Data Infrastructures for AI in Medical Imaging
A report on the experiences of five EU projects

(with a focus on the Deidentification approaches)

Haridimos Kondylakis
FORTH-ICS, Greece

Leader of the Data Management WG,
AI for Health Imaging Network (AI4HI), EU

The AI4HI working group

EuCanImage	A European Cancer Image Platform Linked to Biological and Health Data for Next-Generation Artificial Intelligence and Precision Medicine in Oncology.
CHAIMELEON	Accelerating the lab to market transition of AI tools for cancer management.
INCISIVE	A multimodal AI-based toolbox and an interoperable health imaging repository for the empowerment of imaging analysis related to the diagnosis, prediction and follow-up of cancer.
ProCancer-I	An AI Platform integrating imaging data and models, supporting precision care through prostate cancer's continuum.
PRIMAGE	PRedictive In-silico Multiscale Analytics to support cancer personalized diagnosis and prognosis, Empowered by imaging biomarkers.

Key Dimensions of Data Infrastructures*

	CHAIMELEON	EuCanImage	INCISIVE	ProCAncer-I	PRIMAGE
Cancer types	Lung Colorectal Breast Prostate	Colorectal Liver Breast	Lung Colorectal Breast Prostate	Prostate	Neuroblastoma Diffuse intrinsic pontine glioma
Architecture	Hybrid	Accommodating both decentralized and centralized storage	Hybrid (federated and centralized storage)	Centralized	Centralized
Data models and types of data	DICOM-MIABIS OMOP CDM (terminology IDs mainly) Structure of the eCRF	DICOM-MIABIS FHIR (and terminologies supported by FHIR + extensions)	FHIR SNOMED-CT LOINC DICOM	DICOM-RT OMOP CDM with extensions	DICOM-MIABIS OMOP CDM
Deidentification process	Pseudonymized initially for curation and then fully anonymized data at the central repository	Pseudonymized data	Pseudonymized data	Fully anonymized data	Pseudonymized data
Curation tools	Data completeness and consistency tools, image quality checking, Image anonymization, annotation, segmentation and harmonization	Image anonymization/pseudonymization, quality control and annotation, non-imaging data anonymization and homogenization	Image de-identification tools, quality control, and annotation tools	Image quality control, anonymization, motion-correction, co-registration & annotation.	Image labelling, quality checking, annotation, denoising, motion correction, registration
Number of potential subjects	13,000 full cases (images + clinical data), 34,000 image only cases	25,000	8,850	17,115	1,500

*Kondylakis, Haridimos, et al. "Data infrastructures for AI in medical imaging: a report on the experiences of five EU projects." *Eur. Rad. Exp.* 2023

CHAIMELEON De-Id process

- **Step 1:** Pseudonymization
  - DICOM images are pseudonymized "E. Attribute Confidentiality Profiles" (a) clean descriptors, (b) retain longitudinal full dates, and (c) retain patient characteristics direct identifiers are either removed or replaced by a randomly generated pseudonym.
  - A table of correspondence is kept secure within the hospital
  - Clinical data are associated with images using the same pseudonym.
  - All dates, including exam dates and dates of birth, are kept at this stage - for handling discrepancies during the curation process, allowing back-and-forth discussions.
- **Step 2.** Anonymization
  - Data curators perform a quality check
  - A new patient identifier is generated (no table of correspondence kept)
  - All dates are shifted to keep longitudinal information.
  - Data is sent to the central repository.

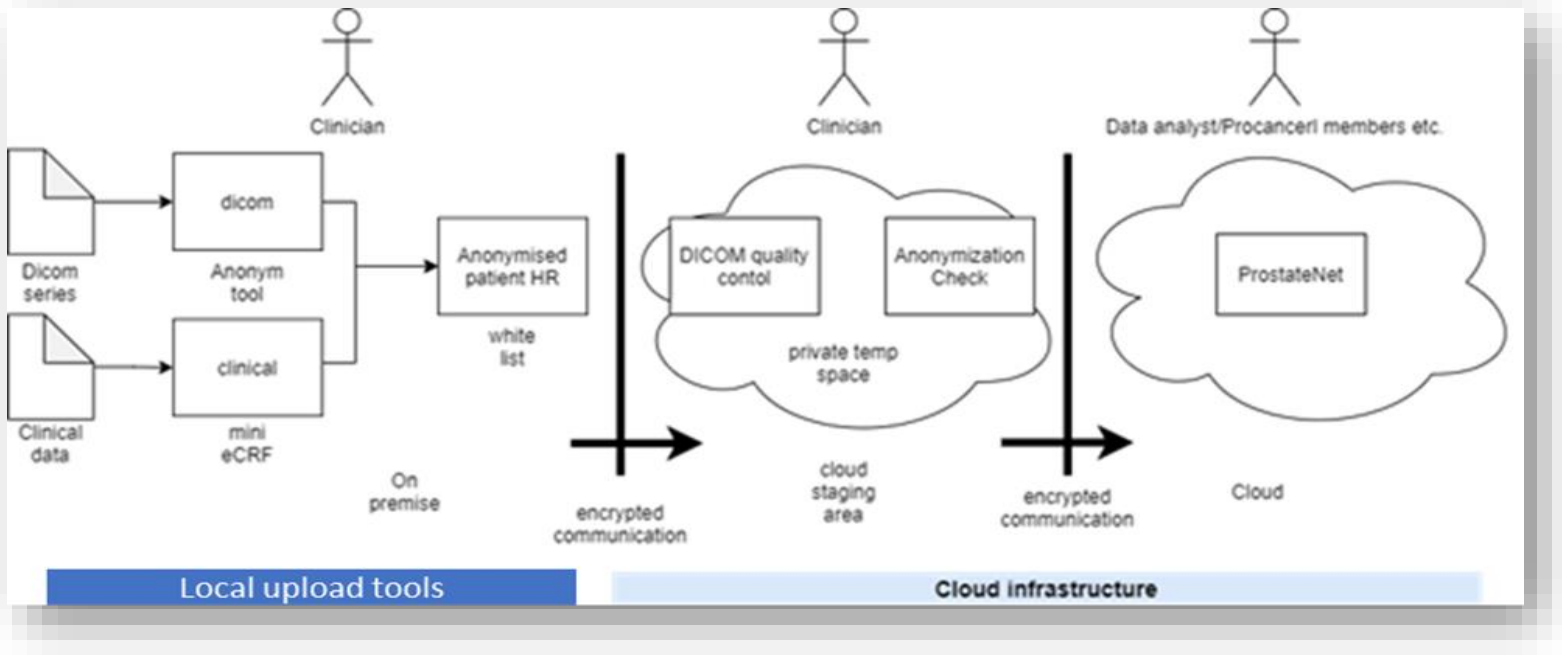
EuCanImage De-Id process

- **Step 1:** DICOM Images Pseudonymization
  - One single deidentification step to encrypt the patient's medical record ID
  - Performed in collective minds radiology (CMRAD), a cloud-based GDPR-compliant platform
    - The standard hash algorithm adopted by the platform is SHA512/256, the truncated version of SHA512
    - To generate the hash key the input information is a secret key (unique for each hospital) concatenated together with the personal patient's ID.
    - The final patient's hashed ID is a 64 alphanumeric characters code.
- **Step 2:** Clinical Data Anonymization
  - Clinical data is collected using an electronic case report form (eCRF) REDcap
  - The data attributes collected do not contain any direct identifier, only patient pseudonyms.
  - Additional measures are enforced to decrease the risk of identifiability
    - e.g., replacing the date of diagnosis by age at diagnosis, substituting the collection of dates of starting and ending of procedures by periods of time when possible, or the use of arbitrary dates.

INCISIVE De-Id process

- **Step 1:** Anonymization through CTP Anonymizer
  - The CTP Anonymizer was used on a custom protocol for anonymizing DICOM images.
  - The protocol established collaboratively between data providers, AI developers and legal partners, finding the balance between the usability of the data by the AI developers and privacy.
    - For the name and identifier of the patient, a naming convention was proposed.
    - For other different identifiers, a hash function would be applied using the de-identified patient id as a seed.
    - For the dates, it was mandatory to keep the original offset between consecutive examinations of the patient the same after the de-identification process.
    - Other DICOM fields that might contain information leading to patient identification and were not useful for the AI developers were either removed completely or replaced with a zero-length value.
  - All of the transformations were provided through the CTP Anonymizer tool and were configured in a single de-identification script shared with the Data Providers.
- **Step 2:** Implement own de-identification tool.
  - Based solely on the NEMA protocol and gives the Data Providers the option to select the level of privacy they want to apply to their data.
  - This is achieved through various options inside the tool which the Data Provider can use to remove or de-identify different DICOM fields.

ProCAncer-I De-Id process



- DICOM Committee Supplement 142-Clinical trial de-identification profile was used
- RSNA CTP Anonymizer by a set of rules in the designated script file
  - An offset was defined from the original examination date for the dates
  - Other extreme accuracy fields such as time of acquisition were obscured

PRIMAGE De-Id process

- Access is granted to several registries and clinical trial databases for secondary use of available clinical data.
  - Already Pseudonymized Data for model development
    - SIOPEN-r-net (International Society of Pediatric Oncology European Neuroblastoma Research Network)
    - GPOH clinical trials database (German Society of Pediatric Oncology and Hematology).
  - External hospitals for external models validation
    - EUPID was also included as the pseudonymization tool in the PRIMAGE project.

- **Step 1.** When a new patient is incorporated in the PRIMAGE database, a new and unique pseudonym is given for its pseudonymization.

- **Step 2.** All the DICOM tags with sensitive information as stated in the DICOM standards PS3.15 are removed or emptied from the uploaded files.

The AI4HI working group

EuCanImage	A European Cancer Image Platform Linked to Biological and Health Data for Next-Generation Artificial Intelligence and Precision Medicine in Oncology.
CHAIMELEON	Accelerating the lab to market transition of AI tools for cancer management.
INCISIVE	A multimodal AI-based toolbox and an interoperable health imaging repository for the empowerment of imaging analysis related to the diagnosis, prediction and follow-up of cancer.
ProCancer-I	An AI Platform integrating imaging data and models, supporting precision care through prostate cancer's continuum.
PRIMAGE	PRedictive In-silico Multiscale Analytics to support cancer personalized diaGnosis and prognosis, Empowered by imaging biomarkers.

EUCAIM	European Federation for Cancer Images