**Project title:** European Federation for Cancer Images

**Project acronym:** EUCAIM

**Grant Agreement:** 101100633

**Call identifier:** DIGITAL-2022-CLOUD-AI-02

# D7.3: Definition of a set of use cases

**Responsible partner:** GUMED

**Other partners:** UMU, EATRIS, Erasmus MC, REgVB, KI/KS, LIU, HULAFE, APHP, PSD, TUM, PHILIPS, UNIPI, FCRB, HCS, GOC, CF, MUW, RadboudUMC, UKA, BBMRI, QUIBIM, IQVIA, MAT, FPG, IFOM, CNR, CHUP, SERMAS, EUBI(NEUROMED), EUBI(UDC), BBMRI(PORT))

**Author(s):** Michał Kosno (GUMED), Maciej Bobowicz (GUMED), Stefanie Charalambous (MAG), Eirini Kaldeli (MAG), Francesca Pia Caputo (UNIPI), Philipp Seeböck (MUW)

**Contributors:** Alessandra Viale (EUBI), Katrine Riklund (UMU), Sara Zullino (EATRIS), Maria Gonzalez López (SAS), Krystian Brzozowski (GUMED), José Almeida (CF), Linda Chaabane (CNR), Ignacio Blanquer (UPV), Celia Martin Vicario (QUIBIM), Gianna Tsakou (MAG), Valia Kalokyri (FORTH), Federica Cruciani (IFOM)

**Reviewers:** Mirna El Ghosh (LIMICS), Kurt Majcen (BBMRI-ERIC)

**Date of delivery:** 30/06/2025

**Version:** 1

# Table of Contents

# Executive summary

The European Federation for Cancer Images (EUCAIM) project seeks to build a secure, integrated, and wide-reaching platform that enables reliable scientific research. To ensure the platform functions effectively, it undergoes rigorous testing through the implementation of carefully selected use cases identified through calls and AI for Health imaging (AI4HI) network. Detailed descriptions of the calls are available in D7.1 and D7.2. This document compiles use cases that have been identified through the previous projects within the AI4HI network, two rounds of internal calls and an open call. A number of use cases have been identified, both related to data sharing and data usage. This document also details the present implementation plan for pilot use cases, which have been rigorously selected to thoroughly assess the functionality of the platform developed within the EUCAIM project. These pilot use cases are implemented in accordance with the Data Holders EUCAIM Federation Handbook. The use cases described herein pertain to data sharing.

# Introduction

The European Federation for Cancer Images (EUCAIM) project aims to establish a comprehensive, secure, and federated platform that enables trusted scientific research through a robust and interoperable infrastructure. By facilitating the secure sharing, processing, and analysis of multimodal cancer imaging and associated clinical data, EUCAIM seeks to accelerate advancements in artificial intelligence (AI) for cancer research and personalized medicine.

The overarching goal of the project is to ensure that researchers and clinicians across Europe can effectively collaborate while adhering to ethical, legal, and technical standards that prioritize patient privacy and data security. To achieve this ambitious objective, EUCAIM has undertaken an extensive process of identifying, gathering, and curating use cases that will serve as critical benchmarks for evaluating the platform's capabilities. These use cases, which have been sourced through contributions from the AI for Health Imaging (AI4HI) projects, internal calls and open calls are carefully designed to test and validate the platform’s performance under real-world research conditions. The AI for Health Imaging (AI4HI) projects contributed to EUCAIM from its proposal phase, laying the groundwork for early use case development. Building on this foundation, EUCAIM later launched two additional types of calls. Internal and open calls that expanded the scope of use case contributions. As a result, the compiled use cases can be categorized into three distinct groups: those originating from AI4HI, those submitted through internal calls, and those received via the open call process. They encompass a diverse range of challenges that hold significant scientific and clinical relevance, reflecting the evolving needs of the European research and healthcare communities. By engaging a broad spectrum of stakeholders, including hospitals, research institutions, technology providers, and policymakers, EUCAIM ensures that the platform remains aligned with the priorities of the European Health Data Space (EHDS) and the wider digital health landscape. A fundamental aspect of the EUCAIM project is the rigorous assessment of its infrastructure through a series of controlled pilot studies and real-world evaluations. These evaluations will provide crucial insights into the platform’s utility, effectiveness, scalability, and overall readiness to support cutting-edge scientific research. The findings from these assessments will directly feed into the platform’s iterative development, ensuring continuous improvements in interoperability, security, and user experience. Moreover, the EUCAIM framework is designed to uphold the principles of Findability, Accessibility, Interoperability, and Reusability (FAIR), thereby fostering an ecosystem where medical data can be effectively leveraged to drive innovation in AI-driven healthcare solutions. This document provides a detailed account of the identified use cases, outlining their objectives, scope, and relevance to the broader research and

clinical community. This document plays a crucial role in the validation of the EUCAIM platform within Task 7.4 of the project. The subsequent chapters present a comprehensive list of use cases, along with implementation examples to ensure seamless integration, interoperability, and adherence to FAIR principles.

# Definition of Use Cases

This section outlines the use cases selected to validate the EUCAIM platform. Sourced from internal initiatives, open calls, and AI4HI projects, these use cases address key challenges in medical research and clinical practice, including early diagnosis, prognosis, personalized treatment planning and monitoring. Each project requires secure, federated access to medical imaging and clinical data and utilizes core components of the EUCAIM infrastructure, including data ingestion, harmonization and AI tools, privacy-preserving mechanisms, and user-friendly interfaces. As such, these projects are essential to assess the EUCAIM platform's capacity to support large-scale, collaborative research efforts in a real-world context. By incorporating this wide range of functionalities, the use cases enable a thorough and multifaceted evaluation of the platform's technical capabilities, interoperability, and compliance with ethical and legal standards, particularly in relation to GDPR (General Data Protection Regulation) and data governance. A detailed description of each use case, including their specific objectives, methodologies, data requirements, and expected outcomes, is provided in **Annex 1**. This annex serves as a key reference for understanding the scientific and technical rationale behind each use case and illustrates how they collectively contribute to the validation and future enhancement of the EUCAIM platform. All of them were already evaluated by the Access Committee in terms of scientific quality, alignment with the call objectives, compliance with technical requirements and compliance with ethical/legal requirements.

# Internal Use Cases – I$^{st}$ round

Use cases submitted in the internal call originate from consortium members and are divided into three main categories: data users, data holders and both (combined Data Users and Holders). One of the use cases was submitted as "other". Following a thorough review, the Access Committee has approved all 14 use cases, which are outlined in this document, along with a high-level summary in the tables below (Table 1). The use cases presented in the Table 1 cover various types of cancers, including: prostate cancer, breast cancer, lung cancer, colon cancer, primary liver cancer, non-hodgkin lymphoma (NHL) and hepatocellular carcinoma. The collected data are also characterized by diverse modalities: magnetic resonance (MR), positron emission tomography (PET), computed tomography (CT), digitized HE slides, Mammography (MG) and dynamic contrast-enhanced magnetic resonance (DCE-MR). The most common modalities are magnetic resonance imaging and computed tomography. Data User use cases refer to scientific projects that will utilize data available through the EUCAIM public catalogue, implemented via reference nodes or through the federated network. Data Holder use cases involve sharing data either by uploading to a reference node (central repository) or by making it accessible through a federated architecture, without transferring the data.

In cases where a consortium member participates both as a Data Holder and a Data User, the use case has been divided into two entries: one describing the scientific investigation and the other detailing the data-sharing process. These dual-role use cases are of particular interest due to their complexity and potential impact. They allow for both contribution to the data ecosystem and access to existing datasets on the platform. This dual engagement supports the development of sophisticated, data-driven solutions, leveraging diverse datasets through the federated model to create robust and interoperable imaging diagnostic tools.

Summarizing the findings, five of the 14 use cases will share data via a federated node, five will store data in a central repository. Furthermore, the submitted use cases exhibit differing levels of alignment with the EUCAIM platform. Specifically, three use cases have been classified as Tier 3 compliant, whereas six have been assessed as meeting Tier 1 compliance criteria. The use cases span a wide range of oncological domains, including prostate cancer, breast cancer, non-Hodgkin lymphoma, lung cancer, hepatocellular carcinoma, primary liver cancer, and brain cancer. The datasets vary in terms of patient numbers, age distribution, and gender, providing a rich and diverse foundation for research and model development. The table below provides a high-level summary of the first round of the internal call. The numbers for the individual use cases were created as follows: I (Internal call) / call number, DH (Data Holder) or DU (Data User) and use case number. For instance, I/1DH1 is an abbreviation for the first data holder use case from the first round of internal call.

Table 1. Use cases collected from the first round of internal calls related to data sharing.

| No. of Use Case | Use Case title | Organization | Data sharing | Tier | No. of dataset | Modality | Age range (years) | Sex | No. of subjects | Cancer type |
|---|---|---|---|---|---|---|---|---|---|---|
| I/1DH1 | SAS Use Case | Servicio Andaluz De Salud | Federated Node | 3 | 1 | Magnetic Resonance | 40 - 90 | Male | 600 | Prostate cancer |
| | | | | | 2 | Positron Emission Tomography / Computed Tomography | 30 - 80 | Female (99.9%) and Male (0.1%) | 100 | Breast cancer |
| | | | | | 3 | Positron Emission Tomography / Computed Tomography | 18 - 80 | Male (60%) and Female (40%) | 125 | Non-Hodgkin Lymphoma (NHL) |
| | | | | | 4 | Computed Tomography | 40 - 90 | Male (80%) and Female (20%) | 600 | Lung cancer |
| I/1DH2 | SCIENSANO Use Case | Sciensano | Federated Node | 1 | 1 | Digitized HE slides | Over 18 | Male and Female | 937 | All |
| I/1DH3 | KI Use Case | Karolinska Institutet | Federated Node | 2 | 1 | Mammography | avg. 57 | Female | 15564 | Breast cancer |
| I/1DH4 | APHP Use Case | Assistance Publique Hôpitaux De Paris | Federated Node | 3 | 1 | Computed Tomography / Magnetic Resonance / Digitized HE slides | Over 18 | Male and Female | 872 | Primary Liver Cancer |
| | | | | | 2 | Magnetic Resonance | Over 40 | Male | 300 | Prostate cancer |

| No. of Use Case | Use Case title | Organization | Data sharing | Tier | No. of dataset | Modality | Age range (years) | Sex | No. of subjects | Cancer type |
|---|---|---|---|---|---|---|---|---|---|---|
| I/1DH5 | SAPIENZA Use Case 1 | Universita Degli Studi Di Roma La Sapienza | Central Repository | 1 | 1 | Magnetic Resonance | 45 - 75 | Male | 200 | Prostate cancer |
| I/1DH6 | MUW Use Case | Medizinische Universitaet Wien | Central Repository | 1 | 1 | Dynamic Contrast-enhanced Magnetic Resonance | 20 - 87 | Female | 100 | Breast cancer |
| I/1DH7 | SAPIENZA Use Case 2 | Universita Degli Studi Di Roma La Sapienza | Central Repository | 1 | 1 | Computed Tomography / Magnetic Resonance | 41 - 89 | Male and Female | 100 | Hepatocellular carcinoma |
| I/1DH8 | SAPIENZA Use Case 3 | Universita Degli Studi Di Roma La Sapienza | Federated Node | 1 | 1 | Computed Tomography | 18 - 90 | Male and Female | 200 | Colon cancer |
| I/1DH9 | CHUP Use Case | Centro Hospitalar Universitario Do Porto Epe | Central Repository | 1 | 1 | Computed Tomography | 30 - 90 | Male and Female | 150 | Lung cancer |
| I/1DH10 | HULAFE Use Case | Fundacion Para La Investigacion Del Hospital Universitario La Fe De La Comunidad Valenciana | Central Repository | 3 | 1 | Magnetic Resonance | 30 - 90 | Male and Female | 400 | Brain cancer |

Table 2. Use cases collected from the first round of internal calls related to data usage.

| No. of Use Case | Use Case title | Intention | Organization | Short description |
|---|---|---|---|---|
| I/1DU1 | OCEANUS | • train/validate AI tools<br>• development of AI tools and solutions<br>• training of AI tools and solutions | IFOM | OCEANUS is a pragmatic project that will establish the clinical utility of I-ROR (Individual Risk Of Resistance), a first- in-kind digital marker predicting the Risk Of Resistance to chemotherapy in individual patients with metastatic colorectal cancer (mCRC). I-ROR will change the paradigm of care for this highly prevalent tumor enabling the personalized management of patients, that will increase their survival and quality of life while optimizing the burden of care and minimizing the socio-economical costs. The project consists in a pragmatic clinical trial representing a step forward towards individualized precision medicine for metastatic colorectal cancer (mCRC) by confirming the clinical utility of the I-ROR diagnostic digital marker. I-ROR is an AI-driven predictive marker that quantifies the responsiveness to standard chemotherapy in individual patients. I-ROR is a cost-effective, widely-applicable, and easy-accessible diagnostic tool that integrates three AI-based signatures derived from easily accessible information already collected per standard of care (FFPE slides and CT-scans) universally available at point-of-care in every European hospital. |
| I/1DU2 | Domain Generalization of AI-based models in Cancer Imaging using Disentanglement Learning | • development of AI tools and solutions<br>• training of AI tools and solutions<br>• validation of AI tools and solutions | Medizinische Universitaet Wien | Main Objectives: Developing novel machine learning methodology to enhance domain adaptation and generalization abilities of AI models, ensuring robustness to technical noise in imaging data related to changes in imaging technology. Expected Results: Novel methodology for effective detection and segmentation of cancer across various scanners, institutes and hospitals. Clinical Impact: Imaging data from different sites or hospitals often vary significantly due to different acquisition techniques, devices, imaging protocols or patient population characteristics. Furthermore, the mentioned factors may change over time as technology advances, and image characteristics evolve. This impedes applicability of models across sites, leads to a reduction of model performance, and in the worst case a need to train several models for an inhomogeneous reality of large-scale multi-center repositories. These challenges lead to models that are currently typically trained and evaluated on single-center datasets, limiting both their potential generalization performance as well as the explanatory power of the evaluation with respect to clinical applicability. The novel methodology developed in this project will serve as the basis for AI models that can be trained on single or multi-institutional datasets and applied on hold-out institutional data with differing imaging appearance or distributions. Methodology: Based on prior work in continual learning, domain generalization and disentanglement, we will develop novel deep learning based methodology which is robust to changes in imaging technology and capable of handling data heterogeneity across scanners, institutes and hospitals. We aim to use multi-centric data to develop and validate prediction algorithms that yield robust prediction accuracy across centers. |
| I/1DU3 | Predictive recurrence location and time to recurrence using IA | • development of AI tools and solutions<br>• training of AI tools and solutions<br>• validation of AI tools and solutions | Fundacion Para La Investigacion Del Hospital Universitario La Fe De La Comunidad Valenciana | General Description: Our project aims to develop a predictive model leveraging artificial intelligence (AI) techniques to forecast the location and time to recurrence in patients with GBM. By harnessing imaging biomarkers such as diffusion, radiomics, and perfusion, along with diagnostic clinical data, we seek to enhance clinical decision-making and improve patient outcomes. Main Objectives: - Develop AI algorithms to analyze imaging biomarkers extracted from MRI scans. - Create a model capable of estimating the time to GBM recurrence based on predefined imaging biomarkers. - Develop an IA model to predict the location of GBM recurrence within the brain using diagnostic imaging data. Expected Results: We anticipate that our predictive models will accurately forecast the time to GBM recurrence and identify the location of recurrent tumors with high precision, leveraging the rich information provided by imaging biomarkers. Expected Clinical Impact: Our project has the potential to revolutionize the approach to GBM management by facilitating personalized treatment planning based on individual patient characteristics and tumor biology. Also, the ability to predict GBM recurrence and its location will enable clinicians to intervene promptly, potentially leading to earlier detection of recurrence, |

| | | | | |
|---|---|---|---|---|
| | | | | improved response to therapy, and better patient outcomes. By providing clinicians with predictive tools to anticipate recurrence and progression, our project may help optimize resource allocation in healthcare settings, ensuring that patients receive timely and appropriate interventions based on their predicted risk of recurrence. Methodology: We will utilize advanced image processing techniques to extract quantitative imaging biomarkers from MRI scans, including diffusion metrics, radiomic features, and perfusion parameters. 4 After, we will employ machine learning and deep learning algorithms to develop predictive models for time to recurrence and location of recurrence based on the extracted imaging biomarkers and diagnostic imaging data. Furthermore, we will validate the predictive models and assess their performance in terms of accuracy, sensitivity, specificity, and clinical utility. |
| I/1DU4 | Developing methods for collaborative research | • Developing methods for collaborative research | Linkopings Universitet | The AIDA Data Hub Sensitive Data Services (SDS 2.0) platform is designed to provide a suite of services tailored to ethically sanctioned research endeavors or activities grounded in a legal and ethical framework. Within this platform, an extensive array of offerings will be made available, encompassing sensitive data processing, data sharing, primary storage, private remote desktop functionalities, as well as provision for private VMs (Virtual Machines), web applications, and PACS (Picture Archiving and Communication System) systems. Additionally, there are plans to incorporate trusted medical imaging import capabilities, facilitating the direct transmission of medical examinations from scanners to the appropriate destination. The platform will accommodate Petabyte object storage, enabling the accumulation, refinement, annotation, and collaborative exploration of data. Furthermore, a comprehensive range of compute services will be provided, incorporating both GPU-enhanced and standard compute resources, utilizing the OpenStack framework. This will be complemented by the integration of a Kubernetes platform, empowering users to deploy and manage their preferred web services securely within private environments. Services such as backed-up long-term primary storage, large-scale project storage, and robust CPU and GPU compute options will be available as supplementary offerings. Crucially, the platform ensures strict segregation of user environments, affording each user autonomy within their designated "bubble" without visibility into other users' activities. Authentication mechanisms facilitate institutional login, enabling users to access the platform using credentials from their home institution, facilitated through the Life Science Login framework. This system will not require VPN connections or specialized account provisioning, streamlining access while enabling connectivity to other private services. SDS 2.0 will provide varied access modes to cater to a broad spectrum of users, including expert AI developers and clinicians, thereby enhancing its appeal and usability. For researchers inclined to share their data, the platform will facilitate this through the REMS resource entitlement management system, a web-based interface empowering researchers to delegate or manage their data-sharing decisions autonomously. SDS 2.0 will provide a data collaboration platform for EUCAIM, and this driver use case aims at providing a tailored environment for IDx Panorama, a comprehensive research initiative focused on multimodal and multi-omic cancer imaging. In IDx Panorama, researchers aim to integrate photon counting computed tomography (PCCT) technology into existing clinical procedures in the context of liver cancer. This involves incorporating PCCT alongside standard care pathways for patients diagnosed with liver cancer. Prior to surgery patients with diagnosed liver cancer are subjected to radiology assessment using CT and/or MRI. After surgical resection, in addition to standard care, the resected tumour tissue is subjected to ex-vivo higher image quality PCCT. This allows for detailed imaging of the tumor, facilitating correlations between its appearance and its biological characteristics. These insights could potentially inform the development of improved imaging diagnostics software. Following tumor removal, the tissue is sent for histological analysis and pathology assessment to confirm the cancer diagnosis. Comparisons are then made between the histological findings and the photon counting data, aiding in the validation of the photon counting technology's accuracy. Ultimately, the goal is to refine multimodal diagnostics for liver cancer, integrating photon counting with genetic analysis and other laboratory analyses. SDS 2.0 will |

| | | | | |
|---|---|---|---|---|
| | | | | function as a collaboration platform for this project, offering the tooling to streamline the transfer of examination data directly. |

## Internal Use Cases – II$^{nd}$ round

Following the initial round of internal calls, which yielded 14 use cases, a second round was conducted, leading to the approval of nine new projects. The call remains open. This is the status as of May 2025. The following section details the use cases that were submitted by the consortium partners in the second round. Of the nine, five use cases focus on data sharing, predominantly through federated mechanisms (one application involved uploading data to a central server). These datasets have been linked to various types of cancer, including: breast cancer, meningioma, thyroid cancer, colorectal cancer and glioblastoma. The collected data are also characterized by diverse modalities, which include: magnetic resonance (MR), scintigraphy, mammography (MG) and computed tomography (CT). The remaining four projects focus on leveraging radiomics, machine learning, and artificial intelligence, with the aim to enhance diagnostic imaging and improve its quality. In terms of data quality and readiness, three of the data-sharing use cases are Tier 3 compliant, indicating a high level of data curation and maturity, while two are Tier 2 compliant, representing intermediate readiness for integration and analysis. A key feature across these projects is their emphasis on interoperability, which plays a critical role in enabling seamless data access and facilitating cross-institutional collaboration. Enhanced interoperability will support more efficient data processing, integration across systems, and broader use of the resulting insights within the clinical and research communities.

Table 3. Use cases collected from the second round of internal calls related to data sharing.

| No. of Use Case | Use Case title | Organization | Data sharing | Tier | No. of dataset | Modality | Age range (years) | Sex | No. of subjects | Cancer type |
|---|---|---|---|---|---|---|---|---|---|---|
| I/2DH1 | HCB Use Case | Hospital Clinic De Barcelona | Federated Node | 3 | 1 | Magnetic Resonance | 18 - 95 | Male and Female | 150 | Meningioma |
| I/2DH2 | AUTH Use Case 1 | Aristotelio Panepistimio Thessalonikis | Federated Node | 3 | 1 | Scintigraphy | Over 18 | Male and Female | 80 - 100 | Thyroid cancer |
| I/2DH3 | AUTH Use Case 2 | Aristotelio Panepistimio Thessalonikis | Federated Node | 3 | 1 | Mammography | Unknown | Female | 72 | Breast cancer |
| | | | | | 2 | Mammography | Unknown | Female | 125 | Breast cancer |
| I/2DH4 | SERMAS Use Case | Servicio Madrileno De Salud | Federated Node | 2 | 1 | Computed Tomography | 34 - 91 | Male (61) and Female (62) | 123 | Colorectal cancer |
| I/2DH5 | UDC Use Case | Universidade De Coimbra | Central Repository | 2 | 1 | Magnetic Resonance | 18 - 85 | Male | > 50 | Glioblastoma |

Table 4. Use cases collected from the second round of internal calls related to data usage.

| No. of Use Case | Use Case title | Intention | Organization | Short description |
|---|---|---|---|---|
| I/2DU1 | Prediction of Treatment Response and Disease Progression in Lung Cancer Patients Using Voxel-Level Radiomics from PET-CT Data | • train/validate AI tools<br>• development of AI tools and solutions<br>• training of AI tools and solutions<br>• validation of AI tools and solutions | Aristotelio Panepistimio Thessalonikis | The main objective of this case is to develop an AI model that predicts treatment response and disease progression in lung cancer patients based on PET-CT data. By extracting radiomics features at the voxel level from PET-CT images and training a machine learning (ML) or deep learning (DL) network, the model is expected to provide accurate prognostic information that will help clinicians in personalized treatment planning. |
| I/2DU2 | Whole gland radiomics based model for prostate cancer classification | • train/validate AI tools<br>• development of AI tools and solutions<br>• training of AI tools and solutions<br>• validation of AI tools and solutions | Aristotelio Panepistimio Thessalonikis | Radiomics, a quantitative method for analysing medical images, and AI models have been developed targeting prostate cancer (PCa) based on the analysis of cancer lesions, which is a significant limitation, given that the manual segmentation of which is both resource-intensive and time-consuming, while relying on AI models for segmentation affects robustness due to potential inaccuracies. Automated delineation of the entire prostate gland demonstrates higher accuracy. Moreover, most AI models proposed in the literature are trained on datasets originating from limited clinical sites, with images acquired using specific protocols from scanners that may have varying calibration parameters. Consequently, the data are inherently inhomogeneous, making harmonization a crucial preprocessing step. In the current model, we will apply and test different harmonization techniques to reduce unwanted variation at both the image and feature levels. The main objective is the development of a federated model which is able to identify clinically significant prostate cancer (csPCa) by combining radiomic analysis of MRI images from the entire prostate gland with basic clinical characteristics of the patient. Additionally, we will investigate the significance of the radiomic features from different prostate regions. The ultimate goal is to develop with enhanced generalizability. |
| I/2DU3 | Breast Density Prediction Algorithm - Application of Multiple Deep Neural Network Based Model for Automated Breast Density Classification | • train/validate AI tools<br>• development of AI tools and solutions<br>• training of AI tools and solutions<br>• validation of AI tools and solutions | Gdanski Uniwersytet Medyczny | The accurate evaluation of breast density from mammography images is critical for the accurate estimation of breast cancer risk, as higher breast density is associated with an increased likelihood of developing the disease. However, the current process for visually assessing breast density remains highly challenging and subjective, with considerable inter-observer variability, even among trained radiologists. The inherent variability and errors in visual assessments underscore the need for a standardized, objective approach to breast density classification. Recent advances in deep learning methodologies offer promising solutions by providing tools capable of objective and reproducible assessment. Our research specifically aims to enhance breast density assessment by leveraging the Tree-structured Parzen Estimator (TPE) algorithm-driven transfer learning to optimize ResNet, DenseNet and EfficientNet convolutional neural networks (CNNs) for this task. These deep learning architectures are widely recognized for their high performance on medical imaging tasks, with DenseNet's connected pathways and EfficientNet's optimized scaling enabling high accuracy in feature extraction and classification. Our ultimate objective is to develop and rigorously validate an artificial intelligence-based (AI) tool that can automatically and accurately classify breasts by their density based on digital mammography. Our methodology ensures the development of a robust and reliable classification tool that performs well across a range of clinical environments. Initial tests have demonstrated that the TPE-driven transfer learning approach improves model performance in breast density classification. However, further validation on a larger and more diverse dataset is essential to confirm these findings and maximize the model's robustness across demographic and device variations. |

| | | | | |
|---|---|---|---|---|
| I/2DU4 | Breast MRI Screening - Background Parenchyma Enhancement (BPE) and lesion localization | • train/validate AI tools | Philips GMBH | Women with high risk have a breast MRI as part of their breast screening. In breast screening, an early diagnosis is paramount. Two challenges for early diagnosis on breast MRI are small lesion detection and BPE. The lesions may be typically small <5mm and easy to miss. An algorithm for small lesion detection could facilitate the workflow and help detect missed lesions. Similarly BPE is defined qualitatively. Imaging protocol variation leads to difficulties in defining a quantitative metric leading to site-specific definitions. A qualitative metric could help on prognostic power for cancer diagnosis. |

# External Use Cases (Open Call)

Following the external open call, 66 institutions applied, of which 19 were incorporated into the consortium. The selection was made by the access committee based on criteria that had been established previously. A ranking list was then created to determine which partners would be included in the consortium. Out of the 19 applications that were approved by the European Commission, 12 were classified as data holders and 11 as data users. The further refinement of the use cases, based on the accepted submissions, was carried out in the same way as for the internal calls. The tables below provide a high-level summary of the use cases resulting from the relevant submissions. The numbers for the individual use cases were created as follows: O (Open call) / DH (Data Holder) or DU (Data User) and use case number. For instance, O/DH1 is an abbreviation for the first data holder use case from the open call. A detailed open call summary can be found in D7.2. The data reported in the open call cover a wide range of cancer types, which are included in the table below. They are also characterized by a variety of modalities, including: magnetic resonance (MR), slide microscopy, computed tomography (CT), positron emission tomography - computed tomography (PET-CT), positron emission tomography (PET), mammography (MG), ultrasound (US), digitized histological slides, X-Ray, digital microscopic image, pathology scanned images and colonoscopy video recording.

Table 5. Use cases collected from the open calls related to data sharing.

| No. of Use Case | Use case title | Organization | Data sharing | Tier | No. of Dataset | Modality | Age range (years) | Sex | No. of subjects | Cancer type |
|---|---|---|---|---|---|---|---|---|---|---|
| O/DH1 | OUS Use Case | Oslo Universitetssykehus HF | Federated Node | 2 | 1 | Magnetic Resonance | 36 - 68 | Female (29.63%) and Male (70.37%) | 27 | Primary brain cancer |
| | | | | | 2 | Magnetic Resonance | 18 - 70 | Female (60%) and Male (40%) | 120 | Brain metastases from lung cancer and malignant melanoma |
| | | | | | 3 | Magnetic Resonance | 0 - 17 | Female (60%) and Male (40%) | 1000 | Brain tumors |
| O/DH2 | LU Use Case | University of Latvia | Federated Node | 1 | 1 | Slide Microscopy | 35 - 87 | Female and Male | ~ 300 - 500 | Gastric cancer |
| O/DH3 | AUMC Use Case | Stichting Amsterdam UMC | Central Repository | 1 | 1 | Magnetic Resonance | Over 18 | Female and Male | 1200 | Diffuse glioma |
| O/DH4 | HM HOSPITALES Use Case | Fundación de Investigación HM Hospitales | Central Repository | 1 | 1 | Magnetic Resonance, Computed Tomography, Positron Emission Tomography-Computed Tomography, Positron Emission Tomography and Mammography | 4 - 102 | Female and Male | 3513 | Liver cancer |
| | | | | | 2 | Magnetic Resonance, Computed Tomography, Positron Emission Tomography-Computed Tomography, Positron Emission Tomography and Mammography | 24 - 102 | Female and Male | 1216 | Pancreatic cancer |
| | | | | | 3 | Magnetic Resonance, Computed Tomography, Positron Emission Tomography-Computed Tomography, Positron Emission Tomography and Mammography | 0 - 102 | Female and Male | 4345 | Lung cancer |
| | | | | | 4 | Magnetic Resonance, Computed Tomography, Positron Emission Tomography-Computed Tomography, | 0 - 102 | Female and Male | 3808 | Breast cancer |

| | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | | | | | Positron Emission Tomography and Mammography | | | | |
| | | | | | 5 | Magnetic Resonance, Computed Tomography, Positron Emission Tomography-Computed Tomography, Positron Emission Tomography and Mammography | 3 - 102 | Female and Male | 4701 | Prostate cancer |
| O/DH4 | CETIR Use Case | CETIR Centre Medic SL | Central Repository | 2 | 1 | Magnetic Resonance, Computed Tomography, Positron Emission Tomography-Computed Tomography, Positron Emission Tomography and Mammography | 19 - 90 | Female and Male | 15600 | Breast, prostate, uterine cancers |
| | | | | | 2 | Magnetic Resonance, Computed Tomography | 19 - 90 | Female and Male | 1200 | Colon, rectal cancers |
| | | | | | 3 | Magnetic Resonance, Computed Tomography | 19 - 90 | Female and Male | 1900 | Pancreatic, urothelial cancers |
| | | | | | 4 | Magnetic Resonance | 19 - 90 | Female and Male | 500 | Brain cancer |
| | | | | | 5 | Computed Tomography | 19 - 90 | Female and Male | 1850 | Lung cancer |
| O/DH5 | PSHYVINVOINTIALUE Use Case | Wellbeing Services County of North Savo | Federated Node | 1 | 1 | X-Ray, Magnetic Resonance, Ultrasound, Computed Tomography, Digitized Histological Slides | 18 - 99 | Female and Male | 5200 | Breast Cancer |
| | | | | | 2 | X-Ray, Magnetic Resonance, Ultrasound, Computed Tomography, Digitized Histological Slides | 18 - 99 | Female and Male | 1900 | Lung Cancer |
| | | | | | 3 | X-Ray, Magnetic Resonance, Ultrasound, Computed Tomography, Digitized Histological Slides | 18 - 99 | Male | 4000 | Prostate cancer |
| | | | | | 4 | X-Ray, Magnetic Resonance, Ultrasound, Computed Tomography, Digitized Histological Slides | 18 - 99 | Female | 2400 | Gynecological cancers (e.g. ovarian cancer and endometrial cancer) |
| O/DH6 | ISABIAL Use Case | Fundacion De La Comunitat Valenciana Para La Gestion Del Instituto Deinvestigacion Sanitaria Y Biomedicade Alicante | Central Repository | 1 | 1 | Digital Microscopic Image | 30 - 90 | Female and Male | 2000 | Colorectal cancer |
| | | | | | 2 | Digital Microscopic Image | 30 - 90 | Female and Male | 900 | Lung cancer |
| | | | | | 3 | Digital Microscopic Image | 15 - 90 | Female and Male | 400 | Brain tumors (e.g. glioblastomas, diffuse and pilocytic astrocytomas, oligodendrogliomas, ependymomas, gangliogliomas) |
| O/DH7 | LPCC-NRC Use Case | Liga Portuguesa Contra o Cancro - Núcleo Regional do Centro | Central Repository | 1 | 1 | Mammography | 50 - 69 | Female | 90000/year | Breast cancer |

| | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| O/DH8 | FISEVI Use Case | Fundación para la Gestión de Investigación en Salud en Sevilla | Federated Node | 2 | 1 | Computed Tomography | 18 - 99 | Female and Male | 5000 | Lung cancer |
| | | | | | 2 | Mammography | 18 - 99 | Female | 4500 | Breast cancer |
| | | | | | 3 | Computed Tomography | 18 - 99 | Female and Male | 5000 | Colorectal cancer |
| | | | | | 4 | Computed Tomography | 18 - 99 | Female and Male | 3500 | Bladder cancer |
| | | | | | 5 | Computed Tomography | 18 - 99 | Male | 3000 | Prostate Cancer |
| O/DH9 | UOA Use Case | National and Kapodistrian University of Athens | Federated Node | 2 | 1 | Magnetic Resonance, Computed Tomography | 18-90 | Male and Female | ~80 | Sarcomas |
| | | | | | 2 | Magnetic Resonance | 18-90 | Male and Female | ~240 | Brain tumors (e.g. Glioma, primary cerebral lymphomas, meningiomas and schwannomas) |
| | | | | | 3 | Magnetic Resonance | Over 50 | Male | ~50 | Prostate |
| O/DH10 | NHRF Use Case | National Hellenic Research Foundation | Central Repository | 3 | 1 | Ultrasound | 15 - 70 | Female | 1000 | Ovarian and Endometrial cancer |
| | | | | | 2 | N/A | 20 - 70 | Female and Male | 200 | Breast, Lung, Colorectal |
| O/DH11 | IACS Use Case | Instituto Aragonés de Ciencias de la Salud | Federated Node | 3 | 1 | Pathology Scanned Images | 31 – 98 | Male | ~100 | Prostate cancer |
| | | | | | 2 | Colonoscopy Video Recording | 3 – 97 | Female and Male | ~500 | Colorectal cancer |

Table 6. Use cases collected from the open calls related to data usage.

| No. of Use Case | Use Case title | Intention | Organization | Short description |
|---|---|---|---|---|
| O/DU1 | TumorTrace | • development, validation or training of AI tools considered for medical devices | Stichting Amsterdam UMC | Gliomas show a non-linear growth pattern, which can be influenced by therapy. Monitoring of gliomas occurs with MRI. It would be of exceptional value to anticipate glioma growth activity to be able to time and choose therapies. For the radiologist, it is hardly possible to tell from the MRI images if a tumor will remain stable or, instead, will rapidly expand in the following weeks. However, the information of growth over time may be present in the images and can be decipherable with AI. To develop, test, and validate an AI tool that can predict imminent acceleration in tumor growth or stable disease. Such a tool can be used during multidisciplinary meetings to drive decision making in favor of the patient's health. |
| O/DU2 | CLEAR-AI: Enhancing High-Resolution Image Segmentation Precision through Collaborative Learning of Deep Neural Networks for Accurate Assessment of Axillary Lymph Node Metastasis based on Full-Field Digital Mammography Analysis | • development of AI tools and solutions<br>• training of AI tools and solutions<br>• validation of AI tools and solutions | Gdańsk University of Technology | Currently, the breast cancer diagnosis is based on clinical examination and various imaging techniques. These examinations are followed by invasive breast tumor and lymph node biopsy procedures and histopathological examination of specimens to verify the clinical diagnosis. There is no faster, less invasive and cheaper diagnostic approach so far, which could reduce the number of necessary radiological, surgical and pathology tests to confirm the presence of cancer and its potential spread to axillary lymph nodes (ALN). Doctors need all this information to facilitate the selection of appropriate treatment and its sequence. The project will try to identify the set of information embedded in the full-field digital mammography (FFDM) images and the minimal set of clinical information representative for predicting ALN metastasis. The current state-of-the-art has no such an option for clinicians and their patients. The challenge is to efficiently take advantage of the AI tools to extract this knowledge hidden in medical data. Our ambition is to develop theoretical foundations and AI-driven methods for automated early diagnosis of breast cancer, paying special attention to ALN status. The system will be designed to analyze medical data, extract information from it, and provide clinicians with the suggestions for diagnoses regarding breast lesion detection, type of lesion (benign versus malignant), and ALN status. Special attention will be paid to problems related to the shortage of training data, especially annotated at the pixel level, the extraction and identification of medically important features to support the diagnosis, clear explanations of decisions undertaken by AI-based systems, and mechanisms to enable synergetic cooperation between different AI models and their end-users, doctors (Human-in-the-Loop approach). The main objective of the use case is to develop collaboratively learned deep neural networks (DNNs) for the classification and segmentation of breast lesions, utilizing image and pixel-level annotations in high- resolution medical image settings, specifically FFDM. The system will be able to differentiate lesion types and predict ALN status based on information from mammography images. Emphasis will be placed on ensuring that classification results can be explained, in accordance with academic standards, clinical requirements and regulatory requirements, including EU AI Act legislation. Expected results include the development of an AI-based breast tumor segmentation tool suitable for high-resolution medical images. Moreover, the system should be able to differentiate tumor types and assess metastatic ALN with a similar or higher level of sensitivity and the same specificity as the current 'gold standard' methods, while providing visual reasons for prediction, thus increasing the system's trustworthiness. The expected clinical impact of this tool will be to reduce the number of necessary surgical and pathological tests required to confirm the presence of cancer and its potential spread to the ALN. This knowledge of the tumor characteristics and ALN status will assist clinicians in selecting the appropriate treatment and its sequence. Moreover, the system for automatic masks generation learned through collaboration can be used for studies on more comprehensive cancer analysis. Methodology: The methodology for achieving the objectives of the use case will involve supervised learning of a single |

| | | | | |
|---|---|---|---|---|
| | | | | DNN for a classification task using high-resolution FFDM images and their labels. This model will not only classify breast cancer but will also be capable of providing visual explanation maps. Concurrently, another DNN will be trained for a segmentation task. Afterwards, models will be collaboratively trained with skillful data-sharing between them. An additional model, the discriminator, will be employed during the training. This model being concurrently trained with the aim to match the two generated masks to their source will refine the collaboration process. This combination is expected to lead to improved classification and segmentation outcomes. |
| O/DU3 | Algorithms for predicting the genetic profile of different neoplasms from microscopic imaging | • development of AI tools and solutions<br>• training of AI tools and solutions<br>• validation of AI tools and solutions<br>• development, validation or training of AI tools considered for medical devices | Fundacion De La Comunitat Valenciana Para La Gestion Del Instituto Deinvestigacion Sanitaria Y Biomedicade Alicante | Our main objective is to optimize our customized AI algorithms and deep learning procedures thanks to increasing the number of cases in the type of cancers in which we are already data holders: colorectal, lung and brain tumors, with the possibility to extend the study to other subtypes not fully contemplated in our own datasets (for example, pediatric glioblastomas, brain metastatic cancer).<br>We expect to improve the power of our own AI tools. This will speed up the implementation of AI-based services for diagnosis and prognosis of the referred cancers, incorporating the main advantages of AI:<br>reduction of time, error and costs of diagnosis, personalized medicine and tailored life planning for intractable cancers. |
| O/DU4 | Lung cancer detection from longitudinal LDCT and CT data (CONTACT) | • training of AI tools and solutions<br>• validation of AI tools and solutions | Fundacion Centro De Tecnologias De Interaccion Visual Y Comunicaciones | Lung cancer (LC) is the leading cause of cancer deaths worldwide. Early detection significantly reduces LC mortality by shifting diagnoses from late-stage, often incurable, to early- stage, which offers more curative treatment options, improves quality of life, and reduces economic impact. Currently, the main imaging modalities for managing LC are Computed Tomography (CT) and Low-Dose Computed Tomography (LDCT). CT identifies lung abnormalities and monitors treatment response, while LDCT is used for screening. Radiologists use the Lung CT Screening Reporting and Data System (Lung- RADS) to standardize the management of detected nodules. However, Lung-RADS involves image measurements that are not consistently and systematically performed, leading to variability in clinical practice. Image interpretation is also time-consuming and prone to errors. Research in computational radiology and computer-aided detection (CADe) and diagnosis (CADx) systems has been prominent in recent decades. Deep learning holds significant potential for automation and enhancement of AI-enabled image analysis pipelines, benefiting future screening programs, as we see in the LUCIA project, part of the Understanding Cancer cluster of projects funded by HEU under Cancer Mission. Vicomtech leads image analysis tasks within LUCIA, collaborating with CHUL (Centre Hospitalier Universitaire De Liege, Belgium), Osakidetza (Basque Health Service, Spain), and SAS (Servicio Andaluz de Salud, Spain). |
| O/DU5 | Validation of an AI algorithm to detect and classify pre-cancerous lesions in the pancreas | • development, validation or training of AI tools considered for medical devices | Sycai Technologies, S.L. | Cancer in upper-abdomen organs causes 1.4 million deaths worldwide every year. Pancreatic cancer is detected only in advanced stages, showing an extremely low survival rate of only 9%. Standard practice tries to identify tumors through diagnosis by medical imaging (computerized tomography scans, CT and magnetic resonance images, MRI). Previously to developing a malignant tumor, the organ can be affected by a focal lesion; this is an abnormal area or spot that can be identified on imaging tests. When focal lesions are identified in the early stages, the chances of successful treatment and positive outcome increases significantly. Remarkably, a non-negligible proportion of focal lesions are found incidentally in imaging scans, ranging between 35% and 65%, by radiologists. Some reasons are the millimetric size of lesions exceeding the limits of the human eye, lack of symptoms to request a CT/MRI test and increased workload in the radiology departments (2.5x images for the same staff). As mentioned above, Sycai Medical is a tool that address this critical issue: the lack of early-stage, affordable, patient-centric and automated cancer screening methods for the upper abdomen organs. The main objective of the project is to validate |

| | | | | |
|---|---|---|---|---|
| | | | | the tool we have developed and patented for the pancreas. Specifically, we aim to assess the scalability of the algorithm we have trained and ensure that it operates in an unbiased manner. This involves evaluating its performance on a previously unseen dataset to confirm its robustness and generalizability across diverse data samples. We expect the obtained metrics to be satisfactory. In case any issues are detected or if the algorithm appears to be biased, we will work on modifying and improving it accordingly. |
| O/DU6 | Deep learning-based detection and assessment of NSCLC tumors on chest CTs | • development of AI tools and solutions<br>• training of AI tools and solutions<br>• validation of AI tools and solutions | Universitätsklinikum Heidelberg | The currently available AI-based algorithms are limited to segmenting peripherally localized lung nodules and do not apply to patients with large tumor masses or tumors centrally localized around large blood vessels. However, patients with non-small cell lung cancer (NSCLC) are usually diagnosed at more advanced stages when the tumor is unresectable but is usually suitable for chemotherapy. Accurate segmentation of these advanced-stage tumors could allow automated extraction of radiomics features that correlate with genetic mutations, paving the way for radiomics-based "virtual biopsies". Moreover, AIbased segmentation can also play an important role in patient follow-up assessing response to chemotherapy.<br><br>The primary aim of our study was to train an AI algorithm to segment these advanced-stage NSCLCs. The proposed version of our AI tool can segment the primary lung tumor on contrast-enhanced CT scans with promising accuracy. The next phase would be to finetune and validate it on datasets in the EUCAIM database. Depending on the number and quality of CT scans and the genetic data provided by the data holders joining the EUCAIM project, our study also aims to extend our segmentation algorithm by automatically extracting radiomics features to identify imaging biomarkers of targeted genetic mutations<br><br>(e.g. EGFR, ALK). The identified predictive radiomics features would be used to build a machine learning model to predict the mutational status of tumors. A secondary aim of our project is to further extend our AI tool to segment target lung lesions to assess response to treatment at follow-up scans based on tumor volume changes. |
| O/DU7 | Enhancing Breast Cancer Diagnosis with AI: Deep Learning-Based Detection and Discrimination of Mammographic Findings | • training of AI tools and solutions<br>• validation of AI tools and solutions<br>• development, validation or training of AI tools considered for medical devices | Istituto Europeo di Oncologia | We developed a deep learning-based tool aimed at improving the detection and discrimination of breast microcalcifications on mammography. This tool is designed to enhance the accuracy and efficiency of breast cancer diagnosis, with plans to extend its application to other mammographic suspicious findings such as radiopacities and distortions. Additionally, the tool will be adapted for use with tomosynthesis, providing enhanced 3D imaging capabilities beyond traditional 2D mammographic projections.<br><br>Main Objectives:<br><br>• Validate AI Models: Conduct rigorous validation using diverse datasets to ensure robustness and reliability.<br>• Enhance Detection Accuracy: Improve sensitivity and specificity in detecting breast microcalcifications and other mammographic findings. |
| O/DU8 | Diagnostics in gynaecological lung and collorectal cancers | • development of AI tools and solutions<br>• training of AI tools and solutions<br>• validation of AI tools and solutions | National Hellenic Research Foundation | Major sources of clinical and imaging data that could serve as input information for the development of decision-making tools are ultrasound images of either antenatal pregnancy surveillance or gynecological pathologies. We might estimate that over 3,000 ultrasound obstetric images and over 1,000 gynecological ultrasound images might be available on an annual basis. Furthermore, continuous performance of laparoscopy as surgical approach of various gynecological pathologies might avail approximately 100 recorded video procedures of at least 1 hour duration, including at least 50 of gynecological malignancy. Furthermore, stored epidemiological data of antenatal obstetrical screening (approximately 1,000 cases on annual basis), as well as clinical and histopathological data of endometrial, cervical and ovarian cancer |

| | | | | |
|---|---|---|---|---|
| | | | | (approximately over 500 cases for last 3 years) might also serve on the level of input data as primary sources of decision-making tools.<br><br>Regarding ultrasound images of antenatal surveillance, they might avail data on two major domains, the first concerning screening of congenital abnormalities and chromosomal deficiencies and the second concerning growth of fetus. Specifically, nuchal translucency, Doppler measurement of uterine arteries, detection of echogenic biomarkers such as echogenic bowel, single umbilical artery and short femural length are considered the main representative markers of congenital and chromosomal abnormalities. Regarding surveillance of fetal growth, measurement of head circumference, abdominal circumference and femural length along with Doppler examination of umbilical and middle cerebral artery are the most representative ones.<br><br>Regarding gynecological ultrasound images, they rather adhere to the IOTA and IETA criteria developed by the relative international Consortium. Biomarkers related with U/S images might be measurement of CA 125 and potentially CA 19-9, while standard evaluation of ultrasound images especially for ovarian masses included measurement of tumor length, detection of solid or multiform part inside the tumor, presence of acoustic shadow, papillary injection and presence of ascites. Evaluation for endometrial tumors relates with endometrial thickness and pattern of echogenicity, presence and intensity of vascularization based on Color Doppler, position of uterine masses based on FIGO Classification, presence of cystic parts, in an effort to identify images suspicious for endometrial cancer, myomas or sarcomas.<br><br>Finally, regarding the stored data of cancer patients, these might be divided into three main domains. The first concerns epidemiological data regarding age, comorbidities, obstetrical and gynecological history. The second concerns histopathological data, namely histopathological type, grade, presence of LVSI in endometrial cancer patients, nodal status based on final surgical staging as well as potential expression of POLE, MMR and p53 mutations in EC patients. The third domain of data concerns prognostic outcomes of patients, namely disease-free survival, overall survival, recurrence of tumor, Kind of recurrence and treatment of recurrence. All relative data are consistently registered in ESGO related databases in a continuous effort of Clinical audit as well as development of clinical and scientific outcomes.<br><br>On the molecular level, we will use whole exome and next generation sequencing data as well as data on mRNA expression (transcriptomics) and metabolomics from melanoma, lung, breast and colorectal cancer patients (n=200) from the Athens Comprehensive Cancer Center and its collaborating hospitals and cancer treatment centers running since 2017 in close collaboration with its German counterpart in Heidelberg running from the German Research Center on Cancer (DKFZ). Enhanced multi-omics bioinformatics algorithms supported by machine- and deep-learning approaches developed provide an extended dataset that has already been used to identify potential biomarkers of effect that could be used as early prognostic signals for disease onset and development and in terms of successful treatment and/or remission.<br><br>We consider that all relative data and biomarkers and validation throughout decision-making tools will contribute in daily clinical practice, with new sophisticated individualized algorithms of diagnosis and treatment. |
| O/DU9 | eXplainable Artificial Intelligence for Breast Cancer -XAI_Breast_Cancer | • development, validation or training of AI tools considered for medical devices | Università degli Studi di Bari Aldo Moro | XAI_Breast_Cancer will provide the technological enablers for experimentation of AI-based solutions to improve prediction, diagnosis and contributing to a more precise and personalized management of cancer. |

| | | | | |
|---|---|---|---|---|
| | | | | In particular augmenting the interpretability of Machine Learning approaches making predictions to be sufficiently understandable or interpretable to humans. This project will provide the users explanations<br><br>for patient-specific predictions as well as to peer into a model and understand how predictions are made. XA_Breast_Cancer will grade general views of which prediction features are essential to assign a patient to a particular clinical outcome providing other more detailed and graphical depictions of evidence relationships underpinning individual predictions. Furthermore, XAI_Breast_Cancer will increase the acceptability of Machine/Deep Learning (ML/DL) in the clinic, and it will support the clinicians with the generation of new hypotheses and in understanding the mechanisms underlying particular pathological states for a better decision making. |
| O/DU10 | Developing multi cancer AI | • development of AI tools and solutions<br>• training of AI tools and solutions<br>• validation of AI tools and solutions<br>• development, validation or training of AI tools considered for medical devices | Better Medicine OU | We are doing a full-body solution to speed up oncology workflow in radiology. For that we are building automated AI models that can detect/classify and measure lesions in all organs of the abdominal cavity. We have already achieved substantial progress with models for kidney and lung, have PoC models working for liver and pancreas, while models for lymph nodes and bones are next in line. This sort of work requires rich datasets of CT scans with lesions located in different organs. |

# AI4HI Use Cases

The EUCAIM platform was built on the "AI for Health Imaging" Network (AI4HI), a cluster consisting of six largest EU-funded projects (CHAIMELEON, EuCanImage, INCISIVE, ProCAncer-I, PRIMAGE and RadioVal). The (AI4HI) projects also reported many use cases or pilot studies aimed at validating and determining the scientific utility of the platforms developed within the above-mentioned use cases. All these studies can also successfully serve to validate the EUCAIM platform, which is the continuation of the use cases implemented within AI4HI. The following list details all use cases and pilot studies that have been implemented as part of AI4HI.

## CHAIMELEON

The CHAIMELEON project is focused on creating a well-organized repository that brings together high-quality medical imaging and related clinical data for Europe's most common cancers: lung, breast, prostate, and colorectal. This repository, designed to be interoperable across the European Union, is intended to significantly support the advancement and validation of artificial intelligence tools aimed at enhancing cancer care, while promoting data harmonization and reproducibility across centres. The main goals of the CHAIMELEON project include:

- Ensuring legally and ethically compliant access to extensive medical datasets.
- Building a pan-European, interoperable platform with high-quality imaging data for AI development and validation in cancer treatment.
- Leveraging existing infrastructures to create a distributed data-sharing network.
- Investigating innovative data harmonization techniques and offering online tools for image harmonization.
- developing online workflows to improve the reliability and clarity of AI applications.
- Conducting internal and external evaluations of the repository's performance.
- Testing AI-driven tools in real-world clinical settings from an early stage.
- Promoting the long-term sustainability of the platform and fostering a broad, engaged user community.

In order to test the usability and validation of the repository created within the project, the use cases presented in the table below were performed.

Table 7. Use cases implemented within the CHAIMELEON project and their descriptions.

| No. | Use Case title | Description |
|---|---|---|
| 1 | Prostate cancer: Patient risk (Low or High risk) prediction at baseline | To identify imaging features that distinguish low-grade from high-grade cancers to minimize unnecessary aggressive treatments in low-grade cases and guide optimal advanced therapy for high-grade tumors |
| 2 | Lung cancer: Patient Overall Survival (OS) prediction at baseline | To evaluate baseline imaging and radiomic features in a diverse cohort of non-small cell lung cancer (NSCLC) patients undergoing immunotherapy, and correlate these features with progression-free survival |
| 3 | Breast cancer: Histology subtype prediction | To determine the optimal combination of multimodal imaging features that enables accurate diagnosis of breast tumors |
| 4 | Colon cancer: Pathological TNM prediction | To identify imaging features on baseline CT scans that can predict the likelihood of disease progression. |
| 5 | Rectal cancer: Extramural Vascular Invasion and Mesorectal Fascia Invasion (imaging findings) prediction | To assess baseline MRI features to predict treatment response and tumor regression grade (TRG). |

## EuCanImage

EuCanImage aims to develop and demonstrate a scalable, GDPR-compliant platform that enables the use of extensive, high-quality, and interoperable cancer imaging datasets, effectively connected with relevant biological and clinical cancer data. The platform incorporates cutting-edge tools and emerging standards to support the creation and validation of integrative decision-support systems for precision oncology, promoting greater clinical confidence and uptake. The EuCanImage project pursues the following core objectives:

- Creating a FAIR-compliant imaging platform ensuring cancer imaging data is Findable, Accessible, Interoperable, and Reusable, and integrated with biological and health repositories for AI in clinical oncology.
- Offering robust tools and training providing user-friendly resources for data curation, annotation, and management to support future data contributions and platform scalability.
- Developing a multi-centre AI development environment leveraging the consortium's expertise in radiomics, distributed learning, and explainable AI to advance cancer imaging solutions.
- Establishing a comprehensive AI evaluation framework supporting multidisciplinary benchmarking and assessment of imaging-based AI tools for oncology care.
- Implementing legal and ethical frameworks enabling responsible data sharing and fostering Open Science across EuCanImage and the broader cancer research community.
- Addressing critical clinical needs advancing personalized cancer care by targeting unmet challenges through AI-driven insights.
- Engaging the research and development community expanding adoption by building a robust network of data contributors and AI developers via the consortium’s outreach efforts.

The use cases implemented within the EuCanImage project are listed in the table below.

Table 8. Use cases implemented within the EuCanImage project and their descriptions.

| No. | Use Case title | Description |
|---|---|---|
| 1 | Can AI increase the diagnostic sensitivity of liver MRI, for detection of small hepatocellular carcinoma lesions with kept high specificity? | To enhance liver MRI sensitivity for detecting small hepatocellular carcinoma lesions while maintaining high specificity, using a dataset of 1,283 contrast-enhanced MRI exams, including benign and malignant cases |
| 2 | Can AI identify liver metastases in colorectal cancer from pre- and post-operative CT? | To detect liver metastases from colorectal cancer on pre- and post-operative CT, using 3,057 scans including metastatic, other hepatic lesions, and normal cases across multiple contrast phases |
| 3 | Can AI identify mesorectal lymph node metastases in pelvic MRI? | To identify mesorectal lymph node metastases in pre-treatment pelvic MRI of rectal cancer patients, using 992 annotated exams with T2-weighted and DWI sequences |
| 4 | Can AI predict the level of response to neoadjuvant radio(chemo)therapy based on primary MRI in rectal cancer for local staging and restaging? | To predict response to neoadjuvant radio(chemo)therapy in rectal cancer from baseline MRI, using 992 pre-treatment scans with detailed clinical data and balanced response categories |
| 5 | Can AI distinguish five molecular subtypes of invasive ductal breast carcinoma on mammograms? | To classify five molecular subtypes of invasive ductal breast carcinoma using 7,357 mammograms from diagnosed patients, with CC and MLO views and detailed clinical metadata from both screening and symptomatic cohorts |
| 6 | Could AI tools enable de-escalate neoadjuvant systemic therapy (NST) in patients highly likely to achieve a pathological complete response (pCR)? | To predict pathological complete response (pCR) to neoadjuvant systemic therapy in breast cancer, using 1,581 diffusion contrast-enhanced MRIs and detailed clinical data, aiming to enable therapy de-escalation in likely responders |
| 7 | Can AI improve the assessment of screening and non-screening mammograms by automatically differentiating benign from malignant tumours? | To differentiate benign, malignant, and normal findings in 11,693 mammograms from both screening and clinical settings, using a balanced dataset with CC and MLO views to enhance diagnostic accuracy |

## INCISIVE

The project's main goal is to design and validate an AI-powered toolbox that improves the precision, sensitivity, specificity, interpretability, and cost-efficiency of current cancer imaging techniques. Additionally, INCISIVE will incorporate an automated machine learning (ML)-driven annotation system and create a pan-European, interoperable federated repository. This repository will securely host clinical data and medical images, supporting data donation and sharing in full compliance with ethical, legal, and privacy standards. The objective is to enhance data accessibility and streamline experimentation with AI-based tools, thereby promoting their broader application in cancer detection, prognosis and monitoring. Throughout the project, INCISIVE will utilize a variety of data types, including imaging data, biological data, and Electronic Health Records (EHRs).

This project is based on the following five pillars:

- Analysis of AI challenges related to cancer imaging while striving to achieve a highly acceptable solution.
- Incorporation of AI features enhancing cancer imaging and enabling effective decision making.
- Implementation of an interoperable pan-European federated repository of clinical data and medical images, including secure data sharing mechanisms.
- System validations, technology assessment and proof-of-concept demonstration.
- Active engagement of stakeholders, improving user acceptance and resulting in measurable impacts.

Table 9. Use cases implemented within the INCISIVE project and their descriptions.

| No. | Use Case Title | Description |
|---|---|---|
| 1 | INCISIVE Lung Cancer AI Services | To improve lung cancer detection, staging, and risk assessment by leveraging X-rays and CT scans through AI pipelines, including explainable classification, lesion segmentation, and metastasis risk prediction, using annotated imaging data. |
| 2 | INCISIVE Colorectal Cancer AI Services | To support colorectal cancer diagnosis and prognosis by applying AI models to MRI and histopathology images for lesion and cell segmentation, patient prioritization, localization assistance, and survival prediction, using multimodal imaging data. |
| 3 | INCISIVE Breast Cancer AI Services | To support early detection and diagnostic precision in breast cancer by applying AI models to mammography and MRI scans, offering prioritization, lesion localization, segmentation, Breast Imaging-Reporting and Data System (BIRADS) and breast density classification with explainability, based on annotated imaging data. |
| 4 | INCISIVE Prostate Cancer AI Services | To improve prostate cancer diagnosis and severity assessment using MRI-based AI services for prostate and lesion segmentation, patient prioritization, International Society of Urological Pathology (ISUP) score classification with clinical data integration and explainability, delivered as a diagnostic pipeline. |

## ProCAncer-I

Prostate cancer (PCa) is the second most common cancer and the third leading cause of cancer death among men in Europe. Current diagnostic practices often result in overdiagnosis and overtreatment of indolent tumours, highlighting the need for advanced AI tools that can identify subtle imaging patterns to better distinguish between indolent and aggressive disease, predict recurrence,

detect metastases, and assess therapy effectiveness. However, existing efforts are fragmented and rely on limited, non-generalizable datasets. The ProCAncer-I project brings together 20 leading institutions, AI experts, and SMEs to create a secure, cloud-based European infrastructure for PCa imaging data and AI development. The platform is host to the world's largest collection of anonymised multi-parametric MRI data (>14,000 cases), and supports robust, vendor-neutral AI models across nine clinical scenarios. The project emphasises fairness, safety, explainability and reproducibility, while also working closely with regulators to define a certification roadmap, ensuring clinical trust and paving the way for the adoption of AI in PCa care.

Table 10. Use cases implemented within the PROCANCERI project and their descriptions.

| No. | Use Case Title | Description |
|---|---|---|
| 1 | Use Case 1 | Detection of prostate cancer with high accuracy both in peripheral and transitional zones to identify which men have cancer and those with no cancer. From a clinical point of view UC1 will help stratifying men with prostate hypertrophy or inflammation despite the high PSA values (>4 ng/ml) and those who should undergo additional diagnostic tests (e.g. biopsy) to identify if suspicious lesions identified on MRI are clinically significant or if there is an indolent disease with no harm for the patient |
| 2 | Use Case 2 | Characterization of cancer according to its biological aggressiveness into clinically significant and non-significant disease. UC2 aims to stratify men with suspicious findings on MRI into high-risk cases, which need radical treatments to ensure that cancer will not grow and spread to remote parts of the body becoming a deadly disease, from low-risk cases which could be safely followed-up with active surveillance, avoiding comorbidities of treatment and ensuring the highest possible quality of life for patients |
| 3 | Use Case 3 | Identification of patients with metastatic prostate cancer as early as possible among cases with high-risk PCa. Clinically, UC3 will help to adjust treatment strategies and mitigate metastatic spread that will finally kill the patient, as well as adjust follow up frequency to patients with high metastatic risk. AI models will provide early indications whether a patient belongs to the metastatic subtype that needs a different therapeutic approach, mining tumor characteristics that probably are related to its biological differences |
| 4 | Use Case 4 | Radiologic – Histopathologic correlation to provide biology-based validation of AI models to compare side by side pathologic data with AI results to improve understanding of the features that AI models are making use to reach specific decisions. Moreover, UC4 will help correlating imaging phenotype derived from MRI to microscopic findings from pathology and predicting cancer presence and/or its biology characteristics from radiologic imaging |
| 5 | Use Case 5 | Prediction of the risk of disease recurrence after radical prostatectomy, based on imaging data and AI techniques. In UC5 post-surgery findings, such as positive surgical margins and extracapsular extension, will be considered in a nomogram comprising also radiomics and clinical variables to predict disease recurrence. UC5 will help clinicians to choose between different treatment techniques (conventional, nerve sparing, laparoscopic, robot-assisted radical prostatectomy), tailoring treatment to the predicted risk of disease recurrence |
| 6 | Use Case 6 | Prediction of treatment response in case of radiation therapy, assessing the risk of disease recurrence to promptly adjust therapeutic strategy at an early stage and avoid patient discomfort and non-optimal distribution of medical resources. UC6 is similar to UC5, but refers to radiotherapy recurrence and it will help radiation oncologist to tailor treatments |

| | | |
|---|---|---|
| 7 | Use Case 7 | Prediction of post radical prostatectomy and/or radiation-induced urinary toxicity, in order to consider additional or alternative measures to alleviate therapy-induced undesired effects. Using UC7 results, patients with high risk for toxicity can be thoroughly informed on the side effects and alternative possibilities for therapy. This could help balancing the risk-benefit ratio related to whole gland treatments, in particular in patients with no life-threatening PCa. UC7 will take into consideration urinary incontinence, irritative/obstructive bowel, sexual/erectile dysfunction, and hormonal domains |
| 8 | Use Case 8 | AI-powered patient stratification for enrolment in Active Surveillance programs, to develop a more efficient patient stratification program based on AI decision-making from MRI lesion phenotype. The risk of disease progression in patients who are undergoing active surveillance will be assessed with longitudinal MRI data (combined with biopsy) to reach specific clinical indications (either repeat PSA test, MRI, biopsy or a combination of them). UC8 aims also to predict the time-to-progress to provide a follow-up strategy and stratify patients in those who could safely remain in the active surveillance group and those who will ultimately need treatment |
| 9 | Use Case 9 | Prediction of the best option for patients needing treatment ensuring the lowest possible side effects/toxicity. Results from all previous Use Cases are expected to merge into a holistic model suggesting presence/non-presence of PCa, stratification into clinically significant/insignificant cases and a decision support system suggesting the best treatment option (radical prostatectomy, radiation therapy, active surveillance), considering also the lowest toxicity/side effects to ensure the best possible quality of life |

## PRIMAGE

The PRIMAGE project focuses on developing advanced computational tools for analyzing medical images to support the clinical management of childhood cancers. The objective of the project is to build a cloud-based platform that offers AI-powered predictive tools to aid diagnosis, prognosis, therapy selection, and treatment monitoring. These tools will use novel imaging biomarkers, in-silico tumour growth models, and visualisations with confidence scores. The platform is being validated for two high-impact pediatric cancers: Neuroblastoma (NB) and Diffuse Intrinsic Pontine Glioma (DIPG). PRIMAGE brings together key partners, including the European Society for Paediatric Oncology, major imaging biobanks, and leading pediatric oncology centers, providing access to rich retrospective datasets (imaging, clinical, molecular, and genetic) for training and testing. Robust solutions for data pseudonymisation, extraction, quality control, and secure storage are also being developed, with applicability to prospective data. The end goal is a validated cloud-based prototype that improves clinical decision-making in pediatric oncology through predictive imaging and simulation tools.

Table 11. Use cases implemented within the PRIMAGE project and their descriptions.

| No. | Use Case title | Description |
|---|---|---|
| 1 | Overall survival for Neuroblastoma | Overall Survival for Neuroblastoma: Predict and evaluate the overall survival probability of pediatric patients diagnosed with neuroblastoma using multi-omics, imaging, and clinical data. |
| 2 | Event Free Survival for Neuroblastoma | Event-Free Survival for Neuroblastoma: Estimate the probability of event-free survival (absence of relapse, progression, or death) in neuroblastoma patients based on predictive modeling. |
| 3 | Overall survival for DIPG | Overall Survival for DIPG: Assess the overall survival chances for children affected by Diffuse Intrinsic Pontine Glioma (DIPG), integrating radiomic features and clinical outcomes. |
| 4 | Event Free Survival for DIPG | Event-Free Survival for DIPG: Predict event-free survival for DIPG patients by identifying early indicators of progression or relapse using imaging and clinical follow-up data. |

## RadioVal

RadioVal is the first international clinical study focused on validating radiomics-based models for predicting response to neoadjuvant chemotherapy in breast cancer using MRI. The project's objective is to establish a standardised, comprehensive framework for evaluating radiomics tools, guided by the FUTURE-AI principles of Fairness, Universality, Traceability, Usability, Robustness, and Explainability. Furthermore, RadioVal will develop new mechanisms for transparent, ongoing assessment and monitoring of these tools. The study employs a multi-stakeholder approach, integrating clinical, healthcare, ethical, and regulatory perspectives from the outset. The main objectives of the projects are listed below:

- Implement the very first international, multi-faceted clinical validation study for radiomics-based prediction of response to neoadjuvant therapy in multiple developed and developing countries.
- Introduce a holistic, standardised methodological framework for multi-faceted and trustworthy evaluation of radiomics AI, taking into account multiple technical, clinical as well as ethical criteria.
- Implement a multi-stakeholder, inclusive approach to improve awareness, acceptance and promotion of radiomics AI in future breast cancer care.
- Develop the very first traceability tool for radiomics AI, which will enable transparent monitoring and continuous evaluation of radiomics tools during their lifetime.
- Evaluate wider impacts of clinical deployment of radiomics AI, including associated cost-benefits, socio-ethical implications and regulatory aspects.

Table 12. Use cases implemented within the RadioVal project and their descriptions.

| No. | Use Case title | Description |
|---|---|---|
| 1 | Prediction of response to NAC Outcome | To predict response to neoadjuvant chemotherapy (NAC), The use case classifies outcomes into complete response, partial response, stable disease, or progressive disease using multiclass or binary classification methods. |
| 2 | Deviations from original treatment | To predict deviations from the original treatment plan, UC2 uses binary classification with clinical non-imaging variables as ground truth to determine a Yes/No outcome. |
| 3 | Prediction of 3-yr overall survival or Prediction of overall survival | To predict overall survival or 3-year survival. The use case uses binary classification for Yes/No outcomes and regression analysis with censoring to estimate survival probability over time. |
| 4 | Automated lesion segmentations | To differentiate lesion from background on a per-voxel basis, this task involves segmentation using overlapping, volume, and distance-based metrics. |

# Examples of the proposed implementation of the data use cases

The following section provides examples of the proposed implementation of the several aforementioned use cases. The presented examples describe preparation of the data and the subsequent steps that are taken to share data within data holder use cases. These are the most complex examples and require the most steps. For instance, in order to make data available at the Tier 3 level, it is necessary to adapt the metadata to the EUCAIM Common Data Model (EUCAIM CDM) and structure it according to the hyperontological model. This is merely the initial step in a broader process. In the subsequent steps, a range of tools will be employed to guarantee the highest data quality and alignment with the FAIR principles. In the case of federated data sharing, it is essential to deploy a local federated node and integrate the data at the federated search level. For this reason, this chapter focuses on use cases related to data sharing. The framework for implementing use cases is outlined in a separate document[1].

## Use Case example 1

**Title:** Dataset Lung cancer - CT imaging and clinical data for treatment optimization

**Organization:** Servicio Andaluz De Salud

### Dataset description

The dataset employed in this use case comprises retrospectively collected longitudinal CT imaging studies and clinical data of lung cancer patients treated with radiotherapy. It includes diagnostic and treatment follow-up CT scans, as well as comprehensive clinical information such as pathology and radiology reports, treatment details, and patient demographics. The dataset supports the development of AI methods to differentiate between NSCLC and SCLC and to predict treatment response, contributing to personalized lung cancer care.

### Intended Purpose

[1] https://docs.google.com/document/d/1Yppj9hubJ80cELR3A92k3_EOdrivSuZi68OQ1_yaHfM/edit?tab=t.0

The main objective of this dataset is the distinction between small cell lung cancer (SCLC) and non-small cell lung cancer (NSCLC), as well as the prediction of patient response to radiation therapy, both to optimize lung cancer treatment planning.

**Basic dataset characteristic**

Imaging modality: Computed Tomography (DICOM tag (0008, 0060) = CT)

Imaging body part: Chest, Lung, Abdomen, Pelvis

Age range: 40 – 90 years

Sex: Male (80%), Female (20%)

Number of subjects: 600

Number of DICOM studies: 3,600

De-identification: Personal data is fully anonymized

Declared tier: 3C+

Detailed information on the dataset's metadata will be available at: https://docs.google.com/spreadsheets/d/1660GVsvUPsHncSCBiih59VvKpnblOV9L/edit?gid=2023442309#gid=2023442309

**Dataset structure**

The shared data adheres to the EUCAIM CDM standards and is structured to facilitate harmonization and advanced querying. The data structure is presented in the following link: https://docs.google.com/spreadsheets/d/12PE6kXgj39Skm_GOJff5rpvsuez7ER3Z/edit?gid=705976904#gid=705976904

**The tools that will be utilized**

- DIQCT tool
- DICOM file integrity checker
- DICOM tag extractor
- ETL
- Wizard tool

**The data sharing process flow**

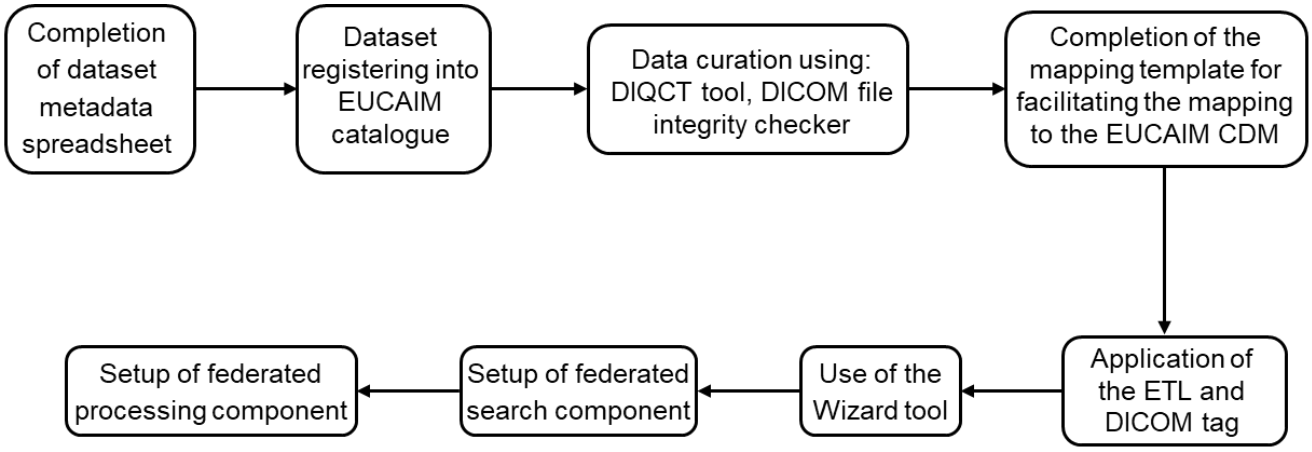


Fig 1. Steps taken to share data by Servicio Andaluz De Salud at the tier 3C+ compliance level.

## Use Case example 2

**Title:** Sarcomas

**Organisation's name:** University of Athens

### Description

This imaging dataset will contain a collection of patients with sarcomas (mainly soft tissue sarcomas). The dataset comprises pre-radiotherapy images (CT and/or MRI) with tumor segmentation on the CT images. Post-radiotherapy MRI scans (T1w and T2w images) performed 4-6 weeks after a 25x2 Gy irradiation scheme may also be provided, if available. A few cases of local currency, following irradiation and surgery, may also be included.

### Intended Purpose

This dataset may serve the identification of imaging biomarkers (in the post-irradiation dataset) for the prognosis of local recurrence. In addition, the pre-irradiation imaging dataset can be used for the extraction of radiomic features which may facilitate the differentiation between different sarcomas and their radiosensitivity.

### Basic dataset characteristic

Imaging modality: Computed Tomography, Magnetic Resonance Imaging

Imaging body part: Head and neck, extremities, trunk, abdomen, pelvis, long bones

Age range: 18 – 90 years

Sex: Male, Female

Number of subjects: up to 80

Number of DICOM studies: up to 160

De-identification: Anonymisation will be performed using the EUCAIM anonymisation tool

Declared tier: 2

Detailed information on the dataset's metadata will be available at: https://docs.google.com/spreadsheets/d/15mf0PQGc7AdCoWbO0cm6Sz9J4R80257e/edit?gid=1904417992#gid=1904417992

**Dataset structure**

The structure of metadata is described in the excel file available at the link: https://docs.google.com/spreadsheets/d/1mIIFgVzX_8c1K19RxWeFtLwEOaNT0QJJ/edit?usp=drive_link&ouid=101508161895429784085&rtpof=true&sd=true

**The tools that will be utilized**

- EUCAIM anonymization tool
- DICOM file integrity checker
- DIQCT tool
- DICOM tag extractor
- ETL
- Wizard tool

**The data sharing process flow**

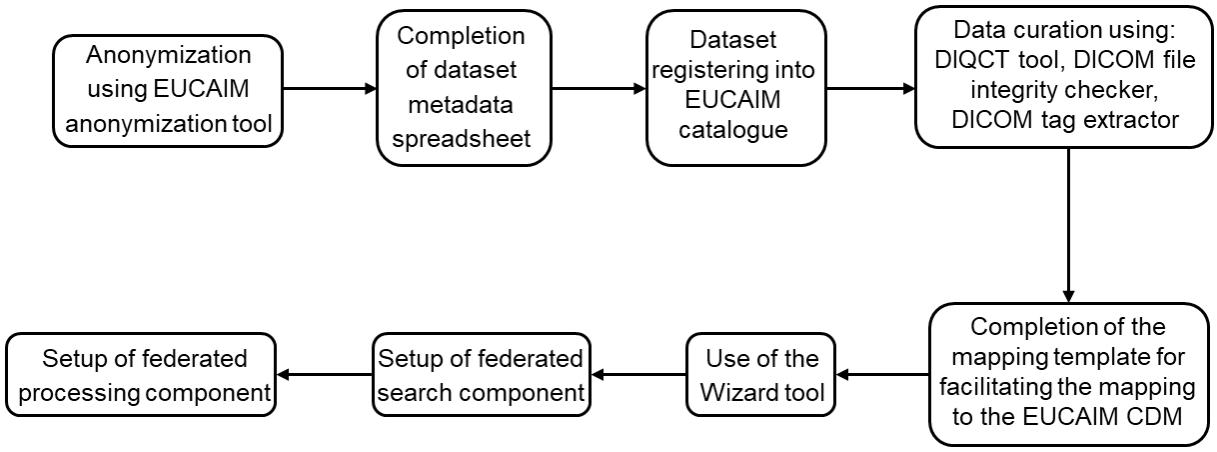


Fig 2. Steps taken to share data by University of Athens at the tier 2 compliance level.

## Use Case example 3

**Title:** Sciensano imaging dataset

**Organisation's name:** Sciensano

**Description**

Clinical, genomic and imaging data of a collection of patients who have undergone NGS test on tissue biopsies.

**Intended Purpose**

The primary objective of the GeNeo study is to evaluate the added value of the CGP over targeted NGS test using smaller gene panels.

**Basic dataset characteristic**

Imaging modality: digitized hematoxylin-eosin (HE)-slides

Imaging body part: Various

Age range: > 18 years

Sex: Male, Female

Actual number of subjects: 475

Number of DICOM studies: unspecified

De-identification: Personal data is pseudonymized

Declared tier: 1

Detailed information on the dataset's metadata will be available at: https://docs.google.com/spreadsheets/d/1B9LAHBTy8u_iQ3X-QgvfYzvsgPe8UeNY/edit?gid=472806603#gid=472806603

**Dataset structure**

The data is Tier 1 compliant and does not adhere to the EUCAIM CDM or Hyperontology standards.

## Use Case example 4

**Title:** VAIB(Validation of AI for Breast imaging) Core-dataset

**Organisation's name:** Karolinska Institutet

**Description**

VAIB core-dataset consists of cancer output, radiologist assessment and also clinical information from 3 different regions of Sweden. In addition, VAIB core-dataset is targeting to validate AI for Breast imaging in various perspectives.

**Intended Purpose**

The primary purpose of the VAIB core dataset is to validate AI models for breast imaging systems by providing globally assessed data enriched with diverse clinical information. This enables evaluation of the models' robustness and reliability across varied clinical scenarios.

**Basic dataset characteristic**

Imaging modality: Mammography

Imaging body part: Breast

Age range: ~ 57 years

Sex: Female

Actual number of subjects: 15,564

Number of DICOM studies: ~ 17,131

De-identification: Personal data is fully anonymized

Declared tier: 2

Detailed information on the dataset's metadata will be available at: https://docs.google.com/spreadsheets/d/1HJDFzF2cQf6vIXhAnDlpIp3nw7AH0mdU/edit?gid=2023442309#gid=2023442309

**Dataset structure**

The structure of metadata is described in the excel file available at the link: https://docs.google.com/spreadsheets/d/1VwUKMOg6dgkEiM2jOv1Z3AWq9Zok2wvk/edit?gid=357336097#gid=357336097

**The tools that will be utilized**

Trace4Medical Image tool

DIQCT tool

DICOM file integrity checker

DICOM tag extractor

ETL

Wizard tool

**The data sharing process flow**

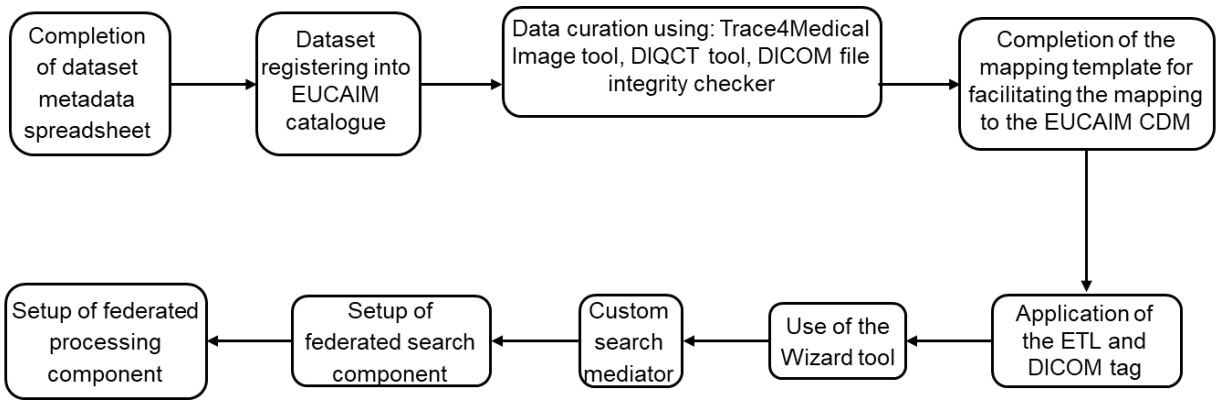


Fig 3. Steps taken to share data by Karolinska Institutet at the tier 2 compliance level.

## Use Case example 5

**Title:** PAIMRI dataset - prostate cancer detection

**Organization's name:** Assistance Publique Hôpitaux De Paris

**Description**

PAIMRI dataset consists of 300 prostate MR (DICOM) with segmentation and label as well as clinical and histological data (including a subset of pathological slides).

**Intended Purpose**

The primary objective of this use case is to develop, train and validate AI tools for the detection of prostate cancer.

**Basic dataset characteristic**

Imaging modality: Magnetic Resonance

Imaging body part: prostate

Age range: > 40 years

Sex: Males

Actual number of subjects: 300

Number of DICOM studies: 300

De-identification: Personal data is pseudonymised

Declared tier: 3

Detailed information on the dataset's metadata will be available at: https://docs.google.com/spreadsheets/d/10dDFiSIGOm-6dY0Mv9xkj8JftM5V4kWy/edit?gid=2023442309#gid=2023442309

**Dataset structure**

The structure of metadata is described in the excel file available at the link: https://docs.google.com/spreadsheets/d/1uVIHIIrVrCVME5bNJQzLDlS1Ibe4NfzS/edit?gid=357336097#gid=357336097

**The tools that will be utilized**

EUCAIM anonymizer

DIQCT tool

DICOM tag extractor

DICOM file integrity checker

ETL

Wizard tool

**The data sharing process flow**

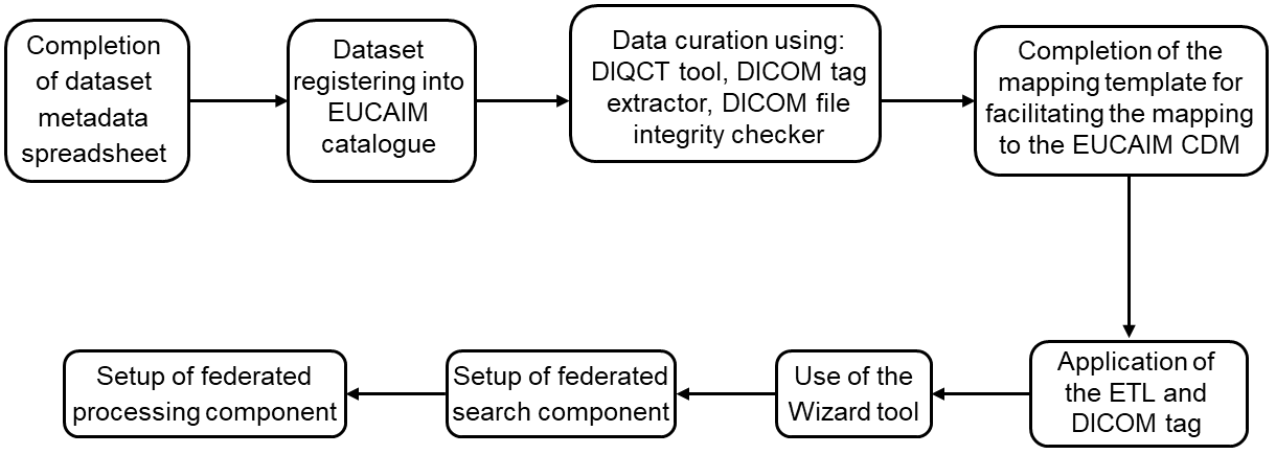


Fig 4. Steps taken to share data by Assistance Publique Hôpitaux De Paris at the tier 3 compliance level.

## Use Case example 6

**Title:** Thyroid Scintigraphy

**Organization's name:** Aristotelio Panepistimio Thessalonikis

### Description

This dataset includes retrospective scintigraphy imaging and clinical data from three key stages of thyroid cancer treatment, offering a comprehensive view of patient response. The imaging tracks residual thyroid tissue and metastases both before and after I-131 therapy, with follow-up scans assessing long-term outcomes. Clinical data, including demographics, medical history, histological details, TNM classification and treatment timelines, complements the imaging insights. By combining these data, we can develop AI models that predict therapy success, support personalized treatment planning, and address clinical questions related to diagnosis and prognosis. This ultimately improves patient care and treatment effectiveness.

### Intended Purpose

The primary purpose of this dataset is to support the development, validation, and evaluation of AI models aimed at improving the diagnosis, prognosis, and treatment planning for thyroid cancer patients undergoing I-131 radioiodine therapy.

### Basic dataset characteristic

Imaging modality: Scintigraphy

Imaging body part: Whole-body scans specifically focused on examining the thyroid

Age range: over 18

Sex: Males and Females

Actual number of subjects: 80 - 100

Number of DICOM studies: N/A

De-identification: Personal data is pseudonymized

Declared tier: 3

Detailed information on the dataset's metadata will be available at: https://docs.google.com/spreadsheets/d/1ahnAvh_quAswku08ubKnyGFUFulP5kFj/edit?gid=1572455913#gid=1572455913

**Dataset structure**

The structure of metadata is described in the excel file available at the link: https://docs.google.com/spreadsheets/d/1pJedcuGB9HJSdiJI_YylPaOsNedzG0on/edit?gid=357336097#gid=357336097

**The tools that will be utilized**

EUCAIM anonymizer

DIQCT tool

DICOM tag extractor

DICOM file integrity checker

ETL

Wizard tool

**The data sharing process flow**

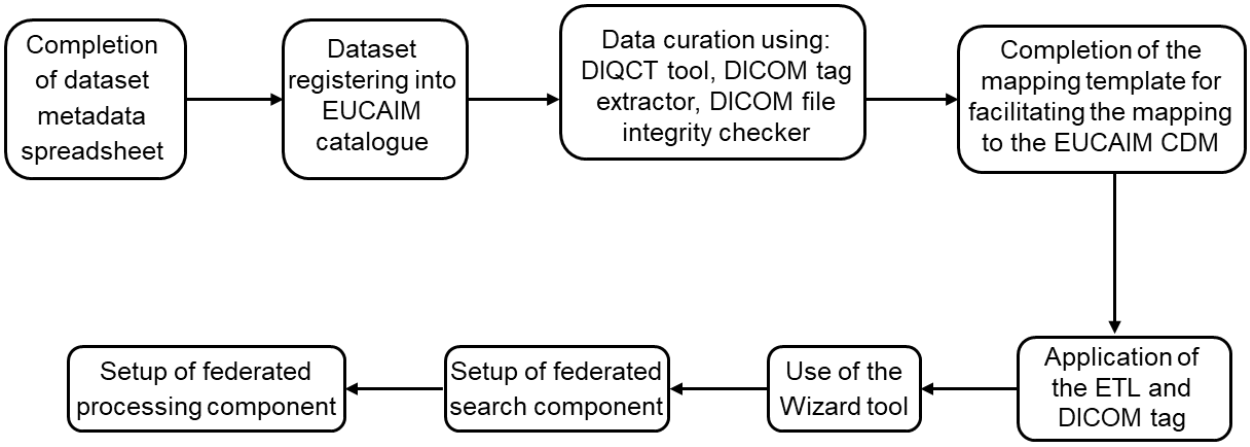


Fig 5. Steps taken to share data by Aristotelio Panepistimio Thessalonikis at the tier 3 compliance level.

# Summary and discussion

The wealth and variety of the submitted projects emphasises the necessity for a platform such as EUCAIM. The implementation of carefully selected pilot use cases demonstrates the complexity of projects submitted under the internal and external open calls, which are so diverse in nature. The implementation process of the pilots revealed several issues and challenges that needed to be addressed. These findings will contribute to the establishment of transparent and unambiguous procedures for implementation of future projects. Furthermore, this will result in the increased automation of services for future users of the platform. The elements that presented a challenge are outlined below:

1. Accurate mapping of local information to the hyper-ontology and CDM. Some ambiguous non-standardized information requires the intervention of experts to ensure a clear and consistent interpretation and classification of data.
2. Minor issues regarding software licensing and installation.
3. Several steps that had to be performed manually were quite time-consuming (e.g. related to the completion of mapping templates, standardization and classification of some ambiguous or incomplete data).

The piloting of the onboarding process, with the participation of Data Holders (DHs) across various Tiers, has helped streamlining the multiple steps involved in the overall workflow. It became clear that different variations of the general onboarding workflow needed to be adopted, depending on the specific characteristics of each DH such as the current status of their data and the technologies they already employ.

For instance, in cases where clinical data preparation was still underway, DHs were advised to begin representing their data in the format required by the EUCAIM Common Data Model (CDM) from the outset. In another case, Tier 2 DHs with clinical data already stored in a database were advised to implement a custom mediator to integrate directly with the search component, rather than converting their data to the EUCAIM CDM.

As part of the pilots, we prepared or adapted various supporting resources, including templates and instructional documents, to help DHs organise and accelerate the setup of their nodes. The piloting process also highlighted several limitations, gaps, and special cases or needs, leading to appropriate revisions and mitigation actions.

For example, the review of the diverse local data models and formats used by DHs, along with the new cancer types introducing new types of information for inclusion in the hyper-ontology, led to several updates and extensions of the Common Data Model to support additional encoding formats and types of information.

Overall, the pilots demonstrated that data transformation to the CDM is a time-consuming process - as expected - that requires thorough inspection of local data structures, a deep understanding of the underlying semantics, significant manual effort, and the involvement of various domain experts, including semantic, clinical, and technical experts. A key document developed and refined during the piloting phase was a template to guide DHs in mapping their local data structure to the CDM, thereby helping to systematise the ETL process.

Additionally, the piloting phase revealed licensing issues with certain tools. These concerns were escalated to both the tool providers and the legal team to ensure the tools could be used without barriers by the DHs. Testing also identified minor limitations in some tools, which were communicated to the providers, prompting improvements in tool deployment and documentation.

# ANNEX 1 – definition of Use Cases: Comprehensive summary

The annex contains tabulated, detailed information from the internal and external call application forms that have been approved by the Access Committee and the European Commission. The annex serves as a supplementary resource to the use case descriptions provided in deliverable 7.3.

## Internal call – I round

- Data Holders:

Table 1. Description of Use Case no. 4 from Servicio Andaluz De Salud

| Author: María González López | Data sharing: Federated Node | Organisation's name: Servicio Andaluz De Salud | Organization's Acronym: SAS | Tier: 3 |
|---|---|---|---|---|
| **General description of the potential use and clinical impact of the shared data:**<br>The shared datasets hold significant potential for advancing research, development, training, and validation of AI algorithms and tools aimed at optimizing the stratification of cancer patients for personalized treatment. This, in turn, promises a profound clinical impact, enhancing treatment planning precision, efficacy, and accuracy through early and refined patient characterization. By effectively identifying the most suitable treatments, cases of under- and over- treatment, as well as adverse side effects and comorbidities could be significantly reduced. Furthermore, the burden on healthcare professionals could be alleviated and the patient waiting times could decrease. All together, this could improve the efficiency and personalization of healthcare targeted at cancer patients, resulting in enhanced health outcomes and quality of life, as well as mitigated costs associated with cancer and its management. Along this line, particularly in the first dataset "Prostate Cancer: Magnetic Resonance (MR) Imaging and clinical data", the characterization of prostate tumors according to their biological aggressiveness is crucial for selecting the treatment. Patients would be stratified into low-risk cases, for which active surveillance is implemented to follow-up the tumor until the need of treatment, or high-risk cases, for which early treatment is urgently needed to ensure that cancer will not spread and progress towards a lethal phenotype. In the second dataset "Locally Advanced Breast Cancer: Positron Emission Tomography (PET) / Computed Tomography (CT) Imaging and clinical data" and the third dataset "Non-Hodgkin Lymphoma (NHL): Positron Emission Tomography (PET) / Computed Tomography (CT) Imaging and clinical data" the stratification of patients according to the predicted response to an specific treatment would allow avoiding treatments that do not provide value to patients due to their lack of necessity, inefficacy, secondary effects, and/or comorbidity, as well as their associated healthcare and economic expenditure. Finally, in the fourth dataset "Lung Cancer: Computed Tomography (CT) Imaging and clinical data", the characterization of lung tumors according to their type is crucial for treatment selection and prognosis. Moreover, the stratification of patients according to the predicted response to radiation therapy would allow minimizing the associated toxicity and comorbidities. | | | | |
| **Dataset 1**<br><br>**Cancer Type:** Prostate Cancer<br><br>**Dataset name:** Prostate Cancer: Magnetic Resonance (MR) Imaging and clinical data<br><br>**Dataset description:**<br>The dataset contains a set of retrospectively collected cases of patients diagnosed with prostate cancer, treated at the Virgen del Rocio University Hospital between 2017 and 2023. It includes, for each patient, Magnetic Resonance imaging studies corresponding to the moment of diagnosis, along with clinical data: radiology and histopathology report information, lab results, tumour markers, and diagnostic procedures (date of image/biopsy acquisition, prostate size, prostate morphology, signs of hemorrhage, shape, edges, zonal location, size and PI-RADS score of each prostate observed lesion, Gleason score, TNM staging, level/density of Prostate-Specific Antigen (PSA) in blood, digital rectal examination, date of diagnosis, and tumor topography and morphology), as well as demographics and other information (subject's date of birth, gender, weight, height, and family history of cancer, information regarding tumor progression and/or recurrence, and date of death if applicable). Encompassing aspects relevant to the aggressiveness evaluation of prostate cancer tumors, the primary purpose of the dataset is to facilitate research and development of AI methods for the estimation of Gleason score and consequent stratification of prostate cancer tumors according to their biological aggressiveness. This is expected to serve as an invaluable tool for researchers, clinicians, and healthcare professionals striving to personalize the management of prostate cancer, significantly contributing to the optimization of treatment planning, by advancing in the characterization of patients into low-risk and high-risk cases.<br><br>**Dataset Collection Method:** Cohort, Disease-specific<br><br>**Dataset Type:** Original Dataset, Annotated Dataset | | | | |

**Dataset Terms of Use:** Access restricted. Access by request

**Dataset Intended Purpose:** Estimate Gleason score in prostate cancer patients for stratification based on biological aggressiveness, to optimize treatment planning

**Imaging Modality:** Magnetic Resonance (DICOM tag (0008, 0060) = MR)

**Imaging body part:** Pelvis, Prostate

**Age range:** 40 / 90 years

**Sex:** Male (100%)

**Number of subjects:** 600

**Number of DICOM studies:** 600

**Image size in GB:** 72

**De-identification:** Personal data is fully anonymized

**Dataset 2**

**Cancer Type:** Breast Cancer

**Dataset name:** Locally Advanced Breast Cancer: Positron Emission Tomography (PET) / Computed Tomography (CT) Imaging and clinical data

**Dataset description:**
The dataset contains a set of retrospectively collected cases of patients diagnosed with locally advanced breast cancer (with axillary lymph node involvement) who underwent primary systemic treatment in the axilla at the Virgen del Rocío University Hospital between 2018 and 2023. It includes, for each patient, Positron Emission Tomography (PET) / Computed Tomography (CT) imaging studies corresponding to the moment of diagnosis and to the assessment of treatment response, along with clinical data: imaging test results and histopathology report information, lab results, treatments, and diagnostic procedures (date of image/biopsy acquisition, primary lesion size, number of affected lymph nodes, metabolic parameters such as SUVmax, SUVmean, ΣTLG and ΣMTV, TNM staging, biochemical parameters, type of treatment, number of cycles and response assessment, date of diagnosis, and tumor topography and morphology), as well as demographics and other information (subject's date of birth, gender, weight, height, and family history of cancer, information regarding tumor progression and/or recurrence, and date of death if applicable). Encompassing images and an extensive set of clinical variables, the primary purpose of the dataset is to facilitate research and development of AI methods for the prediction of the response to primary systemic treatment in the axilla among patients diagnosed with locally advanced breast cancer. This is expected to serve as an invaluable tool for researchers, clinicians, and healthcare professionals striving to personalize the management of breast cancer, significantly contributing to the optimization of treatment planning, by advancing in the identification of patients poised to derive maximum benefit from primary systemic treatment in the axilla.

**Dataset Collection Method:** Cohort, Disease-specific

**Dataset Type:** Original Dataset, Annotated Dataset

**Dataset Terms of Use:** Access restricted. Access by request

**Dataset Intended Purpose:** Predict patient response to neoadjuvant primary systemic treatment in the axilla in locally advanced breast cancer, to optimize treatment planning

**Imaging Modality:** Positron Emission Tomography / Computed Tomography (DICOM tag (0008, 0060) = PET) / (DICOM tag (0008, 0060) = CT)

**Imaging body part:** Head, Neck, Chest, Breast, Abdomen, Pelvis, Extremity

**Age range:** 30 / 80 years

**Sex:** Female (99.9%), Male (0.1%)

**Number of subjects:** 100

**Number of DICOM studies:** 250

**Image size in GB:** 37

**De-identification:** Personal data is fully anonymised

**Dataset 3**

**Cancer Type:** Non-Hodgkin Lymphoma (NHL)

**Dataset name:** Non-Hodgkin Lymphoma (NHL): Positron Emission Tomography (PET) / Computed Tomography (CT) Imaging and clinical data

**Dataset description:**
The dataset contains a set of retrospectively collected cases of patients diagnosed with Non-Hodgkin Lymphoma (NHL) who underwent chimeric antigen receptor T-cell therapy (CAR-T) at the Virgen del Rocío University Hospital between 2019 and 2023.
It includes, for each patient, Positron Emission Tomography (PET) / Computed Tomography (CT) imaging studies pre-treatment and post-treatment (after 1, 3, and 6 months), along with clinical data: imaging test results and histopathology report information, lab results, treatments, and diagnostic procedures (date of image/biopsy acquisition, primary lesion size, number of affected lymph nodes, metabolic parameters such as SUVmax, SUVmean, ΣTLG and ΣMTV, TNM staging, biochemical and hematological parameters, assessment of treatment response, date of diagnosis, and tumor topography and morphology), as well as demographics and other information (subject's date of birth, gender, weight, height and family history of cancer, information regarding tumor progression and/or recurrence, and date of death if applicable). Encompassing images and an extensive set of clinical variables, the primary purpose of the dataset is to facilitate research and development of AI methods for the prediction of the response to CAR-T therapy among patients diagnosed with NHL. This is expected to serve as an invaluable tool for researchers, clinicians, and healthcare professionals striving to personalize the management of NHL, significantly contributing to the optimization of treatment planning, by advancing in the identification of patients poised to derive maximum benefit from CAR-T therapy.

**Dataset Collection Method:** Cohort, Longitudinal, Disease-specific

**Dataset Type:** Original Dataset, Annotated Dataset

**Dataset Terms of Use:** Access restricted. Access by request

**Dataset Intended Purpose:** Predict patient response to chimeric antigen receptor T cell therapy (CAR-T) in Non-Hodgkin Lymphoma (NHL), to optimize treatment planning.

**Imaging Modality:** Positron Emission Tomography / Computed Tomography
(DICOM tag (0008, 0060) = PET) / (DICOM tag (0008, 0060) = CT)

**Imaging body part:** Head, Neck, Chest, Breast, Abdomen, Pelvis, Extremity

**Age range:** 18 / 80 years

**Sex:** Male (60%), Female (40%)

**Number of subjects:** 125

**Number of DICOM studies:** 450

**Image size in GB:** 67

**De-identification:** Personal data is fully anonymised

**Dataset 4**

**Cancer Type:** Lung cancer

**Dataset name:** Lung Cancer: Computed Tomography (CT) Imaging and clinical data

**Dataset description:**
The dataset contains a set of retrospectively collected cases of patients diagnosed with lung cancer who underwent radiation therapy at the Virgen del Rocio University Hospital between 2017 and 2023. It includes, for each patient, Computed Axial Tomography imaging studies corresponding to the moment of diagnosis, to a treatment simulation for its planification, and to post-treatment follow-ups (after 3, 6, 9, and 12 months), along with clinical data: radiology and histopathology report information, lab results, treatments, and diagnostic procedures (date of image/biopsy acquisition, TNM staging, type of treatment and dosimetric data, date of

| |
|---|
| diagnosis, toxicity levels, and tumor topography and morphology), as well as demographics and other information (subject's date of birth, gender and smoking habit, information regarding tumor progression and/or recurrence, and date of death if applicable). Encompassing images and an extensive set of clinical variables, the primary purpose of the dataset is to facilitate research and development of AI methods for, on one hand, the distinction between non-small cell lung cancer (NSCLC) and small cell lung cancer (SCLC), and consequent stratification of lung cancer tumors according to their type. And, on the other hand, the prediction of the response to radiation therapy among patients diagnosed with NSCLC or SCLC. This is expected to serve as an invaluable tool for researchers, clinicians, and healthcare professionals striving to personalize the management of lung cancer, significantly contributing to the optimization of treatment planning, by advancing in the characterization of patients into NSCLC and SCLC cases, and the identification of those poised to derive maximum benefit from radiation therapy.<br><br>**Dataset Collection Method:** Patient-based, Longitudinal, Disease-specific<br><br>**Dataset Type:** Original Dataset, Annotated Dataset<br><br>**Dataset Terms of Use:** Access restricted. Access by request<br><br>**Dataset Intended Purpose:** Differentiate between non-small cell lung cancer (NSCLC) and small cell lung cancer (SCLC) for patient stratification based on lung cancer type. Predict patient response to radiation therapy in each lung cancer type (NSCLC or SCLC). Both to optimize treatment planning.<br><br>**Imaging Modality:** Computed Tomography (DICOM tag (0008, 0060) = CT)<br><br>**Imaging body part:** Chest, Lung, Abdomen, Pelvis<br><br>**Age range:** 40 / 90 years<br><br>**Sex:** Male (80%), Female (20%)<br><br>**Number of subjects:** 600<br><br>**Number of DICOM studies:** 3,600<br><br>**Image size in GB:** unspecified<br><br>**De-identification:** Personal data is fully anonymised |

Table 2. Description of Use Case no. 7 from Sciensano

| Author: Emilie Cauët | Data sharing: Federated Node | Organisation's name: Sciensano | Organization's Acronym: Sciensano | Tier: 1 |
|---|---|---|---|---|
| **General description of the potential use and clinical impact of the shared data:** There's an increased interest and use of AI-tools to better identify select patients for genomic testing, considering the increased number of testing for patients with advanced cancer and the associated health care costs. Currently, there's no pre-screening to select patients for genomic analysis aiming to identify targetable genomic alterations for which the patient can be offered specific treatments. As most of these actionable genomic alterations are very rare (with a prevalence of 1-5% maximum in most solid tumor types) there is an interest to identify patients more accurately before a genomic analysis is requested. This will avoid unnecessary genomic analysis, at least in some tumor types and unnecessary expenditures. These should be avoided, and this cannot be considered separately from correct monetary incentives for comprehensive genomic analysis for our patients who are expected to derive the most benefit from it. An appropriate infrastructure to analyze and discuss these genomic findings in National Molecular Advisory Boards, as they exist for example today within the PRECISION-program of the Belgian Society of Medical Oncology (https://doi.org/10.1016/j.esmoop.2022.100524) is integrally related to the above. PRECISION initiative encompasses four studies (PRECISION 1 & 2, BALLETT and GeNeo) that aim to boost genomic and clinical knowledge with the ultimate goal to offer patients with metastatic solid tumors molecularly guided treatment. In these studies we compared patients with advanced and solid cancers, including CGP analysis in blood, local testing using limited gene-panels (with a maximum of 75 genes), with comprehensive genomic profiling using large-scale panels such as that of Illumina and Foundation Medicine. The aims are:<br><br>1. To determine the added value of comprehensive and agnostic NGS versus "real-world" practice (i.e. local testing, no reimbursement for local testing and/or no accessible metastatic lesion) in providing patients with | | | | |

advanced/metastatic solid tumors access molecular guided therapy and/or immunotherapy based on genomic results.

2. To describe the landscape of genomic alterations detected by reimbursed NGS or by comprehensive

panel testing.

3. To assess the technical success of comprehensive panel testing.

4. To describe the uptake of treatments recommended by the molecular tumor board guided by the genomic testing.

The GeNeo and BALLETT protocols provide access to broad genomic profiling tests enabling the detection of single-nucleotide variants, small indels, copy number variations, gene rearrangements/fusions, RNA splice variants and mutational signatures. For these studies, digitized HE-slides of thousands of patients (>2000) with a variety of solid tumors will be available. All molecular data, demographics of the patients and clinical information can be retrieved. These are stored centrally in the Precision Belgium section of the Healthdata database, a national platform hosted by Sciensano: (https://www.sciensano.be/en/about-sciensano/sciensanos-organogram/healthdatabe). In this use case, we propose to use the unique set of data from, first, GeNeo then BALLETT studies to investigate and demonstrate the development of AI-enhanced pathology tools to predict, directly from cancer pathology slides, the genetic alterations and gene expression. The prediction of genetic alterations and gene expression from routine pathology slides is regarded as one of the most promising future applications of AI in pathology. Since we have catalogued comprehensively all actionable genomic alterations currently in use (including gene fusions, TMB, MSI, HRD, specific mutations (EGFR, ALK, etc...), we have probably the largest nationally acquired real-world comprehensive genomic database in advanced cancer patients. We suggest leveraging this dataset for AI-applications within EUCAIM. Finally, at the national level, we also have a Government-led initiative for use of Gene-Expression Profiles like Mammaprint and OncotypeDx in daily practices. The genomic data of both assays, including all clinical data, is also stored centrally. We aim to digitize all HE-slides of this project for AI-validations that may complement the use of expensive genomic assays.

**Dataset 1**

**Cancer Type:** All

**Dataset name:** GeNeo

**Dataset description:**
Molecular Profiling data (Roche FMI Test): Tissue biopsy or Liquid Biopsy Imaging data.

**Dataset Collection Method:** Patient-based

**Dataset Type:** Original Dataset

**Dataset Terms of Use:** unspecified

**Dataset Intended Purpose:**
1. To determine the added value of comprehensive and agnostic NGS versus “real-world” practice (i.e.

local testing, no reimbursement for local testing and/or no accessible metastatic lesion) in providing patients with advanced/metastatic solid tumors access molecular guided therapy and/or immunotherapy based on genomic results.

2. To describe the landscape of genomic alterations detected by reimbursed NGS.

3. To describe the landscape of genomic alterations detected by comprehensive panel testing.

4. To assess the technical success of comprehensive panel testing.

5. To describe the uptake of treatments recommended by the molecular tumor board guided by the genomic testing. In this use case, we propose to use the unique set of data from GeNeo study to investigate and demonstrate the development of AI-enhanced pathology tools to predict, directly from cancer pathology slides, the genetic alterations and gene expression.

**Imaging Modality:** Digitized HE-slides

**Imaging body part:** Head, Neck, Chest, Breast, Abdomen, Pelvis, Extremity

**Age range:** >18 years

| |
|---|
| **Sex:** Male and Female<br><br>**Number of subjects:** 937<br><br>**Number of DICOM studies:** N/A<br><br>**Image size in GB:** 2<br><br>**De-identification:** Personal data is pseudonymized |

Table 3. Description of Use Case no. 6 from Karolinska Institutet

| Author: Taeyang Choi | Data sharing: Federated Node | Organisation's name: Karolinska Institutet | Organization's Acronym: KI | Tier: 2 |
|---|---|---|---|---|

**General description of the potential use and clinical impact of the shared data:** Karolinska Institutet's contribution would be providing access to the core dataset of the VAI.B (Validation of AI for Breast Cancer) validation platform to EUCAIM partners. In our on-premises database, we have images and data for the validation of AI for cancer detection in screening mammography (AI-CADe), originating from four regions of Sweden: Stockholm, Västmanland, Östergötland, and Södermanland. More specifically, we use the following data:

1. Mammography images from the breast cancer screening programs, extracted from PACS (Picture Archiving and Communication System)
2. Annotations of cancer signs in a selection of the mammography images
3. Human radiologist assessments extracted from RIS (Radiology Information System)
4. Cancer outcomes and characteristics extracted from NKBC (Swedish National Breast Cancer Quality Register)

In addition to the above data, we need AI inference data from the AI model which should be validated. With these data components, the validation output will be generated as below:

- Description of validation population
- Number of individuals and exams
- Number of cancer and healthy
- Radiologist assessments
- Age
- Time from exam to diagnosis
- Description of AI system inferences
- Distribution of exam-level scores overall, subgrouped for cancer and healthy
- Determining thresholds for each scenario below and matched to standard-of-care sensitivity and
- specificity Performance measures (preliminary)

Scenarios (for cancer detection):

One radiologist
Two radiologists
AI alone
AI plus one radiologist
AI plus two radiologists

Metrics:

ROC curves including AUC
Precision-recall curves including AUC
Confusion matrix (inferences vs ground truth) for each threshold
Sensitivity (or Cancer Detection Rate) & Specificity (or False positives) for each threshold
VAI.B enables AI-CADe developers (EUCAIM partners) to obtain validation results showing the accuracy and robustness of their system in a multi-center multi-equipment setting. The data that we will provide as an output to the AI-CADe contributor are the 'Metrics' above.

The pilot project might proceed as follows:

1. The AI model developer registers the AI model in the central hub, and selects where to validate (e.g., the KI node) if there are other data holders; they may also state if they want to include or exclude specific

mammography equipment.

2. The central hub sends the AI model to the KI local federated node.

3. KI local federated node makes the AI model process the images held at KI on-premises.

4. KI local federated node calculates the metrics KI local federated node exports summary results to the central hub.

5. The central hub sends the summary results to the AI model developer.

To accomplish this pilot project, the following specifications are required (will be provided by WP4~6):

1. Specification of the communication protocol between the central hub and the local node for federated validation.

2. Specification of the establishment of the local node for federated validation.

3. Specification of AI models (e.g., manuscript for AI, API, hardware and software requirements, etc.).

**Dataset 1**

**Cancer Type:** Breast cancer

**Dataset name:** VAI.B Dataset

**Dataset description:**
Mammography images from four regions of Sweden (Stockholm, Västmanland, Östergötland, and Södermanland):
DICOM standard
Pseudonymized

Radiologist assessments from hospital records:

Initial read by two radiologists (flag or not flag for consensus discussion)
Recall decisions from the consensus discussion

Cancer outcomes:

Reference standard for determining whether an examination represents a cancer case or not based on a three-year follow-up time linked to the national quality register of breast cancer (NKBC).
Cancer characteristics, e.g., histological origin, invasive/in situ component, tumor size, lymph node metastasis, receptor status.

**Dataset Collection Method:** Patient-based and Cohort

**Dataset Type:**Annotated Dataset

**Dataset Terms of Use:**
1. Patient-level of core data (VAI.B dataset) isn't searchable by users, but EUCAIM user's AI will be validated with core dataset.

- All data will be held at the local node and will be used in the validation process performed on-premises at KI.

2. Authorization to download the validation output through the central hub.

**Dataset Intended Purpose:** Validate breast cancer detection AI models

**Imaging Modality:** Mammography

**Vendor:** GE, Philips, Hologic, Siemens

**Imaging body part:** Breast

**Age range:** avg. 57

**Sex:** Female

**Number of subjects:** Cancer cases (6,527), Control (9,037), Total: 15,564

| |
|---|
| **Number of DICOM studies:** ~ 17 131<br><br>**Image size in GB:** unspecified<br><br>**De-identification:** Personal data is fully anonymized |

Table 4. Description of Use Case no. 8 from Assistance Publique Hôpitaux De Paris

| Author: Valérie Vilgrain | Data sharing: Federated Node | Organisation's name: Assistance Publique Hôpitaux De Paris | Organization's Acronym: APHP | Tier: 3 |
|---|---|---|---|---|
| **General description of the potential use and clinical impact of the shared data:** Primary liver cancers (PLC) define an heterogeneous group of malignant proliferations, including mainly hepatocellular carcinoma (HCC), a common complication of chronic liver diseases and cirrhosis, intrahepatic cholangiocarcinomas (iCCA) and combined tumors (cHCC-CCA) that display both HCC an iCCA components within the same nodule. PLCs are one of the leading causes of cancer-related death with an increased incidence worldwide. The accurate diagnosis of each type of tumor is crucial as specific management is considered, ranging from liver transplantation for HCC to systemic therapies for iCCA. While the diagnosis of HCC is mainly based on imaging in the context of cirrhosis, the gold standard for diagnosing iCCA and cHCC-CCA is histological analysis. Non-invasive imaging diagnosis is based on CT and MR. Tumor biopsy is highly informative providing diagnostic but also prognostic information. However, the histological diagnosis may be challenged, especially in poorly differentiated tumors and cHCC-CCA if biopsy does not sample both tumor components. This reality highlights the need to integrate complementary multiscale imaging approaches (macroscopy/radiology and microscopy/pathology), to provide a deep multidimensional analysis of the tumor for better patient stratification, more precise therapeutic choices, and ultimately, an improvement in the survival rate of patients with PLC. | | | | |
| **Dataset 1**<br><br>**Cancer Type:** Primary Liver Cancer (PLC)<br><br>**Dataset name:** MOSAIC<br><br>**Dataset description:**<br>The retrospective and cross-sectional MOSAIC dataset includes 872 patients (1989-2022) resected from one of three primary liver cancers (PLC) : (a) 451 hepatocellular carcinoma (HCC), (b) 283 intrahepatic cholangiocarcinoma (iCCA), (c) 138 combined hepatocellular carcinoma - cholangiocarcinoma (cHCC- CCA). MOSAIC is a multi-scale dataset centered on imaging: radiology/macroscopic imaging (CT and MR) and pathology/microscopic imaging (digital whole slide imaging). The CT and MR imaging correspond to a preoperative contrast-injected exam. All patients have a preoperative CT and half have a preoperative MR which standardized the names of sequence labels. The minimum bio-clinical information is collected in a structured way in dedicated databases. All patients are associated with at least one hematein eosin saffron (HES) stained digital slide (biopsies or surgical specimens). MOSAIC is based on 4 existing cohorts [PREVISION, PREDIAC, CARAMEL, CLASSTHEP] used and in other research projects in liver, some of which have been published. CLASSTHEP cohort and a quarter of PREVISION cohort are already present in CDW@AP-HP (EDS) research environment. Each of these cohorts already has all the regulatory approvals (although these will have to be updated as part of EUCAIM).<br>- PREVISION (HCC) : CSE20-85_MOSAIC-EDS (IRB00011591 CSE of CDW@AP-HP) - N°CNIL 1980120<br>- CLASSTHEP (PLC) : CSE22-20_MOSAIC-EDS (IRB00011591 CSE of CDW@AP-HP) - N°CNIL 1980120<br>- PREDIAC (iCCA) : CER-2022-168 (IRB00006477 CER Paris Nord) - N°CNIL ongoing<br>- CARAMEL (cHCC-CCA) : CER-2022-169 (IRB00006477 CER Paris Nord) - N° CNIL 20220613182206<br>MOSAIC may be extended in a second phase to other categories of data, such as larger clinical databases that may include longitudinal follow-up, molecular analyses (genomic and transcriptomic), other immuno-histochemical stainings (e.g. CD31, glypican, CK7, CK19, nestin ...), or micro/macroscopic imaging annotations/segmentation. MOSAIC is open to supplementing its data (e.g. new patients, additional data) according to projects and their needs.<br><br>**Dataset Collection Method:**Disease-specific, Cohort<br><br>**Dataset Type:** Original Dataset<br><br>**Dataset Terms of Use:**<br>MOSAIC would like to integrate into Tier 3 of the EUCAM Data Federation Framework and share its data via federated node : data cannot be exported (for federated or local use only) without special authorization from French regulatory bodies (e.g. CNIL, the GDPR French authority) and AP-HP committees. So, access to its data will therefore be restricted and requested. The request for a data-sharing project will be forwarded to a restricted scientific, legal and operational steering committee specific to MOSAIC, in order to process the application rapidly | | | | |

| |
|---|
| and with all AP-HP players. This MOSAIC steering committee will be composed of members of Beaujon Hospital, AP-HP.Nord<br>- Université Paris Cité, i.e. the principal investigator Pr Valérie VILGRAIN (Head of Medical Imaging Department), Dr Jules GREGORY (radiologist), Pr Valérie PARADIS (Head of Pathology Department), Dr Aurélie BEAUFRERE (pathologist), Dr Kévin MONDET (project manager of MOSAIC), Dr Daniel CHRISTEL.<br>This committee is supported by technical contractual and regulatory referents from AP-HP, i.e. Jean NEMBO (AP-HP/EUCAIM IT research department project manager), Aurélien MAIRE & Marzieh KOHANDANI TAFRESH (IT research imaging department), Linda THIEULON (PDO/GDPR referent), Corentin D'HONDT (contractualization referent). The data-sharing submission file will be worked on with the EUCAIM consortium to ensure that it is as harmonized and simple as possible for data users. After approval by the steering committee and all necessary discussions with the data user project team, the following agreements and related documentation will be necessary: the signature of a data sharing agreement (DSA) and a collaboration agreement (CA) per project. Also, the use of data will be subject to CDW@APHP's terms of use and to current ethical and data protection regulations. A certificate of ethical approval by an ethics committee is required for the principle of data sharing within the EUCAIM consortium, and a simplified amendment will be drawn up for each case of data user.<br><br>**Dataset Intended Purpose:**<br>MOSAIC dataset aims to support projects for the detection, segmentation or characterization of primary liver cancers, for both diagnostic and prognostic purposes. It may also be useful for training, validating or qualifying tools based on medical imaging (AI tools, medical devices, imaging analysis ...).<br><br>**Imaging Modality:** CT + MR + digital slide<br><br>**Vendor:** unspecified<br><br>**Imaging body part:** Liver<br><br>**Age range:** >18 years<br><br>**Sex:** Male or Female<br><br>**Number of subjects:** 872<br><br>**Number of DICOM studies:** 872 CT, 400 MR (+ 872 digital in .svs or .ndpi format)<br><br>**Image size in GB:** 200 Mb per CT, 500 Mb per MR, 1 Gb per digital slide<br><br>**De-identification:** Personal data is fully anonymized |

Table 5. Description of Use Case no. 1 from Sapienza

| Author: Valeria Panebianco | Data sharing: Central Repository | Organisation's name: Universita Degli Studi Di Roma La Sapienza | Organization's Acronym: Sapienza | Tier: 1 |
|---|---|---|---|---|
| **General description of the potential use and clinical impact of the shared data:** Prostate cancer the most common cancer in men over 50 years of age. It is highly heterogeneous both biologically and clinically, ranging from indolent disease to very aggressive tumors. Accurate non-invasive early detection and prediction of aggressive tumors is essential to direct biopsy and allow tailored therapeutic planning. Despite magnetic resonance imaging (MRI) and prostate imaging reporting and data system (PI-RADS) scores show optimal diagnostic performance in experienced readers, there is potential for improvement, particularly in equivocal cases (PI-RADS 3 scores). The integration of artificial intelligence (AI) in MRI interpretation may hold the potential of enhancing accuracy and efficiency in early detection and diagnosis, especially considering equivocal lesions identified on MR images. One of the key advantages of MRI in prostate cancer detection is its ability to detect clinically significant tumors. Research has shown that MRI can improve the detection of aggressive tumors while reducing unnecessary biopsies and overdiagnosis of indolent lesions. By accurately identifying high-risk tumors, An MRI helps to guide targeted biopsies to the areas of greatest concern, improving biopsy yield and minimizing patient discomfort. Moreover, MRI plays a crucial role in the staging and risk stratification of prostate cancer, providing valuable information about tumor extent, with potential extra prostatic extension, and involvement of adjacent structures. This information is essential for treatment planning and for predicting patient prognosis. With MRI, clinicians can assess the risk of tumor progression and tailor treatment strategies accordingly, whether it involves active surveillance, surgery, radiation therapy. The integration of AI in MR images evaluation has the potential to further improve prostate cancer detection and diagnosis. AI algorithms can analyze large volumes of MRI data rapidly, and may assist radiologists in detecting and characterizing lesions with greater efficiency. AI-based tools can identify subtle changes in imaging patterns that may indicate the presence of cancer, improving sensitivity and reducing the likelihood of false-negative results. | | | | |

| |
|---|
| Furthermore, AI may help standardize MRI interpretation, reducing interobserver variability and ensuring consistent and reliable diagnosis across different healthcare settings. By learning from vast datasets of annotated MRI images, AI algorithms can continuously improve their performance and adapt to evolving clinical needs. This has the potential to democratize access to high-quality prostate cancer diagnosis, particularly in regions with limited resources or expertise. One of the most interesting applications of AI in MRI-based prostate cancer detection is the development of radiomics – the extraction and analysis of quantitative imaging features from medical images. Radiomics leverages AI techniques to identify subtle imaging biomarkers that may be indicative of tumor presence, aggressiveness, or treatment response. By integrating radiomic features with conventional imaging and clinical data, such as PI-RADS score, PSA levels, PSA-Density and biopsy results, AI models can generate personalized risk assessments and treatment recommendations for individual patients. Additionally, AI-driven decision support systems can aid radiologists and clinicians in interpreting MRI findings and making informed treatment decisions. These systems can provide real-time feedback, highlight areas of concern, and suggest appropriate follow-up actions based on established guidelines and best practices. This not only improves diagnostic accuracy but also enhances workflow efficiency and reduces interpretation time. MRI plays a vital role in early detection and management of prostate cancer, offering superior imaging capabilities compared to other diagnostic methods. The integration of AI in MRI analysis holds interesting promise in further improving diagnostic accuracy, standardizing interpretation, and facilitating personalized treatment approaches. With continued advancements in imaging technology and AI algorithms, MRI-based prostate cancer detection is poised to make significant strides in improving patient outcomes and reducing the burden of this prevalent disease. |
| **Dataset 1**<br><br>**Cancer Type:** Prostate Cancer<br><br>**Dataset name:** Sapienza_Prostate_Cancer<br><br>**Dataset description:**<br>Demographic and clinical data (Excel file) will include: age; family history of prostate cancer; total PSA level; PSA -Density (calculated thanks to MR images); final histopathological result (when available), including Gleason Score, Grade Group, percentage of malignant tissue within cores, length of positive tissue within cores. MRI data will include the entire exams in DICOM files and additional information (Excel file) including prostate volume, number of lesions, lesions' dimension, PI-RADS score, eventual positive lymph nodes or metastasis.<br><br>**Dataset Collection Method:** Disease-specific<br><br>**Dataset Type:** Processed Dataset<br><br>**Dataset Terms of Use:**<br>Our institution owns this data<br><br>**Dataset Intended Purpose:**<br>Characterization of MRI foci suspicious for prostate cancer and correlation with histopathology/clinical/imaging data<br><br>**Imaging Modality:** MRI<br><br>**Vendor:** Siemens and GE<br><br>**Imaging body part:** Pelvis, with focus on prostate gland<br><br>**Age range:** 45-75 years<br><br>**Sex:** Male<br><br>**Number of subjects:** 200<br><br>**Number of DICOM studies:** 200<br><br>**Image size in GB:** unspecified<br><br>**De-identification:** Personal data is pseudonymized |

Table 6. Description of Use Case no. 13 from Medizinische Universitaet Wien

| Author: Philipp Seeböck | Data sharing: Central Repository | Organisation's name: Medizinische Universitaet Wien | Organization's Acronym: MUW | Tier: 1 |
|---|---|---|---|---|

**General description of the potential use and clinical impact of the shared data:** The dataset is comprised of breast DCE-MRI scans of a screening cohort with a high risk for developing breast cancer, with corresponding pixel-wise lesion annotations. This dataset will foster the development of novel AI-based models for (1) segmentation of small breast lesions that require a biopsy, (2) early detection of breast cancer and (3) machine learning related methodological advances with respect to domain adaptation and generalization, as the imaging data was acquired with two different MR scanner types. The development of such models is expected to have a substantial clinical impact. First, breast cancer is the most common type of cancer in women (24.5% of all cancer cases). Second, early detection is crucial for patient prognosis and survival. Third, there is still an unmet need for AI models to exhibit robust generalization performance across institutes with different scanners or image acquisition protocols.

**Dataset 1**

**Cancer Type:** Breast Cancer of patients with a high risk for developing breast cancer

**Dataset name:** CIR High-Risk Breast Cancer DCE-MRI Dataset

**Dataset description:**
The patient cohort consists of about 150 visits from about 100 patients with a high risk for developing breast cancer and were recruited at the Vienna General Hospital (AKH). The patients were included in the study cohort if one of the following conditions applied and patient consent was given: (1) Previous case of breast cancer before the age 36, (2) Previous ovarian cancer before the age of 41, (3) Confirmed mutation in the genes BRCA-1 or BRCA-2, (4) Family anamnesis with a cumulative risk of developing breast cancer before the age of 79 > 20%. The patients participated in regular DCE-MRI screening visits at the AKH. Every patient visit was examined by a trained radiologist who assigned a BI-RADS score and requested a biopsy if a suspicious change in breast tissue was detected (BI-RADS 4 and 5). In the majority of visits (≈ 91%) no suspicious tissue changes were detected (BI-RADS 1, 2 and 3). Suspicious lesions (BI-RADS 4) were detected in ≈ 7.6% and highly suspicious lesions (BI-RADS 5) in less than ≈ 0.2% of the visits. In ≈ 1% of the visits the imaging data was insufficient (BI-RADS 0). Manual lesion segmentation was obtained for the final dataset with the help of two radiologists at the Vienna General Hospital. The segmentation process involved a pixel/voxel vise delineation of the lesion in the first post contrast MRI volume using ITK-SNAP [1]. For each of the visits the following data is available: (1) T1 weighted pre contrast images and at least 3 T1 weighted post contrast images, (2) the information about the MRI scanner, (3) Voxel wise lesion annotations and (4) BI-RADS score. [1] Paul A. Yushkevich et al. User-guided 3D active contour segmentation of anatomical structures: significantly improved efficiency and reliability. NeuroImage, 2006. doi: 10.1016/j.neuroimage.2006.01.01

**Dataset Collection Method:** Disease-specific, Cohort

**Dataset Type:** Annotated Dataset

**Dataset Terms of Use:** This must be defined and set up together with the responsible unit at the Medical University of Vienna (e.g. non-commercial use only)

**Dataset Intended Purpose:** We intend to publish the data for scientific use only. The intended purpose is to foster the development and validation of AI-models which are able to conduct segmentation and early detection of small breast lesions reliably across institutions with varying technical image acquisition standards

**Imaging Modality:** DCE-MRI

**Vendor:** Siemens Mammomat Inspiration, Siemens Avanto

**Imaging body part:** Breast/Thorax

**Age range:** 20 - 87 years

**Sex:** Female

**Number of subjects:** about 100

**Number of DICOM studies:** acquisitions from about 150 visits

**Image size in GB:** unspecified

**De-identification:** Personal data is pseudonymized

Table 7. Description of Use Case no. 3 from Universita Degli Studi Di Roma La Sapienza

| Author: Carlo Catalano | Data sharing: Central Repository | Organisation's name: Universita Degli Studi Di Roma La Sapienza | Organization's Acronym: Sapienza | Tier: 1 |
|---|---|---|---|---|
| **General description of the potential use and clinical impact of the shared data:** Liver cancer is the fifth most common cancer and the second most frequent cause of cancer-related death globally. Hepatocellular carcinoma (HCC) represents about 90% of primary liver cancers and constitutes a major global health problem. Medical imaging is an essential part of hepatocellular carcinoma diagnosis, contributing to primary liver tumour classification and HCC staging, playing a crucial role in clinical settings, with the capacity to non- invasively provide multi-parameter, multi-dimensional, and multi-modality structural and functional information on lesion and peri-tissues on computed tomography (CT) and magnetic resonance imaging (MRI). Conventional imaging methods, however, provide limited information, particularly considering some crucial aspects such as the high heterogeneity and diverse biological behaviors of HCC, which directly affect the prognosis and survival of patients, that remain a key concern and a crucial need to be addressed. Recent studies have proved radiomics and deep learning as an emerging, powerful, and non-invasive tool based on various imaging modalities. They can help detect intrinsic features from conventional radiological image that are not entirely visible to the sole human eye; their use and application are progressively increasing, especially in the research field. These features may provide information regarding heterogeneity and invasiveness of the disease that can be of predictive and/or prognostic value, thus leading to the selection of the best possible treatment. Moreover, these tools can enable comprehensive insightful data mining that has achieved favorable performance in the detection and classification, diagnosis and differentiation, staging and grading, aggressive behavior, treatment responses, prognosis, and survival rates of HCC. Nevertheless, although positive results have been reported in the literature for radiomics, its drawbacks have limited its clinical translation, and the wide implementation of radiomics and deep learning in actual routine clinical practice requires sustainable validation and optimization. According to the Barcelona Clinic Liver Cancer (BCLC) staging system, several interventional procedures are available for the treatment of hepatocellular carcinoma in early and intermediate stage, including thermal ablation and trans-arterial chemoembolization (TACE, drug eluting bead-TACE, balloon-occluded- TACE, and degradable starch microspheres-TACE), although response to treatment is variable across patients. Radiomics and deep learning tools could be applied also to this novel field, expanding the potential role of imaging in the evaluation of response to therapy and more generally in the management of this type of patient. The goal of this use case is to identify innovative imaging biomarkers derived not only from conventional CT and MR images, but also after application of AI-based softwares, with the aim of predicting response to treatment, in order to appropriately select patients and improve clinical outcomes. | | | | |
| **Dataset 1**<br><br>**Cancer Type:** Hepatocellular carcinoma<br><br>**Dataset name:** Sapienza_HCC<br><br>**Dataset description:**<br>Demographic data (Excel file): age, sex. Comorbidities (Excel file). Pre-treatment data (Excel file): Child-Pugh score, MELD and MELDNa scores, BCLC, alpha fetoprotein, AST, ALT, gammaGT, bilirubin, albumin, blood count, encephalopathy, ascites, esophageal varices; imaging data: tumor's location and dimension. After treatment data (Excel file): AST, ALT, gammaGT, bilirubin, albumin, blood count. Imaging (CT and MRI) data will include the entire exams (DICOM files) and additional information (Excel file) including number of lesions, lesions' dimension, eventual positive lymph nodes or metastasis.<br><br>**Dataset Collection Method:** Disease-specific, Cohort<br><br>**Dataset Type:** Annotated dataset<br><br>**Dataset Terms of Use:** Our institution owns this data<br><br>**Dataset Intended Purpose:** Identification of imaging biomarkers from CT or MRI studies of the liver in patients with HCC predicting response to interventional procedures (thermal ablation, trans-arterial chemoembolization)<br><br>**Imaging Modality:** MRI and CT<br><br>**Vendor:** Philips, GE, Siemens<br><br>**Imaging body part:** Abdomen<br><br>**Age range:** 41-89 years<br><br>**Sex:** Male and Female | | | | |

| |
|---|
| **Number of subjects:** 100<br><br>**Number of DICOM studies:** 100<br><br>**Image size in GB:** unspecified<br><br>**De-identification:** Personal data is pseudonymized |

Table 8. Description of Use Case no. 2 from Universita Degli Studi Di Roma La Sapienza

| Author: Andrea Laghi | Data sharing: Federated Node | Organisation's name: Universita Degli Studi Di Roma La Sapienza | Organization's Acronym: Sapienza | Tier: 1 |
|---|---|---|---|---|
| **General description of the potential use and clinical impact of the shared data:** Identification of imaging biomarkers to give the best approximation to Grade/Stage colon cancer at baseline. To characterize biological processes in a voxel-wise high-spatial resolution approach, extracting first, second, and third-order radiomics features from radiological datasets of patients with colorectal tumors. Imaging biomarkers will be subsequently computed by AI- model and integrated with clinical data to generate a radiogenic signature for the personalized management of patients with colon tumors. | | | | |
| **Dataset 1**<br><br>**Cancer Type:** Colon cancer<br><br>**Dataset name:** Colon cancer<br><br>**Dataset description:**<br>Dataset of colon cancer with baseline CT, radiological TNM, histology data, survival data and basic clinical data (age, sex, therapy for the cancer type, survival information)<br><br>**Dataset Collection Method:** Disease-specific, Patient-based<br><br>**Dataset Type:** Original Dataset<br><br>**Dataset Terms of Use:** unspecified<br><br>**Dataset Intended Purpose:** Identify imaging biomarkers to predict survival/progression<br><br>**Imaging Modality:** CT<br><br>**Vendor:** unspecified<br><br>**Imaging body part:** unspecified<br><br>**Age range:** 18-90 years<br><br>**Sex:** Male and Female<br><br>**Number of subjects:** 200<br><br>**Number of DICOM studies:** 200<br><br>**Image size in GB:** unspecified<br><br>**De-identification:** Personal data is pseudonymized | | | | |

Table 9. Description of Use Case no. 5 from Centro Hospitalar Universitario Do Porto Epe

| Author: Manuela Franca | Data sharing: Central Repository | Organisation's name: Centro Hospitalar Universitario Do Porto Epe | Organization's Acronym: CHUP | Tier: 1 |
|---|---|---|---|---|
| **General description of the potential use and clinical impact of the shared data:** Data and imaging sharing for fostering research of AI tools for answering clinical questions provided by consortium partners. | | | | |
| **Dataset 1**<br><br>**Cancer Type:** Lung cancer | | | | |

**Dataset name:** Lung Cancer Santo Antonio

**Dataset description:**
CT images of baseline and follow up studies of patients with lung cancer (small cells and non-small cells), with the respective clinical data (basic demographics and histological type)

**Dataset Collection Method:** Longitudinal

**Dataset Type:** Original dataset

**Dataset Terms of Use:** unspecified

**Dataset Intended Purpose:** Data and imaging sharing for fostering research of AI tools for answering clinical questions provided by consortium partners

**Imaging Modality:** CT

**Vendor:** unspecified

**Imaging body part:** unspecified

**Age range:** 30-90 years

**Sex:** Male and Female

**Number of subjects:** 150

**Number of DICOM studies:** unspecified

**Image size in GB:** unspecified

**De-identification:** Personal data is pseudonymized

- Data Users:

Table 10. Description of Use Case no. 9 from Ifom-Istituto Fondazione Di Oncologia Molecolare Ets

| Author: Silvia Marsoni | Intention:<br>- development of AI tools and solutions<br>- training of AI tools and solutions | Organisation's name: Ifom-Istituto Fondazione Di Oncologia Molecolare Ets | Organization's Acronym: IFOM |
|---|---|---|---|
| **Title of the use case:** OCEANUS | | | |
| **General description of the use case:**<br>OCEANUS is a pragmatic project that will establish the clinical utility of I-ROR, a first- in-kind digital marker predicting the Risk Of Resistance to chemotherapy in individual patients with metastatic colorectal cancer (mCRC). I-ROR will change the paradigm of care for this highly prevalent tumor enabling the personalized management of patients, that will increase their survival and quality of life while optimizing the burden of care and minimizing the socio-economical costs . The project consists in a pragmatic clinical trial representing a step forward towards individualized precision medicine for metastatic colorectal cancer (mCRC) by confirming the clinical utility of the I-ROR diagnostic digital marker. I-ROR (Individual Risk Of Resistance) is an AI-driven predictive marker that quantifies the responsiveness to standard chemotherapy in individual patients. I-ROR is a cost-effective, widely-applicable, and easy-accessible diagnostic tool that integrates three AI-based signatures derived from easily accessible information already collected per standard of care (FFPE slides and CT-scans) universally available at point-of-care in every European hospital.<br><br>**Expected timeline for the realization of the use case:** 2 years<br><br>**Description of the intended use and expected benefit related to the use of the EUCAIM data:**<br>The OCEANUS I-ROR offers a realistic insight of the trade-offs between true efficacy and true toxicity. Such previously unavailable knowledge will foster a truly informed shared decision-making process about treatment choices between mCRC patients and their caregivers. A high probability of response conveyed by I-ROR metrics can convince a patient to choose a more intensive regimen balancing the threat of higher toxicity with a higher probability of long-lasting benefit. Conversely, patients with an I-ROR predicted chemo-refractory tumor could be given the opportunity of an up-front targeted therapy usually delivered in later lines. CRC Patients without actionable targets in their tumor can be offered the choice to receive personalized best supportive care. The | | |

| |
|---|
| deployment of I-ROR as a clinical decision support system will reshape the clinical routine care of this highly prevalent cancer type, optimizing the burden and the costs of care across Europe and beyond. The integrated AI pipeline to extract the OCEANUS I-ROR digital marker will perform radiomics analyses of metastatic CT regions of interest and combine the predictive power of radiomics with pathomics, clinical variables and molecular markers that have already demonstrated predictive power with respect to chemotherapy resistance. Our consortium has already obtained preliminary and promising results in predicting resistance to first line therapy in mCRC using molecular markers, radiomics and pathomics individually. First, we will assess the expression of the proteins encoded by the ATM, RAD51 and RAD51C genes, since these proteins are master regulators in the signaling and/or repair of double-strand DNA breaks, hence critical for modulating cellular sensitivity to DNA-damaging cytotoxic agents. The null expression of these proteins confer sensitivity to oxaliplatin and irinotecan in CRC models both in vitro (cell lines, patients derived organoids)22 and in vivo (patients-derived xenografts . Our group has optimized automated immunohistochemical diagnostic scoring systems for all the proteins. Second, to improve these molecular markers, we will add high-accuracy prognostic information derived from AI-empowered digital pathology analysis based on digitised routine haematoxylin-eosin (H&E) slides. Pathomics has already proven its clinical utility by pinpointing several crucial molecular predictive biomarkers in CRC Accordingly, from an initial analysis of a subset of patients enrolled in the OCEANUS retrospective cohort, the sole pathomics led to identify mCRC patients resistant to chemotherapy with a NPV of 92. Third, Radiomics features will be derived from routine CT scans using a previously tested AI pipeline. The pipeline will also include a model for automatic segmentation of liver metastases which has yielded a detection rate of 98%, for lesions larger than 20 mm in previous studies. Finally, to enhance the performances of individual signatures and molecular markers, we will combine different levels of information (molecular data, radiomics and pathomics) into an integrated signature. This path, already trodden by other authors, has yielded better performances than the individual modality signatures. The novel CRC radiopathomics signatures will be subsequently integrated with the rest of data sources under a clinically driven accurate, robust, and transparent diagnostic Clinical Decision Support System (CDSS). The holistic machine and deep learning AI methodological framework we will propose is based on the extensive and pioneering experience of our partners.<br><br>**Description of the requested data:**<br>Target population: metastatic CRC treated with front line (1st line) chemotherapy for metastatic disease<br>1) a Digitalized H&E stained FFPE slide of the primary tumor of mCRC patients (if WSI not available, clear FFPE slides of 4 micron each)<br>2) whole body or abdominal CT scan including the primary tumor at basal and liver mets prior to chemotherapy start for metastatic disease<br>3) CT scan of liver mets after 3 months of first line chemotherapy for metastatic disease<br>4) CT scan of best response to 1st line chemotherapy<br>5) CT scan of metastatic progression after 1st line chemotherapy<br>6) clinical annotation (demographic (sex, age, age at diagnosis etc) anamnestic (surgery y/n, adjuvant treatment (n/yes specify), PS other), tumor characteristics (Microsatellites and KRAS status), treatments (type of chemotherapy, type of biological therapy e.g. anti EGFR, antiangiogenic, targeted drugs)<br>7) at least 2 years of followup after diagnosis of metastatic disease. |

Table 11. Description of Use Case no.14 from Medizinische Universitaet Wien

| | | | |
|---|---|---|---|
| Author: Silvia Marsoni | Intention:<br>- development of AI tools and solutions<br>- training of AI tools and solutions<br>- validation of AI tools and solutions | Organisation's name: Medizinische Universitaet Wien | Organization's Acronym: MUW |
| **Title of the use case:** Domain Generalization of AI-based models in Cancer Imaging using Disentanglement Learning | | | |
| **General description of the use case:**<br>Main Objectives: Developing novel machine learning methodology to enhance domain adaptation and generalization abilities of AI models, ensuring robustness to technical noise in imaging data related to changes in imaging technology. Expected Results: Novel methodology for effective detection and segmentation of cancer across various scanners, institutes and hospitals. Clinical Impact: Imaging data from different sites or hospitals often vary significantly due to different acquisition techniques, devices, imaging protocols or patient population characteristics. Furthermore, the mentioned factors may change over time as technology advances, and image characteristics evolve. This impedes applicability of models across sites, leads to a reduction of model performance, and in the worst case a need to train several models for an inhomogeneous reality of large-scale multi-center repositories. These challenges lead to models that are currently typically trained and evaluated on single-center datasets, limiting both their potential generalization performance as well as the explanatory power of the evaluation with respect to clinical applicability. The novel methodology developed in this project will serve as the basis for AI models that can be trained on single or multi-institutional datasets and applied on hold-out institutional data with differing imaging appearance or distributions. Methodology: Based on prior work in continual learning, domain generalization and disentanglement, we will develop novel deep learning based methodology which is robust to changes in imaging technology and capable of handling data heterogeneity | | | |

across scanners, institutes and hospitals. We aim to use multi-centric data to develop and validate prediction algorithms that yield robust prediction accuracy across centers.

**Expected timeline for the realization of the use case:** 2 years

**Description of the intended use and expected benefit related to the use of the EUCAIM data:**
In this pilot-study we will advance and evaluate cutting edge continual active machine learning and domain adaptation techniques to optimally exploit data on the scale of the proposed repository for robust, accurate, and trustworthy models. Key foci of the usecase are training robust and accurate models from heterogeneous and imbalanced data, detecting out-of-distribution samples during application time to avoid untrustworthy predictions, conducting centralized and federated continual learning to keep up with advance of technology and changing environments or scanner types, or making efficient use of limited distributed annotation resources for optimal training of models. The use of EUCAIM data from multiple institutions will enable the development and validation of such algorithms. In particular.

**Description of the requested data:**
We will need multi-centric data to develop and validate algorithms that yield robust prediction accuracy across institutions. This data from multiple centers should be relatively consistent in terms of target labels to allow benchmarking and validation of developed models on hold-out institutional data. At the same time, the imaging data may involve a certain level of heterogeneity. Deep learning model development typically requires powerful hardware, including GPUs, CPUs, and fast storage connections. However, since we have the necessary computational resources and local storage available at our institution, we don't need to rely on external providers for these resources (assuming that we can download the data to our local storage).

- Both:

Table 12. Description of Use Case no.10 from Fundacion Para La Investigacion Del Hospital Universitario La Fe De La Comunidad Valenciana

| Author: Luis Martí Bonmatí | Data sharing: Central Repository | Organisation's name: Fundacion Para La Investigacion Del Hospital Universitario La Fe De La Comunidad Valenciana | Organization's Acronym: HULAFE | Tier: 3 |
|---|---|---|---|---|
| Intention:<br>- development of AI tools and solutions<br>- training of AI tools and solutions<br>- validation of AI tools and solutions | | | | |
| **General description of the potential use and clinical impact of the shared data:** Glioblastoma Multiforme (GBM) is the most aggressive primary brain tumor, characterized by rapid growth, infiltrative behavior, and resistance to treatment. The prognosis for GBM patients remains dismal, with a median survival of around 12 to 15 months, despite advances in treatment modalities. Understanding the disease's dynamics across different stages is crucial for improving patient outcomes and developing effective therapeutic strategies. Our dataset formed by 400 GBM patients includes pathological, MRI images, and clinical information, offering invaluable insights into the progression and management of GBM. The shared data provides a wealth of information on the molecular and pathological characteristics of GBM tumors. By analyzing genomic profiles and histopathological features, researchers will be able to identify key molecular signatures associated with tumor aggressiveness, prognosis, and treatment response. This molecular profiling can aid in personalized treatment approaches, such as targeted therapies and immunotherapies, thereby optimizing patient care and outcomes. Also, the dataset offers insights into post-operative outcomes and treatment responses. MRI images coupled with clinical data enable clinicians to monitor disease progression, assess treatment efficacy, and detect early signs of recurrence. By tracking changes in tumor size, location, and morphology, healthcare providers can tailor treatment regimens and interventions to individual patient needs, maximizing therapeutic benefits while minimizing adverse effects. Furthermore, the shared data facilitates collaborative research efforts and data-driven decision-making in clinical practice. By pooling resources and expertise from multiple institutions and research groups, investigators can conduct large-scale analyses, validate findings, and accelerate the translation of research discoveries into clinical applications. This collaborative approach fosters innovation, fosters interdisciplinary collaboration, and enhances the reproducibility and generalizability of research findings, ultimately benefiting patients and advancing the field of neuro-oncology. | | | | |
| **Dataset 1**<br><br>**Cancer Type:** Glioblastoma multiforme (Brain cancer)<br><br>**Dataset name:** Glioblastoma multiforme (Brain cancer)<br><br>**Dataset description:**<br>450 Glioblastoma Multiforme patients. This comprehensive dataset includes pathological, MRI images, and clinical information, offering invaluable insights into the progression and management of GBM.<br><br>**Dataset Collection Method:** Disease-specific<br><br>**Dataset Type:** Original Dataset<br><br>**Dataset Terms of Use:** The access right of the data will be granted on a case-by-case basis<br><br>**Dataset Intended Purpose:** The aim of this dataset is to understand GBM progression and identify biomarkers. It aims to make it possible to predict treatment response and prognosis for personalized care; identify therapeutic targets and pathways for intervention; monitor disease evolution and recurrence patterns and facilitate translational research for innovative therapies<br><br>**Imaging Modality:** Magnetic resonance image<br><br>**Vendor:** multiple vendors<br><br>**Imaging body part:** Brain<br><br>**Age range:** 30-90 years<br><br>**Sex:** Males and Females | | | | |

**Number of subjects:** 400

**Number of DICOM studies:** unspecified

**Image size in GB:** unspecified

**De-identification:** Personal data is de-identified (direct identifiers have been removed)

**Title of the use case:** Predictive recurrence location and time to recurrence using IA

**General description of the use case:**
General Description: Our project aims to develop a predictive model leveraging artificial intelligence (AI) techniques to forecast the location and time to recurrence in patients with GBM. By harnessing imaging biomarkers such as diffusion, radiomics, and perfusion, along with diagnostic clinical data, we seek to enhance clinical decision-making and improve patient outcomes. Main Objectives: - Develop AI algorithms to analyze imaging biomarkers extracted from MRI scans. - Create a model capable of estimating the time to GBM recurrence based on predefined imaging biomarkers. - Develop an IA model to predict the location of GBM recurrence within the brain using diagnostic imaging data. Expected Results: We anticipate that our predictive models will accurately forecast the time to GBM recurrence and identify the location of recurrent tumors with high precision, leveraging the rich information provided by imaging biomarkers. Expected Clinical Impact: Our project has the potential to revolutionize the approach to GBM management by facilitating personalized treatment planning based on individual patient characteristics and tumor biology. Also, the ability to predict GBM recurrence and its location will enable clinicians to intervene promptly, potentially leading to earlier detection of recurrence, improved response to therapy, and better patient outcomes. By providing clinicians with predictive tools to anticipate recurrence and progression, our project may help optimize resource allocation in healthcare settings, ensuring that patients receive timely and appropriate interventions based on their predicted risk of recurrence. Methodology: We will utilize advanced image processing techniques to extract quantitative imaging biomarkers from MRI scans, including diffusion metrics, radiomic features, and perfusion parameters. 4 After, we will employ machine learning and deep learning algorithms to develop predictive models for time to recurrence and location of recurrence based on the extracted imaging biomarkers and diagnostic imaging data. Furthermore, we will validate the predictive models and assess their performance in terms of accuracy, sensitivity, specificity, and clinical utility.

**Expected timeline for the realization of the use case:** 2 years

**Description of the intended use and expected benefit related to the use of the EUCAIM data:**
Our project, leveraging the EUCAIM dataset, aims to develop and validate AI algorithms for predicting GBM recurrence. By utilizing imaging biomarkers of diffusion, radiomics, and perfusion, alongside diagnostic imaging data, our primary goal is to train models capable of predicting both the time to recurrence and the location of recurrent tumors within the brain. Anticipated benefits include improved patient outcomes through early detection and intervention, enhanced treatment planning with personalized therapy selection, resource optimization in healthcare settings, and the advancement of precision medicine in neuro-oncology. We plan to employ machine learning and deep learning algorithms such as Convolutional Neural Networks (CNNs), Random Forest (RF), Gradient Boosting Machines (GBM), and Recurrent Neural Networks (RNNs) as real time detection networks such as YOLO, to analyze imaging data and develop predictive models for GBM recurrence. This initiative aligns with our mission to optimize patient care, advance research in neuro-oncology, and contribute to the development of innovative solutions for improving outcomes in GBM patients.

**Description of the requested data:**
The requested data encompasses patients with GBM, with a minimum of 300 subjects targeted. The data should include MR image encompassing anatomical, diffusion, and perfusion imaging modalities, providing comprehensive insights into tumor characteristics and progression. Additionally, clinical data, captured through standardized case report forms, will complement the imaging data, facilitating comprehensive analysis and interpretation. The recruitment period spans from 2000 to the present, ensuring a diverse and representative cohort reflective of temporal trends in GBM management and outcomes. It is expected that the requested data will be annotated and harmonized to facilitate interoperability and consistency across datasets. Annotation tools and computational resources will be utilized to ensure accurate labeling and analysis of the imaging and clinical data. Temporary storage solutions will be employed to securely store and manage the extensive dataset during processing and analysis, safeguarding patient privacy and data integrity throughout the research process.

Table 13. Description of Use Case no.11 from Assistance Publique Hopitaux De Paris

| Author: Raphaele Renard Penna | Data sharing: Federated Node | Organisation's name: Assistance Publique Hopitaux De Paris | Organization's Acronym: APHP | Tier: 3 |
|---|---|---|---|---|
| Intention:<br>- development, validation or training of AI tools considered for medical devices | | | | |
| **General description of the potential use and clinical impact of the shared data:** With more than 50,000 new cases per year in France and 9,000 deaths, prostate cancer (PCa) is the most common in men and represents the 3rd cause of cancer mortality (1) . The diagnostic strategy has recently evolved with the positioning of prostate mpMRI as a first-line tool for diagnosis (2,3). The European Commission has proposed to evaluate the feasibility and effectiveness of organised prostate cancer screening for men on the basis of prostate specific antigen ( PSA ) testing in combination with MRI scanning as follow-up. The main objective of prostate MRI is to identify clinically significant PCa (csPCa) (i.e., Gleason Score ≥ 3 + 4) while sparring men with benign lesions or indolent prostate from unnecessary interventions or treatment. It has been shown that the NPV of MRI for detection of csPCa mpMRI was up to 90%.The prostate imaging-reporting and data system (PIRADS) was introduced in 2012 and updated twice to standardize prostate MRI acquisition and interpretation. Though the benefits of the PI-RADS have been well recognized over the years, prostate MRI still suffers from intra-reader and inter-reader differences and non-negligible amounts of false-positive and false-negative results. However, prostate interpretation conditions patient care: indication and schedule of biopsy samples, therapeutic strategy and prognostic information. A massive increase in prostate mpMRIs is expected given the frequency of this cancer. Its interpretation requires a high level of expertise, which is rarely available. The result is low accessibility and unequal access to care for patients. Deep learning (DL) has shown remarkable performance on a broad spectrum of medical imaging tasks in recent years, with prostate cancer diagnostics no exception. However, despite the promises of DL technology in PCa, most of the proposed DL models have been trained on small single-center proprietary data and are not publicly shared, hindering the reliability and wide-spread adaptation. In order to reap the benefits of these AI models in clinical care, there remains the need for developing publicly available anonymized large patient data sets, to train and test AI models with external validation, to assess their performance. Here, the primary objective of the data sharing will be to enable researchers to design, train and test publicly available DL models for identifying cs PCa on a multicenter multivendor database with radiological and histological ground truth. The development of a diagnostic aid tool would make it possible to achieve greater efficiency through a qualitative improvement in results and an increase in access to care.<br>(1) Carioli G, Bertuccio P, Boffetta P, Levi F, La Vecchia C, Negri E, et al. European cancer mortality predictions for the year 2020 with a focus on prostate cancer. Ann Oncol. 2020;31(5):650-8.<br>(2) Kasivisvanathan V, Rannikko AS, Borghi M, Panebianco V, Mynderse LA, Vaarala MH, et al. MRI- Targeted or Standard Biopsy for Prostate-Cancer Diagnosis. N Engl J Med. 2018;378(19):1767-77.<br>(3) .Rouviere O, Puech P, Renard-Penna R, Claudon M, Roy C, Mege-Lechevallier F, et al. Use of prostate systematic and targeted biopsy on the basis of multiparametric MRI in biopsy-naive patients (MRI-FIRST): a prospective, multicentre, paired diagnostic study. Lancet Oncol. 2019;20(1):100-9. | | | | |

| |
|---|
| **Dataset 1**<br><br>**Cancer Type:** Prostate<br><br>**Dataset name:** PAIMRI<br><br>**Dataset description:**<br>300 prostate MR (DICOM) with segmentation and label, clinical and histological data (including a subset of pathological slides).<br><br>**Dataset Collection Method:** Cohort<br><br>**Dataset Type:** Annotated datasets<br><br>**Dataset Terms of Use:** Restricted access; by request. A certificate of ethics approval by an ethics committee is required for sharing data with EUCAIM<br><br>**Dataset Intended Purpose:** Development, training and validation of AI tools for prostate cancer detection<br><br>**Imaging Modality:** Magnetic Resonance<br><br>**Vendor:** multiple vendors<br><br>**Imaging body part:** prostate<br><br>**Age range:** > 40 years<br><br>**Sex:** Males<br><br>**Number of subjects:** 300<br><br>**Number of DICOM studies:** 300<br><br>**Image size in GB:** unspecified<br><br>**De-identification:** Personal data is pseudonymized |
| **Title of the use case:** Prostate Cancer Detection on MRI |

**General description of the use case:**
The main objective is the development of an AI algorithm for the detection of prostate cancer lesions from mpMRI images with automated prostate segmentation and volume estimation. Data :The data set include: - images of prostate mpMRI (multiparametric)( n=300) in DICOM format, from multiple vendors (GE Healthcare and Siemens) and different fields (1.5 and 3T MRI scanners) , - the corresponding structured report (localization, dimension, PI-RADS V2.1 score per lesion). All MRI scans are acquired according to the PIRADS technical recommendations - the corresponding structured histological data (tumor aggressiveness score or ISUP score) from prostate biopsies (standard and target) , - additional critical information such as demographic (age) and biological (PSA level) data in text format. The variables collected are only those necessary for the training, validation and testing of the algorithms as well as for the description of the main characteristics of the sample used. Methodology : The T2 modality will be used to train the model for prostate gland segmentation, the DWI and ADC images will be adjusted to match with the T2 in terms of size and resolution for prostate cancer detection. Clinically critical information sur as PSA will be used as an additional input. Pathology csPCa is defined as GS>=34 in any core biopsy. Quality control of the data will be carried out by evaluating the quality and completeness of images, the adequacy between the histological and clinical data, the representativeness of all types of PI-RADS Expected clinical impact : The tool will be used to automatically segment the prostate, and identify suspicious lesions ( detection and localization) . The goal is to enable radiologists to provide faster and more accurate reading, while increasing the diagnostic accuracy in detection of highly suspicious lesions on prostate MRI studies , reducing inter-reader variability and reading time. Such DL software solutions may provide valuable assistance in the interpretation of prostate MRI studies in light of an increasing demand.

**Expected timeline for the realization of the use case:** 2 years

**Description of the intended use and expected benefit related to the use of the EUCAIM data:**
We plan to train an AI algorithm to detect clinically significant prostate cancer lesions corresponding to radiological lesion aggressiveness with a PIRADS score equal or greater to 3. The input of the algorithm will be multiple MR sequences (typically 3: T2, ADC and DWI) and the output will be a map of the location of significant lesions in the images. The algorithm will proceed by first resampling the 3 sequences to a common image resolution and then normalizing the intensities of the 3 sequences. In a second stage, a prostate segmentation algorithm [1] will isolate the main prostate regions and sectors in the image, typically the peripherical and transition zones of the prostate. Indeed, these two regions are important information that influence the nature and the appearance of the prostate lesions. In the final stage, a lesion detection algorithm [2] will outline the potential cancer lesions in the prostate that are considered to be significatively at risk for the patient (PIRADS ≥ 3). The prostate segmentation and lesion detection algorithms are deep learning algorithms that can be either trained or tested on multiparametric MR datasets in a federated or centralized manner. We plan to start with a validation phase of already trained algorithms in order to test the suitability of existing dataset on our preprocessing pipeline. We will then experiment a training and validation phase in a multicentric manner if possible, where a subset of the available data will be used as a test set.
[1] Dimitri Hamzaoui, Sarah Montagne, Raphaele Renard-Penna, Nicholas Ayache, and Hervé Delingette. Automatic Zonal Segmentation of the Prostate from 2D and 3D T2-weighted MRI and Evaluation for Clinical Use. Journal of Medical Imaging, 9(2):024001, March 2022
[2]Dimitri Hamzaoui, Raphaële Renard-Penna, Sarah Montagne, Sébastien Molière,Nicholas Ayache, Hervé Delingette, Weak and Mixed supervision for prostate cancer detection through radiological reports, submitted
[3] Dimitri Hamzaoui. AI-based diagnosis of prostate cancer from multiparametric MRI. Theses, Université Côte d'Azur, June 2023

**Description of the requested data:**
Target population : Adult male who underwent Prostate MRI
For prostate cancer detection (clinical suspicion of prostate cancer)
Type of data (image modalities, case report forms):
Prostate MRI / DICOM format
Prostate MRI / standardized prostate report v 2.1 including
“Prostate volume,
PSA density,
for each target ( max 4 targets) : dimension ( maximal axial diameter / or volume), location and PI-RADS score “

Histological Data:
- Systematic and target cores in cases where both are performed
- Total number of cores sampled , number of positive cores
with Gleason score , percent high grade, and Grade groups,
- Tumor length ( mm) or percent core involvement
- Presence of Perineural invasion, intraductal carcinoma, and/or cribriform pattern
Datasets: filtering criteria, recruitment period)
patient who underwent prostate MRI for prostate cancer detection due to an elevated PSA, abnormal DRE, ethnic or family context (clinical suspicion of prostate cancer) from 2015-2024

| |
|---|
| MRI field strength and manufacturer: all accepted but should be specified Optional but highly desirable: a segmentation of each target on T2W and DWI, with annotation (Target number, location and PIRADS score) |

- Other:

Table 14. Linkopings Universitet

| Author: Caroline Bivik Stadler | Organisation's name: Linkopings Universitet | Organization's Acronym: LIU |
|---|---|---|
| Intention:<br>- developing methods for collaborative research | | |
| **General description of the use case:**<br>The AIDA Data Hub Sensitive Data Services (SDS 2.0) platform is designed to provide a suite of services tailored to ethically sanctioned research endeavors or activities grounded in a legal and ethical framework. Within this platform, an extensive array of offerings will be made available, encompassing sensitive data processing, data sharing, primary storage, private remote desktop functionalities, as well as provision for private VMs (Virtual Machines), web applications, and PACS (Picture Archiving and Communication System) systems. Additionally, there are plans to incorporate trusted medical imaging import capabilities, facilitating the direct transmission of medical examinations from scanners to the appropriate destination. The platform will accommodate Petabyte object storage, enabling the accumulation, refinement, annotation, and collaborative exploration of data. Furthermore, a comprehensive range of compute services will be provided, incorporating both GPU-enhanced and standard compute resources, utilizing the OpenStack framework. This will be complemented by the integration of a Kubernetes platform, empowering users to deploy and manage their preferred web services securely within private environments. Services such as backed-up long-term primary storage, large-scale project storage, and robust CPU and GPU compute options will be available as supplementary offerings. Crucially, the platform ensures strict segregation of user environments, affording each user autonomy within their designated "bubble" without visibility into other users' activities. Authentication mechanisms facilitate institutional login, enabling users to access the platform using credentials from their home institution, facilitated through the Life Science Login framework. This system will not require VPN connections or specialized account provisioning, streamlining access while enabling connectivity to other private services. SDS 2.0 will provide varied access modes to cater to a broad spectrum of users, including expert AI developers and clinicians, thereby enhancing its appeal and usability. For researchers inclined to share their data, the platform will facilitate this through the REMS resource entitlement management system, a web-based interface empowering researchers to delegate or manage their data-sharing decisions autonomously. SDS 2.0 will provide a data collaboration platform for EUCAIM, and this driver use case aims at providing a tailored environment for IDx Panorama, a comprehensive research initiative focused on multimodal and multi-omic cancer imaging. In IDx Panorama, researchers aim to integrate photon counting computed tomography (PCCT) technology into existing clinical procedures in the context of liver cancer. This involves incorporating PCCT alongside standard care pathways for patients diagnosed with liver cancer. Prior to surgery patients with diagnosed liver cancer are subjected to radiology assessment using CT and/or MRI. After surgical resection, in addition to standard care, the resected tumor tissue is subjected to ex-vivo higher image quality PCCT. This allows for detailed imaging of the tumor, facilitating correlations between its appearance and its biological characteristics. These insights could potentially inform the development of improved imaging diagnostics software. Following tumor removal, the tissue is sent for histological analysis and pathology assessment to confirm the cancer diagnosis. Comparisons are then made between the histological findings and the photon counting data, aiding in the validation of the photon counting technology's accuracy. Ultimately, the goal is to refine multimodal diagnostics for liver cancer, integrating photon counting with genetic analysis and other laboratory analyses. SDS 2.0 will function as a collaboration platform for this project, offering the tooling to streamline the transfer of examination data directly.<br><br>**Expected timeline for the realization of the use case:** 2 years | | |

# Internal calls – II round

- Data Holders:

Table 15. Description of Use Case no. 2 from Hospital Clinic De Barcelona

| Author: Josep Puig | Data sharing: Federated Node | Organisation's name: Hospital Clinic De Barcelona | Organization's Acronym: HCB | Tier: 2 |
|---|---|---|---|---|
| **General description of the potential use and clinical impact of the shared data:**<br>Meningiomas, tumors arising from meningothelial cells, are among the most prevalent intracranial neoplasms, with a 0.9% occurrence rate in routine brain MRI scans. These lesions account for nearly one-third of primary intracranial tumors. The World Health Organization (WHO) classifies meningiomas into three grades: benign (grade I), atypical (grade II), and anaplastic (grade III). This histological grading system helps predict the biological behavior and prognosis of meningiomas. Higher-grade meningiomas (II and III) are characterized by increased recurrence risk, invasiveness, and aggressiveness. MRI plays a crucial role in meningioma diagnosis, characterization, surgical planning, treatment decisions, and monitoring. Typical meningiomas appear sessile or lentiform, with sharp boundaries and broad dural attachments. They exhibit strong contrast enhancement and are usually isointense to hyperintense on T2-weighted and fluid-attenuated inversion recovery (FLAIR) images. Apparent diffusion coefficient (ADC) values can vary significantly among meningiomas but are often isointense to normal brain tissue. Some cases may present with peritumoural edema in the brain parenchyma, particularly in larger tumors. Atypical and anaplastic meningiomas tend to have larger volumes and faster growth rates compared to benign meningiomas. However, no definitive radiological criteria currently exist to reliably differentiate between grade I and II meningiomas. Grade III anaplastic meningiomas often display irregular shapes on MRI. Given the typically slow, multifocal, and multidirectional progression of meningiomas, automated detection could enhance image interpretation. Studies have shown that three-dimensional volumetric assessments of meningiomas in MRI offer increased sensitivity for detecting tumor progression compared to two-dimensional methods. While volumetric evaluation is superior to traditional diameter measurements for assessing tumor growth, it is time-intensive. Volumetric analysis of brain MR images is frequently performed in routine assessments and is essential for various neurological conditions, including brain tumors. | | | | |
| **Dataset 1**<br><br>**Cancer Type:** Meningioma<br><br>**Dataset name:** Meningioma-HCB<br><br>**Dataset description:**<br>This dataset includes a cohort of 150 patients with meningioma who have been treated and followed up over time at the Hospital Clínic de Barcelona. Demographic, clinical, histopathological, therapeutic, and evolutionary data are available. All cases have been confirmed through histopathological analysis, and tumor segmentation (lesional volumetry) have been performed by a radiologist with over 20 years of experience in both clinical practice and research. Lesion volumetries at prequirurgical stage were obtained for pre-contrast T1, T2WI, FLAIR, ADC, T2GE/SWI, and post-contrast T1 sequences. Therefore, this dataset sets up a matrix with almost 1000 tumor volumes. Additionally, a series of radiological characteristics related to qualitative variables have been assessed by neuroradiologists from the HCB, which will also be available in the dataset.<br><br>**Dataset Collection Method:** Disease-specific, Longitudinal<br><br>**Dataset Type:** Annotated Dataset<br><br>**Dataset Terms of Use:** unspecified<br><br>**Dataset Intended Purpose:**<br>1. Implementation of automatic brain tumour segmentation algorithms. Automated detection and segmentation of meningiomas in MRI studies may be performed as pre-processing before reading the images, possibly allowing for a more detailed analysis of tumour volumes and further multiparametric image analysis. This development may lead to an increased robustness and reliability due to reduced interreader bias. | | | | |

| |
|---|
| 2. Implementation of radiomics tool to extract imaging features from preoperative multiparameter MRI.<br>3. Developing a machine learning classifier for its noninvasive prediction using preoperative MRI from lesion volumetry, qualitative and quantitative data, including radiomics. While most are benign (WHO grade 1) and have a favorable prognosis, up to one-fourth are classified as higher-grade, falling into WHO grade 2 or 3 categories.<br><br>**Imaging Modality:** Magnetic resonance imaging<br><br>**Vendor:** Siemens<br><br>**Imaging body part:** Brain<br><br>**Age range:** 18 - 95 years<br><br>**Sex:** Males and Females<br><br>**Number of subjects:** 150<br><br>**Number of DICOM studies:** unspecified<br><br>**Image size in GB:** unspecified<br><br>**De-identification:** Personal data is included in the images. In this case EUCAIM can support you with specific tools and guidance for de-identification once the application is accepted. |

Table 16. Description of Use Case no. 7 from Aristotelio Panepistimio Thessalonikis

| Author: Ioanna Chouvarda | Data sharing: Federated Node | Organisation's name: Aristotelio Panepistimio Thessalonikis | Organization's Acronym: AUTH | Tier: 3 |
|---|---|---|---|---|
| **General description of the potential use and clinical impact of the shared data:**<br>The dataset encompasses both scintigraphy imaging and clinical data, collected retrospectively at three crucial stages during thyroid cancer treatment. By including these key data points, the dataset aims to provide a thorough perspective on each patient's response to the therapy. Imaging data in particular offer critical metrics that allow for the assessment of thyroid function, identification of residual thyroid tissue, and detection of metastatic sites. Complementary clinical data, such as patient demographics and detailed histological information, enhance the capacity to analyze and understand individual treatment responses comprehensively. The first scintigraphy scan performed post-surgery but before I-131 administration, serves as a baseline. This pre-therapy scan evaluates any remaining thyroid tissue, helping to identify residual cells that may require further treatment. The second scan, conducted immediately following I-131 administration, is aimed at visualizing the distribution of the isotope within the thyroid or metastatic regions. This early imaging approach provides an initial indication of how effective the treatment is. The third and final scintigraphy scan takes place several months post-therapy to evaluate long-term treatment outcomes. This follow-up imaging is essential for determining the success of the radioiodine therapy and monitoring any changes that may indicate the need for additional intervention. In terms of clinical data, the dataset includes comprehensive general patient information, such as age, sex, and medical history. Histological information encompasses data such as the date of the surgery that was done before the I-131 treatment, cancer subtype, the maximum diameter of the lesion measured, the number of the lesions, and also whether cancer affected the right lobe, the left lobe, both the lobes, or the isthmus of the thyroid. Additionally, the dataset also includes the tumor-node-metastasis (TNM) classification, utilized for precise cancer staging, and also important details of the therapy process, including essential dates. Leveraging the information in this dataset, an AI model can be developed to predict the likelihood of -131 therapy success. Such predictive models would empower healthcare providers to customize treatment plans with greater accuracy, thereby enhancing patient care and improving the overall efficacy of treatment. The dataset can also be used for models/algorithms trained with the aforementioned data and serve clinical questions including accuracy of diagnosis, prognosis etc. | | | | |

**Dataset 1**

**Cancer Type:** Thyroid cancer

**Dataset name:** Thyroid Scintigraphy

**Dataset description:**
The dataset consists of clinical data and scintigraphy DICOM files at three crucial stages during the I-131 treatment.

**Dataset Collection Method:** Disease-specific

**Dataset Type:** Original Dataset

**Dataset Terms of Use:** unspecified

**Dataset Intended Purpose:** AI model training

**Imaging Modality:** Scintigraphy

**Vendor:** AHEPA University Hospital

**Imaging body part:** Whole-body scans specifically focused on examining the thyroid

**Age range:** over 18

**Sex:** Males and Females

**Number of subjects:** 80 - 100

**Number of DICOM studies:** N/A

**Image size in GB:** 0.13 – 1 MB

**De-identification:** Personal data is pseudonymized

Table 17. Description of Use Case no. 5 from Aristotelio Panepistimio Thessalonikis

| Author: Ioanna Chouvarda | Data sharing: Federated Node | Organisation's name: Aristotelio Panepistimio Thessalonikis | Organization's Acronym: AUTH | Tier: 3 |
|---|---|---|---|---|

**General description of the potential use and clinical impact of the shared data:**
Use case: Breast Cancer Detection and Classification

Scenario: The use case focuses on utilizing the European Cancer Imaging Initiative (EUCAIM) to significantly improve the early detection of breast cancer. In this particular application, the focus is on analyzing mammogram images to identify and assess abnormalities in breast tissue. The AI-driven process, that can be developed with this dataset, involves a detailed examination to classify these findings into three categories: normal, indicating no signs of abnormalities; benign, referring to non-cancerous irregularities; or malignant, identifying cancerous tissues. The objective is, by providing this dataset, to contribute in developing a highly reliable and efficient AI algorithm that can support radiologists in making early, accurate diagnoses. By doing so, the algorithm aims to enhance the precision in detecting breast cancer, while simultaneously reducing the occurrence of false positives and minimizing the need for unnecessary biopsies. Additionally, by relying on mammograms, an already widely available and cost-effective imaging tool, the solution has the potential to significantly lower the overall cost of diagnosis,

making advanced breast cancer screening more accessible to a broader population.This approach seeks to empower medical professionals with advanced tools for decision-making, improving overall patient outcomes through earlier intervention and more targeted treatment strategies. Steps:

1. Patient Screening: Select patients eligible for this case

2. Data collection: Export raw data from the hospital's system

3. Data preprocessing: Prepare the dataset prior to data provision: de-identification, quality check.

Expected outcomes:

1. Improved early detection of breast cancer through screening procedures.

2. Enhanced efficiency in radiology workflows, reducing the time required for manual review and the need for double reading (evaluation from two radiologists).

3. Increased accuracy in distinguishing between benign and malignant lesions, reducing false positives and unnecessary interventions.

**Dataset 1**

**Cancer Type:** Breast cancer

**Dataset name:** AUTHBreast1

**Dataset description:**
72 Mammograms of 33 benign, 10 malignant and 29 normal cases. The dataset contains imaging data and some information needed for the use case: Case {benign, malignant, normal}, biopsy {yes, no}, and pathology {Masses / Microcalcifications / Calcifications / Structural Asymmetries / Focal Distortions}

**Dataset Collection Method:** Disease-specific

**Dataset Type:** Annotated Dataset

**Dataset Terms of Use:** unspecified

**Dataset Intended Purpose:** The dataset can be used for classification between normal, benign and malignant cases.

**Imaging Modality:** Mammography

**Vendor:** Hologic Selenia

**Imaging body part:** Breast

**Age range:** Unknown

**Sex:** Female

**Number of subjects:** 72

**Number of DICOM studies:** 72

**Image size in GB:** ~14GB

**De-identification:** Personal data is pseudonymised

**Dataset 2**

**Cancer Type:** Breast

**Dataset name:** AUTHBreast2

**Dataset description:**
125 Mammograms of 70 benign, 5 malignant and 50 normal cases. The dataset contains annotated imaging data and the minimum clinical information: Gender, Age at Diagnosis, Ethnicity, Current Survival Status, Case {benign, malignant, normal}, Pathology {Mass Shape, Mass Margin, Mass orientation, Mass echo pattern, Posterior features, Calcifications, Associated Features, Breast composition, Parenchymal enhancement level, Parenchymal enhancement symmetry, Presence of foci Pathological lymph-nodes}, BIRADS, metastasis location, Biopsy Findings {Cancer type, Molecular type, Grade}, Biomarkers (if available) {HER2 (c-erB2), ER, PR, Ki67, BRCA 1 mutation, BRCA 2 mutation, VI /LVI}.

**Dataset Collection Method:** Disease-specific

**Dataset Type:** Annotated Dataset

**Dataset Terms of Use:** unspecified

**Dataset Intended Purpose:** The dataset can be used for classification between normal, benign and malignant cases.

**Imaging Modality:** Mammography

**Vendor:** Fujifilm

**Imaging body part:** Breast

**Age range:** Unknown

**Sex:** Female

**Number of subjects:** 125

**Number of DICOM studies:** 125

**Image size in GB:** ~16Gb

**De-identification:** Personal data is pseudonymised

Table 18. Description of Use Case no. 1 from Servicio Madrileno De Salud

| Author: Javier Blázquez Sánchez | Data sharing: Federated Node | Organisation's name: Servicio Madrileno De Salud | Organization's Acronym: SERMAS | Tier: 2 |
|---|---|---|---|---|

**General description of the potential use and clinical impact of the shared data:**
The dataset comprises 123 annotated CT studies from patients with colorectal cancer (CRC). These images were acquired within 30 days prior to the surgical resection of the primary tumor and associated lymph nodes (LNs) and were collected retrospectively. Both the primary tumor and the most malignant-appearing LN were segmented by radiologists. In addition to basic clinical information, the dataset includes data on radiological staging, adjuvant treatment, surgical procedures, pathological findings, post-surgical follow-up (including relapse and death), carcinoembryonic antigen (CEA) measurements, and technical parameters of image acquisition. The dataset was initially compiled to investigate the potential of radiomics for predicting LN involvement preoperatively. In CRC, accurately determining LN involvement is crucial for treatment planning. The current standard practice involves removing all regional LNs during surgery to ensure any potentially malignant nodes are excised. However, this approach can lead to overtreatment, with unnecessary lymphadenectomies performed on nodes that are not malignant. This can result in increased surgical morbidity without providing additional therapeutic benefit. The primary aim of this retrospective observational study was to develop a radiomics-based model to predict LN involvement based on the tumor's radiomic features. The model

demonstrated promising results, with an area under the curve (AUC) of 0.88. Additionally, a separate model was developed to differentiate between malignant and benign LNs based on their radiomic features, achieving an AUC of 0.94. A notable limitation of the dataset is that, while we know the total number of malignant LNs per patient, specific information about which nodes were malignant or their exact locations is lacking. Despite this limitation, expanding the dataset is crucial for enhancing the robustness and clinical applicability of these predictive models. With additional imaging data and thorough validation, this work could lead to a reliable tool capable of reducing unnecessary lymphadenectomies. Beyond its initial purpose, this dataset could be leveraged for other applications, such as the development of automated detection and segmentation tools for tumors and LNs. The annotated data also offer opportunities to explore correlations between radiomic features and various clinical outcomes, such as treatment response or long-term survival. We believe this dataset represents a valuable resource for advancing imaging-based approaches in CRC. While the models developed using this dataset have shown strong initial results, further expansion and validation are needed to create clinically reliable tools. This dataset has the potential to contribute to more personalized treatment strategies in CRC, optimizing surgical decision-making and improving patient outcomes.

**Dataset 1**

**Cancer Type:** Colorectal cancer

**Dataset name:** GL-CCR

**Dataset description:**
The dataset comprises 123 annotated abdominal contrast-enhanced CT studies from patients with confirmed colorectal cancer, 57 with lymph node involvement and 66 without nodal metastasis. The studies were acquired within 30 days prior to surgical resection and are provided in DICOM format. Both the primary tumor and the most malignant-appearing lymph node were segmented by radiologists. Three radiologists with varying levels of experience participated, with each study segmented by a single radiologist. The segmentations are also provided in DICOM format. In addition to the images and segmentations, the dataset includes the additional data described above, although some (non-essential) clinical information has a significant number of missing values. This additional data does not currently adhere to any specific nomenclature or Common Data Model (CDM), but the team has the resources to adapt the dataset to comply with EUCAIM's CDM. The dataset is currently being revised, and the research team is reviewing the PACS and EHR to identify any patients who may have been inadvertently omitted during the initial data collection, so the total number of patients included may increase slightly.

**Dataset Collection Method:** Cohort

**Dataset Type:** Annotated Dataset

**Dataset Terms of Use:** The terms of use for the dataset are as outlined in the protocol approved by the Institutional Review Board (IRB). The approved use of the images and data is for developing a clinical-radiomic model to assess preoperative lymph node involvement in patients with colorectal cancer. Any sharing of data with other centers or additional uses must be disclosed and approved by the IRB.

**Dataset Intended Purpose:** To train and internally validate a machine learning model using radiomic and clinical data to predict preoperative lymph node involvement in colorectal cancer patients.

**Imaging Modality:** Computed Tomography

**Vendor:** Philips, Siemens, Toshiba

**Imaging body part:** Abdomen

**Age range:** 34 - 91

**Sex:** Male (61) and Female (62)

**Number of subjects:** 123

| |
|---|
| **Number of DICOM studies:** 123<br><br>**Image size in GB:** aprox. 600MB / study<br><br>**De-identification:** Personal data is pseudonymised |

Table 19. Description of Use Case no. 9 from Universidade De Coimbra

| Author: Sulaiman Abuhaiba | Data sharing: Central Repository | Organisation's name: Universidade De Coimbra | Organization's Acronym: UDC | Tier: 2 |
|---|---|---|---|---|
| **General description of the potential use and clinical impact of the shared data:**<br>Glioblastoma (GBM) is an aggressive and common malignant brain tumor with poor prognosis despite treatment advances. Imaging plays a crucial role in GBM diagnosis, treatment planning, and monitoring, making a well-curated dataset essential for research and clinical applications. This project will benefit from a comprehensive glioblastoma dataset comprising MRI data from multiple sequences, including T1, FLAIR, Diffusion-Weighted Imaging (DWI), Arterial Spin Labeling (ASL), T2, Dynamic Susceptibility Contrast Perfusion-Weighted Imaging (DSC-PWI), and T1 post-contrast. The integration of these imaging modalities offers valuable insights into tumor biology, progression, and response to treatment, supporting AI-driven analysis and predictive modeling. Each MRI sequence contributes a unique diagnostic value. T1-weighted imaging provides high-resolution anatomical details, while T1 post-contrast highlights blood-brain barrier disruptions, aiding in tumor delineation. FLAIR imaging is essential for detecting peritumoral edema and tumor infiltration. DWI and its associated Apparent Diffusion Coefficient (ADC) maps assess tumor cellularity and necrosis, distinguishing progression from pseudoprogression. ASL provides non-invasive perfusion imaging, offering insights into tumor vascularity and hypoxia. T2-weighted imaging enhances visualization of tumor mass effects, and DSC-PWI enables quantification of cerebral blood volume, a key biomarker for differentiating tumor progression from treatment effects. The shared dataset has significant clinical applications. Machine learning models trained on this multimodal imaging data can improve tumor segmentation, classification, and radiomics-based analysis, enhancing diagnostic accuracy and reducing inter-observer variability. AI-driven models could assist radiologists in more precise tumor margin identification, aiding surgical planning and radiation therapy. Additionally, integrating multiparametric imaging features can improve risk stratification and personalize treatment approaches. This dataset is also valuable for assessing treatment response, particularly in differentiating true tumor progression from pseudoprogression and radiation necrosis, which remain major challenges in GBM management. By leveraging AI and imaging biomarkers from DSC-PWI, DWI, and ASL, researchers can refine response criteria, facilitating early interventions and improved patient monitoring. Beyond clinical use, this dataset supports drug development and clinical trials. Imaging biomarkers across multiple sequences can be correlated with molecular and histopathological profiles, advancing precision medicine. Identifying imaging-based surrogate markers for treatment response could accelerate drug development and improve trial efficiency. Longitudinal analysis of imaging data may also provide deeper insights into GBM progression and recurrence patterns, guiding the development of novel therapeutic strategies. Furthermore, sharing this dataset contributes to standardization efforts in neuro-oncology imaging. The inclusion of diverse MRI sequences ensures computational models trained on this dataset are robust and generalizable, facilitating their integration into clinical workflows. This resource supports the broader scientific community in validating algorithms, refining imaging-based grading systems, and developing consensus guidelines for GBM imaging assessment. In summary, this dataset of glioblastoma patients, comprising multiple MRI sequences, holds immense potential for improving GBM diagnosis, prognosis, and treatment response assessment. By enabling AI-driven tools, enhancing clinical decision-making, supporting drug development, and promoting imaging standardization, it represents a critical resource for advancing neuro-oncology research and ultimately improving patient outcomes. | | | | |
| **Dataset 1**<br><br>**Cancer Type:** Glioblastoma<br><br>**Dataset name:** Glioblastoma UDC | | | | |

**Dataset description:**
The dataset consists of MRI imaging data from glioblastoma (GBM) patients, encompassing multiple imaging sequences critical for tumor assessment and clinical decision-making. These sequences include: T1-weighted (T1): High-resolution anatomical imaging used for structural assessment of the brain. T1 post-contrast: Enhances visualization of blood-brain barrier disruptions, aiding in tumor delineation. Fluid-Attenuated Inversion Recovery (FLAIR): Suppresses cerebrospinal fluid signals, making it useful for detecting peritumoral edema and tumor infiltration. Diffusion-Weighted Imaging (DWI): Evaluates tumor cellularity and necrosis, with Apparent Diffusion Coefficient (ADC) maps distinguishing progression from pseudoprogression. Arterial Spin Labeling (ASL): Provides non-invasive perfusion imaging, offering insights into tumor vascularity and hypoxia. T2-weighted (T2): Highlights tumor mass effects and associated edema. Dynamic Susceptibility Contrast Perfusion Weighted Imaging (DSC-PWI): Measures cerebral blood volume (rCBV), a key biomarker for distinguishing tumor progression from treatment effects. The dataset is structured to facilitate multimodal analysis, enabling AI-driven research, tumor segmentation, radiomic feature extraction, and treatment response assessment. It includes both raw and
processed imaging data, ensuring compatibility with deep learning models and statistical analyses. By providing a diverse set of imaging biomarkers, the dataset supports efforts in early diagnosis, prognosis prediction, and personalized treatment planning for glioblastoma patients.

**Dataset Collection Method:** Disease-specific

**Dataset Type:** Processed Dataset

**Dataset Terms of Use:** unspecified

**Dataset Intended Purpose:** The primary purpose of this dataset is to advance research in glioblastoma diagnosis, prognosis, and treatment response assessment using advanced imaging techniques and artificial intelligence (AI).

**Imaging Modality:** Magnetic Resonance Imaging

**Vendor:** unspecified

**Imaging body part:** Brain

**Age range:** 18 - 85

**Sex:** Male and Female

**Number of subjects:** Over 50

**Number of DICOM studies:** unspecified

**Image size in GB:** unspecified

**De-identification:** Personal data is included in the images. In this case EUCAIM can support you with specific tools and guidance for de-identification once the application is accepted.

- Data Users:

Table 20. Description of Use Case no. 8 from Aristotelio Panepistimio Thessalonikis

| | | | |
|---|---|---|---|
| Author: Ioanna Chouvarda | Intention:<br>- train/validate AI tools (Data User)<br>- development of AI tools and solutions<br>- training of AI tools and solutions<br>- validation of AI tools and solutions | Organisation's name: Aristotelio Panepistimio Thessalonikis | Organization's Acronym: AUTH |
| **Title of the use case:** Prediction of Treatment Response and Disease Progression in Lung Cancer Patients Using Voxel-Level Radiomics from PET-CT Data | | | |
| **General description of the use case:**<br>Objective: The main objective of this case is to develop an AI model that predicts treatment response and disease progression in lung cancer patients based on PET-CT data. By extracting radiomics features at the voxel level from PET-CT images and training a machine learning (ML) or deep learning (DL) network, the model is expected to provide accurate prognostic information that will help clinicians in personalized treatment planning.<br>Expected results: The AI model is expected to accurately predict treatment response and cancer progression in lung cancer patients based on PET-CT images. Expected performance metrics include high accuracy, sensitivity, specificity and an AUC value above current benchmarks. Expected clinical impact: The proposed AI model aims to overcome the current limitations of methods such as the RECIST criteria by providing an objective, quantitative analysis of tumor characteristics. Clinical benefits include assisting experts in creating personalized treatment plans, increasing efficiency through automation, improving accuracy and optimizing healthcare resources.<br><br>Methodology:<br>• Data Collection and Processing:<br>o Collect PET-CT images and clinical data from collaborating institutions and public databases such as TCIA.<br>o Target population: Patients diagnosed with NSCLC.<br>o Incorporate patient demographics, smoking history, and tumor stage for improved model predictions.<br>• Segmentation and Feature Extraction:<br>o Use provided tumor segmentation and integrate lung segmentation if available.<br>o Extract voxel-level radiomic features: intensity, texture, shape, and potentially peritumoral features.<br>• Data Harmonization and Feature Selection:<br>o Implement image harmonization methods and intensity normalization techniques.<br>o Apply dimensionality reduction and focus on relevant ROIs.<br>• Model Development and Validation:<br>o Develop a deep learning network tailored for voxel-based radiomic data.<br>o Train the model with robust data augmentation and validate with cross-validation and independent test sets.<br>**Expected timeline for the realization of the use case:**<br>Project Timeline:<br>a. Duration: ~1.5 years.<br>b. Milestones:<br>i. data preprocessing, and radiomics M4<br>ii. A simple DL model M8<br>iii. Refined models M12 and preliminary report ready<br>iv. All performance metrices calculated and final report M16<br><br>**Description of the intended use and expected benefit related to the use of the EUCAIM data:**<br>Expected Results:<br>The AI model is expected to predict treatment response and cancer progression in lung cancer patients based on PET-CT images. These will be treated initially as binary problems. Using radiomics features at the voxel level, the model will capture details of tumor heterogeneity, metabolic activity and spatial patterns associated with treatment outcomes.<br>Expected performance metrics include:<br>o Accuracy: High overall correctness of predictions.<br>o Sensitivity: Effective identification of patients who will respond to treatment.<br>o Specificity: Accurate detection of patients who may not respond to treatment.<br>o Area Under the ROC Curve (AUC): Aim for values above the current benchmarks in this range. In addition, standard fairness metrics will be calculated.<br>Expected Clinical Impact:<br>The proposed AI model addresses limitations reported in the know literature by providing an objective, quantitative analysis of tumor characteristics at the voxel level. By analyzing detailed radiomic features, the model can detect subtle changes in tumor biology and provide a quantitative digital biomarker approach.<br>Clinical benefits include:<br>• Support for experts: Gain predictive insights into treatment efficacy, enabling them to create personalized treatment plans.<br>• Increased efficiency: Reduce time spent on image analysis by automating feature extraction and prediction. | | | |

| |
|---|
| • Improve accuracy: More accurate predictions than traditional methods, which can lead to better patient outcomes.<br>• Resource optimization: Enables more efficient allocation of healthcare resources by identifying patients who are likely or unlikely to benefit from certain treatments.<br><br>**Description of the requested data:**<br>Requested Data:<br>• Target Population: Patients diagnosed with NSCLC across different stages.<br>• Type of Data: PET-CT imaging modalities, clinical data, and other relevant data.<br>Datasets:<br>• Selection criteria include PET-CT data with complete annotations.<br>• Number of Cases: the whole availiable dataset(s) donated on the platform.<br>• Annotations: Tumor segmentation and, if available, lung segmentation data. |

Table 21. Description of Use Case no. 6 from Aristotelio Panepistimio Thessalonikis

| | | | |
|---|---|---|---|
| Author: Ioanna Chouvarda | Intention:<br>- train/validate AI tools (Data User)<br>- development of AI tools and solutions<br>- training of AI tools and solutions<br>- validation of AI tools and solutions | Organisation's name: Aristotelio Panepistimio Thessalonikis | Organization's Acronym: AUTH |
| **Title of the use case:** Whole gland radiomics based model for prostate cancer classification | | | |
| **General description of the use case:**<br><br>Main Objectives:<br>Radiomics, a quantitative method for analysing medical images, and AI models have been developed targeting prostate cancer (PCa) based on the analysis of cancer lesions, which is a significant limitation, given that the manual segmentation of which is both resource-intensive and time-consuming, while relying on AI models for segmentation affects robustness due to potential inaccuracies. Automated delineation of the entire prostate gland demonstrates higher accuracy. Moreover, most AI models proposed in the literature are trained on datasets originating from limited clinical sites, with images acquired using specific protocols from scanners that may have varying calibration parameters. Consequently, the data are inherently inhomogeneous, making harmonization a crucial preprocessing step. In the current model, we will apply and test different harmonization techniques to reduce unwanted variation at both the image and feature levels. The main objective is the development of a federated model which is able to identify clinically significant prostate cancer (csPCa) by combining radiomic analysis of MRI images from the entire prostate gland with basic clinical characteristics of the patient. Additionally, we will investigate the significance of the radiomic features from different prostate regions. The ultimate goal is to develop with enhanced generalizability.<br><br>Expected Results:<br>The AI model is expected to detect clinically significant PCa using T2-weighted (T2w) axial, diffusion-weighted imaging (DWI), and apparent diffusion coefficient (ADC) images. This model will build on an already trained model developed using data from an open dataset consisting of 1,500 patients, which demonstrated acceptable performance in identifying csPCa (balanced accuracy = 80%). By integrating data from the EUCAIM repository, which includes images from additional clinical sites and imaging protocols, and by improving data harmonization while considering different regions of the prostate gland, we expect to enhance the model accuracy, robustness, and generalizability.<br><br>Expected Clinical Impact:<br>Histopathological analysis (Gleason score) remains the gold standard for stratifying PCa patients. In contrast, PI-RADS is based on the evaluation of cancer lesions, the detection of which is challenging and requires a specialized radiologist. Identifying non-invasively clinically significant PCa through the analysis of the entire prostate gland has the potential to be more robust. This approach could assist clinical experts in identifying csPCa, improving decision-making, and enhancing patient management.<br><br>Methodology:<br>The core of the AI model will be based on the radiomic analysis of the prostate gland. Additionally, routinely collected clinical characteristics will be integrated. T2w, DWI and ADC will be utilized for model development and validation. The pipeline will build upon an existing methodology, with all its components revisited and improved to leverage the availability of a large dataset from the EUCAIM repository. State-of-the-art AI models will be used for delineating prostate gland. Radiomic features will be extracted not only from the whole gland but also from specific prostate's anatomical regions, such as transition and peripheral zone, using established segmentation algorithms with acceptable performance. While pixel-level harmonization of images in a federated learning context is relatively straightforward, harmonizing radiomic features under these conditions poses significant challenges. To address this, we will investigate various approaches to feature harmonization to address vendor batch effects. Different feature selection methods will be employed, and federated approaches will be applied | | | |

| |
|---|
| using various machine learning models, such as Support Vector Machines and Balanced Random Forests. The model will be trained and tested on distinct data subsets. Data from approximately 10,000 patients, available through the ProCAncer-I project will be used for training and validation. This dataset includes multiparametric MRI images from different planes. Additionally, it contains clinical information such as Gleason score, age and PSA, , depending on the specific use case to which each patient belongs. Segmentation masks are also available.<br><br>**Expected timeline for the realization of the use case:**<br>January 2025 till June 2026 (18 months in total), given the availability of the relevant datasets in EUCAIM platform<br><br>**Description of the intended use and expected benefit related to the use of the EUCAIM data:**<br>Our team has developed an AI model for characterizing prostate cancer based on its clinical significance, utilizing data from open datasets comprising approximately 1,000 patients. The entire pipeline—including prostate gland delineation, radiomic feature extraction, feature harmonization and selection, as well as model training—has already been established. Validation of the model demonstrated a balanced accuracy of around 80%. The entire pipeline developed for this model has been described in the 21st IEEE International Symposium on Biomedical Imaging (ISBI 2024) [1] and will be the basis for the AI model that we aim to develop. The data from bpMRI, collected through the ProCAncer-I project, which includes approximately 10,000 patients and is stored in the EUCAIM repository, will serve as a valuable resource for retraining and validating the already developed model. In this respect, the whole pipeline will be evaluated and modifications will be made, if required, especially related to the harmonization of the data that originate from different clinical sites and acquired from different vendor machines. The ML model will be retrained using both the data from the open dataset and the EUCAIM repository whereas a portion of the latter will be used for the evaluation of the model. The inclusion of additional data from approximately<br>10,000 patients in the EUCAIM repository holds the potential to enhance the robustness of the pipeline, leading to improved AI performance and increased trustworthiness of the model.<br>[1] Filos, D., Fotopoulos, D., Rouni, M. A., & Chouvarda, I. (2024, May). Machine Learning-Based Whole Gland Radiomics Analysis for Prostate Cancer Classification. In 2024 IEEE International Symposium on Biomedical Imaging (ISBI) (pp. 1-5). IEEE.<br>**Description of the requested data:**<br>The requested data include T2-weighted (T2W), diffusion-weighted imaging (DWI) in axial projection, and ADC images from bpMRI. Additionally, clinical characteristics at initial diagnosis, such as PSA levels, Gleason score, and the patient's age, are required. The target population is prostate cancer patients of diverse ages, with the dataset expected to include cases of both clinically significant and clinically insignificant prostate cancer. The requested datasets are ProCAncer-I Use Case 1 and Use Case 2, which are anticipated to contain similar information. To ensure the robustness of the final model, data from scanners by various vendors and diverse clinical sites are expected. Information about the acquisition site of the data would also be appreciated. The requested dataset is expected to include approximately 10,000 patients. Since the model relies on radiomics analysis, segmentation masks of the whole prostate gland would be beneficial. However, for analytical purposes, deep learning algorithms can be applied for gland segmentation, meaning the absence of segmentation files in the dataset will not impede its use. For CPU requirements, a multi-core processor (8+ cores) is recommended to enable parallel processing during feature extraction and model operations. Additionally, 16GB of RAM will be sufficient to handle feature matrices. Regarding storage, only radiomic features will be stored, with the size depending on the dataset's volume; approximately 50–100GB is anticipated. |

Table 22. Description of Use Case no. 3 from Gdanski Uniwerytet Medyczny

| | | | |
|---|---|---|---|
| Author: Michał Kosno | Intention:<br>- train/validate AI tools (Data User)<br>- development of AI tools and solutions<br>- training of AI tools and solutions<br>- validation of AI tools and solutions | Organisation's name: Gdanski Uniwerytet Medyczny | Organization's Acronym: GUMed |
| **Title of the use case:** Breast Density Prediction Algorithm - Application of Multiple Deep Neural Network Based Model for Automated Breast Density Classification | | | |
| **General description of the use case:**<br><br>Main objectives:<br>The accurate evaluation of breast density from mammography images is critical for the accurate estimation of breast cancer risk, as higher breast density is associated with an increased likelihood of developing the disease. However, the current process for visually assessing breast density remains highly challenging and subjective, with considerable inter observer variability, even among trained radiologists. The inherent variability and errors in visual assessments underscore the need for a standardized, objective approach to breast density classification. Recent advances in deep learning methodologies offer promising solutions by providing tools capable of objective and reproducible assessment. Our research specifically aims to enhance breast density assessment by leveraging the Tree-structured Parzen Estimator (TPE) algorithm-driven transfer learning to optimize ResNet, DenseNet and EfficientNet convolutional neural networks (CNNs) for this task. These deep | | | |

learning architectures are widely recognized for their high performance on medical imaging tasks, with DenseNet's connected pathways and EfficientNet's optimized scaling enabling high accuracy in feature extraction and classification. Our ultimate objective is to develop and rigorously validate an artificial intelligence-based (AI) tool that can automatically and accurately classify breasts by their density based on digital mammography. Our methodology ensures the development of a robust and reliable classification tool that performs well across a range of clinical environments. Initial tests have demonstrated that the TPE-driven transfer learning approach improves model performance in breast density classification. However, further validation on a larger and more diverse dataset is essential to confirm these findings and maximize the model's robustness across demographic and device variations.

Expected results:
The expected outcome of our research is a fully validated, AI-driven tool capable of accurately classifyingbreast density in mammographic images across multiple clinical settings. The tool will be adaptable to data from various scanners, institutions, and hospitals, ensuring generalizability and robustness. A key criterion for success is the model's ability to classify breast density consistently regardless of the imaging equipment or patient demographics involved, enabling widespread applicability. The validation process, conducted with an expansive dataset, will provide conclusive evidence of the tool's reliability, paving the way for its clinical integration as a supplementary tool for radiologists.

Clinical impact:
Once validated, our AI-based breast density classification tool could have a transformative impact on clinical workflows and outcomes. By providing radiologists with a reliable, automated solution, the tool will support more efficient interpretation of mammography images and minimize observer-dependent errors. This improvement in workflow efficiency could also lead to faster diagnosis and treatment planning, improving patient outcomes. In addition, an objective, automated breast density assessment tool could significantly contribute to personalized risk assessment models, in line with the movement toward individualized cancer prevention strategies.

Methodology:
The study will utilise data collected in previous research projects (EuCanImage, INCISIVE, CHAIMELEON) and aggregated on the EUCAIM research platform. The data collected in the EUCAIM platform will enable us to validate our algorithm on a large dataset, ensuring the correct functioning of our tool and its independence from any constraints imposed by patient ethnicity or the devices used. So far we have built an ensemble model which employs ResNet50, DenseNet121 and Efficientnet_b0 with the following metrics: AUC (0.99), accuracy (0.91), F1-score (0.91), and recall values (0.90) for the test dataset. We intend to improve the effectiveness and reliability of our model by training it on a large dataset containing normal, benign, and malignant cases and testing it on a diverse dataset containing images with different stages of lesions.

Detailed steps:
- Selection of datasets that will contribute as additional training data
- Refining an ensemble model
- Selection of datasets that will serve as a external validation
- Results analysis in the context of FAIRNESS and robustness

**Expected timeline for the realization of the use case:**
1 year
M1-M3: data processing and preparation
M4-M6: refinement of the algorithm
M7-M9: validation
M10-M12: results analysis

**Description of the intended use and expected benefit related to the use of the EUCAIM data:**
The expected benefit of the EUCAIM platform is to demonstrate the usability of the aggregated datasets for clinical practice. An intended use of the algorithm is to enhance the ability to correctly predict breast density, the critical step for further diagnosis. It will serve to show the user pathway through the project towards task 7.2 objectives and report.

**Description of the requested data:**
In order to achieve the project's objectives, it is essential to obtain a comprehensive and robust dataset comprising multi-centric mammographic examinations. This dataset will primarily consist of DICOM data. The inclusion of mammographic data from multiple centres and a diverse array of patient backgrounds is crucial to capture the inherent variability in populations across different ethnicities, geographic locations, and clinical environments. By doing so, we can construct a dataset that is both clinically relevant and sufficiently representative to support meaningful analysis, model development, and validation. The target population for this dataset is adult women of diverse ethnic backgrounds, aged 18 and above. By focusing on a broad age range, it is possible to obtain data relevant to varying risk levels, breast densities, and cancer incidences. By targeting multiple ethnic groups, the aim is to address healthcare disparities and ensure that predictive models are applicable across demographics. Additionally, the study will encompass both healthy individuals and those with

| |
|---|
| various stages of breast disease, including benign and malignant conditions, to ensure a comprehensive representation. In summary, the dataset should include normal, benign and malignant cases - a real life distribution or a distribution with equal numbers of patients in each class would be acceptable. To build a statistically meaningful dataset, we aim to collect at least 10,000 cases, with a minimum of 1,000 cases per major ethnic group. Each case will ideally represent various stages of breast disease ranging from normal to malignant. To process such a complex dataset we intend to employ a federated learning approach provided by the EUCAIM platform. It is essential that the imaging data contains labels representing breast density. |

Table 23. Description of Use Case no. 4 from Philips GMBH

| Author: Jose Alejandro Matute Flores | Intention:<br>- train/validate AI tools | Organisation's name: Philips GMBH | Organization's Acronym: Philips |
|---|---|---|---|
| **Title of the use case:** Breast MRI Screening - Background Parenchyma Enhancement (BPE) and lesion localization | | | |
| **General description of the use case:**<br><br>Description: Women with high risk have a breast MRI as part of their breast screening. In breast screening, an early diagnosis is paramount. Two challenges for early diagnosis on breast MRI are small lesion detection and BPE. The lesions may be typically small <5mm and easy to miss. An algorithm for small lesion detection could facilitate the workflow and help detect missed lesions. Similarly BPE is defined qualitatively. Imaging protocol variation leads to difficulties in defining a quantitative metric leading to site-specific definitions. A qualitative metric could help on prognostic power for cancer diagnosis. Main Objectives: Quantitative Metric for BPE and algorithm small lesion localization. Expected Clinical<br>Impact: Standardization of BPE would lead to better prognosis. And Small lesions would lead to a faster workflow.<br>Methodology: AI/ML<br><br>**Expected timeline for the realization of the use case:**<br>"Initial data collection - Month: 0-3<br>Initial BPE quantification - Month: 2-5<br>2nd Wave data collection + initial lesion annotation - Month: 5-8<br>Initial lesion model training - Month: 7-10<br>3rd data wave collection + lesion annotation: Month - 9-13<br>Federated training of lesion models – Month: 12-15<br>Statistic analysis – Month: 14-16<br>Final report - Month: 15-18 months"<br><br>**Description of the intended use and expected benefit related to the use of the EUCAIM data:**<br>BPE: Imaging protocols may vary widely from institution to institution. EUCAIM can help complement with enough variability.<br>Small Lesion: During screening, small lesions would be the minority in ~10% of cases. EUCAIM may help expand the data needed for training a segmentation algorithm.<br>AI Algorithm: Unet variations (F-net, nnUnet). For BPE combination of tissue segmentation and regression approaches.<br><br>**Description of the requested data:**<br>Data: From HR&C... MRI images + Case Reports, Population: Adult Females with High risk ( family history \| existence of genetic mutation) with exclusion of personal breast cancer. The study will be prospective. Mid-2024 onwards. Tools/Computational Resources/Temporary Storage: Tooling as provided by EUCAIM. Initial Model to be trained internally, fine-tuned within EUCAIM. | | | |

# External open calls

- Data Holders:

Table 24. Description of Use Case no. 22 from Oslo Universitetssykehus HF

| Author: Kyrre E. Emblem | Data sharing: Federated Node | Organisation's name: Oslo Universitetssykehus HF | Organization's Acronym: OUS | Tier: 2 |
|---|---|---|---|---|
| **General description of the potential use and clinical impact of the shared data:**<br>We provide a federated node for hosting local data using available resources in accordance with EUCAIM's interfaces and specifications. This will also include registering prepared datasets of specific projects in the EUCAIM's Metadata catalogue, as well as offering local processing capabilities (with available resources). We have designed our own database management system with S3, RDBMS, Graph and Document capabilities. MedQuery is designed to optimize SQL and Cypher queries, making use of scalable analysis components written in Golang and Python. As a result, we do not need to create data marts. MedQuery is cloud native and will be deployed to an on-prem K8s distribution. Our data warehouse infrastructure as-is focuses on imaging (radiology). The data policy is disease agnostic. We frequently add new de-identified data to the database. This data includes processed data, as well as data from newly available sources. Q&A is performed on scans to assess data corruption, identify distortion, and remove low-quality data. We store image segmentations as byte arrays of int8. Data can be exported to any required format (DICOM, nifti, etc). We structure and describe our data according to the principles and nomenclature of the Brain Imaging Data Structure (BIDS) initiative. The data warehouse reporting and visualization infrastructure is developed by us on-site. NeoSeg - our in- house developed medical image and segmentation viewer equivalent to online tools like ITK-Snap. The platform has a data viewer with a search engine. You can also trigger AI processing from the platform and run data analysis on selected data. You can generate structured analysis reports from the platform as well. We will ensure data quality and integrity and comply with privacy and security regulations to facilitate collaborative data sharing and analysis within the federated system. If our institution becomes a EUCAIM federated node, we will sign the required Data Sharing Agreement (DSA). As a federate node, we will retain autonomous control over the data and the ability to oversee it. Most users obtain access to the data only through the platform. Their data access is governed by a granulated access control system with three levels, which includes access control based on research projects. You are only allowed to see the de- identified data that belongs to your system administration level and the research project(s) you are part of. The activities presented in this section reflect key visions at Oslo University Hospital. These include conducting medical research of very high international caliber, translational research to directly increase medical competence, bridge the gap between basal and clinical milieus and high priority to research on non-invasive diagnostic methods. The latter is considered more cost effective and patient friendly compared to invasive alternatives following surgery, especially for long-term follow-up. The knowledge gained from these projects will have translational value to most forms of solid cancer. High-quality biomarkers are critical for early patient response assessment and correct therapeutic 7 decision-making. Our focus on method standardization and start-to-end automatic procedures as an alternative to laborious and user-dependent methods will benefit cancer patients today, and tomorrow. Our projects benefit from the truly multidisciplinary expertise at Oslo University Hospital, including medicine, physiology, bioengineering, optics, mathematics, physics and surgical oncology. The image analysis tools are a direct aid to the treating physician and radiologist, allowing pioneering personalized and longitudinal (i.e. visit-by-visit) comparisons of our advanced imaging data. We have a tight connection with national and international oncological and imaging milieus to help expand new and relevant knowledge on cancer imaging biomarkers. Here, project members are used extensively for advisory boards, education programs, invited key-note presentations and as hosts for international seminars. | | | | |
| **Dataset 1**<br><br>**Cancer Type:** Primary brain cancer<br><br>**Dataset name:** SAILOR<br><br>**Dataset description:**<br>This longitudinal dataset contains human treatment data of 27 patients with high-grade glioma: Diagnostic MRI exams prior to surgery and right after, followed by exams from three to 19 time points during chemoradiotherapy | | | | |

(CRT) and chemotherapy (temozolomide: TMZ). Patients are in the age range 32–68 years, median age 56 years, female/male ratio of 8/19, and with 19 months median overall survival.

**Dataset Collection Method:** Longitudinal

**Dataset Type:** Processed dataset

**Dataset Terms of Use:** The dataset is accompanied by suitable metadata elements to ensure they meet the project's objectives and align with best practices in cancer research and data sharing. The dataset is de-identified (for federated data nodes).

**Dataset Intended Purpose:** A potential use of this data is to study both treatment effects and disease progression on an individual patient level and group level, as seen through clinical data, histology, and multi-sequence MRI, using both structural/anatomical and functional/physiological MRI.

**Imaging Modality:** MR

**Vendor:** Philips

**Imaging body part:** Neuro (head)

**Age range:** 36 / 68 years

**Sex:** female/male ratio of 8/19

**Number of subjects:** 27

**Number of DICOM studies:** 300

**Image size in GB:** unspecified

**De-identification:** Personal data is de-identified (direct identifiers have been removed)

**Dataset 2**

**Cancer Type:** Brain metastases from lung cancer and malignant melanoma

**Dataset name:** TREATMENT

**Dataset description:**
TREATMENT is an observational study addressing the need for knowledge and adequate diagnostic biomarkers in the response assessment of patients with brain metastasis. Reliable response assessment will be highly relevant in the coming years given the introduction of next-generation cancer drugs, including immunotherapy. This project uses advanced Magnetic Resonance Imaging (MRI) and Vessel Architecture Imaging (VAI) to better understand the response to traditional stereotactic radiosurgery (SRS) and immunotherapy.
This longitudinal dataset contains life-long longitudinal imaging data every third month of 70 patients with brain metastases from lung cancer and malignant melanoma, prior to, during, and after stereotactic radiosurgery (SRS) and chemotherapy and/or immunotherapy. Study also includes 50 control patients receiving baseline scans (prior to therapy) and follow-up at 6 months and one year. Patients are in the age range 18–70 years.
Please see study details: https://www.clinicaltrials.gov/study/NCT03458455

**Dataset Collection Method:** Longitudinal

**Dataset Type:** Processed Dataset

**Dataset Terms of Use:** The dataset is accompanied by suitable metadata elements to ensure they meet the project's objectives and align with best practices in cancer research and data sharing. The dataset is de-identified (for federated data nodes).

**Dataset Intended Purpose:** A potential use of this data is to study both treatment effects and disease progression on an individual patient level and group level, as seen through clinical data, histology, and multi-sequence MRI, using both structural/anatomical and functional/physiological MRI.

**Imaging Modality:** MRI

**Vendor:** Siemens, Philips, GE Healthcare

**Imaging body part:** Neuro (Head)

**Age range:** 18 / 70 years

**Sex:** Approx. 60% (female) / 40% (male)

**Number of subjects:** 120

**Number of DICOM studies:** 300

**Image size in GB:** unspecified

**De-identification:** Personal data is de-identified (direct identifiers have been removed)

**Dataset 3**

**Cancer Type:** Brain tumors

**Dataset name:** BROADEN

**Dataset description:**
Brain tumor genotyping and AI-based imaging phenotyping to optimize service and augment discovery in pediatric neuro-oncology. This is a national project aiming to include all relevant pediatric patients diagnosed with a brain tumor from 2005 to 2022. Imaging is performed at baseline and then up until every third month depending on tumor type and patient status for pediatric patients suspected of a brain tumor and if possible, a biopsy is performed using methylation arrays to classify the tumor according to the WHO classification of 2021. In BROADEN we want to establish associations between genetic markers and serial brain MRI metrics to predict targets or to predict more about the tumor biology. This could be especially important in cases where biopsy is not possible due to location, tumor size, or age of the child. There are several aspects relevant for AI models, including predicting relapses earlier than by traditional means, or to predict those not at risk of relapse. For patients with high-grade malignant tumors, relapse treatment could start earlier to reduce mutilating surgery and allow for smaller radio-therapeutic fields.

**Dataset Collection Method:** Longitudinal

**Dataset Type:** Processed dataset

**Dataset Terms of Use:** The dataset is accompanied by suitable metadata elements to ensure they meet the project's objectives and align with best practices in cancer research and data sharing. The dataset is de-identified (for federated data nodes).

**Dataset Intended Purpose:** A potential use of this data is to study both treatment effects and disease progression on an individual patient level and group level, as seen through clinical data, histology, and multi-sequence MRI, using both structural/anatomical and functional/physiological MRI.

**Imaging Modality:** MRI

**Vendor:** Philips, Siemens, GE Healthcare

| |
|---|
| **Imaging body part:** Neuro (head)<br><br>**Age range:** 0 / 17 years<br><br>**Sex:** Male and Female<br><br>**Number of subjects:** Up to 1000<br><br>**Number of DICOM studies:** 1500<br><br>**Image size in GB:** unspecified<br><br>**De-identification:** Personal data is de-identified (direct identifiers have been removed) |

Table 25. Description of Use Case no. 6 from University of Latvia

| Author: Marcis Leja | Data sharing: Federated Node | Organisation's name: University of Latvia | Organization's Acronym: LU | Tier: 1 |
|---|---|---|---|---|
| **General description of the potential use and clinical impact of the shared data:** AI is expected to support the pathologist in the future either by screening images with normal morphology or by identifying lesions (such as dysplasia) still possessing high interobserver variability among the pathologists. This could have a particular importance with limited pathology resources where the expertise in digestive pathology is insufficient. At the same time, this has to be emphasized, that precise detection of the high-risk lesions is of particular importance, since the intervals in surveillance strategies would depend on proper classification of the lesions (see MAPS II, DOI: 10.1055/a-0859-1883). As a result, AI is expected to change the pathology practice substantially in the future, and the current project is expected to contribute to this change in upper digestive pathology. The data set will consist of whole slide images (WSI) of gastric tissue collected either during surgery (mainly cancer tissue) or by taking biopsies during endoscopies. The images will cover normal tissue and various changes in gastric mucosa (atrophy, intestinal metaplasia, dysplasia) that can lead to the development of cancer, as well as cancer itself. Therefore, this data set can be used to train AI models for various tasks related to detection of high-risk lesions (these patients in clinical settings are then followed and monitored to detect gastric cancer development early) and cancer. Additional information will be provided on the status of Helicobacter pylori (a known carcinogen) and use of PPI which can affect the states. Some examples of potential use cases: • AI models that detect cancer and precancerous lesions and assist pathologists by pointing their attention to regions where the pathologies are detected;<br>• AI models that not only detect the pathologies but also quantify the severity, e.g., stages of atrophy, intestinal metaplasia, dysplasia.<br>• AI based pathologist training systems that evaluate the work of students/pathologists and explain the reason for marking a region and the corresponding label by applying methods of explainable AI. The images have high resolution and are large in size, therefore the usual practice when training AI models is to use smaller tiles, and this data set provides several hundred to thousands of training samples, depending on the size of tiles. However, it might be useful to incorporate the context (surrounding areas) of the tiles during the training process for better results in some cases. | | | | |
| **Dataset 1**<br><br>**Cancer Type:** Gastric cancer<br><br>**Dataset name:** Gistar histopathology (gastric cancer and precancerous lesions) | | | | |

| |
|---|
| **Dataset description:**<br>Histopathology is a diagnostics gold standard for gastric cancer and precancerous conditions. Traditionally it is performed by trained professionals, who jointly evaluate the same slides to form a consensus opinion. For borderline and high risk cases consensus of more than one expert is optimal due to possible interobserver disagreement (i.e., see Lage et al. 2016, DOI: 10.1016/j.bpg.2016.09.004), therefore AI-based digital pathology tools that suggest the location and type of pathology may be helpful for guiding the pathologists. Furthermore, for gastric cancer prevention, it is important to accurately determine precancerous conditions, because it is preceded by a cascade of precancerous lesions detectable in histopathology, i.e., atrophy, intestinal metaplasia, dysplasia (Correa & Piazuelo 2012, DOI: 10.1111/j.1751- 2980.2011.00550.x). Therefore, the proposed data set will include not only images of cancer but also these mentioned precancerous lesions. The data set will include the following images, their corresponding metadata and annotations:<br>1. 100 slides with gastric cancer (including intestinal and diffuse cancer),<br>2. 100 slides with dysplasia,<br>3. 100 slides with intestinal metaplasia,<br>4. 100 slides with atrophy,<br>5. 100 slides with normal tissue.<br>Therefore, this collection will cover the whole cascade of gastric cancer development and would allow to train different AI models for precancerous lesions and cancer detection.<br><br>**Dataset Collection Method:** Disease-specific<br><br>**Dataset Type:** Annotated Dataset<br><br>**Dataset Terms of Use:** For training/validation of AI within the EUCAIM consortium. The source of the original data will have to be referenced in publications and similar results.<br><br>**Dataset Intended Purpose:** Training and validation of AI to detect gastric precancerous lesions and gastric cancer to assist pathologists (as intended within EUCAIM framework).<br><br>**Imaging Modality:** Slide Microscopy (SM)<br><br>**Vendor:** 3DHistech<br><br>**Imaging body part:** Stomach mucosa<br><br>**Age range:** 35 / 87 years<br><br>**Sex:** Male and Female<br><br>**Number of subjects:** ~300-500 (there are cases with more than one slide per person; the collection will include the best quality slides and it is difficult to predict the number of subjects)<br><br>**Number of DICOM studies:** unspecified<br><br>**Image size in GB:** ~1 GB per image<br><br>**De-identification:** Personal data is fully anonymized |

Table 26. Description of Use Case no. 8 from Fundación de Investigación HM Hospitales

| | | | | |
|---|---|---|---|---|
| Author: Adrián Pelaéz Laderas | Data sharing: Central repository | Organisation's name: Fundación de Investigación HM Hospitales | Organization's Acronym: HM Hospitales | Tier: 1 |
| **General description of the potential use and clinical impact of the shared data:**<br>The potential use and clinical impact of the shared data will not only improve the sample size for training artificial intelligence (AI) tools, but also enrich the clinical context by incorporating images from multiple sites and diverse populations. 1. Augmented AI Training and Validation By sharing our medical imaging data within the EUCAIM | | | | |

framework, we substantially increase the sample size available for training and validating AI tools. The diversity of our data, which includes images from multiple sites and a broad range of patient demographics, will ensure that the AI models are robust and generalizable. This enhanced dataset will allow for the creation of more accurate predictive models, reducing biases and improving the reliability of AI applications in real-world clinical settings. 2. Improved Diagnostic Accuracy The integration of our data into EUCAIM's data lake will enable the development of AI-driven diagnostic tools that can identify subtle patterns and anomalies in medical images with greater precision. These tools have the potential to assist radiologists and oncologists in detecting cancer at earlier stages, where treatment outcomes are typically more favorable. Early and accurate diagnosis can significantly impact patient prognosis, leading to increased survival rates and better quality of life for cancer patients. 3. Personalized Treatment Plans The comprehensive dataset resulting from our collaboration will facilitate the creation of AI models that can predict individual responses to various treatments. By analyzing a wide array of imaging data alongside clinical and genomic information, these models can help tailor personalized treatment plans. This precision medicine approach ensures that patients receive the most effective therapies based on their unique characteristics, minimizing adverse effects and maximizing therapeutic efficacy. 4. Data Lake Infrastructure for Future Projects In addition to immediate research and clinical benefits, this project will aid in the development of a robust data lake infrastructure. This infrastructure will be instrumental for future research endeavors, allowing for seamless integration and analysis of new datasets as they become available. The data lake will support ongoing and future AI initiatives, fostering continuous innovation and improvement in cancer care.

**Dataset 1**

**Cancer Type:** Liver cancer

**Dataset name:** Liver images

**Dataset description:**
Set of images of patients with liver cancer between 2015-09-23 and 2024-05-20 collected in our centers.

**Dataset Collection Method:** Disease-specific

**Dataset Type:** Original Dataset

**Dataset Terms of Use:** Unspecified

**Dataset Intended Purpose:** The purpose of this dataset is to support the development, validation, and implementation of AI tools for cancer diagnosis and treatment

**Imaging Modality:** MRI,CT, PET-CT, PET and MRI, Mammography

**Vendor:** Unspecified

**Imaging body part:** Stomach mucosa

**Age range:** 4 / 102 years

**Sex:** Male and Female

**Number of subjects:** 3513

**Number of DICOM studies:** 21266

**Image size in GB:** ~1 GB per image

**De-identification:** Personal data is pseudonymized

**Dataset 2**

**Cancer Type:** Pancreatic cancer

**Dataset name:** Pancreas images

**Dataset description:**
Set of images of patients with pancreas cancer between 2015-12-10 and 2024-05-14 collected in our centers.

**Dataset Collection Method:** Disease-specific

**Dataset Type:** Original Dataset

**Dataset Terms of Use:** Unspecified

**Dataset Intended Purpose:** Clinical and radiological file

**Imaging Modality:** MR, CT, PET-CT, PET-MR, Mammography

**Vendor:** Unspecified

**Imaging body part:** Unspecified

**Age range:** 24 / 102 years

**Sex:** Male and Female

**Number of subjects:** 1216

**Number of DICOM studies:** 9877

**Image size in GB:** unspecified

**De-identification:** Personal data is pseudonymized

**Dataset 3**

**Cancer Type:** Lung cancer

**Dataset name:** Lung  images

**Dataset description:**
Set of images of patients with lung cancer between 2015-09-23 and 2024-05-21 collected in our centers.

**Dataset Collection Method:** Disease-specific
**Dataset Type:** Original Dataset

**Dataset Terms of Use:** Unspecified

**Dataset Intended Purpose:** Clinical and radiological file

**Imaging Modality:** MR, CT, PET-CT, PET-MR, Mammography

**Vendor:** Unspecified

**Imaging body part:** Unspecified

**Age range:** 0 / 102 years

**Sex:** Male and Female

**Number of subjects:** 4345

**Number of DICOM studies:** 28131

**Image size in GB:** unspecified

**De-identification:** Personal data is pseudonymized

**Dataset 4**

**Cancer Type:** Breast cancer

**Dataset name:** Breast images

**Dataset description:**
Set of images of patients with breast cancer between 2015-12-29 and 2024-05-20 collected in our centers.

**Dataset Collection Method:** Disease-specific

**Dataset Type:** Original Dataset

**Dataset Terms of Use:** Unspecified

**Dataset Intended Purpose:** Clinical and radiological file

**Imaging Modality:** MR, CT, PET-CT, PET-MR, Mammography

**Vendor:** Unspecified

**Imaging body part:** Unspecified

**Age range:** 0 / 102 years

**Sex:** Male and Female

**Number of subjects:** 3808

**Number of DICOM studies:** 8090

**Image size in GB:** unspecified

**De-identification:** Personal data is pseudonymized

**Dataset 5**

**Cancer Type:** Prostate cancer

**Dataset name:** Prostate images

**Dataset description:**
Set of images of patients with prostate cancer between 2015-12-30 and 2024-05-20 collected in our centers.

**Dataset Collection Method:** Disease-specific

**Dataset Type:** Original Dataset

| |
|---|
| **Dataset Terms of Use:** Unspecified<br><br>**Dataset Intended Purpose:** These reports may be accompanied by any biochemical or pathological anatomy results taken at our center.<br><br>**Imaging Modality:** MR, CT, PET-CT, PET-MR, Mammography<br><br>**Vendor:** Unspecified<br><br>**Imaging body part:** Unspecified<br><br>**Age range:** 3 / 102 years<br><br>**Sex:** Males<br><br>**Number of subjects:** 1259<br><br>**Number of DICOM studies:** 4701<br><br>**Image size in GB:** unspecified<br><br>**De-identification:** Personal data is pseudonymized |

Table 27. Description of Use Case no. 48 from CETIR Centre Medic SL

| Author: Jorge Pastor Peidro | Data sharing: Central repository | Organisation's name: CETIR Centre Medic SL | Organization's Acronym: CETIR | Tier: 2 |
|---|---|---|---|---|
| **General description of the potential use and clinical impact of the shared data:** CETIR Centre Mèdic presents comprehensive cohorts of cases across multiple cancer types, including breast, prostate, lung, pancreatic, brain (gliomas, glioblastomas), kidney, and uterine cancers. This rich dataset is poised to significantly enhance the EUCAIM project by providing valuable insights and fostering advancements in cancer research, diagnosis, and treatment. The clinical impact of sharing this data can be profound, addressing several critical areas:<br>1. Enhanced Diagnostic Accuracy: Integrating diverse datasets from various cancer types will allow EUCAIM to develop and refine AI algorithms, improving diagnostic precision. Machine learning models trained on these comprehensive datasets will recognize patterns and anomalies more effectively, leading to earlier and more accurate diagnoses. This is particularly crucial for aggressive cancers like glioblastoma and pancreatic cancer, where early detection can significantly influence patient outcomes.<br><br>2. Personalized Treatment Plans: The detailed clinical and imaging data will facilitate the development of personalized treatment strategies. By analyzing tumor characteristics and patient responses to different therapies, the data can help identify biomarkers predictive of treatment success. This personalization can optimize treatment efficacy, reduce adverse effects, and improve overall patient prognosis.<br><br>3. Advancements in Radiomics and Radiogenomics: Integrating imaging data with genomic information can lead to groundbreaking advancements in radiomics and radiogenomics. This approach can uncover novel correlations between imaging phenotypes and genetic profiles, leading to new diagnostic biomarkers and therapeutic targets. It will also support the development of non-invasive diagnostic techniques, reducing the need for biopsies and invasive procedures.<br><br>4. Facilitating Multi-Center Studies and Clinical Trials: Sharing data across multiple institutions enhances the feasibility of large-scale, multi-center studies and clinical trials. This collaborative approach can accelerate the validation of new diagnostic tools and treatment protocols, ensuring findings are robust, generalizable, and applicable to diverse patient populations. | | | | |

5. Improving Prognostic Models: The comprehensive datasets will enable the development of sophisticated prognostic models that predict disease progression and patient outcomes with high accuracy. These models can guide clinical decision-making, helping clinicians identify high-risk patients and tailor follow-up care accordingly.

6. Supporting Health Policy and Resource Allocation: The insights derived from this data can inform health policy and resource allocation, ensuring that healthcare systems are better prepared to manage the burden of cancer. Policymakers can leverage this information to design effective screening programs, allocate resources efficiently, and implement preventive measures.

7. Educational and Training Resource: The dataset will serve as a valuable educational and training resource for medical professionals. It can be used to develop training modules, case studies, and simulation exercises that enhance the skills and knowledge of healthcare providers, ensuring they are equipped with the latest advancements in cancer diagnosis and treatment.

8. Promoting Innovation and Research: By making this data accessible to the research community, CETIR Centre Mèdic encourages innovation and the development of novel diagnostic and therapeutic approaches. It fosters a collaborative research environment, enabling scientists to build upon existing knowledge and explore new frontiers in oncology. In summary, the data shared by CETIR Centre Mèdic holds the potential to revolutionize cancer care by enhancing diagnostic accuracy, personalizing treatment, advancing research in radiomics and radiogenomics, facilitating large-scale studies, improving prognostic models, informing health policy, serving as an educational resource, and promoting innovation. This comprehensive approach aligns with the goals of the EUCAIM project and promises to make a significant impact on the global fight against cancer.

**Dataset 1**

**Cancer Type:** Breast, prostate & uterine cancers

**Dataset name:** Hormone-related Dataset

**Dataset description:**
A dataset focusing on hormone-related tumors, encompassing prostate, breast, and uterine cancers, including 15,600 cases, providing imaging data for diagnostic and therapeutic research. Dataset Components 1. Prostate Cancer: Number of Cases: 2,000 Imaging Modalities: Multiparametric MRI (mpMRI) and PET-CT scans. Description: The dataset includes mpMRI and PET-CT images capturing prostate anatomy and metabolic activity, essential for accurate staging and treatment planning. 2. Breast Cancer: Number of Cases: 12,000 Imaging Modalities: Mammography: 10,000 cases Dynamic Breast MRI: 2,000 cases Description: This component features a large volume of mammography images and dynamic MRI scans, annotated for various breast cancer types and stages, facilitating robust model training for diagnosis and prognosis. 3. Uterine Cancer: Number of Cases: 1,600 Imaging Modalities: MRI and PET-CT scans. Description: The uterine cancer dataset includes MRI and PET-CT images, offering comprehensive views of tumor morphology and metabolic profiles, aiding in precise assessment and treatment strategies. Data Quality and Annotation Expert Annotations: All images are annotated by specialized radiologists, ensuring high accuracy and reliability. The annotations include tumor characteristics, staging information, and treatment responses. Standardization: The dataset follows standardized imaging protocols and quality control measures to ensure consistency across different modalities and patient groups. Potential Impact Enhanced Diagnostic Accuracy: The diverse imaging data allows for the development of sophisticated AI models that improve the accuracy of cancer detection and staging. Personalized Treatment: Detailed imaging and clinical data support personalized treatment plans, optimizing therapeutic outcomes and minimizing side effects. Research and Innovation: This dataset provides a valuable resource for researchers to explore new imaging biomarkers and validate AI algorithms, driving innovations in oncology. The hormone-related tumors dataset, comprising 15,600 cases across prostate, breast, and uterine cancers, offers high-quality, annotated imaging data. This dataset will contribute to the EUCAIM project by enhancing diagnostic accuracy, supporting personalized treatment, and fostering research innovations.

**Dataset Collection Method:** Disease-specific

**Dataset Type:** Annotated Dataset

**Dataset Terms of Use:** CETIR Centre Mèdic is committed to providing high-quality, fully anonymized datasets for the EUCAIM project. The datasets, which include imaging data and clinical reports from various cancer types, have been meticulously anonymized to ensure patient privacy and compliance with ethical standards. Below are

the detailed terms of use for accessing and utilizing these datasets. Anonymization and Privacy Full Anonymization: All datasets have undergone a rigorous anonymization process. Personal identifiers have been removed, and each entry has been assigned a unique code to ensure that individual patients cannot be identified. Compliance with Standards: The anonymization process adheres to international privacy standards, including the General Data Protection Regulation (GDPR). This ensures that all data sharing and usage comply with legal and ethical requirements, maintaining the highest level of data security and patient confidentiality. Access and Usage Access Permissions: Access to the datasets will be granted to researchers and institutions that are part of the EUCAIM project. All users must agree to the terms of use and demonstrate a legitimate research purpose. Non-Commercial Use: The datasets are provided strictly for non-commercial research purposes. Any commercial use of the data is prohibited unless explicitly authorized by CETIR Centre Mèdic. Research Purpose: The data must be used solely for the purpose of advancing medical research, improving diagnostic tools, and developing personalized treatment strategies. Any other use of the data must be approved in writing by CETIR Centre Mèdic. Data Security and Management Data Security: Users must ensure that the datasets are stored securely and access is restricted to authorized personnel only. Appropriate measures, such as encryption and access controls, must be implemented to protect the data from unauthorized access or breaches. Data Integrity: Users are responsible for maintaining the integrity of the datasets. Any modifications or annotations made to the data must be documented, and the original data must remain unchanged. Ethical Considerations Ethical Compliance: All research conducted using the datasets must comply with ethical standards and guidelines. Researchers must obtain necessary approvals from their institutional review boards (IRBs) or ethics committees before using the data. Publication and Sharing: Any publications or presentations resulting from the use of these datasets must acknowledge CETIR Centre Mèdic as the data source. Additionally, researchers are encouraged to share their findings with the broader scientific community to promote transparency and collaboration.

**Dataset Intended Purpose:** The datasets provided by CETIR Centre Mèdic are intended to advance oncology research, improve diagnostic and prognostic tools, develop personalized treatment strategies, and support multi-center collaborations. They also play a crucial role in training healthcare professionals and informing health policy decisions. By utilizing these high-quality and fully anonymized datasets, the EUCAIM project can significantly enhance cancer research and patient care, ultimately leading to better outcomes for patients worldwide.

**Imaging Modality:** CT, PET, MRI, & Mammograms

**Vendor:** GE, Philips, Siemens

**Imaging body part:** Breast, abdomen, pelvis

**Age range:** 19 / 90 years

**Sex:** Male and Female

**Number of subjects:** 15600

**Number of DICOM studies:** Unspecified

**Image size in GB:** ~1 GB per image

**De-identification:** Personal data is fully anonymized

**Dataset 2**

**Cancer Type:** Colon & rectal cancer

**Dataset name:** Colon & rectal Dataset

**Dataset description:**
The colorectal cancer dataset provided by CETIR Centre Mèdic consists of fully anonymized images and reports from 1,600 cases of colon and rectal cancer. This dataset includes imaging modalities such as MRI and CT scans, capturing a comprehensive range of cases with various follow-ups and diagnoses. Data Acquisition Source: The images and reports were collected from patients who underwent diagnostic imaging at CETIR

Centre Mèdic. The data spans several years, ensuring a diverse and representative collection of colorectal cancer cases. Imaging Protocols: Standard imaging protocols were used for MRI and CT scans, ensuring consistency and high-quality data. This includes high-resolution images for both initial diagnoses and follow-up assessments. Data Anonymization Process: A thorough anonymization process was implemented to protect patient privacy and comply with ethical standards. All personal identifiers were removed, and unique codes were assigned to each dataset entry. Compliance: The anonymization process complies with international standards, including the General Data Protection Regulation (GDPR), ensuring data security and patient confidentiality. Data Components Images: The dataset includes high-resolution MRI and CT images for each colorectal cancer case. These images provide detailed visualization of tumor characteristics, essential for accurate diagnosis and treatment planning. Reports: Accompanying the images are detailed radiology reports prepared by expert radiologists. These reports contain critical clinical information, including tumor size, location, staging, and follow-up observations. Potential Impact Follow-Up Data: The inclusion of follow-up imaging studies allows for the analysis of tumor progression and treatment responses, contributing to more accurate prognostic models and personalized treatment plans. Diverse Diagnoses: The dataset includes a range of diagnoses and stages, making it a valuable resource for developing and validating diagnostic and prognostic tools.

**Dataset Collection Method:** Longitudinal

**Dataset Type:** Original Dataset

**Dataset Terms of Use:** CETIR Centre Mèdic is committed to providing high-quality, fully anonymized datasets for the EUCAIM project. The datasets, which include imaging data and clinical reports from various cancer types, have been meticulously anonymized to ensure patient privacy and compliance with ethical standards. Below are the detailed terms of use for accessing and utilizing these datasets. Anonymization and Privacy Full Anonymization: All datasets have undergone a rigorous anonymization process. Personal identifiers have been removed, and each entry has been assigned a unique code to ensure that individual patients cannot be identified. Compliance with Standards: The anonymization process adheres to international privacy standards, including the General Data Protection Regulation (GDPR). This ensures that all data sharing and usage comply with legal and ethical requirements, maintaining the highest level of data security and patient confidentiality. Access and Usage Access Permissions: Access to the datasets will be granted to researchers and institutions that are part of the EUCAIM project. All users must agree to the terms of use and demonstrate a legitimate research purpose. Non-Commercial Use: The datasets are provided strictly for non-commercial research purposes. Any commercial use of the data is prohibited unless explicitly authorized by CETIR Centre Mèdic. Research Purpose: The data must be used solely for the purpose of advancing medical research, improving diagnostic tools, and developing personalized treatment strategies. Any other use of the data must be approved in writing by CETIR Centre Mèdic. Data Security and Management Data Security: Users must ensure that the datasets are stored securely and access is restricted to authorized personnel only. Appropriate measures, such as encryption and access controls, must be implemented to protect the data from unauthorized access or breaches. Data Integrity: Users are responsible for maintaining the integrity of the datasets. Any modifications or annotations made to the data must be documented, and the original data must remain unchanged. Ethical Considerations Ethical Compliance: All research conducted using the datasets must comply with ethical standards and guidelines. Researchers must obtain necessary approvals from their institutional review boards (IRBs) or ethics committees before using the data. Publication and Sharing: Any publications or presentations resulting from the use of these datasets must acknowledge CETIR Centre Mèdic as the data source. Additionally, researchers are encouraged to share their findings with the broader scientific community to promote transparency and collaboration.

**Dataset Intended Purpose:** The datasets provided by CETIR Centre Mèdic are intended to advance oncology research, improve diagnostic and prognostic tools, develop personalized treatment strategies, and support multi-center collaborations. They also play a crucial role in training healthcare professionals and informing health policy decisions. By utilizing these high-quality and fully anonymized datasets, the EUCAIM project can significantly enhance cancer research and patient care, ultimately leading to better outcomes for patients worldwide.

**Imaging Modality:** MRI, CT

**Vendor:** GE, Philips, Siemens

**Imaging body part:** abdomen & pelvis

**Age range:** 19 / 90 years

**Sex:** Male and Female

**Number of subjects:** 1200

**Number of DICOM studies:** unspecified

**Image size in GB:** unspecified

**De-identification:** Personal data is fully anonymized

**Dataset 3**

**Cancer Type:** Pancreatic & urothelial cancers

**Dataset name:** Pancreatic & urothelial Dataset

**Dataset description:**
This dataset includes fully anonymized images and detailed reports for pancreatic and urothelial cancers, comprising 3,200 cases (1,600 pancreatic and 1,600 urothelial cancer cases). The dataset includes MRI and CT scans, ensuring comprehensive coverage for both cancer types. Data Acquisition Source: The images and reports were collected from patients who underwent diagnostic imaging at CETIR Centre Mèdic. The data spans several years, capturing a wide range of cases to ensure diversity and comprehensiveness. Imaging Protocols: Standard imaging protocols were used for MRI and CT scans, ensuring high-resolution images and consistent quality across the dataset. This includes both initial diagnostic scans and follow- up imaging studies. Data Anonymization Process: A rigorous anonymization process was applied to all collected data to protect patient privacy and comply with ethical standards. All personal identifiers were removed, and each dataset entry was assigned a unique code to ensure anonymity. Compliance: The anonymization process complies with international standards, including the General Data Protection Regulation (GDPR), ensuring that data security and patient confidentiality are maintained. Data Components Images: The dataset includes high-resolution MRI and CT images for each case. These images provide detailed visualization of tumor characteristics, essential for accurate diagnosis, staging, and treatment planning. Reports: Accompanying the images are detailed radiology reports prepared by expert radiologists. These reports contain critical clinical information such as tumor size, location, staging, and follow-up observations, providing a comprehensive overview of each case. Potential Impact Enhanced Diagnostic Accuracy: The diverse imaging data allows for the development of sophisticated AI models that improve the accuracy of cancer detection and staging. Personalized Treatment: Detailed imaging and clinical data support personalized treatment plans, optimizing therapeutic outcomes and minimizing side effects. Research and Innovation: This dataset provides a valuable resource for researchers to explore new imaging biomarkers and validate AI algorithms, driving innovations in oncology.

**Dataset Collection Method:** Disease-specific

**Dataset Type:** Original-Dataset

**Dataset Terms of Use:** CETIR Centre Mèdic is committed to providing high-quality, fully anonymized datasets for the EUCAIM project. The datasets, which include imaging data and clinical reports from various cancer types, have been meticulously anonymized to ensure patient privacy and compliance with ethical standards. Below are the detailed terms of use for accessing and utilizing these datasets. Anonymization and Privacy Full Anonymization: All datasets have undergone a rigorous anonymization process. Personal identifiers have been removed, and each entry has been assigned a unique code to ensure that individual patients cannot be identified. Compliance with Standards: The anonymization process adheres to international privacy standards, including the General Data Protection Regulation (GDPR). This ensures that all data sharing and usage comply with legal and ethical requirements, maintaining the highest level of data security and patient confidentiality. Access and Usage Access Permissions: Access to the datasets will be granted to researchers and institutions that are part of the EUCAIM project. All users must agree to the terms of use and demonstrate a legitimate research purpose. Non-Commercial Use: The datasets are provided strictly for non-commercial research purposes. Any commercial use of the data is prohibited unless explicitly authorized by CETIR Centre Mèdic. Research Purpose: The data

must be used solely for the purpose of advancing medical research, improving diagnostic tools, and developing personalized treatment strategies. Any other use of the data must be approved in writing by CETIR Centre Mèdic. Data Security and Management Data Security: Users must ensure that the datasets are stored securely and access is restricted to authorized personnel only. Appropriate measures, such as encryption and access controls, must be implemented to protect the data from unauthorized access or breaches. Data Integrity: Users are responsible for maintaining the integrity of the datasets. Any modifications or annotations made to the data must be documented, and the original data must remain unchanged. Ethical Considerations Ethical Compliance: All research conducted using the datasets must comply with ethical standards and guidelines. Researchers must obtain necessary approvals from their institutional review boards (IRBs) or ethics committees before using the data. Publication and Sharing: Any publications or presentations resulting from the use of these datasets must acknowledge CETIR Centre Mèdic as the data source. Additionally, researchers are encouraged to share their findings with the broader scientific community to promote transparency and collaboration.

**Dataset Intended Purpose:** The datasets provided by CETIR Centre Mèdic are intended to advance oncology research, improve diagnostic and prognostic tools, develop personalized treatment strategies, and support multi-center collaborations. They also play a crucial role in training healthcare professionals and informing health policy decisions. By utilizing these high-quality and fully anonymized datasets, the EUCAIM project can significantly enhance cancer research and patient care, ultimately leading to better outcomes for patients worldwide.

**Imaging Modality:** MRI, CT

**Vendor:** GE, Philips, Siemens

**Imaging body part:** Chest, Abdomen, pelvis

**Age range:** 19 / 90 years

**Sex:** Male and Female

**Number of subjects:** 1900

**Number of DICOM studies:** unspecified

**Image size in GB:** unspecified

**De-identification:** Personal data is fully anonymized

**Dataset 4**

**Cancer Type:** Brain cancer

**Dataset name:** Brain Dataset

**Dataset description:**
CETIR Centre Mèdic is contributing a specialized dataset focusing on brain tumors, which includes a total of 810 cases. This dataset consists of high-resolution MRI images and detailed clinical reports for three types of brain tumors: gliomas, glioblastomas, and neuroblastomas. Dataset Components 1. Gliomas: Number of Cases: 650 Imaging Modality: Brain MRI Description: This subdataset includes MRI images capturing various grades and stages of gliomas. The images provide detailed insights into tumor morphology and are essential for accurate diagnosis and treatment planning. 2. Glioblastomas: Number of Cases: 150 Imaging Modality: Brain MRI Description: Glioblastomas are among the most aggressive brain tumors. This subdataset includes high-resolution MRI images that highlight the complex nature of glioblastomas, aiding in precise diagnosis and therapeutic strategy development. 3. Neuroblastomas: Number of Cases: 10 Imaging Modality: Brain MRI Description: Neuroblastomas, though rare in the brain, are included to provide a comprehensive view of pediatric brain tumors. The MRI images in this subdataset offer valuable data for understanding and treating these cases. Data Quality and Annotation Expert Annotations: All MRI images have been annotated by specialized radiologists with expertise in neuro-oncology. The annotations include details on tumor size, location, and characteristics, ensuring high-quality and reliable data. Standardization: The dataset follows standardized

imaging protocols and quality control measures, ensuring consistency across all cases. Potential Impact Enhanced Diagnostic Tools: The detailed MRI images enable the development of advanced AI models to improve diagnostic accuracy for brain tumors. Research and Innovation: This dataset supports research into new imaging biomarkers and treatment methods, fostering innovation in neuro-oncology. Personalized Treatment: The clinical reports, combined with imaging data, support personalized treatment plans, optimizing outcomes for patients with brain tumors.

**Dataset Collection Method:** Disease-specific

**Dataset Type:** Annotated Dataset

**Dataset Terms of Use:** CETIR Centre Mèdic is committed to providing high-quality, fully anonymized datasets for the EUCAIM project. The datasets, which include imaging data and clinical reports from various cancer types, have been meticulously anonymized to ensure patient privacy and compliance with ethical standards. Below are the detailed terms of use for accessing and utilizing these datasets. Anonymization and Privacy Full Anonymization: All datasets have undergone a rigorous anonymization process. Personal identifiers have been removed, and each entry has been assigned a unique code to ensure that individual patients cannot be identified. Compliance with Standards: The anonymization process adheres to international privacy standards, including the General Data Protection Regulation (GDPR). This ensures that all data sharing and usage comply with legal and ethical requirements, maintaining the highest level of data security and patient confidentiality. Access and Usage Access Permissions: Access to the datasets will be granted to researchers and institutions that are part of the EUCAIM project. All users must agree to the terms of use and demonstrate a legitimate research purpose. Non-Commercial Use: The datasets are provided strictly for non-commercial research purposes. Any commercial use of the data is prohibited unless explicitly authorized by CETIR Centre Mèdic. Research Purpose: The data must be used solely for the purpose of advancing medical research, improving diagnostic tools, and developing personalized treatment strategies. Any other use of the data must be approved in writing by CETIR Centre Mèdic. Data Security and Management Data Security: Users must ensure that the datasets are stored securely and access is restricted to authorized personnel only. Appropriate measures, such as encryption and access controls, must be implemented to protect the data from unauthorized access or breaches.
Data Integrity: Users are responsible for maintaining the integrity of the datasets. Any modifications or annotations made to the data must be documented, and the original data must remain unchanged. Ethical Considerations Ethical Compliance: All research conducted using the datasets must comply with ethical standards and guidelines. Researchers must obtain necessary approvals from their institutional review boards (IRBs) or ethics committees before using the data. Publication and Sharing: Any publications or presentations resulting from the use of these datasets must acknowledge CETIR Centre Mèdic as the data source. Additionally, researchers are encouraged to share their findings with the broader scientific community to promote transparency and collaboration.

**Dataset Intended Purpose:** The datasets provided by CETIR Centre Mèdic are intended to advance oncology research, improve diagnostic and prognostic tools, develop personalized treatment strategies, and support multi-center collaborations. They also play a crucial role in training healthcare professionals and informing health policy decisions. By utilizing these high-quality and fully anonymized datasets, the EUCAIM project can significantly enhance cancer research and patient care, ultimately leading to better outcomes for patients worldwide.

**Imaging Modality:** MRI

**Vendor:** GE, Philips, Siemens

**Imaging body part:** Brain

**Age range:** 19 / 90 years

**Sex:** Male and Female

**Number of subjects:** 500

**Number of DICOM studies:** unspecified

**Image size in GB:** unspecified

**De-identification:** Personal data is fully anonymized

**Dataset 5**

**Cancer Type:** Lung cancer

**Dataset name:** Lung Dataset

**Dataset description:**
CETIR Centre Mèdic is contributing a comprehensive dataset focusing on lung cancer, which includes a total of 2,700 cases. This dataset consists of high-resolution CT scans and detailed clinical reports, providing an extensive resource for advancing research and improving diagnostic and therapeutic strategies for lung cancer. Dataset Components 1. Lung Cancer Cases: Number of Cases: 2,700 Imaging Modality: Chest CT scans Description: This dataset includes high-resolution CT scans capturing various stages and types of lung cancer. The scans provide detailed insights into tumor morphology, nodal involvement, and metastatic spread, which are essential for accurate diagnosis, staging, and treatment planning. Data Quality and Annotation Expert Annotations: All CT scans have been meticulously annotated by a team of expert radiologists specialized in thoracic oncology. The annotations include detailed descriptions of tumor size, location, characteristics, and any observed metastases, ensuring high-quality and reliable data . Standardization: The dataset follows standardized imaging protocols and stringent quality control measures, ensuring consistency and interoperability across different research and clinical applications. Potential Impact Enhanced Diagnostic Tools: The detailed CT scans enable the development of advanced AI models and algorithms to improve the accuracy and efficiency of lung cancer detection and staging. Personalized Treatment: Comprehensive imaging and clinical data support the creation of personalized treatment plans, optimizing therapeutic outcomes and minimizing side effects for patients. Research and Innovation: This dataset serves as a valuable resource for researchers exploring new imaging biomarkers, treatment methods, and validating AI-driven diagnostic tools, fostering innovation in thoracic oncology.

**Dataset Collection Method:** Disease-specific

**Dataset Type:** Annotated Dataset

**Dataset Terms of Use:** CETIR Centre Mèdic is committed to providing high-quality, fully anonymized datasets for the EUCAIM project. The datasets, which include imaging data and clinical reports from various cancer types, have been meticulously anonymized to ensure patient privacy and compliance with ethical standards. Below are the detailed terms of use for accessing and utilizing these datasets. Anonymization and Privacy Full Anonymization: All datasets have undergone a rigorous anonymization process. Personal identifiers have been removed, and each entry has been assigned a unique code to ensure that individual patients cannot be identified. Compliance with Standards: The anonymization process adheres to international privacy standards, including the General Data Protection Regulation (GDPR). This ensures that all data sharing and usage comply with legal and ethical requirements, maintaining the highest level of data security and patient confidentiality. Access and Usage Access Permissions: Access to the datasets will be granted to researchers and institutions that are part of the EUCAIM project. All users must agree to the terms of use and demonstrate a legitimate research purpose. Non-Commercial Use: The datasets are provided strictly for non-commercial research purposes. Any commercial use of the data is prohibited unless explicitly authorized by CETIR Centre Mèdic. Research Purpose: The data must be used solely for the purpose of advancing medical research, improving diagnostic tools, and developing personalized treatment strategies. Any other use of the data must be approved in writing by CETIR Centre Mèdic. Data Security and Management Data Security: Users must ensure that the datasets are stored securely and access is restricted to authorized personnel only. Appropriate measures, such as encryption and access controls, must be implemented to protect the data from unauthorized access or breaches. Data Integrity: Users are responsible for maintaining the integrity of the datasets. Any modifications or annotations made to the data must be documented, and the original data must remain unchanged. Ethical Considerations Ethical Compliance: All research conducted using the datasets must comply with ethical standards and guidelines. Researchers must obtain necessary approvals from their institutional review boards (IRBs) or ethics committees before using the data. Publication and Sharing: Any publications or presentations resulting from the use of these datasets must

| |
|---|
| acknowledge CETIR Centre Mèdic as the data source. Additionally, researchers are encouraged to share their findings with the broader scientific community to promote transparency and collaboration.<br><br>**Dataset Intended Purpose:** The datasets provided by CETIR Centre Mèdic are intended to advance oncology research, improve diagnostic and prognostic tools, develop personalized treatment strategies, and support multi-center collaborations. They also play a crucial role in training healthcare professionals and informing health policy decisions. By utilizing these high-quality and fully anonymized datasets, the EUCAIM project can significantly enhance cancer research and patient care, ultimately leading to better outcomes for patients worldwide.<br><br>**Imaging Modality:** CT<br><br>**Vendor:** GE, Philips, Siemens, Toshiba<br><br>**Imaging body part:** Chest, abdomen, Pelvis<br><br>**Age range:** 19 / 90 years<br><br>**Sex:** Male and Female<br><br>**Number of subjects:** 1850<br><br>**Number of DICOM studies:** unspecified<br><br>**Image size in GB:** unspecified<br><br>**De-identification:** Personal data is fully anonymized |

Table 28. Description of Use Case no. 28 from Wellbeing Services County of North Savo

| Author: Arto Mannermaa | Data sharing: Federated Node | Organisation's name: Wellbeing Services County of North Savo | Organization's Acronym: Pshyvinvointialue | Tier: 1 |
|---|---|---|---|---|
| **General description of the potential use and clinical impact of the shared data:**<br>Genetically homogenous isolated Eastern Finnish population provides unique opportunities for improving patient care. Genotype data combined with phenotype data offers openings for development of AI solutions aiming at personalized medicine. Currently biobank projects utilize AI solutions including automated digitized tumor microscopic slide analysis and deep machine learning of multimodal data. Identification of new biomarkers that are specifically associated with cancer are useful for prevention, early detection, and treatment. Biobank of Eastern Finland holds vast longitudinal clinical data including imaging, therapy, therapy response and follow-up information as well as high quality longitudinal sample series including tumor and liquid biopsy samples. Cancer research is one of our focus areas and we are experienced in providing histological images and other imaging data e.g. mammography images for researchers. The Biobank of Eastern Finland supports biomedical research that covers promotion of population health, identification of factors contributing to disease mechanisms, prevention of diseases as well as development of products and treatment practices used in healthcare and medical care. | | | | |
| **Dataset 1**<br><br>**Cancer Type:** Breast Cancer<br><br>**Dataset name:** Multimodal breast cancer cohort<br><br>**Dataset description:**<br>Biobank of Eastern Finland has been collecting tumor samples from patients with breast cancer since 2000 and blood samples since 2016. Collections contain patient associated clinical data (EHR), imaging, histopathological data and genotype data. Longitudinal clinical data with time stamps when required: | | | | |

- Imaging data
- Digitized histological slides
- Demographics (age, genders etc.)
- Diagnoses (ICD-10 codes)
- Electronic Health Records
- Histopathological data (histology, organ, grade, TNM stage, tumour markers)
- Laboratory results
- Hospital administered medications and prescriptions
- Operations and procedures
- Radiation therapy
- Resource utilization
- Genomics (GWAS)

Samples:
- Whole blood, Plasma, Serum
- cf-DNA
- DNA
- Tissue (FFPE)

**Dataset Collection Method:** Patient-based

**Dataset Type:** Ongoing collection

**Dataset Terms of Use:** For biobank research promoting health, understanding mechanisms behind diseases disease mechanisms or developing medicines and treatment practices. A person requesting access to the samples or data stored in a biobank must submit a written request addressed to the biobank.

**Dataset Intended Purpose:** Biobank research. The Biobank of Eastern Finland provides samples and / or associated clinical information for medical research and R&D purposes.

**Imaging Modality:** x-ray (mammography), MRI, Ultrasound, CT, PET, digitized histological slides

**Vendor:** Unspecified

**Imaging body part:** Thorax

**Age range:** 18 / 99 years

**Sex:** Male and Female

**Number of subjects:** 5200

**Number of DICOM studies:** Unspecified

**Image size in GB:** Unspecified

**De-identification:** Personal data is fully anonymized

**Dataset 2**

**Cancer Type:** Lung cancer

**Dataset name:** Multimodal lung cancer cohort

**Dataset description:**
Biobank of Eastern Finland has been collecting tumor samples from patients with lung cancer since 2000 and blood samples since 2016. Collections contain patient associated clinical data (EHR), imaging, histopathological data and genotype data. Longitudinal clinical data with time stamps when required:
- Imaging data

- Digitized histological slides
- Demographics (age, genders etc.)
- Diagnoses (ICD-10 codes)
- Electronic Health Records
- Histopathological data (histology, organ, grade, TNM stage, tumour markers, mutations)
- Laboratory results
- Hospital administered medications and prescriptions
- Operations and procedures
- Radiation therapy
- Resource utilization
- Genomics (GWAS)

Samples:
- Whole blood, Plasma, Serum
- cf-DNA
- DNA
- Tissue (FFPE)

**Dataset Collection Method:** Patient-based

**Dataset Type:** Processed Dataset

**Dataset Terms of Use:** For biobank research promoting health, understanding mechanisms behind diseases disease mechanisms or developing medicines and treatment practices. A person requesting access to the samples or data stored in a biobank must submit a written request addressed to the biobank.

**Dataset Intended Purpose:** Biobank research. The Biobank of Eastern Finland provides samples and / or associated clinical information for medical research and R&D purposes.

**Imaging Modality:** x-ray , MRI, Ultrasound, CT, digitized histological slides

**Vendor:** Unspecified

**Imaging body part:** Thorax

**Age range:** 18 / 99 years

**Sex:** Male and Female

**Number of subjects:** 1900

**Number of DICOM studies:** unspecified

**Image size in GB:** unspecified

**De-identification:** Personal data is fully anonymized

**Dataset 3**

**Cancer Type:** Prostate cancer

**Dataset name:** Multimodal prostate cancer cohort

**Dataset description:**
Biobank of Eastern Finland has been collecting tumor samples from patients with prostate cancer since 2000 and blood samples since 2016. Collections contain patient associated clinical data (EHR), imaging, histopathological data and genotype data. Longitudinal clinical data with time stamps when required:
- Imaging data
- Digitized histological slides

- Demographics (age, genders etc.)
- Diagnoses (ICD-10 codes)
- Electronic Health Records
- Histopathological data (histology, organ, grade, TNM stage, tumour markers)
- Laboratory results
- Hospital administered medications and prescriptions
- Operations and procedures
- Radiation therapy
- Resource utilization
- Genomics (GWAS)

Samples:
- Whole blood, Plasma, Serum
- cf-DNA
- DNA
- Tissue (FFPE)

**Dataset Collection Method:** Patient-based

**Dataset Type:** Original Dataset

**Dataset Terms of Use:** For biobank research promoting health, understanding mechanisms behind diseases disease mechanisms or developing medicines and treatment practices. A person requesting access to the samples or data stored in a biobank must submit a written request addressed to the biobank.

**Dataset Intended Purpose:** Biobank research. The Biobank of Eastern Finland provides samples and / or associated clinical information for medical research and R&D purposes.

**Imaging Modality:** x-ray, MRI, Ultrasound, CT, PET, digitized histological slides

**Vendor:** Unspecified

**Imaging body part:** Thorax

**Age range:** 18 / 99 years

**Sex:** Male

**Number of subjects:** 4000

**Number of DICOM studies:** Unspecified

**Image size in GB:** Unspecified

**De-identification:** Personal data is fully anonymized

**Dataset 4**

**Cancer Type:** Gynecological cancers (e.g. ovarian and endometrial cancer)

**Dataset name:** Multimodal gynecological cancer cohort

**Dataset description:**
Biobank of Eastern Finland has been collecting tumor samples from patients with gynecological cancers including ovarian and endometrial cancer since 2000 and blood samples since 2016. Collections contain patient associated clinical data (EHR), imaging, histopathological data and genotype data. Longitudinal clinical data with time stamps when required:
- Imaging data
- Digitized histological slides

| |
|---|
| - Demographics (age, genders etc.)<br>- Diagnoses (ICD-10 codes)<br>- Electronic Health Records<br>- Histopathological data (histology, organ, grade, TNM stage, tumour markers)<br>- Laboratory results<br>- Hospital administered medications and prescriptions<br>- Operations and procedures<br>- Radiation therapy<br>- Resource utilization<br>- Genomics (GWAS)<br>Samples:<br>- Whole blood, Plasma, Serum<br>- cf-DNA<br>- DNA<br>- Tissue (FFPE, fresh frozen)<br><br>**Dataset Collection Method:** Patient-based<br><br>**Dataset Type:** Original Dataset<br><br>**Dataset Terms of Use:** For biobank research promoting health, understanding mechanisms behind diseases disease mechanisms or developing medicines and treatment practices. A person requesting access to the samples or data stored in a biobank must submit a written request addressed to the biobank<br><br>**Dataset Intended Purpose:** Biobank research. The Biobank of Eastern Finland provides samples and / or associated clinical information for medical research and R&D purposes.<br><br>**Imaging Modality:** x-ray, MRI, Ultrasound, CT, digitized histological slides<br><br>**Vendor:** Unspecified<br><br>**Imaging body part:** Unspecified<br><br>**Age range:** 18 / 99 years<br><br>**Sex:** Female<br><br>**Number of subjects:** 2400<br><br>**Number of DICOM studies:** Unspecified<br><br>**Image size in GB:** Unspecified<br><br>**De-identification:** Personal data is fully anonymized |

Table 29. Description of Use Case no. 23 from Fundación para la Gestión de Investigación en Salud en Sevilla

| | | | | |
|---|---|---|---|---|
| Author: Alberto Moreno Conde | Data sharing: Federated Node | Organisation's name: Fundación para la Gestión de Investigación en Salud en Sevilla | Organization's Acronym: FISEVI | Tier: 2 |
| **General description of the potential use and clinical impact of the shared data:**<br>The shared datasets could include all the images collected in routine care in our Hospital for 13 types of cancers since 2011 (lung cancer, breast cancer, head and neck cancer, brain cancers, Hepatobiliary and Pancreatic | | | | |

Cancers, gastric cancers, colorectal cancers, bladder cancer, prostate cancer, kidney cancer, melanoma, lymphoma and gynecological cancers). Nevertheless since the maximum number of datasets that can be detailed in the form is 5, we only detailed the datasets from lung cancer, breast cancer, prostate cancer, bladder cancer and colorectal cancer. The complete dataset includes more than 80000 images from more than 27000 patients. These images are mainly CT scans and were collected in the diagnostic, treatment and follow-up of these patients. Furthermore, the Innovation & Data Analysis Unit has mapped all the Electronic Health Records of the hospital to the OMOP CDM, so these images could be accompanied by the clinical records from these patients and be used to train different predictive models to assist diagnosis, predict toxicity, prognosis, survival, etc.

**Dataset 1**

**Cancer Type:** Lung cancer

**Dataset name:** Lung cancer patients in HUVM

**Dataset description:**
This dataset contains all the thorax and abdomen CT scans of all the more than 5000 Lung Cancer patients that have been treated at the hospital since 2011.

**Dataset Collection Method:** Cohort

**Dataset Type:** Original-Dataset

**Dataset Terms of Use:** Compliance with regional, national and European regulation, Approval of Research Ethical Committee and signature of DSA

**Dataset Intended Purpose:** These dataset was collected for clinical care of the subjects

**Imaging Modality:** CT scan

**Vendor:** Unspecified

**Imaging body part:** thorax and abdomen

**Age range:** 18 / 99 years

**Sex:** Male and Female

**Number of subjects:** 5000

**Number of DICOM studies:** Unspecified

**Image size in GB:** Unspecified

**De-identification:** Personal data is included in the images. In this case EUCAIM can support you with specific tools and guidance for de-identification once the application is accepted.

**Dataset 2**

**Cancer Type:** Breast cancer

**Dataset name:** Breast cancer patients mammograms anc CT scans in HUVM

**Dataset description:**
This dataset contains all the mammography of breast cancer patients that were treated at the hospital since 2011

**Dataset Collection Method:** Cohort

**Dataset Type:** Original Dataset

**Dataset Terms of Use:** Compliance with regional, national and European regulation, Approval of Research Ethical Committee and signature of DSA

**Dataset Intended Purpose:** Clinical care

**Imaging Modality:** Mammography

**Vendor:** Unspecified

**Imaging body part:** Breast

**Age range:** 18 / 99 years

**Sex:** Female

**Number of subjects:** 4500

**Number of DICOM studies:** unspecified

**Image size in GB:** unspecified

**De-identification:** Personal data is included in the images. In this case EUCAIM can support you with specific tools and guidance for de-identification once the application is accepted.

**Dataset 3**

**Cancer Type:** Colorectal cancer

**Dataset name:** Colorectal cancer patients in HUVM

**Dataset description:**
This dataset contains all the CT scans of Colorectal cancer that were treated at the hospital since 2011

**Dataset Collection Method:** Cohort

**Dataset Type:** Original Dataset

**Dataset Terms of Use:** Compliance with regional, national and European regulation, Approval of Research Ethical Committee and signature of DSA

**Dataset Intended Purpose:** Routine Care of patients

**Imaging Modality:** CT scan

**Vendor:** Unspecified

**Imaging body part:** colon and rectum

**Age range:** 18 / 99 years

**Sex:** Male and Female

**Number of subjects:** 5000

**Number of DICOM studies:** Unspecified

**Image size in GB:** Unspecified

**De-identification:** Personal data is included in the images. In this case EUCAIM can support you with specific tools and guidance for de-identification once the application is accepted.

**Dataset 4**

**Cancer Type:** Bladder cancer

**Dataset name:** Bladder cancer patients in HUVM

**Dataset description:**
This dataset contains all the CT scans of all bladder cancer patients that were treated at the hospital since 2011

**Dataset Collection Method:** Cohort

**Dataset Type:** Original Dataset

**Dataset Terms of Use:** Compliance with regional, national and European regulation, Approval of Research Ethical Committee and signature of DSA

**Dataset Intended Purpose:** Clinical care

**Imaging Modality:** CT scan

**Vendor:** Unspecified

**Imaging body part:** abdomen and pelvis

**Age range:** 18 / 99 years

**Sex:** Male and Female

**Number of subjects:** 3500

**Number of DICOM studies:** Unspecified

**Image size in GB:** Unspecified

**De-identification:** Personal data is included in the images. In this case EUCAIM can support you with specific tools and guidance for de-identification once the application is accepted.

**Dataset 5**

**Cancer Type:** Prostate Cancer

**Dataset name:** Prostate cancer patients in HUVM

**Dataset description:**
This dataset contains all the CT scans of all prostate cancer patients that were treated at the hospital since 2011

**Dataset Collection Method:** Cohort

**Dataset Type:** Original Dataset

| |
|---|
| **Dataset Terms of Use:** Compliance with regional, national and European regulation, Approval of Research Ethical Committee and signature of DSA<br><br>**Dataset Intended Purpose:** Clinical care<br><br>**Imaging Modality:** CT scan<br><br>**Vendor: Unspecified**<br><br>**Imaging body part:** abdomen and pelvis<br><br>**Age range:** 18 / 99 years<br><br>**Sex:** Male and Female<br><br>**Number of subjects:** 3000<br><br>**Number of DICOM studies: Unspecified**<br><br>**Image size in GB: Unspecified**<br><br>**De-identification:** Personal data is included in the images. In this case EUCAIM can support you with specific tools and guidance for de-identification once the application is accepted. |

Table 30. Description of Use Case no. 62 from National and Kapodistrian University of Athens

| | | | | |
|---|---|---|---|---|
| Author: Ioannis Seimenis | Data sharing: Federated Node | Organisation's name: National and Kapodistrian University of Athens | Organization's Acronym: UoA | Tier: 2 |
| **General description of the potential use and clinical impact of the shared data:**<br>UoA proposes to participate in EUCAIM with three cancer types (and multiple subtypes). Some of the cancers concerned are regarded as rare cancers (e,g., sarcomas and primary cerebral lymphomas). As detailed below, some imaging data to be provided is accompanied by lesion segmentations, as well as by a set of clinical / biochemical data. As suggested in the relevant intended purposes, many use cases could be implemented to address specific scientific questions. | | | | |
| **Dataset 1**<br><br>**Cancer Type:** Sarcomas and primary cerebral lymphomas<br><br>**Dataset name:** Sarcomas<br><br>**Dataset description:**<br>A dataset of about 80 sarcomas is expected to be provided within the implementation period. The dataset will comprise pre-radiotherapy images (CT and/or MRI) with tumor segmentation, core biopsy results, post-radiotherapy images (MRI performed 4-6 weeks after irradiation with a 25x2 Gy scheme) and histopathological data following surgical excision. Following surgery, patients are monitored for local recurrency.<br><br>**Dataset Collection Method:** Longitudinal<br><br>**Dataset Type:** Annotated Dataset<br><br>**Dataset Terms of Use:** Unspecified | | | | |

**Dataset Intended Purpose:** This dataset may serve the identification of imaging biomarkers (in the post-irradiation dataset) for the prognosis of local recurrency. In addition, the pre-irradiation imaging dataset can be used for the extraction of radiomic features which may facillitate the differentiation between different sarcomas and their radiosensitivity.

**Imaging Modality:** MRI/CT

**Vendor:** Unspecified

**Imaging body part:** Unspecified

**Age range:** Unspecified

**Sex:** Male and Female

**Number of subjects:** up to 80

**Number of DICOM studies:** up tp 160

**Image size in GB:** Unspecified

**De-identification:** Personal data is included in the images. In this case EUCAIM can support you with specific tools and guidance for de-identification once the application is accepted.

**Dataset 2**

**Cancer Type:** Brain tumors (e.g. Glioma, primary cerebral lymphomas, meningiomas and schwannomas)

**Dataset name:** Brain tumors

**Dataset description:**
This imaging dataset contains a collection of patients with brain tumors (mainly primary tumors). The dataset comprises gliomas (IDH-mutant and IDH wild type), meningiomas, schwannomas and primary cerebral lymphomas. Cases of metastatic brain tumors will be included to serve as negative cases, whilst control cases will also be provided. Studies include pre- and post-Gd anatomical series (2D and 3D), as well as DWI, DTI, SWI and DSE series. Relevant quantitative (parametric) maps are also incorporated.

**Dataset Collection Method:** Longitudinal

**Dataset Type:** Annotated Dataset

**Dataset Terms of Use:** Unspecified

**Dataset Intended Purpose:** This dataset may facilitate the differentiation between primary brain tumors and metastatic tumors. In addition, it could be used to facilitate tumor grading, e.g., to differentiate between IDH-mutant and IDH wild type gliomas and/or to discriminate between benign and atypical/malignant meningiomas. This dataset allows for the extraction of imaging biomarkers and radiomic features.

**Imaging Modality:** MRI

**Vendor:** Unspecified

**Imaging body part:** Unspecified

**Age range:** Unspecified

**Sex:** Male and Female

**Number of subjects:** up to 240

**Number of DICOM studies:** up to 240

**Image size in GB:** unspecified

**De-identification:** Personal data is included in the images. In this case EUCAIM can support you with specific tools and guidance for de-identification once the application is accepted.

**Dataset 3**

**Cancer Type:** Prostate cancer

**Dataset name:** Prostate

**Dataset description:**
This imaging dataset contains a collection of patients with increased prostate specific antigen (PSA) and prostatic hyperplasia or neoplasia. Pre-biopsy mpMRI examinations (T2 Axial, DWI, DSE) are provided, whilst quantitative (parametric) maps (e.g. ADC, MTT, TTP, rBF, rBV) are also included.

**Dataset Collection Method:** Longitudinal

**Dataset Type:** Annotated Dataset

**Dataset Terms of Use:** Unspecified

**Dataset Intended Purpose:** This dataset may facilitate the differentiation, based on imaging biomarkers, between prostatic cancer and hyperplasia. In addition, it could allow for the extraction of radiomic features from MRI-identifiable lesions and their correlation to Gleason score (GS). A goal could be the effective use of radiomic features and imaging biomarkers for differentiating between normal (GS<6) vs abnormal (≥6) prostatic tissue and low (1) vs high (>1) ISUP score in cancerous tissue, in an effort to tackle overdiagnosis and overtreatment observed in prostate cancer. Additionally, the difference in the discriminative power between T2w- and DW-extracted features could be examined, especially with regard to the PI-RADS approach (DWI domination for peripheral zone lesions and T2w domination for transition zone lesions).

**Imaging Modality:** MRI

**Vendor:** Unspecified

**Imaging body part:** Unspecified

**Age range:** Over 50 years of age

**Sex:** Males

**Number of subjects:** up to 50

**Number of DICOM studies:** up tp 50

**Image size in GB:** Unspecified

**De-identification:** Personal data is included in the images. In this case EUCAIM can support you with specific tools and guidance for de-identification once the application is accepted.

Table 31. Description of Use Case no. 31 Instituto Aragonés de Ciencias de la Salud

| Author: Juan González-García | Data sharing: Federated Node | Organisation's name: Instituto Aragonés de Ciencias de la Salud | Organization's Acronym: IACS | Tier: 3 |
|---|---|---|---|---|
| **General description of the potential use and clinical impact of the shared data:**<br><br>The IACS is an institution of the health system, providing comprehensive access to all clinical data, including imaging, through the BIGAN platform. BIGAN platform, as the regional health data lake, integrates a vast array of health information, thereby facilitating its secondary use for research, innovation and policy-making. The depth and richness of the integrated data make it a valuable resource for future users of the EUCAIM infrastructure, who will benefit from the detailed and expansive datasets for research, clinical advancements, and enhanced patient care. This seamless integration ensures that the wealth of clinical data can be leveraged to its fullest potential, driving innovation and improving healthcare outcomes across the board. | | | | |
| **Dataset 1**<br><br>**Cancer Type:** Prostate cancer<br><br>**Dataset name:** AI4HealthyAgeing - Prostate Cancer Hospital Universitario Miguel Servet<br><br>**Dataset description:**<br>Scanned images from histological sections of pathological anatomy<br><br>**Dataset Collection Method:** Patient-based<br><br>**Dataset Type:** Original Dataset<br><br>**Dataset Terms of Use:** Available upon request and access permit once the AI4HealthyAgeing is finished (January 2025).<br><br>**Dataset Intended Purpose:** Model training for diagnosis<br><br>**Imaging Modality:** Pathology scanned images<br><br>**Vendor:** Unspecified<br><br>**Imaging body part:** Prostate<br><br>**Age range:** 31-98<br><br>**Sex:** Male<br><br>**Number of subjects:** ~100<br><br>**Number of DICOM studies:** Unspecified<br><br>**Image size in GB:** Unspecified<br><br>**De-identification:** Personal data is pseudonymized | | | | |
| **Dataset 2**<br><br>**Cancer Type:** Colorectal cancer<br><br>**Dataset name:** AI4HealthyAgeing - Colorectal Cancer Hospital Clinico Zaragoza | | | | |

| |
|---|
| **Dataset description:**<br>Collection of frames extracted from colonoscopy video recordings. The frames contain morphological abnormalities such as tumors, polyps or adenomas. There are unlabeled and labelled frames surrounding the abnormalities.<br><br>**Dataset Collection Method:** Patient-based<br><br>**Dataset Type:** Original Dataset<br><br>**Dataset Terms of Use:** Available upon request and access permit once the AI4HealthyAgeing is finished (January 2025).<br><br>**Dataset Intended Purpose:** Model training for colorectal morphological abnormalities diagnosis.<br><br>**Imaging Modality:** Colonoscopy video recording.<br><br>**Vendor:** Unspecified<br><br>**Imaging body part:** Colon/Rectum<br><br>**Age range:** 3-97<br><br>**Sex:** Male and Female<br><br>**Number of subjects:** ~ 5000 *images*<br><br>**Number of DICOM studies:** Unspecified<br><br>**Image size in GB:** Unspecified<br><br>**De-identification:** Personal data is pseudonymized |

- Data Users:

Table 32. Description of Use Case no. 12 from Gdańsk University of Technology

| | | | |
|---|---|---|---|
| Author: Michał Grochowski | Intention:<br>- train/validate AI tools | Organisation's name: Gdańsk University of Technology | Organization's Acronym: GDANSK TECH |
| **Title of the use case:** CLEAR-AI: Enhancing High-Resolution Image Segmentation Precision through Collaborative Learning of Deep Neural Networks for Accurate Assessment of Axillary Lymph Node Metastasis based on Full-Field Digital Mammography Analysis | | | |
| **General description of the use case:**<br>Currently, the breast cancer diagnosis is based on clinical examination and various imaging techniques. These examinations are followed by invasive breast tumor and lymph node biopsy procedures and histopathological examination of specimens to verify the clinical diagnosis. There is no faster, less invasive and cheaper diagnostic approach so far, which could reduce the number of necessary radiological, surgical and pathology tests to confirm the presence of cancer and its potential spread to axillary lymph nodes (ALN). Doctors need all this information to facilitate the selection of appropriate treatment and its sequence. The project will try to identify the set of information embedded in the full-field digital mammography (FFDM) images and the minimal set of clinical information representative for predicting ALN metastasis. The current state-of-the-art has no such an option for clinicians and their patients. The challenge is to efficiently take the advantage of the AI tools to extract this knowledge hidden in medical data. Our ambition is to develop theoretical foundations and AI-driven methods for automated early diagnosis of breast cancer, paying special attention to ALN status. The system will be designed to analyze medical data, extract information from it, and provide clinicians with the suggestions for diagnoses regarding breast lesion detection, type of lesion (benign versus malignant), and ALN status. Special attention will be paid to problems related to the shortage of training data, especially annotated at the pixel level, the extraction and identification of medically important features to support the diagnosis, clear explanations of decisions undertaken by AI-based systems, and mechanisms to enable synergetic cooperation between different AI models and their end-users, doctors (Human-in-the-Loop approach). The main objective of the use case is to | | | |

develop collaboratively learned deep neural networks (DNNs) for the classification and segmentation of breast lesions, utilizing image and pixel-level annotations in high- resolution medical image settings, specifically FFDM. The system will be able to differentiate lesion types and predict ALN status based on information from mammography images. Emphasis will be placed on ensuring that classification results can be explained, in accordance with academic standards, clinical requirements and regulatory requirements, including EU AI Act legislation. Expected results include the development of an AI-based breast tumor segmentation tool suitable for high-resolution medical images. Moreover, the system should be able to differentiate tumor types and assess metastatic ALN with a similar or higher level of sensitivity and the same specificity as the current 'gold standard' methods, while providing visual reasons for prediction, thus increasing the system's trustworthiness. The expected clinical impact of this tool will be to reduce the number of necessary surgical and pathological tests required to confirm the presence of cancer and its potential spread to the ALN. This knowledge of the tumor characteristics and ALN status will assist clinicians in selecting the appropriate treatment and its sequence. Moreover, the system for automatic masks generation learned through collaboration can be used for studies on more comprehensive cancer analysis. Methodology: The methodology for achieving the objectives of the use case will involve supervised learning of a single DNN for a classification task using high-resolution FFDM images and their labels. This model will not only classify breast cancer but will also be capable of providing visual explanation maps. Concurrently, another DNN will be trained for a segmentation task. Afterwards, models will be collaboratively trained with skillful data-sharing between them. An additional model, the discriminator, will be employed during the training. This model being concurrently trained with the aim to match the two generated masks to their source will refine the collaboration process. This combination is expected to lead to improved classification and segmentation outcomes.

**Expected timeline for the realization of the use case:**
RT1. Classification module (M1-M6) The objective of this research task is to develop a classifier for breast cancer lesions and potential axillary lymph node metastases. The work will involve the use of methods that are capable of processing high- resolution images without image resizing in order to preserve full image features. The research will comprise the learning of state-of-the-art deep neural networks in the Multiple Instance Learning (MIL) approach, with a focus on analyzing various pooling mechanisms, including Attention-based MIL (AMIL), Gated-attention-based (GAMIL), Dual Stream MIL (DSMIL) and Clustering-constrained Attention MIL (CLAM). Furthermore, this research task will investigate the strengths of other foundational models, including the utilization of Vision Transformers (ViT), State Space Models (SSM) and enhanced SSMs with added elements of the traditional transformer architecture. Additionally, eXplainable Artificial Intelligence (XAI) algorithms will be analyzed, including Grad-Cam, Grad-Cam++ and Integrated Gradients. The preliminary works, in the form of schematics and illustrations, are presented in Appendix A (Figures A1-A5), which can be found in the section "Supporting Documentation".
RT2. Segmentation module (M5-M12) The objective of this research task is to develop a strategy for segmenting high-resolution images. This will involve an analysis of both the methods for segmenting image fragments from a large image and then combining the fragments in a manner that minimizes edge effects, and advanced algorithms for enhancing feature extraction.
RT3. Collaborative learning (M9-M20) Once these two models have been developed, they will be engaged in a collaborative learning process, thereby refining the segmentation and classification of breast lesions. The system will assist in the identification of imaging features related to cancer spread to the axillary lymph nodes. The concept of the proposed collaborative framework is presented in Appendix A (Figures A8-A9), which can be found in the section "Supporting Documentation".
RT4. Integration of the tool into the EUCAIM platform (M18-M24) This task includes the integration of the developed algorithms with the user interface software into the consortium tool repository. The objective is to adapt the tool's code to integrate seamlessly into the consortium's platform, as well as refining its functionality and presentation of results to enhance explainability and thus trustworthiness for different end-users. The tool will be designed in order to satisfy a variety of end-users, ensuring versatility in its applications. A standard graphical user interface (GUI) application will be provided as a pre-prototype, offering an intuitive interface for researchers, clinicians and non-technical users. Additionally, a more developer-oriented command-line interface will be made available, facilitating automatic segmentation and classification for advanced users and developers. This dual approach ensures that the tool can be used effectively in a variety of scenarios by consortium members.

**Description of the intended use and expected benefit related to the use of the EUCAIM data:**
The intended use of the EUCAIM data is to train and validate AI algorithms/tools for the automatic segmentation and classification of breast lesions in high-resolution medical images. In our use case, the availability of requested data could help in the identification of meaningful imaging features related to ALN metastasis. Such a tool could be successfully used with other consortium's biobank datasets containing imaging data annotated at the image and pixel levels, thus facilitating comprehensive analysis of cancer imaging features, for example research studies on identifying imaging biomarkers. The intended use of the EUCAIM data will benefit from its extensive and geographically diverse collection of medical images to enhance FAIR and unbiased research and development in cancer imaging. Moreover, the validation of the algorithms for other compatible biobanks could be of particular value due to the large variety of imaging modalities and its broad geographical coverage, which includes contributions from multiple stakeholders across Europe. The entire system will comprise three distinct AI models, a classification network, a segmentation network and a mask discriminator. Initially, the classification module will undergo training solely with image-level annotations. Given the

incompatibility of pretrained networks for such high image resolution, the images will be utilized in a fragmented way - in the form of a bag of instances. As a foundational architecture, we propose the utilization of deep neural networks (DNNs) trained within a Multiple Instance Learning (MIL) framework. The MIL approach will incorporate attention aggregation mechanisms for the instances' features. Other foundational models include transformer-based networks featuring self-attention mechanisms, cutting-edge of state-of-the-art (SoTA) mixtures of structured state space models (SSMs) such as Mamba and hybrid-transformers like Jamba, which are designed to improve long-range dependency in vision tasks. Such an approach will yield a model proficient in tumor classification while simultaneously generating coarse but accurate attention maps derived from aggregation weights. Furthermore, by leveraging eXplainable Artificial Intelligence (XAI) methodologies, we expect to obtain enhanced attribution mapping, later referred to as a 'pseudo-mask' (obtained after smoothing and thresholding). Furthermore, a neural network will be developed for segmentation directly using pixel-level annotations, although it will operate in a non-standard way, using image fragments. The foundational models within this module will be based on the U-Net encoder-decoder architecture, which will be responsible for generating the segmentation mask, later referred to as the 'weak-mask' (obtained after aggregation and smoothing). The core aspect of the methodology involves the collaborative learning of described DNNs to classify and segment breast lesions, using annotation at the image-level as well as at the pixel-level. In collaborative learning, pre-trained models will share the generated weak/pseudo masks. The data exchange will enable semi-supervised learning of the segmentation model, while the weak-masks will be used to label individual instances of the MIL classifier. In addition, a mask discriminator will be implemented, making the two models compete between themselves, whose task is to recognize the model used to generate both types of masks. For the implementation of the mask discriminator, we consider two options. The first option is to use a normal convolutional neural network utilizing weak and pseudo mask crops resized to compatible sizes, while the second option would introduce patch-based discrimination. During the learning process, the segmentation DNN will continuously attempt to 'outsmart' the discriminator, thereby improving the quality of the generated masks. At the same time, the discrimination process will optimize the thresholding value of the generated attentional maps in order to efficiently generate binary masks in the classification module. The planned approach assumes that simultaneous, joint learning of the classifier and the segmentation mask generator will lead to their optimal use under the given conditions.

**Description of the requested data:**
The target population for the breast cancer study should include women aged 18 and older who have been diagnosed with breast cancer, at various stages of breast cancer progression, covering both early and advanced stage cases as well as lesion free patients (healthy controls). This should also include information on metastasis to axillary lymph nodes (ALN). Ideally, the population should include patients from different medical facilities, varying medical devices used for examination, and diverse demographic backgrounds, including different ethnicities and geographic locations.

TYPE OF DATA

1.The anonymized full-field digital mammography (FFDM) images in DICOM format.
2.Segmentations in DICOM-RT-STRUCT/DICOM-SEG/NIfTI format segmentations.
3.Clinical information, including ground-truth (CSV/XLSX/JSON files):
1)Age
2)Cancer/benign/no cancer
3)Molecular subtype
4)Axillary lymph node metastasis Yes/No
5)Breast density
6)Optional and good to have information: Unilateral/bilateral breast cancer; Unifocal/multifocal lesions (To determine the method's efficacy in real-world data with multiple lesions); Other findings than the segmented cancer lesion in images Yes/No; Other findings than the segmented cancer lesion in segmentations Yes/No

DATASETS

Inclusion criteria for our study on breast cancer classification require female patients aged 18 or over with primary invasive breast cancer, and available FFDM imaging in DICOM format, including both breast CC and MLO views depending on availability, and all parameters defined in the proposal. Additionally, biopsy- confirmed diagnosis is essential. The exclusion criteria are as follows: males, minors, pregnant or lactating individuals, multifocality, carcinoma in situ without the invasive component, other cancers outside the breast, and incomplete data. From our perspective, the criteria for the recruitment period are of minimal significance. Nevertheless, the sole prerequisite is that the images be generated using FFDM technology.

NUMBER OF CASES

The collection of approximately 2,000 individuals would provide a solid foundation for the development of our algorithms. However, we would be grateful for as much data as possible. During development, it would be preferable if the stated number of cases were to take into account the preference for a balance of data sets. It is our preference to gather datasets that are balanced with respect to the labels/classes assigned to patients and across various demographic and clinical factors listed in the 'types of data' section. However, in order to test the tool in a real-life scenario, the number of patients would need to be even higher, as we know that only about 2-7% of all screening tests turn out to be cancerous.

| |
|---|
| ANNOTATIONS<br><br>Both image-level and pixel-level annotations are required. The algorithm developed can use only pixel-level annotations or image-level annotations, or both. Image-level annotations should indicate whether the breast lesion is present, and if so, whether it is benign or malignant (cancer), along with the ALN status. The dataset should include pixel-level annotation masks corresponding to the specific image-level class/label.<br><br>TOOLS<br><br>We would be pleased to employ any of consortium partners tools that would be advantageous for the aforementioned use case. Leveraging your technology could significantly enhance our project's outcomes. Furthermore, we are eager to cooperate and possibly, through the utilization and/or testing of your tools, we may contribute to their improvement. We see this as an opportunity to provide valuable insights and feedback that could enhance the functionality and effectiveness of your tools. COMPUTATIONAL RESOURCES AND TEMPORARY STORAGE<br><br>The data access method will determine the extent to which computational resources and storage will be required. If a data download option is available, we will provide the required computational resources. In the context of federated learning, it was estimated that the use of two to four GPUs would be beneficial, with a preference for cards with 48(24) GB VRAM and a CPU with 128(64) GB RAM and 4(2) TB memory storage. |

Table 33. Description of Use Case no. 5 from Fundacion Centro De Tecnologias De Interaccion Visual Y Comunicaciones

| Author: Iván Macía | Intention:<br>- train/validate AI tools | Organisation's name: Fundacion Centro De Tecnologias De Interaccion Visual Y Comunicaciones | Organization's Acronym: VICOMTECH |
|---|---|---|---|
| **Title of the use case:** Lung cancer detection from longitudinal LDCT and CT data (CONTACT) | | | |
| **General description of the use case:**<br>GENERAL DESCRIPTION: Lung cancer (LC) is the leading cause of cancer deaths worldwide. Early detection significantly reduces LC mortality by shifting diagnoses from late-stage, often incurable, to early- stage, which offers more curative treatment options, improves quality of life, and reduces economic impact. Currently, the main imaging modalities for managing LC are Computed Tomography (CT) and Low-Dose Computed Tomography (LDCT). CT identifies lung abnormalities and monitors treatment response, while LDCT is used for screening. Radiologists use the Lung CT Screening Reporting and Data System (Lung- RADS) to standardize the management of detected nodules. However, Lung-RADS involves image measurements that are not consistently and systematically performed, leading to variability in clinical practice. Image interpretation is also time-consuming and prone to errors. Research in computational radiology and computer-aided detection (CADe) and diagnosis (CADx) systems has been prominent in recent decades. Deep learning holds significant potential for automation and enhancement of AI-enabled image analysis pipelines, benefiting future screening programs, as we see in LUCIA project, part of the Understanding Cancer cluster of projects funded by HEU under Cancer Mission. Vicomtech leads image analysis tasks within LUCIA, collaborating with CHUL (Centre Hospitalier Universitaire De Liege, Belgium), Osakidetza (Basque Health Service, Spain), and SAS (Servicio Andaluz de Salud, Spain).<br>OBJECTIVES: Our goal is to advance in a computational radiology framework for lung nodule characterization and risk analysis by developing AI models for lung cancer detection in screening and diagnosis scenarios, applicable to CT and LDCT data. We aim to train and validate deep learning models for detecting, segmenting, characterizing (potentially with radiomics), and following up on lung nodules using multicentric longitudinal imaging data. To achieve this, we will use data from the National Lung Screening Trial (NLST) (26,000 longitudinal LDCT scans), already granted access to Vicomtech, retrospective and prospective data from LUCIA's clinical centers (over 25,000 longitudinal CT scans from clinical sites In Spain, Belgium, Latvia), published datasets (LIDC, LUNA), and imaging data from EUCAIM as part of this application.<br>METHODOLOGY: We will build on state-of-the-art techniques for nodule detection and segmentation using convolutional neural networks (CNN) with attention mechanisms. Weak supervision will be employed to use data without detailed annotations, and specific architectures will address domain shift, including self- supervised pretraining and contrastive learning. We will also explore CNNs and graph neural networks for nodule reidentification from longitudinal data. Our models will be trained, validated, and tested on NLST, LUCIA, LIDC, and LUNA datasets, with external validation using EUCAIM data (CHAIMELION and INCISIVE datasets).<br>EXPECTED RESULTS: The result will be a set of AI-based computational radiology tools and image analysis pipelines highly relevant for LC patient management and for quantitative image research. We will demonstrate our algorithms are applicable to both CT and LDCT data coming from multiple sources, proving their robustness and generalization to be translatable to clinical practice.<br>EXPECTED CLINICAL IMPACT: Integrating deep learning-based computational radiology tools in lung cancer detection is expected to significantly impact clinical practice by improving screening, diagnosis, treatment, and patient outcomes. These algorithms can analyze radiological images with high precision, potentially detecting early-stage lung cancer and subtle lesions that human radiologists might miss, enabling earlier intervention and better prognosis. They provide consistent evaluations, reducing variability and standardizing image analysis with | | |

quantitative metrics to aid clinicians in decision-making. These tools can monitor tumor changes over time, assessing treatment effectiveness and allowing adjustments as needed. Additionally, automated image analysis reduces radiologists' interpretation time, enabling them to focus on complex cases and handle more scans, improving workflow and reducing diagnosis time. Leveraging large datasets, our models aim to minimize false positives and negatives, enhancing patient safety by reducing unnecessary procedures and missed diagnoses.

**Expected timeline for the realization of the use case:**
We will need 24 months for training and validating our solutions within the EUCAIM framework. Shortly, the following tasks are anticipated:
T1 (M1-M4). Analysis of EUCAIM platform requirements, including model format requirements with regards to deployment into the federated framework, and architecture for data access, processing and AI model development. Fulfillment of ethics obligations.
T2 (M3- M6). Analysis of dataset characteristics and available annotations to ensure appropriateness of the data for training/validating each model and to define best model training approaches. Perform specific analyses to avoid data poisoning and biases.
T3 (M5-M9). Design of model architectures and image processing pipelines for nodule detection, segmentation, characterization (risk assessment) and reidentification based on existing state-of-the-art approaches.
T4(M8-M20). On-site training and validation of AI models using LUCIA, NLST and published datasets. We aim to train and validate deep learning models for detecting, segmenting, characterizing (potentially with radiomics), and following up lesions. To ensure trustworthiness, we will implement methods and procedures to analyze the performance of the models to identify potential biases and mitigate them during model development, and to interpret model outcomes. We will work on model explainability and follow a human-in-the-loop methodology.
T5 (M18-M24). Configuration and deployment of models in the EUCAIM platform (using Docker containers) for federated evaluation (using Flowers for example). Validation and results analyses using INCISIVE AND CHAIMELION datasets. Proposal of licensing model to transfer AI algorithms to the EUCAIM platform.

**Description of the intended use and expected benefit related to the use of the EUCAIM data:**
We plan to use EUCAIM data to validate our lung nodule detection, segmentation, and follow-up algorithms with external cohorts. Specifically, we aim at using the following collections: CHAIMELEON - Lung Cancer Imaging and clinical Data and INCISIVE Lung, which includes longitudinal, annotated data. We cannot find more specific information on the content of the datasets, but we could expect including data coming from other demographic locations and acquisition machines and protocols, which should be beneficial to test the robustness and generalizability of our models

**Description of the requested data:**
We would like to request data from the CHAIMELEON and INCISIVE lung datasets from the catalog. According to the provided information, there are approximately 4,000 cases, all presumably with available CT scans. For INCISIVE, longitudinal data appears to be available. We seek access to all available imaging studies from CT and LDCT modalities, regardless of the acquisition device, protocol, or originating center. We need information on patient/subject inclusion and exclusion criteria to ensure compatibility with our model requirements (e.g., primary lung cancer). Access to specific DICOM metadata is necessary to ensure applicability of our models, evaluate fairness, bias, and generalizability, and perform error analysis. Essential metadata includes patient sex, age, study date, manufacturer, model name, slice thickness/spacing between slices, pixel spacing, and radiation dose information if available. Additionally, we need information on the presence or absence of cancer for each imaging study and any available annotations (e.g., bounding boxes, segmentations, nodule matching information) to conduct validation studies.

Table 34. Description of Use Case no. 9 from SYCAI TECHNOLOGIES, S.L.

| Author: Júlia Rodríguez Comas | Intention:<br>- train/validate AI tools | Organisation's name: SYCAI TECHNOLOGIES, S.L. | Organization's Acronym: Sycai |
|---|---|---|---|
| **Title of the use case:** Validation of an AI algorithm to detect and classify pre-cancerous lesions in the pancreas. | | | |
| **General description of the use case:**<br>Cancer in upper-abdomen organs causes 1.4 million deaths worldwide every year. Pancreatic cancer is detected only in advanced stages, showing an extremely low survival rate of only 9%. Standard practice tries to identify tumors through diagnosis by medical imaging (computerized tomography scans, CT and magnetic resonance images, MRI). Previously to developing a malignant tumor, the organ can be affected by a focal lesion; this is an abnormal area or spot that can be identified on imaging tests. When focal lesions are identified in the early stages, the chances of successful treatment and positive outcome increases significantly. Remarkably, a non-negligible proportion of focal lesions are found incidentally in imaging scans, ranging between 35% and 65%, by radiologists. Some reasons are the millimetric size of lesions exceeding the limits of the human eye, lack of symptoms to request a CT/MRI test and increased workload in the radiology departments (2.5x images for the same staff). As mentioned above, Sycai Medical is a tool that addresses this critical issue: the lack of early-stage, affordable, patient-centric and automated cancer screening methods for the upper abdomen organs. The main objective of the project is to validate the tool we have developed and patented for the pancreas. Specifically, we | | | |

aim to assess the scalability of the algorithm we have trained and ensure that it operates in an unbiased manner. This involves evaluating its performance on a previously unseen dataset to confirm its robustness and generalizability across diverse data samples. We expect the obtained metrics to be satisfactory. In case any issues are detected or if the algorithm appears to be biased, we will work on modifying and improving it accordingly. This project aims to have a significant impact on clinical practice for healthcare centers, radiologists, and patients. For healthcare centers:
- Improve efficiency in healthcare center processes, saving diagnosis time, reducing waiting lists, and prioritizing critical cancer treatments.
- Deals with the exponential increase of imaging tests, that does not match with the few numbers of professionals.
For radiologists:
- Deals with the exponential increase of imaging tests, that does not match with the few numbers of professionals.
- Expands the visual limits of the human eye.
- Reduces mistakes on diagnosis thanks to more time used per test assessment when necessary.
- Alleviates workload by automating routinary assessments of CT/MRI exams, allowing professionals to expend time in high-value activities.
- Empowerment of healthcare professionals with a valuable decision support tool for more informed clinical decisions.
For patients:
- Saves time when it makes the difference between life and death.
- Provides quick diagnosis when a focal lesion exists, eliminating long-time undiagnosed cases.
- Avoids lifelong monitoring protocols through CT/MRI.
- Avoids unnecessary invasive and risky tests and surgeries.
- Improves quality of life and well-being by reducing stress and anxiety.

First, we will integrate with the EUCAIM platform by making the necessary modifications to ensure compatibility with one of the federated learning platforms within Cancer Image Europe. Following this, we will conduct thorough data collection to prepare a balanced dataset. The next step will involve running our algorithm on this dataset and obtaining comprehensive results. These results will be meticulously analyzed, focusing on metrics such as accuracy and sensitivity in detection and classification, as well as the accuracy of co-registration. This will help us evaluate our capability to assess the evolution of lesions effectively. If any issues or biases are detected during this process, we will refine and improve the algorithm accordingly.

**Expected timeline for the realization of the use case:**
The expected timeline for our project is presented below:

Activity 1. Integration into EUCAIM Platform
Conduct the necessary modifications to ensure compatibility with one of the federated learning platforms within Cancer Image Europe. This activity will be carried out by Javier García, Josep Julià, and Sergi Rojas from the technical department. The integration process will last 6 months.

Activity 2. Data Collection
Prepare a balanced dataset that meets all the specified requirements. This task will be developed by
Javier García, Miertixell Riera, Daniel Cañadas, Elena Martín, and Josep Julià from the technical team. The scientific team, consisting of Júlia Rodríguez Comas and Juan Moreno, will prepare the dataset demographics and provide clinical support. This activity is scheduled to last 4 months.

Activity 3. Validation of the AI Model
Run our model on the newly prepared dataset and obtain comprehensive results. This activity will be performed by Javier García, Miertixell Riera, Daniel Cañadas, Elena Martín, and Josep Julià from the technical team. The validation process will take 4 months.

Activity 4. Results and Analysis
Meticulously analyze the results, focusing on metrics such as accuracy and sensitivity in detection and classification, as well as the accuracy of co-registration. This analysis will help us evaluate our capability to assess the evolution of lesions effectively. If any issues or biases are detected, we will refine and improve the algorithm accordingly. This activity will be a collaborative effort between the technical team and the scientific team, lasting 4 months.

Throughout the project, the management department (Sara Toledano and Clàudia Saiz) will monitor progress to ensure the correct completion of all activities. They will also handle administrative tasks, legal matters, and project justification aspects. Furthermore, all activities will be supervised by the PI and CSO of the company, Júlia Rodríguez.

**Description of the intended use and expected benefit related to the use of the EUCAIM data:**
Validating our algorithm using the EUCAIM database ensures an entirely external validation process, promoting impartiality and fairness. By using data from a diverse range of European sources, we can confidently assess our algorithm's performance without internal biases. Leveraging this initiative aligns with our commitment to

ethical practices. Utilizing the full potential of the data acquired by EUCAIM allows us to refine and optimize our algorithm effectively. This collaborative effort ensures that our solution meets the highest standards and contributes positively to pancreatic cancer.

Our AI-based algorithm can be divided in 4 modules:

- DETECTION: CNN techniques and image processing algorithms are used to process a vast number of images to detect any focal lesion. A 3D reconstruction from the 2D images allows the software to perform a localization with pixel-wise accuracy. An own and unique interpretation of the Swint-Unet architecture for neural networks has been implemented. This architecture aims to encode and encode each CT scan in different overlapping patches to be able to locate with a pixelwise precision where the organs and their lesions are.

- FEATURE EXTRACTION: A morphological analysis is performed to extract the most important parameters characterizing the lesions. Our patent (WO2023030734) for pancreatic cystic tumors surpasses the SoA (EP3646240(A1), WO2021096991(A1), CN108898152(B), US20210012505(A1) by using innovative feature extraction techniques, achieving exceptional accuracy and speed as shown in results of our clinical validation. It is based on the approved European Clinical Guidelines. Computer vision is used to analyze the lesion candidates and to perform feature extraction using image radiomics as well as radiological features to characterize them, such as mean Hounsfield Units (HU), relative position (concerning the organ), among other factors. Image radiomics is used to analyze the grey distribution and image texture, checking the frequency distribution of pixels and the relationship between them.

- CO-REGISTRATION: We achieve a precise and automatic method for the tracking of single lesions by applying a unique approach conformed by an affine organ registration between pairs of available imaging studies of a patient, followed by a flexible registration between the generated masks to adapt the borders, and finalizing with a dilation operation to cover corner cases. This method is able to provide a unique ID to all lesions detected at a single time point and perform a spatial tracking of each one of them, precisely identifying where each of the lesions finds itself in consecutive imaging tests. This unique technique has been filed as an international patent with the Number 24382427.3 in April 2024.

- CLASSIFICATION: To prognose the probable evolution of the lesion, a ML-based predictor is applied using not only morphological characteristics but also its past evolution. SYCAI's current patented diagnostic method focuses on classifying the lesions, distinguishing those with malignant potential from those likely to remain benign. According to the European Clinical Guidelines, there are different worrisome features that indicate malignancy of the lesions. Some of them are related to morphological features that lesions present at the time of analysis and some are related to their evolution through time. After parametrizing the detected lesions through several feature extraction methodologies (radiological and radiomics features), our method firstly finds the characteristics that have a more significant impact on the final classification and, secondly, trains a classifier with these more relevant features that can quickly compute the malignant potential of the analyzed lesion. For the latter, several classification models are currently used and combined to give the best results: gradient boosting, random forest, CNN, and support vector machines. Together with this, to define which of all the extracted features have more considerable relevance in the final classification, we currently apply a well-known method named LASSO.

**Description of the requested data:**
- Ct scans belonging to thorax, abdomen, pelvis or combination of those or uro-CT.
- At least two consecutive CT scans belonging to each patient shall be included in the database, if possible - With portal-venous or arterial contrast.
- Target population: over 18 years old. Not pregnant. Around 40% of the data shall belong to "healthy patient" (with no PCL diagnosed). Around 50% of the data shall belong to patients diagnosed with PCL. 10% of the data shall belong to patients diagnosed with pancreatic adenocarcinoma or neuroendocrine tumor.
- Type of data: Modality CT scan, DICOM files with SeriesDescription, StudyDescription, Age, Sex, PixelSpacing, InstanceNumber and ImageOrientationPatient metadata.
- Datasets: between 2010 and 2023
- Number of cases: 400
- Annotations: pancreatic cystic lesion (IPMN, SCA, MCN or presudocyst), pancreas, pancreatic adenocarcinoma, neuroendocrine tumor. The annotations shall be semantic segmentation annotations. - Computational resources: at least 10GB of HDD, GPU with at least 16GB of RAM. At least 4GB of CPU
- RAM.

Table 35. Description of Use Case no. 42 from Universitätsklinikum Heidelberg

| Author: Hans-Ulrich Kauczor | Intention:<br>- train/validate AI tools | Organisation's name: Universitätsklinikum Heidelberg | Organization's Acronym: UKHD |
|---|---|---|---|

**Title of the use case:** Deep learning-based detection and assessment of NSCLC tumors on chest CTs

**General description of the use case:**

The currently available AI-based algorithms are limited to segmenting peripherally localized lung nodules and do not apply to patients with large tumor masses or tumors centrally localized around large blood vessels. However, patients with non-small cell lung cancer (NSCLC) are usually diagnosed at more advanced stages when the tumor is unresectable but is usually suitable for chemotherapy. Accurate segmentation of these advanced-stage tumors could allow automated extraction of radiomics features that correlate with genetic mutations, paving the way for radiomics-based "virtual biopsies". Moreover, AI based segmentation can also play an important role in patient follow-up assessing response to chemotherapy.

The primary aim of our study was to train an AI algorithm to segment these advanced-stage NSCLCs. The proposed version of our AI tool can segment the primary lung tumor on contrast-enhanced CT scans with promising accuracy. The next phase would be to finetune and validate it on datasets in the EUCAIM database. Depending on the number and quality of CT scans and the genetic data provided by the data holders joining the EUCAIM project, our study also aims to extend our segmentation algorithm by automatically extracting radiomics features to identify imaging biomarkers of targeted genetic mutations

(e.g. EGFR, ALK). The identified predictive radiomics features would be used to build a machine learning model to predict the mutational status of tumors. A secondary aim of our project is to further extend our AI tool to segment target lung lesions to assess response to treatment at follow-up scans based on tumor volume changes.

In our initial study for developing the proposed AI tool, patients diagnosed with NSCLC at the Thoraxklinik, Universitätsklinikum Heidelberg between 2009-2019 were retrospectively identified from the Data Warehouse of the German Center for Lung Research (DZL-DWH). The final cohort consisted of 392 eligible patients prominently with advanced-stage disease. The baseline CT scans and the respective histological mutation status of the tumors (ALK, EGFR, or wild) were exported anonymously from the biobank. In each case, only the primary lung tumor was segmented using the ITK-SNAP reviewed by two board-certified radiologists, the results were then discussed in consensus with an expert thoracic radiologist. The CT images and the corresponding segmentation masks were exported in NIfTI file format, no personal data was used during the development of our AI tool.

The study cohort was divided into a training (229 patients), a validation (81 patients), and a test dataset

(82 patients). The nnU-Net trained with 5-fold cross-validation showed promising results with Dice scores of 0.69-0.72. Due to the relatively small number of patients and the unbalanced patient cohort, radiomics based mutation analysis was not feasible in a single-center study design, however, we believe that a radiomics-based tumor mutation prediction model can be built using our data combined with external datasets of the EUCAIM database.

The radiomics analysis will be carried out according to current best practice guidelines using the pyRadiomics package and widely used data-mining, machine learning, and statistics packages coded in Python. The final output of our “virtual biopsy” algorithm would be the segmentation of the primary lung tumor and a score on the probability of its mutation status. We also aim to further develop our tool for segmenting target lung lesions to automatically assess their volume changes. However, these sub-projects are dependent on the available genetic data and follow-up CTs of NSCLC cases in the EUCAIM database.

After successful validation, we believe that our AI tool could facilitate patient care in the future by automatically evaluating baseline and follow-up CT scans of patients with advanced-stage NSCLC by detecting the tumors, assessing the probability of having target mutations, and evaluating their response to chemotherapy.

**Expected timeline for the realization of the use case:**

After being successfully selected to join the EUCAIM project, the first 1-2 months would be spent preparing the necessary documentation and adapting the docked AI tool to make it suitable for use in the federal learning environment. The next 2-4 months (depending on the datasets made available by data holders) would be spent testing the current version of our AI tool on the available EUCAIM datasets, including fine-tuning and validation. The following steps would be mainly dependent on the datasets provided by the data holders who join the EUCAIM project. Depending on the number and quality of CT datasets and the genetic data available in the EUCAIM database, the next 3-4 months would be devoted to radiomics analysis of tumors, including the identification of imaging biomarkers correlated with target mutation status, and the construction of a machine learning-based model predicting tumor mutation status by providing a risk score. In this study, depending on the number of suitable cases available in the EUCAIM project, the internal dataset of our institution would be used as a training dataset, and the constructed radiomics model would be tested on the EUCAIM datasets, but if a considerable number of NSCLC patients would be available, we would also use the EUCAIM datasets for training and validation. If the radiomics analysis would not be feasible due to the lack of data or if the analysis can be finished in time without difficulties, and if follow-up CT scans will also be provided by the data holders who join the EICAIM project, a later step would be to further develop our segmentation AI-tool for the identification of target lung lesions and quantitatively assess their volume changes which would take approximately 8-10months.

**Description of the intended use and expected benefit related to the use of the EUCAIM data:**

As an AI tool provider, our primary aim is to validate and finetune our proposed lung tumor segmentation algorithm on the lung cancer image data of the EUCAIM repository. The EUCAIM repository currently includes

| |
|---|
| two available lung cancer datasets, namely the CHAIMELEON – Lung Cancer Imaging and Clinical Data dataset, and the INCISIVE Lung dataset. Access is requested primarily to these existing datasets, but as it is likely that the EUCAIM repository will be expanded in the near future to include datasets from data holder applicants, additional access to future lung cancer datasets will also be requested where possible. We would require access to these databases for federated processing. Our AI tool was trained for the segmentation of the primary lung tumor on contrast-enhanced chest CT scans of advanced-stage NSCLC patients. Trained on a single-center patient cohort it was able to achieve promising results in the segmentation of the primary lung tumor. However, the validation and finetuning of the algorithm on independent external cases would increase its reliability and generalizability. The radiomics-based non-invasive prediction of tumor mutation status was not feasible in our single-center study design due to the unbalanced patient cohort, however, we believe that this issue could be solved by incorporating external datasets of the EUCAIM repository and continuing this study with a multi-centric study design. In order to identify imaging biomarkers predictive of EGFR and ALK mutational status, we need imaging and genetic data from NSCLC patients in the EUCAIM dataset, as well as a clear indication of which tumor lesion was sampled by biopsy to evaluate the mutation status. In this study, depending on the number of suitable cases available in the EUCAIM project, the internal dataset of our institution would be used as a training dataset, and the constructed radiomics model would be tested on the EUCAIM datasets, but if a considerable number of NSCLC patients would be available, we would also use the EUCAIM datasets for training and validation. We would also further develop our segmentation algorithm to be able to recognize and segment the target lesions besides the primary tumor and compare the follow-up scans of patients to evaluate the volume changes of the tumor lesions. Therefore, we require the baseline and follow-up contrast-enhanced axial CT scans of patients with advanced-stage NSCLC who were treated with targeted therapies. Besides the CT scans, the basic demographic data, the TNM staging, and the genetic data, we also require the segmentation masks of the primary tumor as well as the target lesions. The segmentation masks of the lesions should be labeled separately either by using different voxel values (multiclass segmentation) or by exporting the binary label maps of the lesions into different files. The data should be made available in an anonymized format removing any personal data that would allow patient identification, however, each case should be assigned with an ID that enables the comparison of the baseline and the follow-up scan. DICOM format CT scans are appropriate, but the input file format of our nnU-Net-based algorithm is NIfTI, therefore NIfTI (.nii.gz) format CT scans are also acceptable<br><br>**Description of the requested data:**<br>The main interest of our study is contrast-enhanced axial chest CT scans of patients diagnosed with NSCLC. The EUCAIM repository currently includes two available lung cancer datasets, namely the CHAIMELEON – Lung Cancer Imaging and Clinical Data dataset, and the INCISIVE Lung dataset. Access is requested primarily to these existing datasets, but as it is likely that the EUCAIM repository will be expanded in the near future to include datasets from data holder applicants, additional access to future lung cancer datasets will also be requested where possible. We would require access to these databases for federated processing. We require access to the baseline and follow-up contrast-enhanced axial CT scans of patients with advanced-stage NSCLC who were treated with targeted therapies. Besides the CT scans, the basic demographic data, the TNM staging, and the genetic data (target mutation status), we also require the segmentation masks of the primary tumor as well as the target lesions. The segmentation masks of the lesions should be labeled separately either by using different voxel values (multiclass segmentation) or by exporting the binary label maps of the lesions into different files. No filtering for sex, age, or time of examination is required.<br>The data should be made available in an anonymized format removing any personal data that would allow patient identification, however, each case should be assigned with an ID that enables the comparison of the baseline and the follow-up scan. DICOM format CT scans are appropriate, but the input file format of our nnU-Net-based algorithm is NIfTI, therefore NIfTI (.nii.gz) format CT scans are also acceptable. The computational resources should allow a Linux system, and sufficient RAM and GPU to train and test the nn-UNet algorithm on NSCLC patients currently available in the EUCAIM catalogue and on NSCLC patients provided by data holders joining the EUCAIM project. According to the original documentation of the nnU-Net, the recommended workstation Hardware configuration for training is: CPU: Ryzen 5800X - 5900X or 7900X; GPU: RTX 3090 or RTX 4090; RAM: 64GB; Storage: SSD (M.2 PCIe Gen 3 or better!). Example Server configuration for training is: CPU: 2x AMD EPYC7763 for a total of 128C/256T. 16C/GPU are highly recommended for fast GPUs such as the A100! GPU: 8xA100 PCIe, RAM: 1 TB, Storage: local SSD storage (PCIe Gen 3 or better) or ultra-fast network storage. The radiomics analysis itself and the construction of the prediction model require CPU and RAM. |

Table 36. Description of Use Case no. 11 from Istituto Europeo di Oncologia

| Author: Filippo Pesapane | Intention: - train/validate AI tools | Organisation's name: Istituto Europeo di Oncologia | Organization's Acronym: IEO |
|---|---|---|---|
| **Title of the use case:** Enhancing Breast Cancer Diagnosis with AI: Deep Learning Based Detection and Discrimination of Mammographic Findings | | | |
| **General description of the use case:**<br>We developed a deep learning-based tool aimed at improving the detection and discrimination of breast microcalcifications on mammography. This tool is designed to enhance the accuracy and efficiency of breast cancer diagnosis, with plans to extend its application to other mammographic suspicious findings such as | | | |

radiopacities and distortions. Additionally, the tool will be adapted for use with tomosynthesis, providing enhanced 3D imaging capabilities beyond traditional 2D mammographic projections.
Main Objectives

- Validate AI Models: Conduct rigorous validation using diverse datasets to ensure robustness and reliability.
- Enhance Detection Accuracy: Improve sensitivity and specificity in detecting breast microcalcifications and other mammographic findings.

Expected Results

- Comprehensive AI Validation: Strong evidence supporting the efficacy and safety of the AI tools.
- Improved Diagnostic Performance: Higher accuracy in detecting and classifying breast anomalies.
- Reduced Diagnostic Errors: Decrease in false positives and negatives.
- Enhanced Clinical Decision-Making: More reliable diagnostic information for radiologists.
- Expected Clinical Impact
- Early Detection: Better ability to detect breast cancer at earlier stages.
- Operational Efficiency: Streamlined diagnostic processes reduce radiologist workload.
- Patient Outcomes: Improved diagnostic accuracy and early detection lead to better overall patient outcomes.

Methodology

Model Training:

Use deep learning frameworks to train AI models, focusing on detecting and discriminating microcalcifications, radiopacities, and distortions.

Implement advanced techniques such as convolutional neural networks (CNNs) and transfer learning.

Validation and Testing:

Conduct rigorous validation with separate test datasets to evaluate accuracy, sensitivity, specificity, and overall performance.

Perform cross-validation and independent external validation to ensure robustness and generalizability.

Integration and Deployment:

Develop protocols to incorporate the validated AI tools into clinical workflows.

Pilot the AI tools in clinical settings to assess practical utility and gather feedback from radiologists and clinicians.

Continuous Improvement:

Monitor AI tool performance in clinical use and update models based on new data and feedback. Engage in ongoing research to refine algorithms and expand diagnostic capabilities.

This use case exemplifies IEO's commitment to leveraging advanced technologies and collaborative efforts to improve breast cancer diagnosis: by validating and integrating these AI tools, the project aligns with the European Cancer Imaging Initiative's goals of enhancing clinical outcomes through the innovative use of AI and medical imaging.

**Expected timeline for the realization of the use case:**
This timeline provides a structured approach to realize the use case within two years, with clear phases for data ingestion and Data analysis, Model Validation, Fine-Tuning and Final Validation. It allows for iterative feedback and adjustments, ensuring the final product meets stakeholder expectations and performs reliably in real-world scenarios.

8 months: Data Ingestion and Data Analysis

- Identification and Ingestion on local storage of the data needed for the validation• Data analysis
- Clean and preprocess the data, ensuring it is ready for analysis and model testing.

4 months: Model Validation

- Execution of the model

- Performance extraction
- Performance analysis

8 months: Fine-Tuning

- Model fine-tuning
- Performance extraction
- Performance analysis

4 months: Final Validation and Documentation

- Final validation of the model performance
- Documentation

**Description of the intended use and expected benefit related to the use of the EUCAIM data:**
The validation process is a fundamental and necessary step that enables the use of AI models in daily practice. Without proper validation, the reliability and accuracy of an AI tool cannot be assured, which is crucial for its application in sensitive fields such as medical diagnostics. Validation ensures that the AI model performs well not only on the data it was trained on but also on new, unseen data. This step helps to identify any potential biases, errors, or limitations of the model, thereby providing confidence in its predictions and decisions.
We intend to make use of the data provided by EUCAIM to carry out this validation step, in order to validate our AI model developed for localizing and characterizing microcalcifications on mammograms external to the institute. EUCAIM's dataset is diverse and representative, making it an ideal source for testing the robustness and generalizability of our model.
The AI algorithm implemented is a convolutional neural network (CNN), which is a type of deep learning model particularly well-suited for image analysis tasks. Our CNN was trained on a dataset that included 1000 patients aged 21–73 years and 1986 mammograms. This dataset was carefully curated to include a variety of cases, with 389 malignant and 611 benign groups of microcalcifications.
The training process involved feeding the CNN of annotated mammography images, allowing the model to learn to localize and characterize microcalcifications. This learning process was iterative, with the model constantly adjusting its parameters to minimize prediction errors.
Once the training was completed, initial validation was conducted using a subset of the original dataset to ensure the model's efficacy. However, to truly assess its performance in real-world scenarios, external validation using independent datasets, such as those provided by EUCAIM, is essential. This external validation will help to confirm that the CNN can accurately localize and characterize microcalcifications across different populations and imaging conditions.
In conclusion, validation is not just a procedural formality but a critical step to establish the credibility and effectiveness of AI models. By leveraging comprehensive and diverse datasets like those from EUCAIM, we can ensure that our AI algorithm is reliable, accurate, and ready for integration into clinical practice, ultimately enhancing diagnostic capabilities and patient outcomes. Additionally, the validation phase with EUCAIM data could be also adopted in the next evolution of our AI tool that aims to use tomosynthesis, providing enhanced 3D imaging capabilities beyond traditional 2D mammographic projections.

**Description of the requested data:**
The dataset that our use case requires includes high-quality, high-resolution mammography images from a diverse cohort of patients. Each image needs to be meticulously annotated to localize the microcalcifications and provide detailed information on their benign or malignant characteristics. These annotations are crucial for training and validating our AI model effectively.
To ensure a balanced and representative dataset, it should include mammography images with both benign lesions and histologically proven breast cancers. This balance is necessary to train the model on a wide spectrum of cases, allowing it to learn the distinguishing features of both benign and malignant microcalcifications. The dataset should comprise at least 1000 cases to provide sufficient data for robust model training and validation.
If the data provided by EUCAIM are not annotated, we will undertake the annotation process ourselves. This involves collecting and organizing the mammography images, ensuring they are in a standardized format and meet our quality and resolution requirements. A team of radiologists and domain experts will then meticulously label each image, identifying and marking the locations of microcalcifications. Each identified microcalcification will be classified as benign or malignant based on histological evidence and clinical guidelines.

Table 37. Description of Use Case no. 59 from Università degli Studi di Bari Aldo Moro

| Author: Sabina Tangaro | Intention: - train/validate AI tools | Organisation's name: Università degli Studi di Bari Aldo Moro | Organization's Acronym: UNIBA |
|---|---|---|---|

**Title of the use case:** eXplainable Artificial Intelligence for Breast Cancer - XAI_Breast_Cancer

**General description of the use case:**

XAI_Breast_Cancer will provide the technological enablers for experimentation of AI-based solutions to improve prediction, diagnosis and contributing to a more precise and personalized management of cancer.
In particular augmenting the interpretability of Machine Learning approaches making predictions to be sufficiently understandable or interpretable to humans. This project will provide the users explanations for patient-specific predictions as well as to peer into a model and understand how predictions are made. XAI_Breast_Cancer will grade general views of which prediction features are essential to assign a patient to a particular clinical outcome providing other more detailed and graphical depictions of evidence relationships underpinning individual predictions. Furthermore, XAI_Breast_Cancer will increase the acceptability of Machine/Deep Learning (ML/DL) in the clinic, and it will support the clinicians with the generation of new hypotheses and in understanding the mechanisms underlying particular pathological states for a better decision making.
In previous studies, our group proposed a CAD scheme based on ROI localization, feature extraction, and classification with machine learning algorithms to recognize cancer lesions and microcalcifications in mammography and breast CT. These algorithms will be trained and tested on a large and multi-site database of this project.
A Computer Aided Detection and Diagnosis (CAD) software system for mammography and breast CT on distributed databases will be developed. The use of automatic systems for analyzing medical images is of paramount importance in screening programs, due to the huge amount of data to check.
Many machine learning models, such as boosted trees, support vector machines and neural network based methods are applied to analyze high-dimensional data for many real-world applications. Usually higher complexity allows higher accuracy, but at the expense of interpretability .
Indeed, there is a trade-off between model interpretability and prediction accuracy of current machine learning models.
In practice, model interpretability is as important as accuracy in many critical applications such as clinical decision-making context, in which the understanding of how the model makes the prediction is the key to facilitate physicians to trust the model and utilize the prediction results.
Interpretable machine learning, or XAI, aims to create a suite of techniques that produce more explainable models while maintaining a high accuracy for several reasons:

- The prime motivation behind model interpretability is building trust in models. If a model i s also trustworthy, found patterns are explainable in supervised manner by experts, while bias and errors can be easily identified and the models can be corrected. Moreover, model interpretability aims to provide accountability to model predictions. Many times, good evaluation metrics don't guarantee good real-world performance.

- Second, an interpretable model helps in understanding causality [Miller, Tim. Explanation in artificial intelligence: Insights from the social sciences. Artificial Intelligence, 2018.]. Indeed, interpretable models can extract the associations between predictors and predictions.

- Interpretable models would help bridging the gap between data scientists and domain experts and enable an effective exchange of data insights and knowledge.

Explanation systems will be applied to ML algorithms in order to provide a clear picture of the relevant features affecting the performance of the models. These steps will help the specialists to perform supervised considerations and to adopt the best strategies in a decision-making clinical process.
By exploiting a wide repository of data sources-collected, anonymized, the power of AI can identify patterns across, and connections between, all the features in the data. The success of this precision medicine potentially represents a game changer for healthcare democratization, where patients will take advantage of a number of benefits, such as time reduction and better accuracy in diagnosis, improved ability to predict which treatments will work best for specific patients, and new approaches to preventing, diagnosing, and monitoring disease.

**Expected timeline for the realization of the use case:**

1. Radiomic Feature extraction and classifications (0-6 months).
2. Deep/Machine learning: training-validation (7-14 months).
3. Explanation systems for Machine learning Interpretability (15-18 months)
4. Decision Support System for breast cancer diagnosis: service implementation (20-24 months).

**Description of the intended use and expected benefit related to the use of the EUCAIM data:**

The UNIBA group has been involved in activities related to the analysis and understanding of images and patterns for many years, and has in general gained an expertise on image processing, pattern recognition, machine learning, complex network analysis and related applications in diagnostic medical imaging. The aim of the research group is to develop and deploy related applications in the field of diagnostic medical imaging, besides a particular attention is given to the "Big Data" approach, distributed computing on grid environments as the European Grid Infrastructure (EGI) and cloud services for large scale.
The XAI (eXplainable Artificial Intelligence) algorithm that will teste on EUCAIM data could be described in the following steps:

1. Data harmonization. If available information of image quality will be used to eliminate bias due the acquisition by different imaging systems. After the integration of different sources (meta-data, demographics, imaging), phenotypic information of subjects will be used to improve standardization algorithms for eliminating site effects and compare the different datasets. A complex framework will be implemented to perform an effective harmonization processing of medical imaging data in order to remove unwanted variation associated with sites and preserve biological associations in data.

2. Preprocessing and Radiomics Feature extraction. Several algorithms have been developed to extract useful information from medical images by using High Throughput Computing Environment. Different techniques have been specifically implemented and integrated to build custom pipelines that formalize the main steps for preprocessing and analysis of 2D and 3D medical images (mammography and CT for breast cancer). The workflow allows the end user to upload medical images which could be analyzed and radiomic descriptors can be extracted.

3. Deep/Machine learning. Many machine learning models, such as boosted trees, support vector machines and neural network-based methods are applied to analyze breast images. Usually higher complexity allows higher accuracy, but at the expense of interpretability. A compressive framework with multiple Machine Learning algorithms has been developed to be applied for each of the hypotheses under investigation, e.g., both linear and nonlinear classification models for controls/patients discrimination or regression models for disease's staging.

4. Explanation artificial intelligent systems (XAI) have been included in such a framework in order to provide a clear picture of the relevant features affecting the performance of the models, their relations with the outcomes and with each other and both their local/global effects on the problem under investigation.

The proponent research group has expertise in medical database integration and management. We developed different solutions to manage big datasets, process them, store results in an efficient manner and make all the pipeline steps available for reproducible data analysis. In particular, data retrieved from different servers are aggregated according to the BIDS and other medical standards and the resulting data sets can be processed by using multiple pipelines that are easily customized and integrated into multiple computational workflows. The processing environment will be optimized for large datasets since it has been designed to reliably preprocess and analyze data for hundreds of subjects. The resulting environment will be scalable to make the analysis pipelines easier to compare and to select the best strategy with the optimal parameters.
Validating a XAI algorithm on a large and heterogeneous dataset is essential to ensure that the model not only performs well on familiar data but is also prepared to tackle challenges in real-world environments. This process enhances robustness, reduces bias, improves performance, and prepares the model for a wide range of practical applications, while simultaneously boosting the trustworthiness of the method.

**Description of the requested data:**

- meta-data, demographics,
- annotated mammographic images and (if available) data related to image quality- annotated Breast CT images and (if available) data related to image quality

Table 38. Description of Use Case no. 39 from Better Medicine OU

| Author: Dmytro Fishman | Intention: - train/validate AI tools | Organisation's name: Better Medicine OU | Organization's Acronym: BM |
|---|---|---|---|
| **Title of the use case:** Developing multi cancer AI | | | |
| **General description of the use case:** Developing and clinically validating a multi organ and multi cancer AI detection and quantification model covering primary tumors and metastatic spread. Organs: | | | |

| |
|---|
| Lymph Nodes<br>Pancreas<br>Liver<br>Bile Ducts<br>Gallbladder<br><br>**Expected timeline for the realization of the use case:** 2 years<br><br>**Description of the intended use and expected benefit related to the use of the EUCAIM data:**<br>We are doing a full-body solution to speed up oncology workflow in radiology. For that we are building automated AI models that can detect/classify and measure lesions in all organs of the abdominal cavity. We have already achieved substantial progress with models for kidney and lung, have PoC models working for liver and pancreas, while models for lymph nodes and bones are next in line. This sort of work requires rich datasets of CT scans with lesions located in different organs.<br><br>**Description of the requested data:**<br><br>● Multiphase CTs that can cover different cancers in different organs.<br>● Regions covered: either one or combined - thorax, abdomen, pelvis.<br>● Additional clinical data: diagnostic proof. It is important to have scans with proven diagnoses - be it histology or a definite finding based on radiological features. |

- Both:

Table 39. Description of Use Case no.20 from Stichting Amsterdam UMC

| Author: Vera Keil | Data sharing: Central Repository | Organisation's name: Stichting Amsterdam UMC | Organization's Acronym: AUMC | Tier: 1 |
|---|---|---|---|---|
| Intention:<br>- development, validation or training of AI tools considered for medical devices | | | | |
| **General description of the potential use and clinical impact of the shared data:**<br>We estimate that the IMAGO glioma dataset, offered as a contribution to the EUCAIM project, is one of the largest and most complete glioma database globally. It contains MRI data of up to 1,200 subjects with a mean of 10 long-term follow-up scans. The MRI data provides a substantial percentage of advanced MRI sequences, including fMRI (functional MRI), SWI (depichts blood and calcium), DTI (shows fiber bundles), DSC (perfusion with contrast agent), ASL (perfusion without exogenous contrast agent), and APT-CEST (a protein-measuring technique). The data is multi vendor and multi-site data as the VUmc is a tertiary hospital. This means that several early (and sometimes follow-up) scans are forwarded to this hospital to be evaluated. Furthermore, we collected corresponding data on therapy, pathology, and survival.<br>The dataset allows the testing of innumerable imaging-based research questions and provides a sample size that is especially relevant for artificial intelligence -based research questions such as segmentation, but also prediction of survival, prognosis, or disease subtyping. Advanced MRI data is particularly rare and can be used to create larger bundles of advanced MRI data to answer dedicated research questions. The presence of follow-up MRI scans allows training of models that do not focus on preoperative scan evaluations but on therapy surveillance and risk stratification, which is unique and has large clinical potential. An example might be the development of models that predict disease status – stable vs progressive disease – after surgery by using DSC. The model then has a theoretical basis, namely that DSC has the same usage in normal clinical practice, and can be used to drive decision-making. | | | | |
| **Dataset 1**<br><br>**Cancer Type:** Diffuse glioma<br><br>**Dataset name:** IMAGO<br><br>**Dataset description:**<br>The dataset is largely monocentric, from Amsterdam UMC, and consists of longitudinal MRI plus clinical data of adult-type diffuse glioma patients starting in 2008 to the present. Some MRI are originally from external sites (data import). | | | | |

Prof. Dr. P. Wesseling has verified the diagnosis centrally according to the latest WHO classification for 90% of cases. The distribution of glioma entities is according to the normal incidence in nature and contains thus more than 50% glioblastoma (WHO grade 4).

Patients gave written informed consent to the use of their data for scientific purposes during their first presentation at the Neuro-oncology unit. Approximately 85-90% of patients are deceased. We are currently cross-checking with the alive patients to see if they still consent to the contribution.

Currently, 1,200 cases are registered (with a raw data estimate of 1,400 cases available). These have an average of 10 MRI scans starting at the pre-operative evaluation MRI. The MRI scans usually follow the generally acknowledged brain tumor imaging protocol (BTIP; Ellingson B et al. 2015) sequences (T1w, T1+contrast, T2w, DWI with ADC map, T2 FLAIR). However, numerous cases contain perfusion maps from either ASL or DSC perfusion and other additional sequences (see above).

We have added corresponding clinical information from the hospital PACS system, as listed above, and gathered/arranged it in the fashion below.
Relevant note:
It is possible to split the dataset into:

1. a tumor-segmented version
2. a BTIP protocol-only version
3. an advanced MRI sequence subcontainer

**Dataset Collection Method:** Longitudinal

**Dataset Type:** Annotated Dataset

**Dataset Terms of Use:** There is currently no DUA or data access request process implemented for the IMAGO dataset. Should we receive funding from EUCAIM, we would formulate both harmoniously with EUCAIM regulations.
Our preliminary concept for a DUA embraced the following pillars:

1. Amsterdam UMC remains the data controller of the dataset, while the user becomes a data processor and, therefore, will process the data only on behalf of Amsterdam UMC.
2. There will be no attempt to establish or retrieve the identity of the study participants.
3. The data will not be redistributed with others unless they have independently applied and been granted access.
4. Secondary and derived data will be shared only at the group level.

The AUMC Publication Policy should be used when publicly presenting any results or algorithms.

**Dataset Intended Purpose:** The data shall be used to advance scientific developments for the benefit of patients with brain tumors, particularly through artificial intelligence approaches, but not limited to these. The dataset shall not be sold or otherwise commercialized. Patients only consented to scientific use.

**Imaging Modality:** Magnetic resonance image

**Vendor:** diverse (Philips, Siemens, GE mainly)

**Imaging body part:** head (brain)

**Age range:** 18+

**Sex:** Male and Female

**Number of subjects:** >1200

**Number of DICOM studies:** about 208000 single sequences

**Image size in GB:** Imago in zipped format is 2TB.

**De-identification:** Personal data is fully anonymized

**Title of the use case:** TumorTrace

**General description of the use case:**

General note:

As EUCAIM currently does not provide any glioma data, we plan to perform the AI use case on our own IMAGO data.

Overview:

Gliomas show a non-linear growth pattern, which can be influenced by therapy. Monitoring of gliomas occurs with MRI. It would be of exceptional value to anticipate glioma growth activity to be able to time and choose therapies. For the radiologist, it is hardly possible to tell from the MRI images if a tumor will remain stable or, instead, will rapidly expand in the following weeks. However, the information of growth over time may be present in the images and can be decipherable with AI.

Main objectives:

To develop, test, and validate an AI tool that can predict imminent acceleration in tumor growth or stable disease. Such a tool can be used during multidisciplinary meetings to drive decision making in favor of the patient's health.

Expected results:

The envisioned result is a set of DL algorithms on TRL 6 that can answer the main objective research questions from 1. preoperative MRI images and 2. longitudinal image input of a patient. More specifically, the hypothesis is that tumors grow and behave differently after treatment. Therefore there will be at least two separate AI models that either have preoperative or postoperative data as input to extract the different tumor behaviors and maximize performance.

Clinical impact:

Clinicians can prioritize patients better on therapy waitlists, which will likely positively impact survival. TumorTrace can also assist in decision-making for wait-and-scan eligibility of a patient with potential lower-grade glioma or to determine MRI follow-up intervals. TumorTrace results can serve as a support tool during interdisciplinary case evaluations (MOT). The application can be extended to non gliomatous tumors, e.g., in brain metastases or cancer primaries elsewhere.

Methodology:

The IMAGO dataset will be the basis for the dataset. We will first volumetrize all tumors on T1w, postcontrast, and T2w images and perform voxel-based measurements of diffusion ADC and perfusion values. We will then produce longitudinal volume change maps incorporating therapy as a cofactor. We can then model therapy and histology-specific growth curves for hypothetical average tumors. Each deviation from this growth curve by an identically categorized glioma will correspond to accelerated (or decelerated) growth.
There is a large subset of patients that receive a biopsy rather than open surgery. It can be argued that each glioma subtype requires its own model; glioblastomas, IDH wildtype, behave differently than lower grade tumors. This knowledge might drastically improve model performance and carry similar diagnostic value as described above.

**Expected timeline for the realization of the use case:**
For clarification: We consider the provision of the IMAGO data as use case 1, and the AI project as use case 2. We expect six months time to provide the updated and curated data for the IMAGO database (use case 1: data provision), which we would like to use as the basis for our AI model (use case 2). The AI use case itself is estimated to necessitate 18 months of development and testing:

Months 1-6: Volumetries and volumetric quantitative measurements.

Month 7: Data correction and quality control.

Month 8-12: Data splitting, model training.

Months 12-14: Testing and updating.

Month 15: Validation (if possible also including external data).

Months 16-18: Data analysis and manuscript preparation.

The total project duration is thus 24 months.

**Description of the intended use and expected benefit related to the use of the EUCAIM data:**
This section may be more suitable if data is demanded from existing EUCAIM-stored databases. However, here, we have the exceptional case of using our own data as an AI use case.

**Description of the requested data:**

| |
|---|
| See above. We will use our own IMAGO data, as there is currently no glioma data in EUCAIM we could use alternatively. |

Table 40. Description of Use Case no.43 from Fundacion De La Comunitat Valenciana Para La Gestion Del Instituto Deinvestigacion Sanitaria y Biomedicade Alicante

| | | | | |
|---|---|---|---|---|
| Author: Cristina Alenda González | Data sharing: Central Repository | Organisation's name: FUNDACION DE LA COMUNITAT VALENCIANA PARA LA GESTION DEL INSTITUTO DEINVESTIGACION SANITARIA Y BIOMEDICADE ALICANTE | Organization's Acronym: ISABIAL | Tier: 1 |
| Intention:<br>● development of AI tools and solutions<br>● training of AI tools and solutions<br>● validation of AI tools and solutions<br>- development, validation or training of AI tools considered for medical devices | | | | |
| **General description of the potential use and clinical impact of the shared data:**<br>Dataset 1 (COLORECTAL CARCINOMA). A discriminator capable of classifying the different regions of the image (background without tissue, stroma, adipose tissue, tumor epithelial tissue, non-tumor epithelial tissue, detritus, lymphocytic infiltrates) was designed, reaching an AUC of 0.98. A majority voting algorithm was finally applied for the MSI state decision reaching an AUC of 0.87. Conclusions: An end-to-end MSI prediction system has been obtained from H&E tumor images using artificial vision techniques that integrates the image preprocessor at multiple magnifications, the regions of interest discriminator, the MSI classifier and a bias control system.<br><br>Dataset 2 (LUNG CANCER). Lung cancer is the tumor with the highest incidence worldwide, ranking first for men. It is also the cancer with the highest mortality rate for both sexes at present. 5-year survival of advanced stage lung cancer did not exceed 5% until recently. Molecular profiling prediction in lung cancer using Hematoxylin and Eosin (H&E) stained images holds great promise for enhancing diagnosis, treatment selection, and patient outcomes in the era of precision medicine. This approach could significantly reduce time required for each step of the process, lower the costs associated with inefficient treatments and would have an enormous impact on the patient's quality of life and survival.<br><br>Dataset 3 (BRAIN TUMORS). Brain tumors are highly heterogeneous cancers with diverse outcomes in patients, from a few months to several years of survival after diagnosis. We are developing an AI-based platform to provide estimates of overall survival and diagnostic recommendations thanks to training procedures from H&E stained images supported by multiomics data (transcriptomics, DNA methylomics). As a result, we expect to improve current patients stratification and personalized clinical management, helping in therapeutic decisions and reducing time and costs of current diagnostics assays. | | | | |
| **Dataset 1**<br><br>**Cancer Type:** Colorectal cancer<br><br>**Dataset name:** EPICOLON<br><br>**Dataset description:**<br>Series of 2000 colorectal carcinomas from two multicenter projects involving 24 hospitals. These are two series of unselected population diagnosed with colorectal carcinoma, whose objective was to determine the prevalence of Lynch syndrome in Spain. The tumors are collected in several tissue microarray (TMA) and complete clinicopathological information and several years of follow-up are available. This series was used to develop an AI model for the prediction of microsatellite instability from digitized microscopic images stained with hematoxylin eosin.<br><br>**Dataset Collection Method:** Cohort<br><br>**Dataset Type:** Annotated Dataset<br><br>**Dataset Terms of Use:** Unspecified<br><br>**Dataset Intended Purpose:** Research project | | | | |

**Imaging Modality:** Digital Microscopic Image

**Vendor:** Unspecified

**Imaging body part:** Unspecified

**Age range:** 30-90 years

**Sex:** Males and Females

**Number of subjects:** 2000

**Number of DICOM studies:** Unspecified

**Image size in GB:** Unspecified

**De-identification:** Personal data is fully anonymized

**Dataset 2**

**Cancer Type:** Lung cancer

**Dataset name:** LUNGMARKER

**Dataset description:**
900 lung carcinomas from Alicante General Hospital, in which we will focus on Non-Small Cell Lung Cancer (NCSLC). The dataset comprises digitized microscopic images of tumors stained with hematoxylin and eosin, along with comprehensive clinicopathological information stored in the REDCap software. This series will be utilized to develop an AI model aimed at predicting the molecular profiling of patients. The model will leverage both the hematoxylin and eosin-stained images and associated clinical data, integrating various techniques such as sequencing, immunohistochemistry, and fluorescence in situ hybridization.

**Dataset Collection Method:** Cohort

**Dataset Type:** Annotated Dataset

**Dataset Terms of Use:** Unspecified

**Dataset Intended Purpose:** Research project

**Imaging Modality:** Digital Microscopic Image

**Vendor:** Unspecified

**Imaging body part:** Unspecified

**Age range:** 30-90 years

**Sex:** Males and Females

**Number of subjects:** 900

**Number of DICOM studies:** Unspecified

**Image size in GB:** Unspecified

**De-identification:** Personal data is fully anonymized

**Dataset 3**

**Cancer Type:** Brain tumors

**Dataset name:** GLIO-IA

**Dataset description:**
Primary brain tumors from the main hospitals in the province of Alicante (Hospital General Universitario Dr. Balmis and Hospital General Universitario de Elche) have been collected in tissue microarray (TMA), stained with hematoxylin eosin and digitized at high resolution. Associated complete clinicopathological information has been compiled. These brain tumors are mainly glioblastomas, but there are also examples of tumors with better

outcome: diffuse and pilocytic astrocytomas, oligodendrogliomas, ependymomas, gangliogliomas, among others. We have generated multiomics datasets from approximately 40% of the FFPE samples to provide genome-wide molecular information. This series was used to develop AI models for the prediction of survival rate and molecular diagnosis.

**Dataset Collection Method:** Cohort

**Dataset Type:** Annotated Dataset

**Dataset Terms of Use:** Unspecified

**Dataset Intended Purpose:** Research project

**Imaging Modality:** Digital Microscopic Image

**Vendor:** Unspecified

**Imaging body part:** Unspecified

**Age range:** 15-90 years

**Sex:** Male and Female

**Number of subjects:** 400

**Number of DICOM studies:** Unspecified

**Image size in GB:** Unspecified

**De-identification:** Personal data is fully anonymized

**Title of the use case:** Algorithms for predicting the genetic profile of different neoplasms from microscopic imaging

**General description of the use case:**
Our main objective is to optimize our customized AI algorithms and deep learning procedures thanks to increasing the number of cases in the type of cancers in which we are already data holders: colorectal, lung and brain tumors, with the possibility to extend the study to other subtypes not fully contemplated in our own datasets (for example, pediatric glioblastomas, brain metastatic cancer).
We expect to improve the power of our own AI tools. This will speed up the implementation of AI-based services for diagnosis and prognosis of the referred cancers, incorporating the main advantages of AI:
reduction of time, error and costs of diagnosis, personalized medicine and tailored life planning for intractable cancers.
We will apply the same methodologies that are specific for each type of dataset.

**Expected timeline for the realization of the use case:** 2 years

**Description of the intended use and expected benefit related to the use of the EUCAIM data:**
We will apply the same AI algorithms that we are applying for each type of dataset (for colorectal carcinoma see https://doi.org/10.3390/biom11121786), keeping in mind that the new rounds of training may impose deviations from our original strategies of AI-based analysis.

**Description of the requested data:**
Requested data:

Colorectal carcinoma: histologic type (WHO), location, age, sex, genetic profile (KRAS, BRAF, NRAS, MSI, NTRK 1/2/3)

Lung carcinoma: non small lung carcinoma, location, age, sex, genetic profile (KRAS, EGFR, PDL-1, RET, MET, NTRK1/2/3, ALK)

Primary brain tumors: glioblastoma, astrocytoma, oligodendroglioma, extended to often neglected rare subtypes (ependymomas, gangliogliomas, xantoastrocytomas, etc.) that will be under the category of “long survival tumors". Our original dataset is mainly focused on adults, but we would like to include pediatric and teenager cases as well.

Brain metastatic cancers (e.g., melanomas) for comparison purposes with primary brain tumors. Image modality in all cases: TIFF from H&E slides at high resolution (x40). Associated clinical data will be required: age, sex, available histological and molecular diagnosis, overall survival after diagnosis, de novo or recidivant tumor.

| |
|---|
| Number of cases: as much as possible, until completing in-house storage capabilities (ISABIAL server). |

Table 41. Description of Use Case no.13 from Liga Portuguesa Contra o Cancro - Núcleo Regional do Centro

| Author: Vitor Rodrigues | Data sharing: Central Repository | Organisation's name: Liga Portuguesa Contra o Cancro - Núcleo Regional do Centro | Organization's Acronym: LPCC-NRC | Tier: 1 |
|---|---|---|---|---|
| Intention:<br>● training of AI tools and solutions<br>● validation of AI tools and solutions | | | | |
| **General description of the potential use and clinical impact of the shared data:**<br>Artificial intelligence (AI) and machine learning (ML) have shown promising results in cancer diagnosis in validation tests involving retrospective patient databases.<br>Training datasets comprise samples used to fit machine learning models under construction, i.e., carry out the actual AI development.<br>Constructing these robust pillars of AI involves following best practices.<br>Image based AI systems have shown significant improvements in breast cancer detection. They have the potential to enhance screening outcomes, reduce false negatives and positives, and detect subtle abnormalities missed by human observers. However, challenges like the lack of standardized datasets, potential bias in training data, and regulatory approval hinder their widespread adoption.<br>We pretend to develop a study aimed to explore the extent of actual use of AI/ML protocols for diagnosing breast cancer in prospective settings. | | | | |
| **Dataset 1**<br><br>**Cancer Type:** Breast Cancer<br><br>**Dataset name:** Picture Archiving and Communication System (PACS) Fujifilm Synpase<br><br>**Dataset description:**<br>The Portuguese League Against Cancer (LPCC-NRC) Breast Cancer Screening Program (BCSP) has a Picture Archiving and Communication System (Fujifilm Synapse PACS) that stores and manages the digital mammograms (DR) carried out every 2 years by the women who attend the Screening (around 90,000 mammograms (360.000 incidences) per year, since 2009).<br>The BCSP also has a Management Information System (MIS) that collects clinical and demographic data for anamnesis, assessment and hospital diagnosis and treatment.<br><br>**Dataset Collection Method:** Cohort<br><br>**Dataset Type:** Original Dataset<br><br>**Dataset Terms of Use:**<br>Participation in the Breast Cancer Screening Program presupposes prior knowledge of the risks and benefits (of participation) and signing the authorization to collect and process personal data and the informed consent form. These documents authorize the Liga Portuguesa Contra o Cancro - Núcleo Regional do Centro to carry out research projects.<br><br>**Dataset Intended Purpose:**<br>Using artificial intelligence (AI) to supplement radiologists' evaluations of mammograms may improve breast-cancer screening by reducing false positives without missing cases of cancer and improving sensitivity and specificity in radiological reading.<br><br>**Imaging Modality:** Mammography<br><br>**Vendor:** FUJIFILM<br><br>**Imaging body part:** Unspecified<br><br>**Age range:** 50-69 years<br><br>**Sex:** Females<br><br>**Number of subjects:** 90.000/year<br><br>**Number of DICOM studies:** 360.000/year | | | | |

| |
|---|
| **Image size in GB:** Unspecified<br><br>**De-identification:** Personal data is fully anonymized |
| **Title of the use case:** Artificial Intelligence in Breast Cancer Screening |
| **General description of the use case:**<br>AI algorithms may make radiologists' workflow far more efficient, and they can provide quantitative analyses that are not subject to human bias, making data-driven calls for questionable mammograms that could be interpreted differently. AI-powered software can automate interpretation of breast mammograms, ultrasounds to get patients their results faster.<br>AI techniques can help radiologists identify breast cancer that would have otherwise been undetectable in its early stages.<br>The study will include the evaluation of performance indicators, including their acceptable and desirable levels, which are associated with breast cancer diagnosis.<br><br>**Expected timeline for the realization of the use case:**<br>Currently, we foresee a project timeline that runs from September 2024 to December 2026<br><br>**Description of the intended use and expected benefit related to the use of the EUCAIM data:**<br>Artificial intelligence (AI) algorithms for interpreting mammograms have the potential to improve the effectiveness of population breast cancer screening programs if they can detect cancers, including interval cancers, without contributing substantially to overdiagnosis.<br>Studies suggesting that AI has comparable or greater accuracy than radiologists commonly employ 'enriched' datasets in which cancer prevalence is higher than in population screening.<br>Routine screening outcome metrics (cancer detection and recall rates) cannot be estimated from these datasets, and accuracy estimates may be subject to spectrum bias which limits generalizability to real world screening.<br>We aim to address these limitations by comparing the accuracy of AI and radiologists in a cohort of consecutive women attending a real-world population breast cancer screening program. Mammograms will be reinterpreted by a commercial AI algorithm (Lunit). AI accuracy will be compared with that of radiologists double-reading.<br><br>**Description of the requested data:**<br>Examination: Mammograms (craniocaudal (CC) and mediolateral oblique (MLO))<br>Women between the ages of 50 and 69<br>Image type: DICOM<br>Number of studies: 30,000 approx.<br>BI-RADS classification<br>Histopathological data |

Table 42. Description of Use Case no. 58 from National Hellenic Research Foundation

| Author: Nicos Maglaveras | Data sharing: Central repository | Organisation's name: National Hellenic Research Foundation | Organization's Acronym: NHRF | Tier: 3 |
|---|---|---|---|---|
| Intention:<br>● training of AI tools and solutions<br>● development of AI tools and solutions | | | | |
| **General description of the potential use and clinical impact of the shared data:**<br>The ultrasound images related to ovarian and endometrial cancers are lately gaining importance both as it concerns the surgical procedure for treatment of the malignancy as well as for the risk stratification and diagnostics.<br>Fusing information from the ultrasound images with anamnesis data and metabolomics, transcriptomics as well as NGS and WGS data can be the future for precision medicine and personalized health in the area of gynecological cancer as well as other cancers such as lung, colorectal and skin as referred to in the previous sections. | | | | |
| **Dataset 1**<br><br>**Cancer Type:** Ovarian and Endometrial<br><br>**Dataset name:** HIPPO_ESGO<br><br>**Dataset description:**<br>Regarding ultrasound images of antenatal surveillance, they might avail data on two major domains, the first concerning screening of congenital abnormalities and chromosomal deficiencies and the second concerning growth of fetus. Specifically, nuchal translucency, Doppler measurement of uterine arteries, detection of echogenic biomarkers such as echogenic bowel, single umbilical artery and short femural length are considered the main representative markers of congenital and chromosomal abnormalities. Regarding surveillance of fetal | | | | |

growth, measurement of head circumference, abdominal circumference and femural length along with Doppler examination of umbilical and middle cerebral arteries are the most representative ones.
Regarding gynecological ultrasound images, they rather adhere to the IOTA and IETA criteria developed by the relative international Consortium. Biomarkers related with U/S images might be measurement of CA 125 and potentially CA 19-9, while standard evaluation of ultrasound images especially for ovarian masses included measurement of tumor length, detection of solid or multiform part inside the tumor, presence of acoustic shadow, papillary injection and presence of ascites. Evaluation for endometrial tumors relates with endometrial thickness and pattern of echogenicity, presence and intensity of vascularization based on Color Doppler, position of uterine masses based on FIGO Classification, presence of cystic parts, in an effort to identify images suspicious for endometrial cancer, myomas or sarcomas.

**Dataset Collection Method:** Disease-specific

**Dataset Type:** Original Dataset

**Dataset Terms of Use:**
Ethical Committee approval from Hippokrateio Hospital, Thessaloniki, Greece

**Dataset Intended Purpose:**
Cancer diagnostics, Risk Stratification for antenatal cases, Extraction of ultrasound digital biomarkers for ovarian and endometrial malignancies and pathologies.

**Imaging Modality:** Ultrasound

**Vendor:** Unspecified

**Imaging body part:** Unspecified

**Age range:** 15 - 70 years

**Sex:** Female

**Number of subjects:** 1000

**Number of DICOM studies:** Unspecified

**Image size in GB:** Unspecified

**De-identification:** Personal data is included in the images. In this EUCAIM can support you with specific tools and guidance for de-identification once the application is accepted.

**Dataset 2**

**Cancer Type:** Breast, Lung, Colorectal

**Dataset name:** NHRF_OMICS

**Dataset description:**
On the molecular level, we will use whole exome and next generation sequencing data as well as data on mRNA expression (transcriptomics) and metabolomics from melanoma, lung, breast and colorectal
cancer patients (n=200) from the Athens Comprehensive Cancer Center and its collaborating hospitals
and cancer treatment centers running since 2017 in close collaboration with its German counterpart in
Heidelberg running from the German Research Center on Cancer (DKFZ).

**Dataset Collection Method:** Disease-specific

**Dataset Type:** Original Dataset

**Dataset Terms of Use:**
See previous information in other sections

**Dataset Intended Purpose:**
Enhanced multi-omics bioinformatics algorithms supported by machine- and deep-learning approaches developed provide an extended dataset that has already been used to identify potential biomarkers of effect that could be used as early prognostic signals both in terms of disease onset and development and in terms of successful treatment and/or remission.

**Imaging Modality:** Unspecified

**Vendor:** Unspecified

**Imaging body part:** Unspecified

**Age range:** 20 - 70 years

**Sex:** Male and Female

**Number of subjects:** 200

**Number of DICOM studies:** Unspecified

**Image size in GB:** Unspecified

**De-identification:** Personal data is included in the images. In this EUCAIM can support you with specific tools and guidance for de-identification once the application is accepted.

**Title of the use case:** Diagnostics in gynecological lung and colorectal cancers

**General description of the use case:**
Major sources of clinical and imaging data that could serve as input information for the development of decision-making tools are ultrasound images of either antenatal pregnancy surveillance or gynecological pathologies. We might estimate that over 3,000 ultrasound obstetric images and over 1,000 gynecological ultrasound images might be available on an annual basis. Furthermore, continuous performance of laparoscopy as surgical approach of various gynecological pathologies might avail approximately 100 recorded video procedures of at least 1 hour duration, including at least 50 of gynecological malignancy. Furthermore, stored epidemiological data of antenatal obstetrical screening (approximately 1,000 cases on annual basis), as well as clinical and histopathological data of endometrial, cervical and ovarian cancer (approximately over 500 cases for last 3 years) might also serve on the level of input data as primary sources of decision-making tools.
Regarding ultrasound images of antenatal surveillance, they might avail data on two major domains, the first concerning screening of congenital abnormalities and chromosomal deficiencies and the second concerning growth of fetus. Specifically, nuchal translucency, Doppler measurement of uterine arteries, detection of echogenic biomarkers such as echogenic bowel, single umbilical artery and short femural length are considered the main representative markers of congenital and chromosomal abnormalities. Regarding surveillance of fetal growth, measurement of head circumference, abdominal circumference and femural length along with Doppler examination of umbilical and middle cerebral artery are the most representative ones.
Regarding gynecological ultrasound images, they rather adhere to the IOTA and IETA criteria developed by the relative international Consortium. Biomarkers related with U/S images might be measurement of CA 125 and potentially CA 19-9, while standard evaluation of ultrasound images especially for ovarian masses included measurement of tumor length, detection of solid or multiform part inside the tumor, presence of acoustic shadow, papillary injection and presence of ascites. Evaluation for endometrial tumors relates with endometrial thickness and pattern of echogenicity, presence and intensity of vascularization based on Color Doppler, position of uterine masses based on FIGO Classification, presence of cystic parts, in an effort to identify images suspicious for endometrial cancer, myomas or sarcomas.
Finally, regarding the stored data of cancer patients, these might be divided into three main domains. The first concerns epidemiological data regarding age, comorbidities, obstetrical and gynecological history. The second concerns histopathological data, namely histopathological type, grade, presence of LVSI in endometrial cancer patients, nodal status based on final surgical staging as well as potential expression of POLE, MMR and p53 mutations in EC patients. The third domain of data concerns prognostic outcomes of patients, namely disease-free survival, overall survival, recurrence of tumor, Kind of recurrence and treatment of recurrence. All relative data are consistently registered in ESGO related databases in a continuous effort of Clinical audit as well as development of clinical and scientific outcomes.
On the molecular level, we will use whole exome and next generation sequencing data as well as data on mRNA expression (transcriptomics) and metabolomics from melanoma, lung, breast and colorectal cancer patients (n=200) from the Athens Comprehensive Cancer Center and its collaborating hospitals and cancer treatment centers running since 2017 in close collaboration with its German counterpart in Heidelberg running from the German Research Center on Cancer (DKFZ). Enhanced multi-omics bioinformatics algorithms supported by machine- and deep-learning approaches developed provide an extended dataset that has already been used to identify potential biomarkers of effect that could be used as early prognostic signals for disease onset and development and in terms of successful treatment and/or remission.
We consider that all relative data and biomarkers and validation throughout decision-making tools will contribute in daily clinical practice, with new sophisticated individualized algorithms of diagnosis and treatment.

**Expected timeline for the realization of the use case:** 18 months

**Description of the intended use and expected benefit related to the use of the EUCAIM data:**

More specifically for the analytics of the ultrasound images and videos we intend to use advanced AI driven incremental learning techniques coupled with bi-directional LSTMs and self-supervised vision transformers shall be used in the extraction of digital biomarkers from the above mentioned types of the ultrasound images.
Further, advanced FL techniques (e.g. pareto-optimal approaches) applied on fused information from images, electronic health records data and clinical and histopathological data shall be used to enhance predictive modeling, diagnosis and risk stratification tasks in the women's cancer cases.
On the molecular level, we will use whole exome and next generation sequencing data as well as data on mRNA expression (transcriptomics) and metabolomics from melanoma, lung, breast and colorectal cancer patients (n=200) from the Athens Comprehensive Cancer Center and its collaborating hospitals and cancer treatment centers running since 2017 in close collaboration with its German counterpart in Heidelberg running from the German Research Center on Cancer (DKFZ). Enhanced multi-omics bioinformatics algorithms supported by machine- and deep-learning approaches developed provide an extended dataset that has already been used to identify potential biomarkers of effect that could be used as early prognostic signals both in terms of disease onset and development and in terms of successful treatment and/or remission.
The analytics capacity on the molecular level can be combined with the imaging data from the types of cancer we referred to ( melanoma, lung, breast and colorectal ) enhancing the translational capacity of the digital biomarkers found in imaging modalities of the above-mentioned cancers.
Further in the case of breast cancer and ovarian cancer, since they are related to mutations of BRCA genes we can correlate the imaging biomarkers with the BRCA genes sequencing characteristics.

**Description of the requested data:**
We intend to acquire data from >500 women with gynecological cancers (ovarian, endometrial, breast, HPV) including ultrasounds as described before, histopathological exams, whole exome, NGS transcriptomics and metabolomics as well as Electronic Health Record data.
These data are expected to be used for testing the AI/ML algorithms developed and validated in the already existing data as described before.