator of said deliverable. In short, it harmonizes the numerical values extracted from medical images. The demo video of the tool is available here: https://www.youtube.com/watch?v=9XPDRm79WPg&list=PL3Q1XjQpjfg_GEmwPDrQeESh6nqCMnYyR&index=28.

3.2.3.4. Feature based harmonization (FORTH)

This tool was already described in D5.4 and included in the demonstrator of said deliverable. In short, the tool aims to reduce the variability in the radiomics features due to different scanners, acquisition protocols and conditions by using empirical Bayesian methods to estimate differences in radiomics values and then expressing them in a common space (location/scale adjustment). The tool offers two methods: (1) ComBat method, which shifts the radiomics features to the overall mean and pooled variance of all centers, and (2) M-ComBat method, which shifts the radiomics features to the mean and variance of the chosen reference center with the most samples. The demo video of the tool is available here: https://www.youtube.com/watch?v=ghi3fW3p62w&list=PL3Q1XjQpjfg_GEmwPDrQeESh6nqCMnYyR&index=28.

3.2.3.5. Feature based Harmonization (NestedComBat) (UPM)

This new tool is also a software that performs feature-level harmonization of radiomic data extracted from medical images, based on the ComBat methodology, which uses an empirical Bayes linear model to correct batch effects by estimating location and scale parameters. This tool offers two harmonization strategies: 1/ the ComBat method, which aligns radiomic features to a common scale based on either all centers (ComBat), or 2/ a user-defined batch grouping (M-ComBat), and the Nested ComBat method, which enables simultaneous correction for multiple batch effects in hierarchical settings. This tool is intended to improve the robustness and comparability of radiomic features in multi-center studies.

3.2.3.6. CT Slice thickness harmonization (Quibim)

This new tool standardizes slice thickness in CT scans by resampling them to a common slice spacing. It is intended as a preprocessing step to harmonize CT data, improving consistency across scans acquired with varying slice thicknesses. A demo video of the tool is available here: youtube.com/watch?v=GL-W59VLZ8M&list=PL3Q1XjQpjfg_GEmwPDrQeESh6nqCMnYyR&index=16&pp=iAQB0gcJCTwKAYcqIYzv.

3.2.3.7. Cluster based harmonization (HULAFE)

This new tool is designed to perform radiomics harmonization on large and heterogeneous datasets, where the risk of over-harmonization is present. Instead of directly applying harmonization based on predefined batch labels, the tool first identifies groups of batches that share similar characteristics through clustering of the radiomics data. It then performs harmonization using these cluster-derived labels. The tool allows the harmonization of radiomics variables using two methods: (1) original ComBat method, where each original batch group is considered for the harmonization process, and (2) cluster-based ComBat method, where batch groups with similar radiomics characteristics form clusters and the latter are being considered for the harmonization process. A demo video of the tool is available here: https://www.youtube.com/watch?v=0GJqNJv-Qf8&list=PL3Q1XjQpjfg_GEmwPDrQeESh6nqCMnYyR&index=31.

3.2.3.8. 2D Digital Mammography Harmonization (HULAFE)

This new tool is designed to standardize and harmonize 2D digital mammograms in DICOM through a configurable pipeline that includes: spatial reorientation, pseudo-3D stacking, isotropic resampling, intensity normalization, optional denoising, contrast enhancement, and mask processing (if available). A demo video of the tool is available here:
https://www.youtube.com/watch?v=IrZhMp2lB7g&list=PL3Q1XjQpjfg_GEmwPDrQeESh6nqCMnYyR&index=13.

3.2.3.9. Extended A Priori Probability (EAPP) (ITI)

The tool provides a semi-supervised metric (EAPP) for binary classification tasks, that considers not only the a priori probability but also some possible bias present in the dataset, as well as other features that could provide a relatively trivial separability of the target classes. Therefore, it allows for evaluating the ease or complexity of the task or bias of the data beyond the well-established baseline for any binary classification. A demo video of the tool is available here: https://www.youtube.com/watch?v=6XbariSG2Qw&list=PL3Q1XjQpjfg_GEmwPDrQeESh6nqCMnYyR&index=35.

3.2.4. Dataset FAIRness

3.2.4.1. FAIR-EVA tool (CSIC-IFCA)

EVA stands for “Evaluator, Validator & Advisor”. The FAIR-EVA tool has been developed to check the FAIRness level of digital objects from different repositories or data portals. It requires the object identifier (preferably persistent and unique identifier) and the repository to check. It also provides a generic and agnostic way to check digital objects. This software started to be developed within IFCA-Advanced-Computing receives funding from the European Union’s Horizon 2020 research and innovation programme under grant agreement No 857647.

A deliverable is dedicated to FAIRness assessment and details the functionalities of the tool (D2.5 “The FAIR Report”). A demo video of the tool is available here: https://www.youtube.com/watch?v=7hSYpC60Q9g&list=PL3Q1XjQpjfg_GEmwPDrQeESh6nqCMnYyR&index=4.

4. Future work

Future work will be dedicated to 1) the continuous monitoring of tools onboarding, 2) the testing of the tools and collection of feedback, and 3) the follow-up on documentation. Because important factoring work was performed in the past weeks and months on the preprocessing tools dedicated to data preparation on premise, some documentation may need to be reviewed and updated, and the end-to-end workflow of data preparation in a local node with all preprocessing tools has yet to be tested. Particular focus will be dedicated to the workflow around the EUCAIM ETL, that has so far been applied to clinical data of pilot sites, but not yet on the imaging data. A survey has been created to capture feedback on the use of the tools by DHs from pilot sites (Annex 3). The survey will be shared with DHs and when applicable, modifications will be applied to the documentation currently available (bio.tools, Handbook, technical documentation, demos). Regarding the preprocessing tools addressed to data users, future work will consist in testing them in the environment of the reference nodes, as well as testing their interaction with the FEM client used for federated processing.

Annex 1: EUCAIM tools' status

The table A1 below describes the status of readiness for each EUCAIM data preprocessing tool.

Tool	Tool owner	1 - Software onboarding	2 - EUCAIM SW evaluation: Technical check	3 - Dockerization and registration in EUCAIM SW Registry	
				Uploaded to registry (harbor)	Demo
Data quality					
Time coherence tool	HULAFE	YES	YES	YES	YES
DICOM File Integrity Checker	HULAFE	YES	YES	YES	YES
Image Quality assessement metrics for XNAT platform	CNR	YES	YES	YES	PENDING
Extended A Priori Probability (EAPP)	ITI	YES	YES	YES	YES
Data Integration Quality Check Tool (DIQCT)	AUTH	YES	YES	YES	PENDING
NLmCED denoising Filter	CNR	YES	PENDING	YES	PENDING
Deep Learning Noise Reduction (DLNR)	FORTH	YES	PENDING	YES	YES
N4 Bias Filter	FORTH	YES	YES	YES	YES
RACLAHE filter	FORTH	YES	YES	YES	YES
IQA DCE tool	FORTH	YES	PENDING	YES	YES
Tabular data curator	FORTH	YES	YES	YES	YES

ML model for MR series categorisation	HULAFE	YES	PENDING	YES	YES
Trace4MedicalImageCleaning	Deep Trace	ON-GOING	PENDING	YES	PENDING
Denoising-Inhomogeneity Correction Tool	HULAFE	YES	YES	YES	YES
Image similarity/duplicate checker	AUTH	YES	PENDING	YES	YES
Image Duplicate check	Champalimaud	ON-GOING	PENDING	PENDING	PENDING
Data annotation					
Breast dense tissue segmentation	ITI	YES	YES	YES	YES
MR-based neuroblastoma tumour detection and segmentation	HULAFE	YES	YES	YES	YES
MR-based DIPG tumour detection and segmentation	HULAFE	YES	YES	YES	PENDING
MR-based glioblastoma tumour detection and segmentation	HULAFE	YES	YES	YES	YES
CT-based neuroblastoma tumour detection and segmentation	HULAFE	YES	YES	PENDING	PENDING
Pediatric Brain Tumor Segmenter	UPM	YES	YES	YES	YES
Brain Metastasis Segmenter	UPM	YES	YES	YES	YES

Intercranial Meningioma Segmenter	UPM	YES	YES	YES	YES
Brain Glioma Segmenter	UPM	YES	YES	YES	YES
Sub-Saharan Africa Brain Glioma Segmenter	UPM	YES	YES	YES	YES
Pancreas and pancreas tumor segmentation (nnU-net)	DKFZ	YES	YES	YES	YES
Colon cancer primaries segmentation (nnU-net)	DKFZ	YES	YES	YES	YES
Abdominal Organ Segmentation (nnU-net)	DKFZ	YES	YES	YES	YES
Liver and liver tumor segmentation (nnU-net)	DKFZ	YES	YES	YES	YES
Kidney and kidney tumor segmentation (nnU-net)	DKFZ	YES	YES	PENDING	YES
Brain Tumor segmentation in PET (nnU-net)	DKFZ	YES	YES	YES	YES
Hippocampus segmentation (nnU-Net)	DKFZ	YES	YES	YES	YES
MITK Workbench Viewer	DKFZ	YES	YES	N/A	PENDING
Annotation Seg Converter (NIFTi/RT-STRUCT to DICOM-seg converter)	SYNLAB	YES	PENDING	YES	YES

HCC segmentation tool for contrast-enhanced abdominal CT	F Champalimaud	ON-GOING	PENDING	PENDING	PENDING
Prostate zone segmentation tool	F Champalimaud	YES	YES	YES	YES
Data harmonization					
Biologically motivated normalization techniques	FORTH	YES	YES	YES	YES
MRI image intensity harmonization	Quibim	YES	YES	YES	YES
Trace4Harmonization	Deep Trace	ON-GOING	PENDING	PENDING	YES
Feature based harmonization (Combat)	FORTH	YES	YES	YES	YES
CT Slice thickness harmonization	Quibim	YES	YES	YES	YES
Cluster based harmonization	HULAFE	YES	YES	YES	YES
Feature based Harmonization (NestedComBat)	UPM	ON-GOING	PENDING	PENDING	PENDING
2D Digital Mammography Harmonization	HULAFE	YES	YES	YES	YES
Data FAIRification					
FAIR EVA	CSIC IFCA	YES	YES	YES	YES
Data de-identification					

EUCAIM DICOM Anonymizer	FORTH	YES	YES	YES	YES
Lethe DICOM Anonymizer	FORTH	YES	YES	YES	PENDING
EUCAIM Wizard tool	FORTH	YES	YES	YES	YES
Mainzelliste	DKFZ	YES	YES	YES	PENDING
Data management/ingestion					
EUCAIM ETL Toolset	BAHIA	ON-GOING	PENDING	PENDING	YES
QP-Insights Uploader	Quibim	YES	YES	YES	YES
DICOM tags Extractor	MEDEX	YES	PENDING	YES	YES

Table A1: Readiness status of each EUCAIM data preprocessing tools

Annex 2: Stepwise guidelines for data holders to prepare their datasets

This section is adapted from the EUCAIM Handbook.

Tier 1 datasets

The preparation of your dataset will follow three steps – de-identification, data quality check, and data transfer (optional) – as described below:

Step 1: De-identification

You must ensure that no identifiable information (direct or indirect) is present in the dataset you will share (Figure A2).

⚠Important points to consider before de-identification!

Annotation (optional): you may want to annotate your imaging data to enrich your dataset. We recommend using the EUCAIM annotation tool **MITK (Medical Imaging Interaction Toolkit) Workbench** to avoid the burden (and the risk) of additional conversion procedures (**Figure A1**).

Format standardization (optional): it is recommended that your imaging raw data are in DICOM format, and if applicable, that your annotations are in DICOM-SEG. If you need to convert annotation files to DICOM-SEG, you may use the EUCAIM **Annotation Seg converter** tool.

If your Tier 1 dataset is not originally anonymized and you plan on transferring your dataset to a reference node, we recommend preparing a tabular file associating StudyUIDs from DICOM images with corresponding clinical "episode" and "timepoint". This can be done using the **DICOM tags extractor** tool. For more information, see further below section Tiers 2 and 3's Step 2 on imaging data preparation.

If your imaging data are not already de-identified, you may use the **Lethe EUCAIM Anonymizer**. In this case, you must ensure the following:

- the patient ID linking clinical and imaging data must be identical and listed as the first variable in the clinical dataset for tabular data;
- your raw imaging data are in DICOM format;
- the tool requires as input the SITE_ID, the unique identifier of the data provider, which you can see in your user profile from the EUCAIM Dashboard (Figure A3). In case your Life Science account is not assigned to a known organization, then this will be empty and so you can create a ticket in the Helpdesk to request one;

Special attention should be given to **embedded text** in images, which may contain patient-identifiable information, as well as **skull and head images** that pose a risk of patient re-

identification. You may need to apply additional de-identification techniques to mitigate this risk. Tools such as the **DICOM defacing anonymisation** tool from the EUCAIM catalogue may be used to remove facial features from your DICOM images. For 2D ultrasounds and mammographies **dataset**, you may use the **Trace4MedicalImage cleaning** tool, that detects and removes encapsulated text in DICOM files.

Re-identification risk assessment (optional): Even if no automatic re-identification risk analysis on a combination of clinical and imaging metadata is possible at this Tier, you should carefully assess that no direct or indirect identifiers are present in your data. For assessing the risk of re-identification of patients based on your **imaging metadata** before sharing your dataset, you may use the EUCAIM **Wizard tool**. Extraction of imaging metadata to feed the wizard tool is possible by using the **DICOM tags extractor** tool.

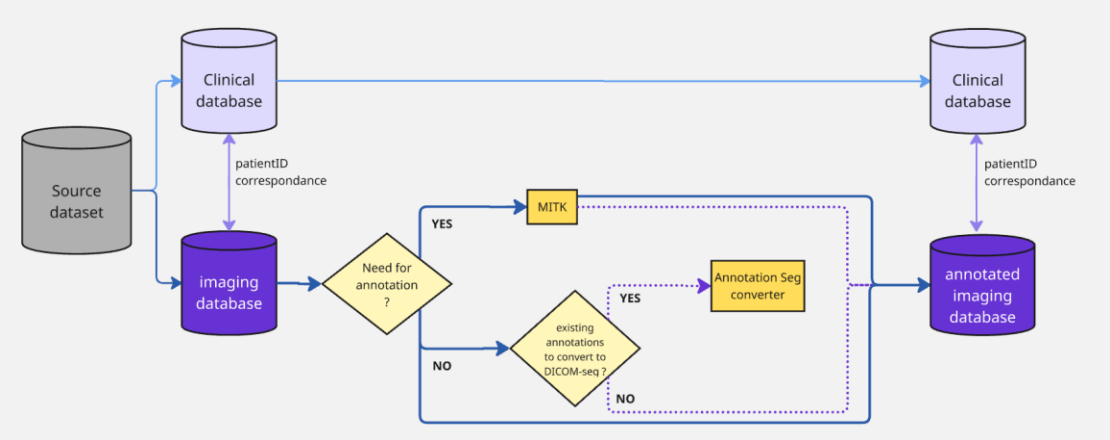

Figure A1: Annotation workflow using EUCAIM tools.

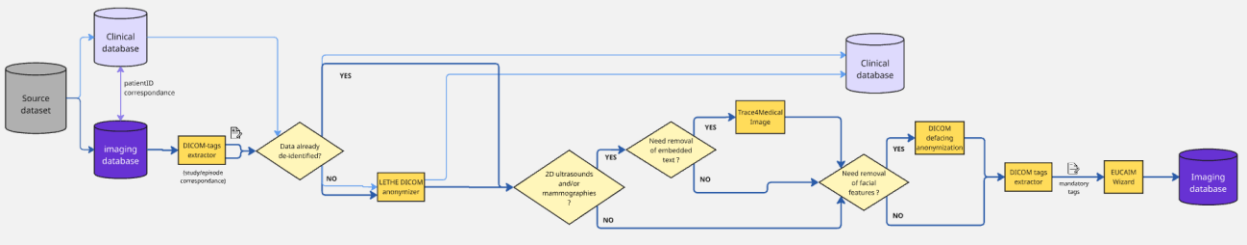

Figure A2: De-identification workflow using EUCAIM tools. Note that the use of EUCAIM tools is not mandatory to achieve all the steps described.

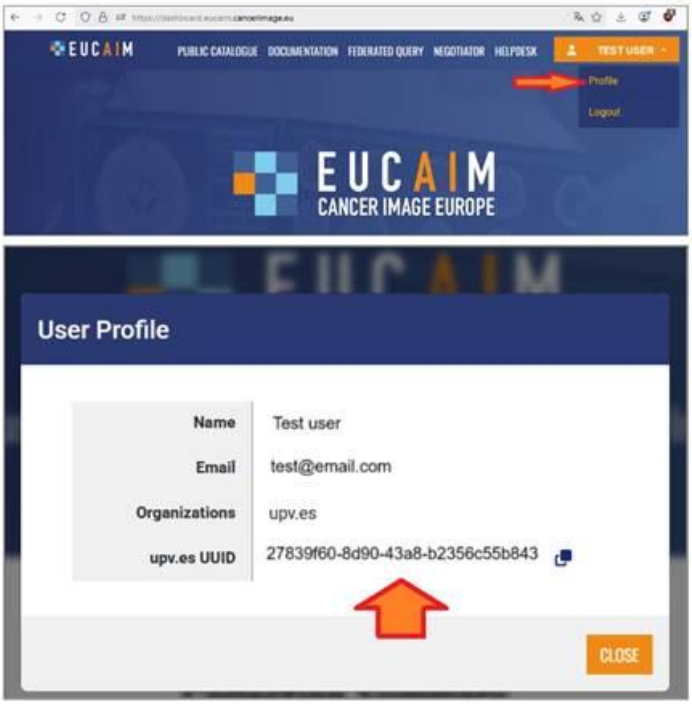

Figure A3: Retrieving SITE ID from the Dashboard.

Step 2: Data quality check:

As per the EUCAIM data quality framework, you must ensure that your dataset is:

- **Complete**: all required data values are present
- **Unique**: no entity exists more than once within the dataset
- **Consistent**: dataset values of two sets of attributes within a record / within a data file / between data files / within a record at different points in time, comply with a rule
- **Accurate**: correspondence between dataset values to real values
- **Showing integrity**: absence of data value loss or corruption

You may use dedicated tools to assess the degree of compliance of your dataset to these principles. Some tools from the EUCAIM catalogue can help you to do so (see **Figure 11**):

- the **accuracy** and **integrity** of your imaging dataset may be assessed using the **DICOM File integrity checker**.
-
- **Uniqueness** can be addressed with two EUCAIM tools that search for image duplicates: the **Image duplicates checker**, capable of detecting duplicate or visually similar DICOM series by that combining metadata analysis, hash-based comparison, and pixel-level similarity metrics; the **Image duplicate check tool**, that detects duplicate DICOM images by analyzing pixel data.

Step 3: Data transfer (optional):

Tier 1 datasets can either be transferred to a reference node (see option 1 in section 6 of this handbook), or remain at your site. Data users interested in your dataset (as per the information found in the EUCAIM catalogue) will be put in direct contact with you.

Tiers 2 and 3 datasets

Data Holders must ensure that their data satisfy the following minimum requirements:

- Image and clinical data must be linked using the same patient ID.
- No entity can exist more than once within your dataset.
- Imaging data should be in DICOM format and associated annotations and segmentations, when available, should be in DICOM-SEG format. Specific cases will be considered if a DH has diagnostic images in different data formats.
- All images should include a minimum set of imaging metadata and comply to the DICOM standard
- Imaging data must be accompanied by a set of minimum clinical metadata.

The minimum metadata requirements for the imaging and clinical data are presented in this document.

Tier 2-3 datasets must ensure their compliance to the EUCAIM's Common Data Model according to the compliance level (the searchable variables for Tier 2 and full Common Data Model in tier 3).

Compliance of your datasets with the EUCAIM **Common Data Model (CDM)** is **recommended for Tier 2** nodes and **mandatory for Tier 3** nodes.

For Tier 2 data holders: if you do not wish to transform your data to the EUCAIM common data model, you can opt to implement a mapping component to interact with the federated search instead. Otherwise, EUCAIM offers dedicated tools that can help you with transforming your data to the EUCAIM CDM as described in Section 5.2.1.

Steps to prepare your Tier 2 or Tier 3 dataset

The preparation of your dataset will follow the steps as described above:

Step 1: Clinical data structuring

In order to have interoperable data that can be queried and processed (Tier 3), we need you to provide us with information on your dataset structure using another tabular template file (EUCAIM_example_file_patients_datasets_CDM_v6) *in addition to* your source dataset.

- How the tabular template file is organized:
    - The "Data elements" tab lists the entities and their corresponding data elements for clinical variables, with definition and data type;
    - The other 3 tabs show an example of how to structure your datasets of positive or diagnostic cases (for negative screening and control groups, please refer to the corresponding template file);
        - the "Overarching Episode" corresponds to the entire course of the patients data collection (example: from diagnosis to death or last contact). All diagnosis information should be in there;
        - each episode recorded in your dataset must be separated from the first tab in another tab in chronological order (example: “Treatment 1”, “Progression”, “Treatment 2”, “Remission”, “Relapse”, “Treatment 3”, “Active Surveillance").

In each tab:

- On line 1 are the names of the variables as they are defined in your own dataset
- On line 2 are the name of the corresponding entity in the EUCAIM CDM, as shown in the "Data elements" tab
- On line 3 are the name of the corresponding data element name in the EUCAIM CDM, as shown in the "Data elements" tab
- On line 4 is the standard used in the dataset
- On line 5 is an example of value

- How to structure your dataset

Your clinical dataset must be structured as a **tabular file**, either xls format, or csv format. As per the ETL requirements, **csv** files must use a full stop “.” as decimal separator, and we also recommend using comma “,” as list separator. If other characters are used (semi-colon, tabs, etc), it should be communicated in advance to the ETL support team.

For datasets with multiple timepoints, we recommend “vertical” datasets, meaning that your dataset has one row per timepoint.

Please give your dataset file a name with the **dataset_ID as first character**.

Example: “Dataset_ID_colon_study_2022.xls”

If you have several datasets, please make sure to store them in separate locations.

- How to complete the template file

Notes before you start: 1/ You may create your own tabular file or use this example file if useful. 2/ The example datasets in this file only contain the mandatory variables; you should provide the full list of variables available in your dataset.

1. We recommend that the name of the template file also contains the dataset_ID as the first character.
2. Please make sure it contains the *exact* variables' names on the first row (matching the variable’s names from your source dataset), and the PatientID as the first variable.
3. Separate all episodes into different tabs as described above, except for Diagnosis that belongs to the Overarching episode.

Note: episodes may correspond to the following: Treatment Episode, Progression, Relapse, Remission, Active Surveillance.

4. For each variable of your dataset, find the corresponding entity and data element name (see data element tab), and add both under the variable name on line 2 and 3, respectively. Important: for several entities, the Code attribute must be accompanied by the Category attribute.

Example 1 with “Imaging acquisition” as Procedure: we need to specify the sequence (CT, MRI) as “Code”, and assign to it “imaging” as Category. See in the Overarching episode tab on this dataset example, columns M-N. Note that the name of the variable is then merged on both columns.

Example 2 with “Smoking Status” as Medical History: we need to specify the status value itself (smoker, non-smoker, etc) as “Code”, and assign to it “Observation” as Category. See in the Overarching episode tab on this dataset example, columns Q-R. Again, the name of the variable must be merged on both columns."

5. For each variable of your dataset, please provide an example value on line 5 (add the value as it is spelled exactly in your dataset)

6. For each variable of your dataset:

    - if the variable follows strictly a specific standard, please provide the name of the standard on line 4

Example: in the Overarching episode tab, column K, the “Histological type” variable strictly follows the SNOMEDCT standard; line 4 specifies “SNOMEDCT”, and an example value is provided on line 5.
Important: both information must be separated by a comma, without space

- if the variable follows specific standard with in-house coding or remaining, please provide the name of the standard on line 4, and provide the correspondence between all possible values from your dataset and the standard values on lines 6 and onwards

Example 1 in the Overarching episode tab: column I, the "Tumor site: Region" variable follows the SNOMEDCT standard using an in-house coding; line 4 specifies "SNOMEDCT", an example value is provided on line 5, and correspondence for all possible values present in the dataset to the SNOMEDCT codes is listed on lines 6-9, separated by a comma.

Example 2 in the Overarching episode tab: column L, the "Histological subtype" variable follows the SNOMEDCT standard using an in-house naming; line 4 specifies "SNOMEDCT", an example value is provided on line 5, and correspondence for all possible values present in the dataset to the SNOMEDCT codes is listed in lines 6-9, separated by a comma.

- if the variable does not follow a specific standard, please state "custom" on line 4, and provide the list of all possible values from your dataset for that variable on lines 6 and onwards

Example in the Overarching episode tab: column J, the "Tumor Site: Laterality" variable does not follow a standard, but only uses the label "Left" or "Right"; in that case line 4 specifies "custom", an example value is provided on line 5, and all possible values present in the dataset (here "Left" and "Right" is listed on lines 6-7.

Step 2: Imaging correspondence with clinical data

First and foremost, you need to make sure that your imaging raw data are in DICOM format, and if applicable, that your annotations are in DICOM-SEG.

In order to successfully link the imaging exams from your dataset with the clinical information you provide, especially the timepoints of each episode, we need to retrieve the correspondence between each imaging study and each clinical episode.

Before de-identification of your dataset*, please create a tabular csv file that contains the following information:

- **PatientID** - the exact one from your DICOM images (attribute (0010,0020))
- **StudyUID** - the exact one from your DICOM images (attribute (0020,000D))

*Note: if your dataset is already anonymized, you can still use the DICOM tags extraction tool to provide the file, proceed with step 2 and skip step 3. It is important that you can still link the (anonymized) PatientID with the episodes and timepoints.

To assist you retrieving all PatientID and StudyUID from your imaging dataset, you may use the **DICOM tags extractor tool** and its "dicom_tags_selection" script. A template csv input file called "imaging_studies_episodes.csv", provided with the tool, allows to retrieve the following attributes from your imaging dataset (cf tool documentation): PatientID, StudyUID, StudyDate, Study description (Table A2).

PatientID (0010,0020)	**StudyUID (0020,000D)**	**StudyDate (0008,0020)**	**StudyDescription (0008,1030)**
ABC-000103	1.2.824.0.2.3886579.08.383.1010.6135	2018-12-11	Whole Body I-131 CT
ABC-000103	1.2.824.0.2.4653289.08.563.1010.4679	2018-12-23	Screening-Bilateral Mammography
ABC-000103	1.2.824.0.2.06135249.08.647.2304.7961	2019-01-13	I131 high dose
ABC-000107	1.2.824.0.2.4862015.07.383.5623.6820	2017-05-17	Bilat Mammography

Table A2: Example output file of the dicom_tags_selection script. The StudyDate, and StudyDescription in Study are provided for indication only, to guide you for the mapping of each study to each episode (see step 2).

You then need to edit the output file by adding the "Episode" and "Timepoint" information for each study (i.e each row) as below:

- **Episode** - The episode information has to match the name of the episode provided in the clinical template file. As per the EUCAIM CDM, possible values are: Diagnosis, Treatment, Progression, Relapse, Remission, Active Surveillance.
- **Timepoint** - As there can be multiple imaging procedures per episode, please number all studies in ascending order (1, 2, 3,…). Note: the numbering only concerns imaging procedures, not any other procedure in between.

PatientID (0010,0020)	StudyUID (0020,000D)	StudyDate (0008,0020)	StudyDescription (0008,1030)	Episode	Imaging Timepoint
ABC-000103	1.2.824.0.2.3886579.08.383.1010.6135	2018-12-11	Whole Body I-131 CT	Diagnosis	1
ABC-000103	1.2.824.0.2.4653289.08.563.1010.4679	2018-12-23	Screening-Bilateral Mammography	Diagnosis	2
ABC-000103	1.2.824.0.2.06135249.08.647.2304.7961	2019-01-13	I131 high dose	Treatment	3
ABC-000107	1.2.824.0.2.4862015.07.383.5623.6820	2017-05-17	Bilat Mammography	Diagnosis	1

Table A3: Example of edited file with correspondence between StudyUID and both Episode and Timepoint. The part in blue corresponds to the part edited manually by the data holder.

Step 3: De-identification

You must ensure that no identifiable information (direct or indirect) is present in the dataset you will share (Figure A2).

Important points to consider before de-identification: **Annotation (optional)**
You may want to annotate your imaging data. We recommend using the EUCAIM annotation tool **MITK (Medical Imaging Interaction Toolkit) Workbench** to avoid the burden (and the risk) of additional conversion procedures. If you have existing annotation files that are not in DICOM-SEG, you may use the EUCAIM **Annotation Seg converter** tool to convert them into DICOM (Figure A1).

If your imaging data are not already de-identified , you may use the **Lethe EUCAIM Anonymizer**. In this case, you must ensure the following:

- the patient ID linking clinical and imaging data must be identical and listed as the first variable in the clinical dataset for tabular data;
- Your raw imaging data are in DICOM format;
- the tool requires as input the SITE_ID (Figure A3), the unique identifier of the data provider, which you can see in your user profile from the EUCAIM Dashboard. In case your

Life Science account is not assigned to a known organization, then this will be empty and so you can create a ticket in the Helpdesk to request one;

Special attention should be given to **embedded text** in images, that may contain patient-identifiable information, as well as **skull and head images** that pose a risk of patient re-identification. You may need to apply additional de-identification techniques to mitigate this risk. Tools such as the **DICOM defacing anonymisation** tool from the EUCAIM catalogue may be used to remove facial features from your DICOM images.

Re-identification risk assessment for imaging data (optional): Even if no automatic re-identification risk analysis on a combination of clinical and imaging metadata is possible at this Tier, you should carefully assess that no direct or indirect identifiers are present in your clinical data. For assessing the risk of re-identification of patients based on your **imaging metadata** before sharing your dataset and further anonymizing your dataset through well-known privacy models, you may use the EUCAIM **Wizard tool**. Extraction of imaging metadata to feed the wizard tool is possible by using the **DICOM tags extractor** tool.

Steps 1 to 3 are illustrated in Figure A2.

Step 4: Data quality assessment

As per the EUCAIM data quality framework, you must ensure that your dataset is:

- **Complete**: all required data values are present
- **Unique**: no entity exists more than once within the dataset
- **Consistent**: dataset values of two sets of attributes within a record / within a data file / between data files / within a record at different points in time, comply with a rule
- **Accurate**: correspondence between dataset values to real values
- **Showing integrity**: absence of data value loss or corruption

You may use dedicated tools to assess the degree of compliance of your dataset to these principles. Some tools from the EUCIAM catalogue can help you to do so (**Figure A4**):

- The **DICOM File integrity checker** can check the **accuracy** and **integrity** of your imaging dataset.
- For 2D ultrasounds and/or mammographies **datasets, validity** assessment is possible using the **Trace4MedicalImage cleaning** tool, that detects and removes encapsulated text in DICOM files.
- **Uniqueness** can be addressed with two EUCAIM tools that search for image duplicates: the **Image duplicates checker**, capable of detecting duplicate or visually similar DICOM series by that combining metadata analysis, hash-based

comparison, and pixel-level similarity metrics; the **Image duplicate check tool**, that detects duplicate DICOM images by analyzing pixel data.

- The **DIQCT** may help you assess various aspects of your dataset's quality, both for imaging and clinical data, such as its **completeness, uniqueness, validity, consistency, integrity.**

.

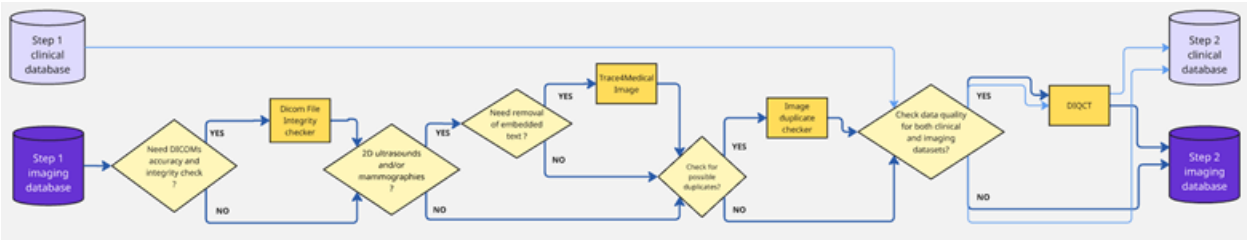

Figure A4: Data quality assessment workflow using EUCAIM tools. Note that the use of EUCAIM tools is not mandatory to achieve all the steps described. The DIQCT is only applicable to Tier 2 and 3 datasets, while all other tools mentioned can apply to any tier's imaging dataset.

Step 5: Data conversion to EUCAIM Common Data Model

Transformation of the clinical and imaging datasets in accordance with the EUCAIM CDM is recommended for Tier 2 nodes and mandatory for Tier 3 nodes. Tier 2 nodes can opt instead to implement a custom mapping component to interact with the federated search service. The transformation step requires:

a) the mapping between the source metadata (clinical and imaging) and the EUCAIM CDM.

b) the actual transformation of all the clinical and imaging data to a format compliant with the EUCAIM CDM through the use of the **EUCAIM ETL**.

For your imaging dataset:

- Fill in a tabular csv file with the correspondence between all the possible values of SeriesDescription to the EUCAIM CDM standard vocabulary entries (Table 5). For all the SeriesDescription that you cannot map, keep the original values. They will serve to enrich the EUCAIM CDM.

- Extract in a tabular csv file all the 75 mandatory attributes (list available in annex 2) present in your dataset. You may already have such file, especially if you used the Wizard tool on step 3 "de-identification" for re-identification risk assessment of imaging data. If not, you may use the **DICOM_tags_extractor** tool now**.**

Finally, share the **two above-mentioned csv files** as well as the **file from step 2 on PatientID/StudyUID correspondence** with the ETL ingestion support team through the EUCAIM helpdesk.

Source series Description	EUCAIM series description
AXIALT2TSE	T2 weighted
axdifb1000	Diffusion weighted
e-THRIVE_BHPERFU	PW
EP2D_DIFF_TRA_B50-1000_TRACEW_DFC_MIX	Diffusion weighted
t2_tse_tra_p2_384ESTRICTO	T2 weighted

Table A4: Example of correspondence between the Series Description from the source images and the Series Description from the EUCAIM standard. The part in blue corresponds to the part edited manually by the data holder. See **Annex 1** for the list of all possible SeriesDescription currently known in the EUCAIM vocabulary.

Step 6: Data transfer (optional)

- If you plan on transferring your dataset to a reference node, next action would be to now proceed with the transfer (QP Insight for the UPV node, XNAT for the Health-RI node). All the next steps will occur directly on the node (Figure 12).

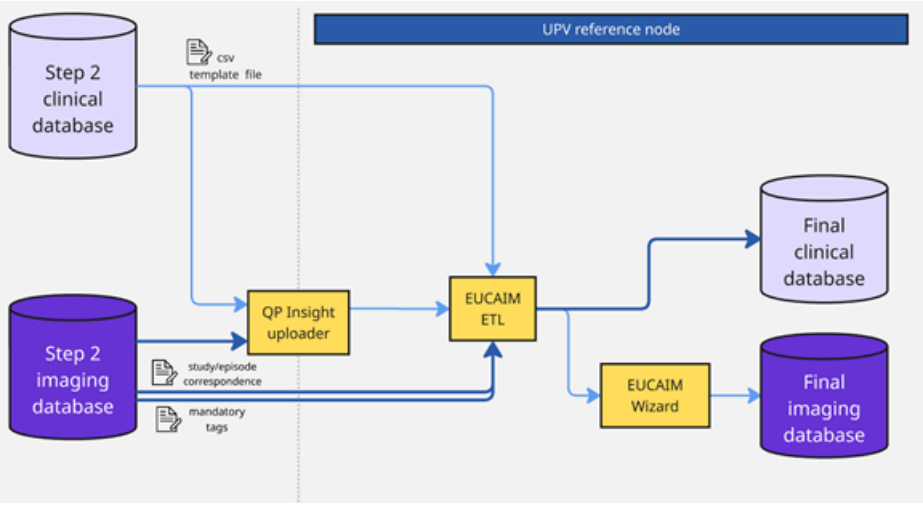

Figure A5: Data mapping and transfer workflow. This workflow only applies to datasets aimed at being transferred to the UPV reference node

- If you aim at storing your dataset in a federated node, make sure it is stored in its final destination, and proceed with the next steps.

The ETL support team will proceed with you with the mapping to EUCAIM CDM at your site.

Re-identification risk assessment (optional): you may want to verify that no direct or indirect identifiers are present in your clinical data. You may apply the Wizard tool to your clinical data file now that it is mapped to the EUCAIM CDM (Figure A6).

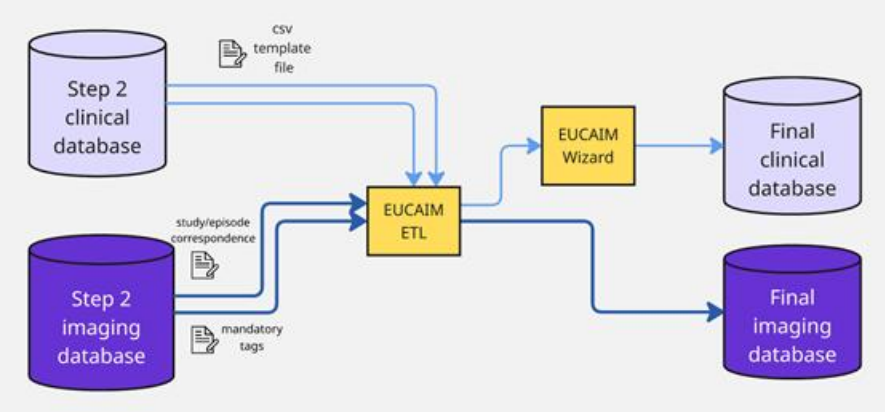

Figure A6: Data mapping workflow for federated nodes

Annex 3: Survey for pilot sites

The survey below was created using Google Form. It is intended for data holders from pilot sites, to provide feedback on their experience using EUCAIM preprocessing tools. The feedback received in this survey will contribute to facilitate the whole preparation workflow and ease DHs experience: the responses will be used to improve the support provided to DHs during their dataset preparation, as well as improve the tools documentation and the description of the preparation workflow.

[EUCAIM] Data preparation tools: survey for pilot sites

* Indicates required question

EUCAIM
CANCER IMAGE EUROPE

Dear Data Holder,
Following your use of our preprocessing tools for dataset preparation, we would like to collect your feedback on your experience with the EUCAIM tools.
Note: If you have tested several tools, we kindly ask you to complete **one survey form per tool**.

Please select the tool that you tested, and for which you are completing this survey : *

- ◯ EUCAIM Anonymizer
- ◯ DCM2SUB-defacing tool
- ◯ Data Integration Quality Check tool (DIQCT)
- ◯ DICOM2seg converter
- ◯ DICOM Files Integrity Checker
- ◯ DICOM tags extractor
- ◯ EUCAIM ETL
- ◯ Image duplicate checker
- ◯ MITK annotation tool
- ◯ QP Insight
- ◯ Trace4MedicalImage Cleaning tool
- ◯ Wizard tool
- ◯ Other

Tool downloading : how was your experience with downloading the tool from the EUCAIM harbor? *

- ◯ Very easy
- ◯ Rather easy
- ◯ Rather difficult
- ◯ Very difficult
- ◯ NA (the tool was not downloaded from the harbor)
- ◯ Other: _____

Any comment on tool downloading?

Your answer

Tool deployment: how was your experience with deploying the tool at your site ? *

- ◯ Very easy
- ◯ Rather easy
- ◯ Rather difficult
- ◯ Very difficult
- ◯ NA (the tool was not deployed)

Any comment on tool deployment?

Your answer

Tool documentation: how did you find the documentation provided ? *

- ◯ Very clear
- ◯ Rather clear
- ◯ Rather unclear
- ◯ Very unclear
- ◯ Documentation not provided
- ◯ I did not look a the documentation

Any comment on tool documentation?

Your answer

Using the tool: how was your experience with running the tool locally ? *

- ◯ Very easy
- ◯ Rather easy
- ◯ Rather difficult
- ◯ Very difficult
- ◯ NA (we did not run the tool)

Any comment on using the tool ?

Your answer

Technical support: how was your experience with the support provided by the EUCAIM technical team on using the tool? *

- ◯ Excellent
- ◯ Good
- ◯ Average
- ◯ Unsatisfying
- ◯ I have not contacted the support for this tool
- ◯ I did not receive any support for this tool

Any comment on technical support?

Your answer

Comments : Please describe any issue you may have encountered with the tool, as well as any additional comments you may have.

Your answer