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

**Project acronym:** EUCAIM

**Grant Agreement:** 101100633

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

# D5.2. The EUCAIM CDM and Hyper-Ontology for Data Interoperability: initial version

**Responsible partner: LIMICS**

**Author(s):** Mirna El Ghosh (LIMICS), Melanie Sambres (LIMICS), Catherine Duclos (LIMICS), Ferdinand Dhombres (LIMICS), Xavier Tannier (LIMICS), Valia Kalokyri (FORTH), Stelios Sfakianakis (FORTH), Manolis Tsiknanis (FORTH), Christel Daniel (APHP-LIMICS)

**Contributors**: Olga Giraldo (DKFZ), Heimo Muller (BBMRI-ERIC), Laure Fournier (APHP), Aurélien Maire (APHP), Jean Nembo (APHP), Kevin Mondet (APHP), Maciej Bobowicz (GUMed), Jean Charlet (LIMICS), Alexandra Kosvyra (AUTH), Ioanna Chouvarda (AUTH), Antonis Aletras (AUTH), Aikaterini Lazou (AUTH), Vasileia Paschaloudi (AUTH), Teresa Garcia Lezana (CRG-CERCA), Michal Kosno (GUMed), Gianna Tsakou (MAG), Celia Martin Vicario (QUIBIM), Laure Saint-Aubert, (BC Platforms), Alejandro Rodríguez Pardavilla (BAHIA), Roberto Romero (CIBER), Jose Tapia (CIBER), Haridimos Kondylakis (FORTH), Maria Christodoulou (HCS), Maria Gonzalez Lopez (SAS), Federica Cruciani (IFOM)

**Reviewers**: Pedro Mallol (HULAFE); Francesco Cremonesi (INRIA)

**Date of delivery:** **29/06/2024**

**Version:** **Final version**

# Table of Contents

# Table of Figures

# List of Tables

# 1. Introduction

This document offers a detailed overview of the key features and contributions of the initial version of the EUCAIM Common Data Model (CDM) and Hyper-ontology.

In this deliverable, we provide a detailed explanation of the various challenges we encountered and our strategy for addressing the heterogeneity in data representation and semantics across various sources of information.

Interoperability in healthcare facilitates the exchange and utilization of health information across diverse systems, improving communication and standardizing patient data sharing. It includes technical, syntactic, and semantic components supported by international standards like HL7's FHIR, and terminologies such as SNOMED CT, which ensure accurate data interpretation and integration.

However, the first step in conducting research in the health domain is finding and requesting access to datasets that fulfill criteria based on the clinical use cases that need to be answered. In order to achieve this, it is essential to appropriately catalog the information held by various data sources and make these catalogues accessible for browsing and querying. EUCAIM has worked on extending DCAT-AP, a Data Catalog Vocabulary Application Profile for data portals in Europe, specifically for health imaging datasets, by establishing mandatory metadata for medical images in the EUCAIM public catalogue, aligning with the on-going efforts of the HealthDCAT-AP specification as well as utilizing the EUCAIM Hyper-ontology specification to define controlled vocabularies for semantic interoperability (section 3).

However, typically, public catalogues are anticipated to include metadata outlining the fundamental and high-level characteristics of the datasets, and as such, the bare minimum metadata required for cataloguing datasets across various cancer types has been included at this level (the set of metadata was extracted after the analysis and the methodology adopted, which is outlined in D5.1[1]). For data users seeking to conduct a more fine-grained search at a *subject-level* based on cancer-specific criteria, the EUCAIM hyper-ontology's concepts and terms shall be used, through the EUCAIM federated query user interface. The EUCAIM hyper-ontology, developed through an iterative and systematic process, integrates diverse clinical and imaging knowledge from projects like CHAIMELEON, ProCAncer-I, EuCanImage, INCISIVE, and PRIMAGE, addressing the semantic and syntactic disparities that exist among diverse data models and standards (section 4).

In the context of EUCAIM, we examined different scenarios for federated processing/analysis and AI model development tasks, guiding decisions regarding the CDM structure and format, with each scenario presenting distinct advantages and challenges regarding data integration, harmonization, and usability (section 5).

The hyper-ontology is semantically represented to ensure alignment with the EUCAIM CDM based on mCode specification. Regarding the integration of the EUCAIM CDM and hyper-ontology, an example of a formalization profile (Primary cancer condition), detailing data elements and their corresponding value sets, is presented in section 6.

[1] EUCAIM D5.1. Early release of the Data Federation Framework, 2023 https://cancerimage.eu/wp-content/uploads/2023/10/D5.1_Early-release-of-the-Data-Federation-Framework_vf.pdf

To finish, four proof of concept scenarios related to prostate and breast cancer, provided by four AI4HI projects: INCISIVE, ProCAncer-I, CHAIMELEON, and EuCanImage, are presented in section 7, in order to demonstrate the feasibility and validation of the EUCAIM hyper-ontology and CDM, based on the clinical/biological and imaging information collected and modeled by the four AI4HI projects.

# 2. Interoperability requirements

Interoperability in healthcare ensures a coherent exchange and use of health information between different systems, applications, and stakeholders. Maintaining interoperability supports communication among various healthcare systems and the sharing of essential patient data in a standardized and meaningful way. There are different components, layers, or levels of interoperability: **technical**, **semantic** and **syntactic** aspects of interoperability.

***Technical*** interoperability ensures basic data exchange capabilities between systems, requiring defined communication channels and protocols for data transmission. Aspects of technical interoperability include interface specifications, interconnection services, data integration services, data presentation and exchange, and the implementation of secure communication protocols. The structured exchange of health data is supported by international standards development organizations (SDOs) such as Health Level Seven International (HL7) or Digital Imaging and Communications in Medicine (DICOM). An emerging standard for the communication of health data is HL7's Fast Healthcare Interoperability Resources (FHIR), which defines common healthcare resources that can be accessed and exchanged using modern web technologies. FHIR is increasingly being adopted by the health industry and supports the development of interoperable health applications that run on different IT systems dedicated to care for research activities.

While standards such as FHIR already define the basic semantics of health data, semantic interoperability is really the domain of medical terminologies, nomenclatures, and ontologies. ***Semantic*** interoperability ensures that the meaning of exchanged data remains preserved and comprehensible across interactions; it strives for a scenario where "what is sent is what is understood". This involves the development of vocabularies to describe data exchanges, ensuring a shared understanding of data elements among all communicating parties, ideally, understandable to humans and machines worldwide. In essence, semantic interoperability revolves around interpreting the meaning of data elements and their relationships. One of the most broadly used clinical terminologies is SNOMED CT, particularly well-suited as a general-purpose language for advancing semantic interoperability in medicine and healthcare, complemented by more domain-specific terminologies such as, for example, International Classification of Diseases (ICD), Logical Observation Identifiers Names and Codes (LOINC) for laboratory data, RxNorm (or the future Identification of Medicinal Products (IDMP)) for medicines, the nomenclature of the HUGO Gene Nomenclature Committee (HGNC) for genes or the Human Phenotype Ontology (HPO) for phenotypic abnormalities. The different protocols (Data collection protocols, Data transfer protocols, Data analysis protocols), formats and terminologies for clinical/biological and imaging data has been introduced in the D5.1, with detailed explanation of each chosen terminologies. Since the D5.1, the hyper-ontology has been enriched based on workshops with medical expert and new use cases, with new concepts (e.g. concept from ICDO3 ). This hyper-ontology will continue to grow through the ongoing evaluation and expert workshop.

***Syntactic*** interoperability specifies the exact format of the information to be exchanged (e.g., XML), conceptual and logical models, and the organization of information, encompassing variable structures, units, data types, transformation and validation rules, etc. The Observational Medical Outcomes Partnership (OMOP) Common Data Model (CDM) is an open community data standard designed to standardize the structure and semantics of observational data that is increasingly adopted and recognized by the healthcare industry. Precisely, the core component of the OMOP CDM is the use of standardized vocabularies such as those

mentioned above, which allows the organization and standardization of medical terms to be used in the various clinical domains. Combined with the standards discussed above through value sets explicitly defined (terminology binding), using these terminologies can ensure that health data have unambiguous semantics.

Addressing interoperability in healthcare with precision requires special attention. EUCAIM, which deals with a considerable amount of diverse data from different repositories/sources, requires defining standards and structures for how the data are modeled and stored to avoid any disambiguation and allow machines, artificial intelligence (AI) systems, or any information tools to deal with the data and metadata. In the following sections, we present, based on our expertise, the main data interoperability requirements and challenges in the context of the EUCAIM project. Towards this purpose, different state-of-the-art interoperability standards have been explored according to the different tiers of the EUCAIM data interoperability framework. Each of them plays a crucial role in the different tiers supported by EUCAIM and stages of the data life cycle: in the publication of datasets, in the data preparation process for federated query purposes, and in the connection between the EUCAIM federated nodes for federated analysis and processing. A quick summary of the tiers described thoroughly in D4.3 is outlined below:

Tier 1: Dataset Metadata Level

- High-level aggregated dataset metadata are registered in the public catalogue of EUCAIM, according to the metadata specification for the datasets. Compliance with the EUCAIM CDM, specific services and node setup at the data holder's side are not required, although the EUCAIM platform functionalities will be limited.

Tier 2: Federated Search Level

- High-level aggregated dataset metadata are registered in the EUCAIM public catalogue as in Tier 1.
- The datasets of the data holder are also integrated into the federated search. This requires the development of a mapping component between the local data structure and the EUCAIM hyper-ontology (semantic interoperability), as well as the installation of a mediator service accessible from the central services of EUCAIM in order to reply to a set of query criteria defined within the project.

Tier 3: Federated Processing Level

- Fulfill requirements of Tier 2.
- The datasets of the data holder should comply with the EUCAIM CDM (semantic, syntactic and technical interoperability). This could be done either directly (through adoption of the EUCAIM CDM) or indirectly (through a mediator component which performs the proper mappings and transformations).

# 3. Data interoperability framework for dataset cataloguing

The first step in conducting research in the health domain is finding and requesting access to datasets that fulfill certain criteria based on the clinical use cases that need to be answered. In order to achieve this, it is essential to appropriately catalog the information held by various data sources and make these catalogues accessible for browsing. Typically, these catalogues are anticipated to include metadata outlining the fundamental and high-level characteristics of the datasets.

In the context of EUCAIM, Tier 1 focuses on achieving interoperability at a dataset metadata level. This entails standardizing the definition, documentation, and exchange of *aggregated* dataset metadata across diverse systems. Key steps for achieving interoperability at this level include adopting widely recognized metadata standards, using controlled vocabularies to prevent ambiguity, and facilitating automatic metadata exchange between systems. By achieving interoperability at the dataset metadata level and standardizing key characteristics of the EUCAIM cancer imaging datasets, we simplify the process for users and applications to find and assess whether a specific EUCAIM dataset meets their needs.

Within the European context, various activities and regulations, notably the EU regulation on the European Health Data Space (Article 55), aim to enhance and promote data sharing. The regulation emphasizes the need for health access bodies to maintain a systematically arranged dataset catalogue accessible online. To fulfill this requirement, a common generic framework is necessary.

As already described and analyzed in D5.1 (section 3.5)[2], the DCAT-AP v3.0.0, along with an extension, has been adopted as the metadata standard for dataset cataloguing and for describing the cancer imaging datasets to be registered into the EUCAIM public catalogue. The extension is necessary for incorporating the *domain-specific* imaging and clinical metadata required for discovering the EUCAIM cancer imaging datasets. Another parallel effort at a European Level, specifically for the health domain, has been the Health-DCAT-AP specification, which aims to extend the general DCAT-AP for describing health-related datasets that also comply with the European Health Data Space regulation. EUCAIM has leveraged and tried to build upon the unofficial Health-DCAT-AP specification[3] currently available, as well as analyze its alignment with the EUCAIM DCAT-AP extension that has been defined within the context of this project for cancer imaging-related information. The following sections describe an updated version of the metadata model described in D5.1 section 3.5, as well as provide detailed mappings of the generic DCAT-AP, the HealthDCAT-AP and the EUCAIM DCAT-AP.

## 3.1 EUCAIM DCAT-AP

Extending DCAT - and creating the so-called *DCAT Application Profiles* for specific domains - comes with a specific set of requirements that should be met:

- The mandatory requirements defined in the DCAT-AP should be respected.
- The controlled vocabularies of the DCAT-AP specification must be respected.

[2] EUCAIM D5.1. Early release of the Data Federation Framework, 2023 https://cancerimage.eu/wp-content/uploads/2023/10/D5.1_Early-release-of-the-Data-Federation-Framework_vf.pdf
[3] https://healthdcat-ap.github.io/

- Recommended and optional properties could become mandatory (have stricter semantics).
- Recommended attributes could become optional (less strict semantics).
- New domain-specific controlled vocabularies could be defined for newly added properties.

Our methodology for extending DCAT-AP, as it was described in D5.1, section 3.5.2, was to establish first the minimum/mandatory information that should accompany the medical images and describe the datasets to be registered in the EUCAIM public catalogue. As a reminder, we adopted a bottom-up approach, gathering the obligatory information mandated by the AI4HI projects for various cancer types considered within these projects. Additionally, we explored the initiatives undertaken by the European Network of Cancer Registries (ENCR), with a focused examination of the Standard Dataset specifications document and the cancer data quality checks proposal. At the same time, we sought to leverage and build upon the work of other European initiatives, such as the BBMRI-ERIC biobank metadata catalogue, the AI4HI project metadata catalogues, as well as the AI interoperability in imaging White Paper which includes a set of required data elements useful for AI model development. Finally, for specifying the semantics and mappings of the clinical terms to be used and therefore defining the set of controlled vocabularies to be used for the newly added properties, we use the EUCAIM Hyper-ontology specification (described in section 4). All the details of the approach and an initial outcome have been described in the deliverable D5.1 on section “3.5. Public Catalogue-Metadata Model”. An updated metadata model that tries to comply with the work of the HealthDCAT-AP is given below in Tables 1 and 2 where the general dataset metadata and the domain-specific EUCAIM dataset metadata are outlined (for the general metadata as these are defined in DCAT-AP v3.0.0 only the mandatory and the recommended properties are outlined. The optional ones are excluded for conciseness).

***Table 1**: General dataset metadata (DCAT-AP specification with stricter semantics in some cases)*

| EUCAIM DCAT-AP | | | | | | |
|---|---|---|---|---|---|---|
| **Property Type** | **Property** | **Description** | **Property IRI** | **Range** | **Cardinality** | Example |
| Mandatory | Title | A clear and concise name for the dataset. | dct:title | rdfs:Literal | 1..n | dct:title "Open Challenge Prostate Cancer V1"@en; |
| Mandatory | Description | A detailed description of the dataset. | dct:description | rdfs:Literal | 1..n | dct:description "This ProCAncer-I project imaging dataset contains a collection of patients with mpMRI examinations (T2ax, DWI, DCE) who have confirmed PCa at biopsy and/or prostatectomy."@en |
| Recommended | Acronym | An acronym that identifies the dataset. | dct:alternative | rdfs:Literal | 0..n | dct:alternative “TCGA"@en |
| Recommended | keyword | A keyword describing the dataset. | dcat:keyword | rdfs:Literal | 0..n | dcat:keyword "prostate cancer"@en, "MRI |

| | | | | | | |
|---|---|---|---|---|---|---|
| | | | | | | performed"@en, "positive histology"@en; |
| Recommended | images creation year | A temporal period that the dataset covers. This corresponds to the year range that the actual (DICOM) images were created/acquired (if this has not been changed in the anonymization process). If this is not available, an estimation can be added. | dct:temporal | dct:PeriodOfTime | 0..n | dct:temporal [<br>a dct:PeriodOfTime;<br>dcat:endDate "2023-12-31"^^<http://www.w3.org/2001/XMLSchema#date>;<br>dcat:startDate "2021-01-01"^^<http://www.w3.org/2001/XMLSchema#date><br>]; |
| Mandatory | contact point | Contact information of the individual/managing organization of the Dataset. | dcat:contactPoint | vcard:Kind | 1..n | dcat:contactPoint [<br>a vcard:Organization;<br>vcard:hasEmail <mailto:access-commitee@procancer-i.com> ]; |
| Recommended | geographical coverage | A geographic region that is covered by the Dataset. | dct:spatial | http://publications.europa.eu/resource/authority/country/<br>OR<br>dct:Location | 1..n | dct:spatial<br><http://publications.europa.eu/resource/authority/country/GRC>; |
| Mandatory | Publisher | An entity (organisation) responsible for making the Dataset available. | dct:publisher | foaf:Organization | 1..1 | dct:publisher [ a foaf:Organization;<br>locn:address [ a locn:Address;<br>foaf:name "FORTH";<br>foaf:mbox <mailto:access-commitee@procancer-i.com>;<br>foaf:homepage<br><https://forth.ics.gr>;<br>];]; |
| Mandatory | Theme | A category of the dataset. | dcat:theme | fixed to: http://publications.europa.eu/resource/authority/data-themeOR<br>subproperty of dct:subject<br>skos:Concept | 1..n | dcat:theme<br><http://publications.europa.eu/resource/authority/data-theme/HEAL>; |
| Mandatory | Identifier | A unique persistent identifier of the dataset (in compliance with the findability aspect of the FAIR principles) | dct:identifier | rdfs:Literal | 1..n | dct:identifier<br>"https://catalogue.eucaim.cancerimage.eu/api/fdp/fdp_Dataset/2081ac523632f434cd5bc4056a30ad5b"^^<http://www.w3.org/2001/XMLSchema#anyURI>; |

| | | | | | | |
|---|---|---|---|---|---|---|
| Mandatory (NSIP) | accessRights | The accessRights of the dataset. | dct:accessRights | fixed to: http://publications.europa.eu/resource/authority/access-right/RESTRICTED | 1..n | dcterms:accessRights <http://publications.europa.eu/resource/authority/access-right/RESTRICTED> ; |
| Mandatory | rights | A statement about the conditions of access and usage of the dataset. | dct:rights | dct:RightsStatement (fixed to a predefined set of values presented in D5.1) | | dct:rights [ a dct:RightsStatement; rdfs:label "Authorisation to access, view and process in-situ the datasets"@en ]; |
| Mandatory | applicableLegislation | The legislation that mandates the creation or management of the Dataset. | dcatap:applicableLegislation | rdfs:Resource | 1..n | dcatap:applicableLegislation <http://data.europa.eu/eli/reg/2022/868/oj>; |
| Recommended | modification date | The most recent date on which the Dataset was changed or modified. | dct:modified | rdfs:Literal typed as xsd:date, xsd:dateTime, xsd:gYear or xsd:gYearMonth | 0..1 | dct:modified "2024-02-05T18:47:54Z"^^<http://www.w3.org/2001/XMLSchema#dateTime>; |
| Recommended | sample | A sample distribution of the dataset. | adms:sample | dcat:Distribution | 0..n | adms:sample [a dcat:Distribution ; dct:description "Synthetic data of the HealthPilot Use Case"@en; dcat:downloadURL <https://github.com/CAVDgit/EHDS2_UC_Sciensano/blob/main/use_case_1_synthetic_data_10K_individuals.csv>; dcat:mediaType <http://www.iana.org/assignments/media-types/text/tab-separated-values> ; ]; |
| Mandatory | provenance | A statement about the lineage of a Dataset, including information about how the data was created, or processed, including methodologies, tools, and protocols used. | dct:provenance | dct:ProvenanceStatement | provenance | dct:provenance [ a dct:ProvenanceStatement; rdfs:label "This data is sourced from several existing datasets, including the Duke dataset, ParcTauli and TCGA datasets. These datasets collectively provide comprehensive demographic and clinical data relevant to the project's objectives"@en ]; |

| | | | | | | |
|---|---|---|---|---|---|---|
| Mandatory | Type | A type of the Dataset. | dct:type | skos:Concept (there is a predefined set of values presented in D5.1). | 1..n | dct:type a skos:Concept ; skos:prefLabel "Annotated Dataset"@en . |
| Mandatory | Version | The version of the dataset. | dcat:version | rdfs:Literal (in SemVer or CalVer format) | 1..1 | dcat:version "20231122" |
| Mandatory | accessURL* | A URL that gives information about accessing the dataset. In EUCAIM, this is the URL of the negotiator service. | dcat:accessURL | rdfs:Resource | 1..1 | dcat:accessURL <https://negotiator.eucaimcancerimage.eu/collection/a96b56cd-59d4-444a-8e59-32a7fb0d7dea> ; |
| Recommended | license* | A license under which the Dataset is made available, assuming there is one license for all Dataset Distributions. If each Distribution has different licenses they should be included at the Distribution level with 1..1 relationship. | dcterms:license | dcterms:LicenseDocument (ideally under CC licenses for interoperability) | 0..* | dcterms:license [ a dcterms:LicenseDocument; dcterms:identifier <http://creativecommons.org/licenses/by/4.0/> ; ]; |
| Recommended | imageSize (in GB)* | The total size of all Distributions in the dataset, which is mainly the image size. | dcat:byteSize | xsd:decimal | 0..1 | dcat:byteSize "325"^^xsd:decimal |
| Recommended | format* | The file format of the Distributions included in the Dataset. | dct:format | dct:MediaTypeOrExtent (IANA Media Types) | 0..n | dct:format <https://www.iana.org/assignments/media-types/application/dicom>; |

*These properties are properties of the “Distribution” Entity. However, they will be included in the metadata catalogue at a dataset level as miltivalued attributes.

***Table 2**: EUCAIM DCAT-AP domain-specific metadata*

| EUCAIM DCAT-AP | | | | | | |
|---|---|---|---|---|---|---|
| **Property Type** | **Property** | **Description** | **Property IRI** | **Range** | **Cardinality** | Example |
| Mandatory | age low | The minimum age of subjects within the dataset. | eucaim:ageLow | rdfs:Integer | 1..1 | eucaim:ageLow "18" ^^xsd:int ; |

| | | | | | | |
|---|---|---|---|---|---|---|
| Mandatory | age high | The maximum age of subjects within the dataset. | eucaim:ageHigh | rdfs:Integer | 1..1 | eucaim:ageHigh "18" ^^xsd:int ; |
| Recommended | age median | The median age of subjects within the dataset. | eucaim:ageMedian | rdfs:Integer | 0..1 | eucaim:ageMedian "45" ^^xsd:int ; |
| Mandatory | birthsex | BirthSex of subjects in the dataset. | eucaim:birthsex | skos:Concept | 1..* | eucaim:birthsex <https://cancerimage.eu/ontology/EUCAIM#COM1000177> |
| Mandatory | number of studies | Total count of DICOM studies. | eucaim:nbrOfStudies | rdfs:Integer | 1..1 | eucaim:nbrOfStudies "8789" ^^xsd:int ; |
| Mandatory | number of subjects | Total count of unique individuals in the dataset. | eucaim:nbrOfSubjects | rdfs:Integer | 1..1 | eucaim:nbrOfSubjects "8237" ^^xsd:int ; |
| Recommended | number of series | Total count of DICOM series. | eucaim:nbrOfSeries | rdfs:Integer | 1..1 | eucaim:nbrOfSeries "24567" ^^xsd:int ; |
| Mandatory | intended purpose | The primary objective for which the dataset was created. | eucaim:intendedPurpose | dpv:Purpose | 1..n | eucaim:intendedPurpose[ a dpv:Purpose ; dct:description " The primary objective of this dataset is the detection of prostate cancer with high accuracy both in peripheral and transitional zones to identify which men have cancer and those with no cancer."@en;] ; |
| Mandatory | collection method | This attribute defines the scope of data aggregation within the dataset. It specifies how data records are organized based on different criteria, allowing users to understand the context in which the data was collected. | eucaim:collectionMethod | subproperty of dct:subject skos:Concept (fixed to a predefined set of values presented in D5.1) | 1..n | eucaim:collectionMethod a skos:Concept ; skos:prefLabel "Only-Image"@en. |
| | | | | | | |
| Mandatory | quality label | A statement related to quality of the Dataset, including rating, quality certificate | dqv:hasQualityAnnotation | dqv:QualityCertificate | 1..1 | dqv:hasQualityAnnotation [a dqv:QualityCertificate ; oa:hasTarget <https://…/dataset/123> |

| | | | | | | |
|---|---|---|---|---|---|---|
| | | as per the EHDS requirements. | | | | ;<br>oa:hasBody <https://…/certificate>;<br>oa:motivatedBy dqv:qualityAssessment]; |
| Mandatory | legal basis | Legal basis used to justify processing of data or use of technology in accordance with a law. | dpv:hasLegalBasis | dpv:LegalBasis | 1..n | dpv:hasLegalBasis [<br>a dpv:LegalBasis ;<br>dct:description "Deliberation no. 21/028 of february 18, 2021, last amended on june 18, 2021, relating to the communication of data to pseudonymized personal character relating to the healthdata of.. , as part of the EUCAIM project and the subsequent processing of personal data pseudonymised by…"@en;<br>dct:source <https://cancerimage.eu/file/view/AXkNfdPml9vUUfvGGfJr?filename=21-028-f212-AFMPS-dataset-modifi%C3%A9e%20le%2018%20juin%202021.pdf> ;<br>] ; |
| Mandatory | condition | The primary cancer condition of individuals in the dataset. | eucaim:hasCondition | skos:Concept (EUCAIM controlled vocabulary based on ICD-03 and SNOMED) | 1..1 | eucaim:condition <https://cancerimage.eu/ontology/EUCAIM#CLIN1000075><br>(Cancer of prostate) |
| Mandatory | image modality | The set of modalities for the images in the dataset. | eucaim:hasImageModality | skos:Concept (EUCAIM controlled vocabulary based on DICOM and Radlex) | 1..n | eucaim:hasImageModality <https://cancerimage.eu/ontology/EUCAIM#IMG1000022><br>(Magnetic Resonance Imaging) |
| Mandatory | image vendor | Manufacturer of the imaging device as it is defined in DICOM tag (0008,0070). | eucaim:hasImageVendor | skos:Concept (EUCAIM controlled vocabulary) | 1..n | eucaim:hasImageVendor<br><https://cancerimage.eu/ontology/EUCAIM#IMG1000047><br>(General Electric) |

| | | | | | | |
|---|---|---|---|---|---|---|
| Mandatory | image body part | Anatomical areas captured in the images. | eucaim:hasImageBodyPart | skos:Concept (EUCAIM controlled vocabulary) | 1..n | eucaim:hasImageBodyPart <https://cancerimage.eu/ontology/EUCAIM#BP1000233><br><br>(Neck and chest) |

The full mappings between DCAT-AP v3.0.0, the current Health-DCAT-AP, and the EUCAIM DCAT-AP are described in: Mappings of DCAT application profiles. In the worksheet, some properties have been highlighted with different colors, to denote either stricter or less strict semantics to the current HealthDCAT-AP specification, as these were discussed in the EUCAIM WP5 related working group.

## 3.2 FAIR principles compliance

For supporting dataset metadata interoperability, it is also crucial to consider the FAIR principles. These principles guide the development of metadata to ensure that datasets are easily discoverable, accessible, interoperable, and reusable across diverse environments. The DCAT-AP, serving as the standard framework for dataset descriptions, aligns seamlessly with the FAIR principles, which introduce another set of requirements that must be met. The Research Data Alliance introduced the FAIR Data Maturity Model, which assesses the level of adherence to the FAIR principles and consists of different maturity levels, typically labeled as F1, F2, F3, and F4, etc. which represent increasing levels of compliance with the FAIR principles. Each level corresponds to specific indicators that can be seen as requirements for the "FAIRification" of the datasets, which should be analyzed and verified with respect to the specific restrictions and requirements of the EUCAIM project.

Finally, interoperability on a dataset metadata level involves facilitating automatic metadata exchange between systems. The concept of the FAIR Data Point (FDP)[4] comes into play as a metadata service that adheres to the FAIR principles and offers a reference implementation (an API) enabling data owners to expose data and metadata in a FAIR manner based on the DCAT metadata standard. Although it is not a requirement for the data holders to have an FDP for exposing their datasets in a machine-readable format for Tier 1 or 2 (although EUCAIM will recommend and facilitate its adoption even on these tiers), EUCAIM will adopt it in the central EUCAIM metadata catalogue in order to not only expose its dataset metadata in an automatic way to other dataset catalogues increasing their visibility and discoverability, but also to harvest dataset metadata from already established infrastructures which have an FDP service on their catalogues.

[4] https://www.fairdatapoint.org/

# 4. Data interoperability framework for federated query

Upon the completion of dataset cataloguing procedures, which involves publishing only aggregated metadata for the datasets, the subsequent interoperability tier is the provision of federated query support.

For enabling federated query, data holders should implement a semantic interoperability layer across their datasets, which includes a) developing a mapping component between their local data structure and the EUCAIM hyper-ontology, and b) installing a mediator service accessible from the central services of EUCAIM.

## 4.1 Why do we need the EUCAIM Hyper-Ontology?

To enable federated querying across established repositories, such as the AI4HI repositories that adopt different data models/standards, the integration of a semantic interoperability layer is required. In this section, we will explore, by using examples, all the challenges that have been identified in querying the OMOP-CDM and FHIR-based AI4HI repositories. These challenges will serve as requirements for the development and application of the hyper-ontology within the context of EUCAIM.

Starting with the example of the PSA (Prostate Specific Antigen), a tumor marker for prostate cancer, its representation varies across repositories; it is represented via the SNOMED-CT standard (Prostate specific antigen measurement (4272032)) or the LOINC standard (Prostate specific Ag [Mass/volume] in Serum or Plasma (LP18192-2)). This variability in the representation across different standards poses a challenge when a user wishes to execute a federated query to “find datasets with ‘PSA’ levels over 20”, raising questions about which standard concept to use for querying, which repository uses which one of the two concepts for PSA and whether these two standard concepts are semantically equivalent or not. Addressing these questions is crucial for enabling different repositories, utilizing different standards, to accurately respond to a query regarding PSA levels.

The EUCAIM hyper-ontology should be designed to specify the relationship between such concepts (see Figure 1 for an example). Despite potentially numerous similar standard concepts for PSA, users will be able to select a specific concept, like the LOINC one, for query execution. In this case, the local mediator or service should be able to understand these concept relationships, accurately map them through the hyper-ontology specification, and return query results.

However, only specifying concept relationships in the hyper-ontology isn't sufficient for queries concerning quantitative variables. These queries must specify not only the variable of interest but also its measurement unit, since repositories might encode the same concept in different units. For instance, two repositories could use the same LOINC concept for PSA, but report values in ng/mL and nmol/L, respectively. Thus, local nodes must convert measurements to match the requested unit, necessitating that the hyper-ontology includes a "units of measure" vocabulary, such as UCUM, and possibly a default or preferred unit of measure.


***Figure 1**: An excerpt of the hyper-ontology (v1.0beta) around representing PSA concepts and their relations*

*Therefore, the hyper-ontology:*

- *should contain a formal representation of medical concepts/terms, and their relationships within the oncology domain.*
- *serves the purpose of providing a comprehensive vocabulary/terminology to cover the source data.*

Beyond semantic interoperability, addressing the syntactic heterogeneity of data models and standards is also crucial for enabling querying. For instance, OMOP-CDM organizes concepts into various domains (e.g., Condition, Measurement, Procedure, etc.), while the FHIR standard categorizes similar concepts within a set of resources (e.g., Observation, Condition, Medication, etc.). Specifically, the PSA concept is represented as a concept within the *Measurement* domain in OMOP-CDM and as a concept in the *Observation* resource type in FHIR. Therefore, the hyper-ontology should also specify the corresponding "class" or "entity" of a concept to facilitate accurate querying. It is also important to recognize that different PSA concepts may correspond to different classes/entities based on their semantics. For example, PSA might refer to a laboratory test with a numerical value (e.g., PSA=20 ng/ml), classified as a "Measurement" in OMOP-CDM or an "Observation" in FHIR. Alternatively, PSA can indicate a "Procedure", denoting whether a patient has undergone a PSA test/procedure (PSA=yes/no), or it can represent a "Condition", reflecting an abnormal PSA level (PSA=normal/abnormal), which implies an elevated PSA without specifying the exact value (see Figure 1). This multifaceted nature of concepts necessitates a comprehensive approach in the hyper-ontology to ensure queries can be accurately executed across different data standards and models.

*Therefore, the hyper-ontology:*

- *should link the concepts from clinical standard terminologies to the corresponding CDM classes of OMOP and FHIR (similar to how OMOP vocabularies specify the Domain of a concept) (see Figure 1 for an example).*

Nonetheless, syntactic heterogeneity remains problematic, even in the same data model. While both OMOP-CDM and FHIR are able to represent a wide range of clinical information, they both allow storing the same piece of information in different ways. To avoid this inconsistency, we need a common way of representation of such concepts. For example, the metastasis cancer staging values of M1 from the “TNM” (Tumor-Node-Metastasis) category could be represented in two different ways: a) as a concept “AJCC/UICC 7th pathological M1a Category”, which is a Cancer Modifier concept of the “Measurement” domain in the OMOP-CDM, or b) as a NAACCR concept “TNM Path M” of the “Measurement” domain with value “pM1a” of the “Meas Value” domain. Therefore, there is the need to decide how to represent the information: by either its complex form or by using a combination of atomic concepts. This

binding information should be attached to the hyper-ontology concepts, linking to its corresponding CDM attribute, by including annotations in the hyper-ontology (see Figure 2 for an example).

Finally, for being able to formulate queries spanning multiple associated classes in the field of oncology (e.g. retrieve number of patients within a dataset that have had prostatectomy), we need a common meta-model that bridges classes/entities between OMOP-CDM and FHIR.

*Therefore, the hyper-ontology:*

- *should abstract concepts over both data models, and act as a common meta-model so that queries can be formulated.*


***Figure 2**: An excerpt of the hyper-ontology (v1.0) around combining atomic concepts (TNM Path M, pM1a) to represent specific concepts (AJCC/UICC 7th pathological M1a Category)*

As an example, the hyper-ontology could define two classes “Cancer Patient” and “Surgical Procedure”, and a relationship “hasUndergone” linking the two classes. (Figure 3) Through the federated query service, a user could formulate a query based on the hyper-ontology targeting the “Cancer Patient” class and based on the “hasUndergone” relationship query the number of patients that have had a “Surgical Procedure” and more specifically a “Prostatectomy”. In this scenario, the local mediator service in each local node should translate the hyper-ontology based query to the local db schema specific query (e.g. SQL query for an OMOP-CDM relational database) based on the semantics and mappings defined in the hyper-ontology (the hyper-ontology should define the mappings of the “Cancer Patient” to the corresponding classes in both OMOP-CDM (e.g. to the ‘Person’) and FHIR (to the ‘Patient’) and how a “Prostatectomy” maps to a procedure record in the two local schemas.


***Figure 3**: An excerpt of the hyper-ontology (v1.0) around Cancer Patient represented in Protege.*

## 4.2 The EUCAIM Hyper-Ontology

The EUCAIM hyper-ontology is a common semantic meta-model that aims to support and maintain semantic interoperability among heterogeneous cancer image data

models/standards. The hyper-ontology model defines a structured and controlled vocabulary permitting disparate and heterogeneous data models/standards to easily and unambiguously communicate and integrate. Using the hyper-ontology, the *real-world* meaning of essential medical and imaging data/metadata is preserved and exchanged in a standardized, consistent, and meaningful way. Therefore, the main challenge of the hyper-ontology is to facilitate integration and interoperability among data stored and modeled using diverse heterogeneous clinical and imaging data models and associated terminologies. EUCAIM's hyper-ontology is not only a domain ontology that reflects the essentials of the oncology domain for the clinical and imaging contexts but also an application ontology that permits the exploration of data collections, federated querying and processing, and image annotation/segmentation.

## 4.3 Data Resources

The main data resources for building the hyper-ontology are the clinical and imaging data/metadata provided by the AI4HI projects CHAIMELEON, ProCAncer-I, EuCanImage, INCISIVE, and PRIMAGE. While the clinical knowledge is provided as standard concepts from various terminologies/ontologies (e.g., SNOMED, LOINC, NAACCR, UCUM, etc.) following the OMOP/FHIR data models/standards, the imaging knowledge is provided either as DICOM tags/values, or standardized concepts from RadLex. The clinical knowledge is collected as use cases (total of 12 UC) organized per cancer type, including information regarding cancer types and subtypes, therapeutic/surgical procedures, cancer staging/grading systems and values, affected body parts, lab tests, etc. Meanwhile, information about image studies, segmentation, or querying is collected from the imaging knowledge. In the following, we outline the diversity of representing data using different models/standards and specifying the minimum common required data among the different projects.

- ***Clinical and biological knowledge***: ProCAncer-I and CHAIMELEON have adopted OMOP as a CDM, and INCISIVE and EuCanImage have adopted FHIR as a data standard. This diversity has affected not only the terminologies/ontologies used to represent clinical data/metadata, but also the syntactic assignment of concepts to OMOP/FHIR entities/classes. We differentiate between projects that adopt 1) the same data models or 2) different data models/standards.

    1. ProCAncer-I and CHAIMELEON have adopted OMOP and OMOP-Like[5] as CDM, respectively, but represented clinical concepts using different terminologies/ontologies (semantic level) and assigned different classes/entities to these concepts (syntactic level). As an example, the metastasis cancer staging values of M1 from the TNM (Tumor-node-metastasis) category are represented using two different ways in these projects: 1) *AJCC/UICC 7th pathological M1a Category*, which is a *Cancer Modifier* concept of the *Measurement* OMOP domain; 2) *TNM Path M*, a *NAACCR* concept of the *Measurement* OMOP domain with value *pM1a* of the *Meas Value* domain.
    2. ProCAncer-I and INCISIVE have adopted OMOP and FHIR, respectively. ProCAncer-I represents the *PSA* (prostate-specific antigen measurement) lab test using *SNOMED* and assigned it to the domain *Measurement*. Meanwhile,

[5] Adopt OMOP conceptual model (terminologies),

in INCISIVE, PSA is represented using *LOINC* and assigned to the resource *Observation*.

- ***Imaging knowledge***: diverse types of imaging data/metadata are provided.

  For instance, in ProCAncer-I, imaging metadata attributes are defined for querying DICOM_SEG (e.g., *study_uid* (0020,000D), *slice_thickness* (0018,0050)). Besides, values are given for imaging attributes, such as *segment label* (e.g., *PZ*, *TZ*, *CZ*, *SV*) and *method* (e.g., *Manual*, *Semiautomatic*, *Automatic*). Standard imaging concepts are also defined in ProCAncer-I, such as *Laterality* (Radlex, RID5821), *Anatomic Region* (Radlex, RID13390), and *Patient Position* (Radlex, RID10420). However, while in CHAIMELEON, DICOM tags are given (e.g., *SeriesDescription* (0008,103E), *BodyPartExamined* (0018,0015), etc.), image annotation labels are provided by the INCISIVE project (e.g., *Suspicious*, *Problematic*, *Malignant*, *Benign*, *lymph node*, etc.).

Given the diversity and disparity of clinical and imaging knowledge on the semantic and syntactic levels, a common semantic meta-model is required to integrate and generalize the different terminologies/ontologies and the associated OMOP/FHIR domains/resources, permitting seamless communication and information exchange among the heterogeneous cancer image data models.

## 4.4 Development Process

We propose an iterative, systematic, formally, and semantically well-founded approach to the hyper-ontology development process. The proposed approach helps to simplify the hyper-ontology construction, facing the complexity and heterogeneity of the application domain and the diversity and disparity of the provided clinical and imaging knowledge. Six main phases are defined in this approach (see Figure 4).

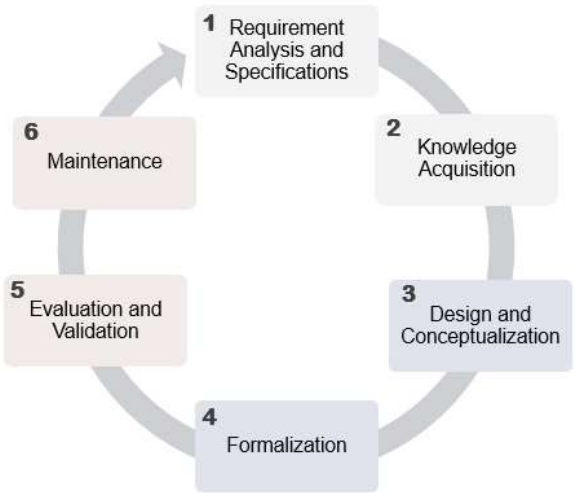


***Figure 4****. An illustration of the Hyper-ontology iterative development process*

### 4.4.1 Requirements Analysis and Specification

After a set of meetings with users and experts from the EUCAIM community, we define the following elements:

- ***Purpose***: To support semantic interoperability by integrating heterogeneous cancer image data models in a common semantic meta-model, which provides the ontology-based standard and structured vocabulary of the oncology domain and the associated semantic relations. Besides, to ensure seamless integration with EUCAIM-CDM, permitting consistent mapping with local nodes, thereby federated querying of data collections.
- ***Scope***: To cover the basics of the oncology domain based on the clinical and imaging knowledge provided by the AI4HI projects, including the following cancer types: prostate, breast, rectum, lung, colon, colorectal, and liver.
- ***Intended uses and users***: To explore data collections through the Public Catalogue[6], Federated Querying (federated search of aggregated data in the collections), and semantic annotation/segmentation of cancer images.
  Hyper-ontology's main users are data users/researchers, persons, or entities that want to explore the public catalog, eventually, request access to data, and process it using the tools available on the platform or their own AI tools.

    - Example of a Data User-Researcher with an experimental lab profile: A Data User-Researcher is leading a project related to prostate cancer. One of the objectives is to allocate treatment based on the analysis of baseline Magnetic Resonance (MR) images at the time of diagnosis. The research team will incorporate AI tools and experience in interpreting the results obtained and applying them in a clinical setting for routine clinical practice.
    - Example of a Data User-Researcher with a Data Scientist profile: A Data Scientist is developing an AI tool to analyze health images and related clinical and molecular data on the most prevalent cancers in Europe. They have an initial model they want to improve with new data. They seek quality and labeled data and do not accept unstructured data or data without a logical folder structure.

- ***Requirements***: Two main types of requirements are defined:
  *Non-Functional Requirements (NFRs):*

    - **NFR1**: To support the English language.
    - **NFR2**: To comply with the FAIR principles.
    - **NFR3**: To align with the General Data Regulation Protection (GDPR).
    - **NFR4**: The terminology in the hyper-ontology must be taken from validated biomedical ontologies and standardized terminologies.
    - **NFR5**: The ontology model should be extensible to handle the periodical updates of semantic standards and to include future ontological aspects and cancer types.

  *Functional Requirements (FRs)*: These are stated as competency questions (CQs) based on the clinical and imaging knowledge provided by the AI4HI projects. We give some examples of FRs and their correspondent CQs/Answers in the following:

[6] https://catalogue.eucaim.cancerimage.eu/

| | |
|---|---|
| **FR1:** To define the basic cancer types. | |
| **CQ1**: What are the leading cancer types? | Prostate cancer, Colon cancer, Breast cancer, Rectal cancer, Lung cancer, Neuroblastoma, Diffuse intrinsic pontine glioma, Colorectal Cancer, Primary malignant neoplasm of liver, Malignant neoplasm of colon and/or rectum, Primary malignant neoplasm of breast. |
| **FR2**: To define the specific cancer types. | |
| **CQ1:** Are there any specific types of breast cancer? | Primary malignant neoplasm of female breast (SNOMEDCT, 363346000), Primary malignant neoplasm of breast with axillary lymph node invasion (disorder) (SNOMEDCT,1082901000112103) |
| **CQ2:** Are there any specific types of prostate cancer? | Benign prostatic hyperplasia (SNOMEDCT, 266569009), Hormone refractory prostate cancer (SNOMEDCT, 427492003), Hormone sensitive prostate cancer (SNOMEDCT, 722103009) |
| **CQ3:** Are there any specific types of liver cancer? | Liver cell carcinoma (disorder) (SNOMEDCT,109841003), Secondary malignant neoplasm of liver (SNOMEDCT, 94381002) |
| **FR3**: To define the main tumor staging methods and values. | |
| **CQ1**: What tumor staging methods are specified for breast cancer? | Edition of American Joint Commission on Cancer, Cancer Staging Manual used for TNM staging (observable entity) (SNOMEDCT, 443941007) |
| **CQ2:** What tumor staging (categorical) values are specified for breast cancer? | American Joint Committee on Cancer clinical T category allowable value (qualifier value) (SNOMEDCT, 1222585009), American Joint Committee on Cancer clinical N category allowable value (qualifier value) (SNOMEDCT, 1222588006), Tumor histopathological grade status values (tumor staging) (SNOMEDCT, 258244004) |
| **FR4**: To define the histology types of cancers. | |
| **CQ1:** Are there any histology types specified for prostate cancer? | Acinar cell carcinoma of prostate gland (ICD-O-3) Intraductal carcinoma, noninfiltrating, NOS, of prostate gland (ICD-O-3), Infiltrating duct carcinoma, NOS, of prostate gland (ICD-O-3), Transitional cell carcinoma, NOS, of prostate gland (ICD-O-3), Adenosquamous carcinoma of prostate gland (ICD-O-3) |
| **FR5:** To define the necessary lab tests for cancer types | |

| | |
|---|---|
| **CQ1:** Are there any lab tests specified for prostate cancer? | Free prostate specific antigen level (SNOMED), Total PSA level (SNOMED), Free:total PSA ratio (SNOMED), Prostate specific antigen normal (SNOMED) |

These requirements, specified as CQs/Answers, are documented in the Ontology Requirements and Specifications Document (ORSD). This document has helped to simplify the Hyper-ontology development process by clarifying the intended content, on which the ontology granularity level depends. In addition, ORSD permits tracking the inconsistencies or lack of information the local nodes provide. The ORSD v1 is available at the following link: https://doi.org/10.5281/zenodo.11109765

Additional functional requirements are defined by EUCAIM experts to help overcome the heterogeneity and disparity of clinical data. We give two examples as follows:

- **FR6:** To ensure the correspondence of concepts to their domains/resources in OMOP and FHIR CDMs. For instance, *Primary malignant neoplasm of breast* and *Chemotherapy* have *Condition* and *Procedure* as FHIR resourceType and OMOP domain, respectively.
- **FR7**: To represent specific concepts by combining atomic-related concepts. For instance, the cancer staging metastasis value of M1 from TNM (Tumor-node-metastasis) category could be represented in two different ways: 1) *AJCC/UICC 7th clinical M1a Category*, which is a concept of the *Measurement* OMOP domain; 2) *TNM Clin M*, a concept of the *Measurement* domain with value cM1a of the Meas Value domain.

### 4.4.2 Knowledge Acquisition

This phase aims to align, or map, the mandatory clinical and imaging knowledge collected from the AI4HI projects and represented in the ORSD document with standard FAIR-compliant terminological and ontological resources. The preferences in terminologies/ontologies (e.g., RadLex[12] for imaging data/metadata) are decided with the help of EUCAIM experts.
Two main types of mappings are performed in this phase:

- ***Hierarchical-based***: to build the hierarchy of the hyper-ontology, we rely on the *is-a* relations extracted from the standard terminologies/ontologies. The extraction process is based on the labels/concepts provided by the AI4HI projects.
- ***Label-based***: to enrich the hyper-ontology with codes from standard terminologies/ontologies, alternative labels, and definitions, an *exact match* similarity approach is applied, considering the clinical and imaging labels of the provided concepts.

The mappings are performed automatically using the following resources, which combine many health and biomedical vocabularies and standards to enable interoperability between computer

systems: OHDSI Athena[7], BioPortal RESTful API[8], and UMLS REST API[9]. Figure 5 depicts examples of mappings.

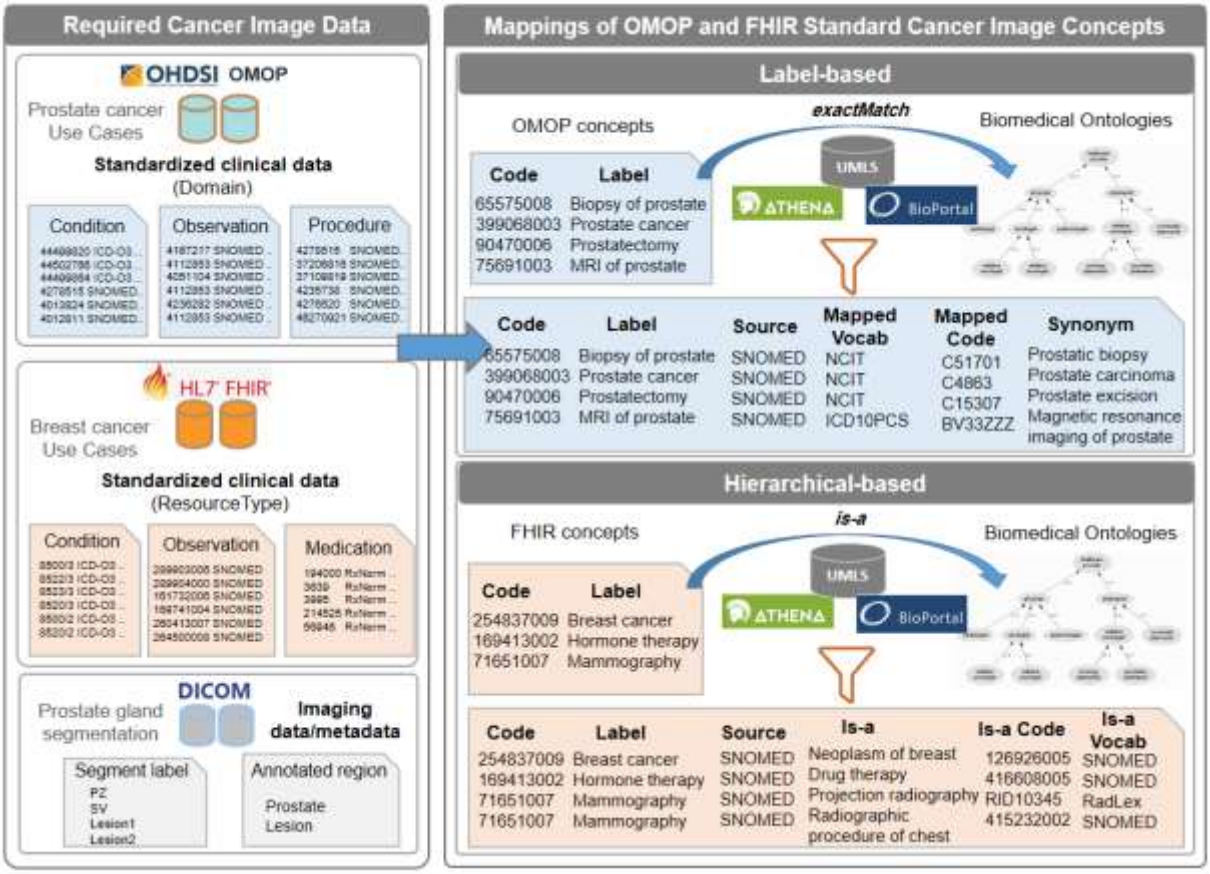


***Figure 5**. An illustration of mappings with Biomedical terminologies/ontologies.*

### 4.4.3 Design and Conceptualization

Faced with the complexity and diversity of clinical/imaging-provided knowledge and the associated mappings with various terminologies and ontologies, we propose dividing the hyper-ontology structure into layers and modules to simplify the building and extension processes. Therefore, four different layers are specified from bottom to top (see Figure 7):

- **Domain-Specific Layer (DSL)**: reflects the granularity level of the hyper-ontology since it includes the domain-specific concepts provided by the OMOP/FHIR projects.
- **Domain Layer (DL)**: includes the concepts obtained from the *is-a* mappings to build the hyper-ontology hierarchy. DL and DSL are maintained using a *bottom-up* strategy relying on the knowledge provided by the AI4HI network.
- **Core Layer (CL)**: defines the core oncology concepts. CL is maintained by considering the conceptual model of mCODE[10]. An ontological analysis is conducted based on well-

[7] https://github.com/OHDSI/Athena
[8] https://data.bioportal.lirmm.fr/documentation
[9] https://documentation.uts.nlm.nih.gov/rest/home.html
[10] https://build.fhir.org/ig/HL7/fhir-mCODE-ig/

known foundational ontologies, such as the Unified Foundational Ontology (UFO)[11], to develop a well-founded ontological model of mCODE. The mCODE core model explicitly defines the real-world entities of the oncology domain and their semantic relations. This approach has helped to clarify or overcome the ambiguity and heterogeneity of how well-known terminologies/ontologies defined essential clinical concepts, such as *Disease* and *Morphology*. Figure 6 depicts an excerpt of the mCODE core ontological model represented using OntoUML around the *Disease* characterization. OntoUML[12] is an Ontology-Driven Conceptual Modeling language where the modeling primitives reflect UFO's ontological distinctions and axiomatization.

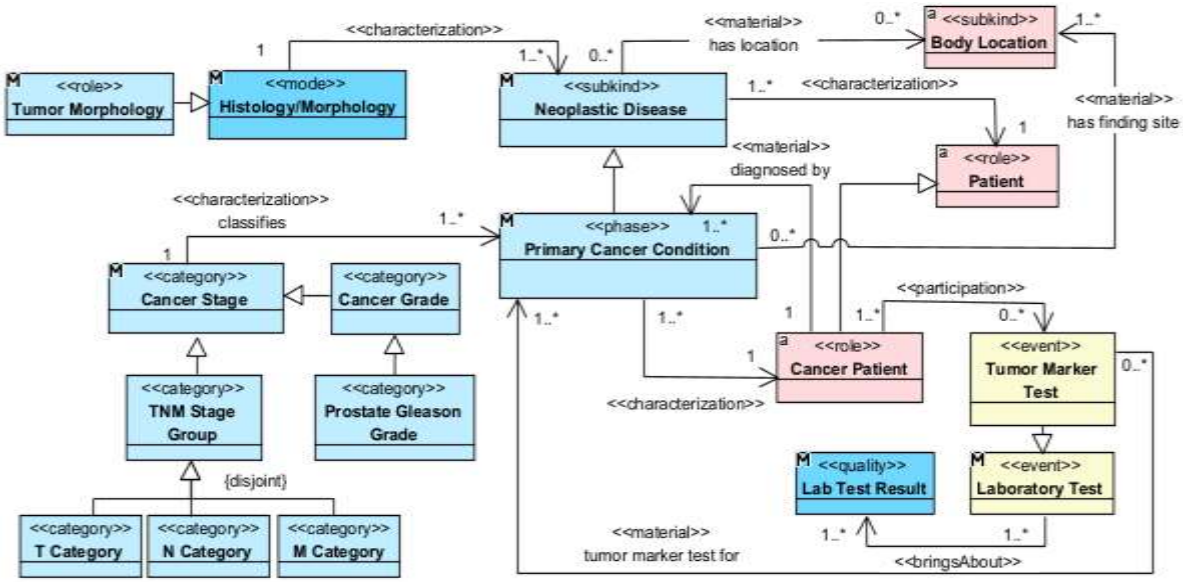


***Figure 6**. An excerpt of the ontological model of mCODE around the Disease characterization represented using OntoUML*

- **Upper Layer (UL)**: This layer is located at the most abstract level and defines the generic concepts of the biomedical domain, such as *Disease*, *Laboratory*, *Surgical Procedure*, *Imaging Procedure*, etc. UL and CL are developed using a *top-down* strategy using OntoUML.

Besides, the hyper-ontology content is divided into three generic modules: *Clinical*, *Imaging*, and *Common* (see Figure 7).

- ***Clinical and Biological* module**: includes the pathological, diagnostic, medical, and biological data/metadata provided by the AI4HI network.
- ***Imaging* module**: includes the modalities, imaging procedures, and attributes such as laterality, orientation, and position. It also defines imaging assessment, such as the *PI-RADS* and *BI-RADS* categories.
- ***Common* module**: specifies mainly the qualifier values required for cancer staging/grading (e.g., *pT1*, *pM2*, *cM3*, *Low histologic grade*, etc.) or image annotation/segmentation (e.g., *benign*, *malignant*, *automatic*, *manual*, etc.). Also, unit measures, such as *millimeter*, *percent*, and *cubic centimeter*, are defined. Besides, *Cancer Patient* and the associated demographics metadata (e.g., *age at diagnosis*, *gender*, and *sex assigned at birth*), are included in this module.

[11] https://nemo.inf.ufes.br/en/projetos/ufo/
[12] https://ontouml.org/

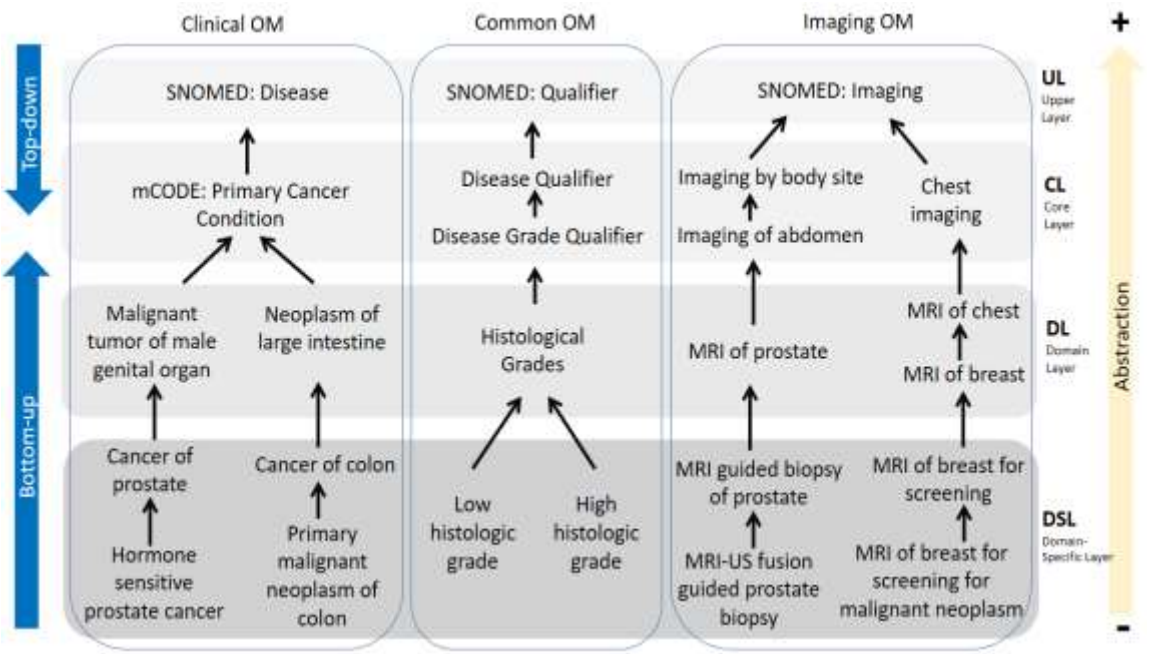


***Figure 7**. An excerpt of the hyper-ontology structure*

### 4.4.4 Formalization

The hyper-ontology model, developed using an iterative approach, is a FIR-compliant ontology model formalized as an OWL[13] (Web Ontology Language) file. Two beta versions (v0.1 and v0.2) of the hyper-ontology have been delivered and shared on Zenodo (https://doi.org/10.5281/zenodo.11109765). Table 3 presents some metrics of hyper-ontology latest version v*1.0* (available at Zenodo at the following https://doi.org/10.5281/zenodo.12583826), including the source and mapping metrics. In table 4, we outline the main terminologies considered in the hyper-ontology. Parts of the formal hyper-ontology model, represented using Protege[14] are depicted and introduced in the following.

***Table 3:** Some metrics of hyper-ontology version 1.0*

| | | | | | |
|---|---|---|---|---|---|
| **Classes** | 2029 | **Mapping to OMOP** | 1755 | **LOINC** | 149 |
| **SubClassOf** | 5395 | **Mapping to FHIR** | 353 | **UCUM** | 25 |
| **Object properties** | 74 | **Mapping to DICOM** | 6 | **RADLEX** | 185 |
| **Equivalence** | 63 | **SNOMEDCT** | 1431 | **DICOM** | 6 |
| **Synonyms** | 2215 | **ICDO3** | 68 | **CPT4** | 9 |
| **Cancer Types/Subtypes** | 148 | **ICD10** | 14 | **Birnlex** | 5 |
| **Histology/Morphology** | 105 | **ICD10PCS** | 9 | **Cancer Modifier** | 158 |
| **Image Modalities types/subtypes** | 35 | **NCIT** | 352 | **UMLS** | 1304 |

[13] https://www.w3.org/OWL/
[14] https://protege.stanford.edu/

| | | | | | |
|---|---|---|---|---|---|
| | | **NAACCR** | 54 | | |

***Table 4**: List of vocabularies supported by the hyper-ontology version 1.0 classified by domain.*

| DOMAIN | TERMINOLOGY |
|---|---|
| **Cancer Types/Subtypes** | SNOMEDCT, ICDO3, ICD10, NCIT |
| **Morphology/Histology** | ICDO3, SNOMEDCT |
| **Body Structure/Topography** | SNOMEDCT, ICDO3, NCIT, RADLEX |
| **Clinical Findings** | SNOMEDCT, NCIT |
| **Family History** | SNOMEDCT |
| **Staging/Grading (e.g., *TNM staging*, *Gleason grading*)** | SNOMEDCT, Cancer Modifier, NAACCR, NCIT |
| **Tumor Marker Test (e.g., *PSA*, *ER*, *PR*)** | LOINC, SNOMEDCT, NCIT |
| **Procedures (surgical, therapeutic, etc.)** | SNOMEDCT, NCIT, CPT4, ICD10PCS |
| **Medication** | RxNorm, SNOMEDCT, ATC, NCIT |
| **Patient Demographics (e.g., *gender*, *sex*, *age at diagnosis*)** | GENDER, SNOMEDCT, LOINC |
| **Absence/Presence Findings (e.g., *negative*, *positive, absent, none*)** | SNOMEDCT, LOINC |
| **Unit of Measure** | UCUM, NCIT, SNOMEDCT |
| **Time Pattern/Time Point (e.g., *start time*, *follow-up*)** | SNOMEDCT, LOINC, NCIT |
| **Image Modalities (e.g., *MRI*, *CT*)** | RADLEX, SNOMEDCT, NCIT |
| **Image Procedures (e.g., *MRI of prostate*)** | SNOMEDCT, RADLEX, NCIT, ICD10PCS |
| **Manufacturer (e.g., *GE*, *Philips*)** | BIRNLEX |
| **Image Assessment (*PI-RADS*, *BI-RADS*)** | RADLEX, SNOMEDCT |

#### *4.4.4.1 Clinical and biological Module*

**Cancer Condition**: Figure 8 depicts the *Primary malignant neoplasm of prostate* (SNOMEDCT:93974005), a cancer condition with associated morphology, the *Malignant neoplasm* (SNOMEDCT), and location, the *Prostate* (SNOMEDCT). The alignment with OMOP is maintained using the semantic relation "*Has correspondence"* and semantic annotation "*OMOP_Domain_ID"*.

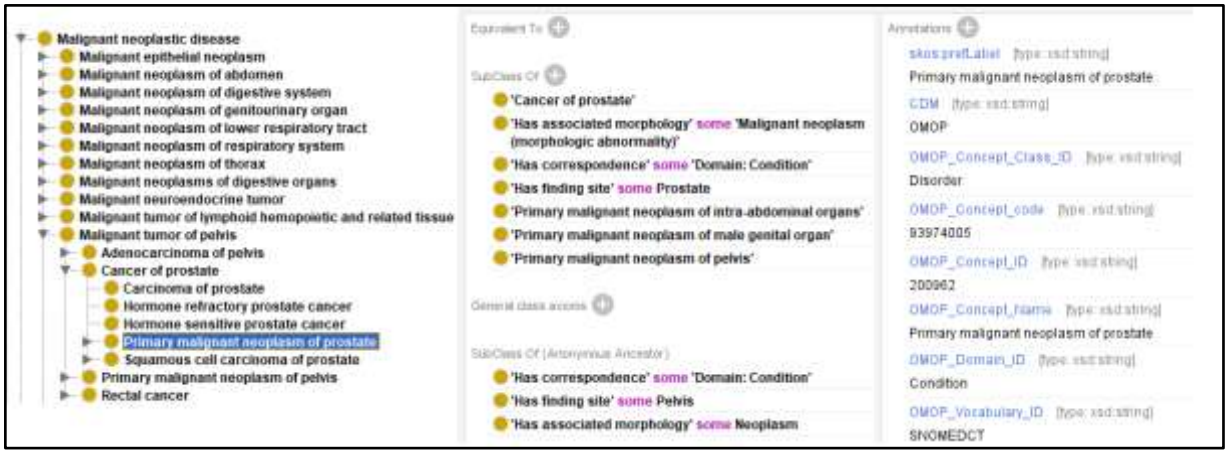


***Figure 8.*** *Part of the hyper-ontology around the concept "Primary malignant neoplasm of prostate" represented using Protege.*

**Morphology**: Figure 9 depicts part of the hyper-ontology around the concept of *Malignant neoplasm (SNOMEDCT:1240414004),* a morphologic abnormality that inheres in "Malignant neoplastic disease" (SNOMEDCT:363346000).

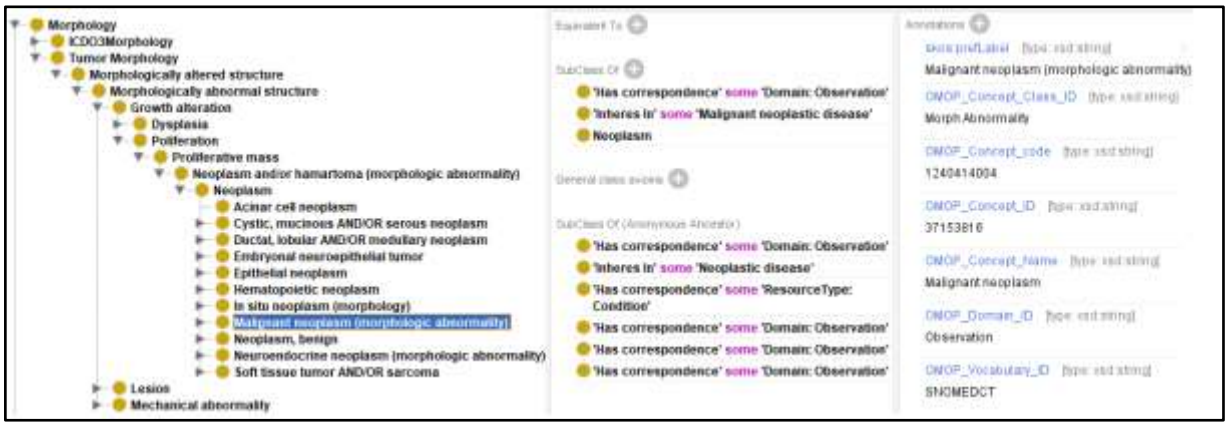


***Figure 9****. Part of the hyper-ontology around the concept "Malignant neoplasm", represented using Protege.*

***Cancer Staging:*** Figure 10 illustrates part of the hyper-ontology around the "*AJCC/UICC 7th clinical M1a Category*" (Cancer Modifier:c-7th_AJCC/UICC-M1a) concept. This concept is represented using other atomic concepts, "*TNM Clin M*" (NAACCR:960) and "*cM1a*" (NAACCR:960@c1A), to solve the disparity problem of representing TNM staging (see **FR7,** section 4.3.1).

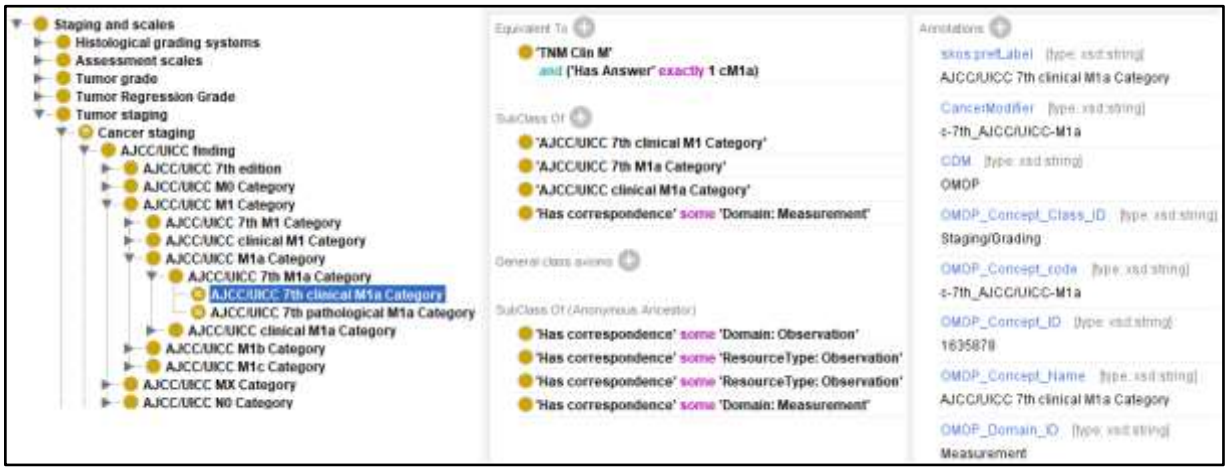


***Figure 10**. Part of the hyper-ontology around the concept "AJCC/UICC 7th clinical M1a Category" represented using Protege.*

**Tumor Marker Test:** Figure 11 depicts part of the hyper-ontology around the concept "*Prostate specific antigen measurement*" (PSA). This concept is defined as *Measurement* in OMOP and *Observation* in FHIR. In the hyper-ontology, this heterogeneity is handled semantically by classifying PSA concepts as *Tumor marker measurement*, which is a specificity *of Measurement of substance* (see **FR6,** section 4.3.1). Meanwhile, the syntactic heterogeneity is maintained by aligning PSA to the corresponding OMOP domain and FHIR resource. PSA is semantically associated with measurement units (*nanogram per milliliter*, *nanogram per deciliter*, etc.), abnormality values (*Normal*, *Abnormal*), and cancer condition (*Cancer of prostate*).


***Figure 11.** Part of the hyper-ontology around the concept of "Prostate specific antigen measurement" represented using Protege.*

### *4.4.4.2 Imaging Module*

***Image Series:*** Figure 12 depicts part of the hyper-ontology around "*Image Series*" (NCIT:C69225), which is part of "*Image Study*" (NCIT:C63859). It *is associated* with the following elements: *Body structure* (SNOMEDCT:123037004), *Imaging modality* (RADLEX:RID10311), *Patient position* (RADLEX: RID10420), and *Laterality* (RADLEX:RID5821). These concepts are also mapped to DICOM by including the corresponding DICOM tags. For instance, *Patient position* and *Laterality* are mapped to the following DICOM tags: *(0018,5100)* and *(0020,0060)*.

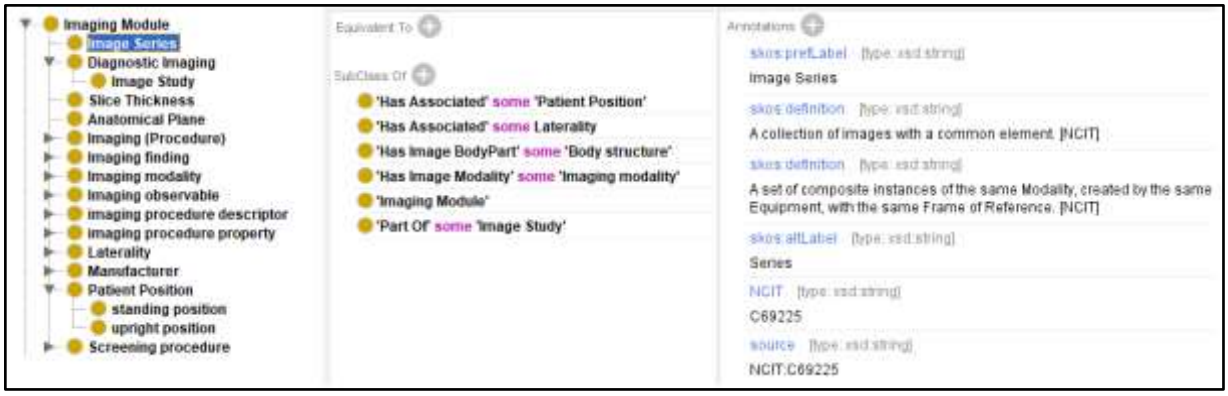


***Figure 12.*** *Part of the hyper-ontology around the concept of "Image series" represented using Protege.*

**Image Modality:** Figure 13 illustrates "*MRI of breast for screening for malignant neoplasm*", a specific concept of "*Imaging Modality*" (SNOMEDCT:360037004) with a *direct procedure site* "*Breast*" (SNOMEDCT:76752008). Also, the *Imaging Modality* general category is aligned to the DICOM tag (0008,0060), permitting a syntactic integration with DICOM.

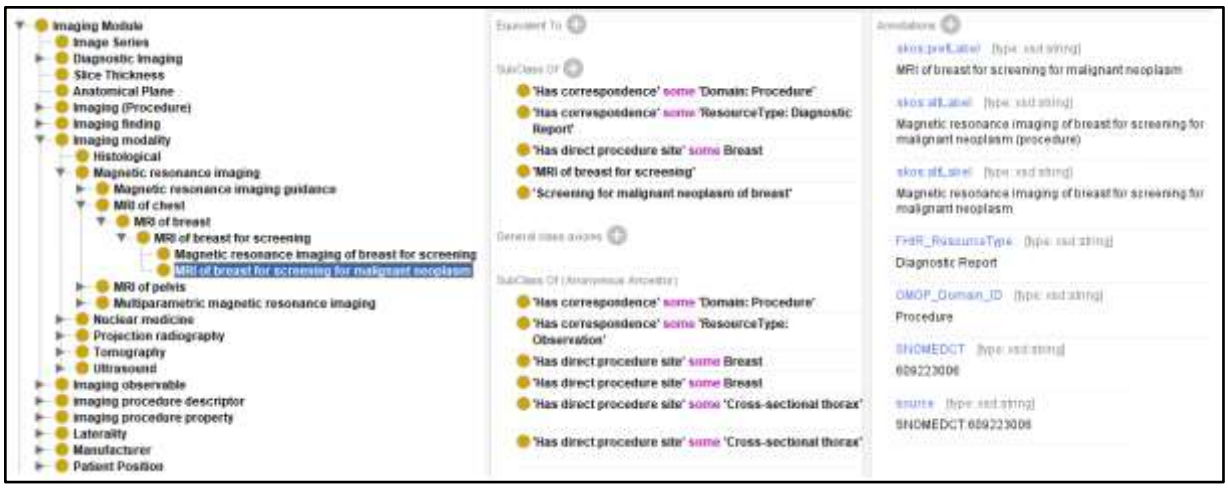


***Figure 13.*** *Part of the hyper-ontology around the concept of "MRI of breast for screening for malignant neoplasm" represented using Protege.*

The hyper-ontology supports the image annotation/segmentation task by considering (standard) specific concepts required as labels/values to annotate the cancer images, permitting a syntactic integration with DICOM SEG. For instance, the image modality label "MRI" or "MR" (Magnetic Resonance Imaging) and the laterality values "Left"/"Right" are defined as specific concepts of *Imaging Modality* and Laterality, respectively. On the other hand, some DICOM tags, such as *segment label (0062,0005)* and *segment algorithm type (0062,0008),* which are provided as imaging metadata, are not defined in standard terminologies/ontologies. Thereby, they are not explicitly specified in the hyper-ontology. However, their associated *values*, which are effectively required for annotation/segmentation tasks, are considered in the hyper-ontology. For instance, the following segment labels, *PZ* (peripheral zone of prostate) (RADLEX:RID347) and *TZ* (transitional zone of prostate) (RADLEX:RID351), are provided by ProCAncer-I as imaging metadata for DICOM SEG querying (see ORSD). They are included in the *Body Structure* category, specifically in the *Region of prostate* (SNOMEDCT:314399000). Similarly, the values associated with segment methods, *Automatic*, *Semi-automatic*, and *Manual,* are defined as modifiers in the *Common Module*. Tables 5 and 6 present the DICOM attributes mapped to the hyper-ontology imaging module and those that are not aligned but whose values are specified.

***Table 5 5**. DICOM tags mapped to the EUCAIM hyper-ontology (version 1.0)*

| DICOM name | DICOM ID | Vocabulary source ID | EUCAIM Concept ID |
|---|---|---|---|
| Patient Position | (0018,5100) | RADLEX: RID10420 | IMG1016605 |
| Body Part Examined | (0018,0015) | SNOMEDCT:52530000 | BP1000024 |
| Manufacturer | (0008,0070) | NCIT:C25392 | IMG1000010 |
| Modality | (0008,0060) | SNOMEDCT:360037004 | IMG1000009 |
| Laterality | (0020,0060) | RADLEX:RID5821 | IMG1016305 |
| Patient Orientation | (0020,0020) | RADLEX: RID10461 | IMG1016610 |
| Slice thickness | (0018,0050) | RADLEX:RID28669 | IMG1016306 |
| Echo time | (0018,0081) | RADLEX:RID12463 | IMG1016641 |

***Table 6 6.** DICOM tags whose values are represented in the EUCAIM hyper-ontology (version 1.0)*

| DICOM name | DICOM ID | Examples of Values | Vocabulary source ID | EUCAIM Concept ID |
|---|---|---|---|---|
| Segment label | (0062,0005) | TZ (Transition Zone of prostate), CZ (Central Zone of prostate), PZ (Peripheral Zone of Prostate) | RADLEX:RID351, RADLEX:RID348, RADLEX:RID347 | BP1000100, BP1000168, BP1000006 |
| Segment method/algorithm type | (0062,0008) | Automatic, Semi-automatic, Manual | SNOMEDCT:8359006, NCIT:C172484, SNOMEDCT:87982008 | COM1000008, COM1000005, COM1000003 |
| Segmentation Type | (0062,0001) | Binary | NCIT:C45969 | COM1000023 |
| Image Type | (0008,0008) | Primary, Axial | SNOMEDCT:63161005, SNOMEDCT:24422004 | COM1000017, COM1000018 |

#### *4.4.4.3 Common Module*

**Cancer Patient**: Figure 14 illustrates the "*Cancer Patient*" concept (NCIT:19700) and the associated semantic relations. Cancer patients are diagnosed with "*Malignant neoplastic disease*" and have undergone some "*Surgical procedure*". "*Gender"* and "Sex assigned at birth" are associated with cancer patients as basic data elements following the mCODE conceptual model.

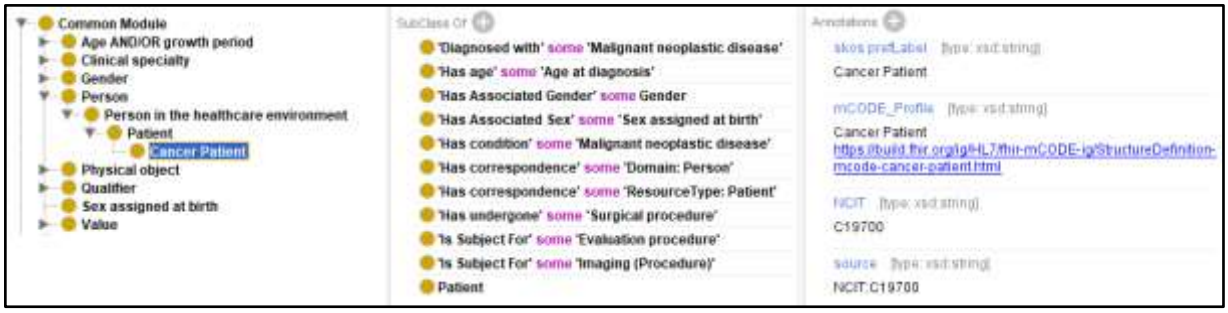


***Figure 14.*** *Part of the hyper-ontology around the concept of "Cancer Patient" represented using Protege.*

**Histologic Grades:** Figures 15 and 16 depict the concepts "*Histological grades*' (SNOMEDCT:370114008) and "*International Society of Pathology histologic grade group"* (ISUP) (SNOMEDCT:1515521000004104). The histological grades are represented in the *Common Module* of the hyper-ontology, specifically in the *Disease Grade Qualifier* category, based on two main reasons: 1) their specification as "Qualifier Value" in OMOP (Concept_Class_ID) and 2) their classification in SNOMEDCT as *Qualifier value* (SNOMEDCT:362981000). Meanwhile, from a clinical expert's perspective, histological grades belong to the *Clinical and Biological Module*, which includes Gleason findings, such as *Gleason grade finding for prostatic cancer* (SNOMEDCT:385377005) having "Clinical Finding" as Concept_Class_ID in OMOP. Both perspectives can be semantically handled and resolved in the hyper-ontology using the *owl:equivalentProperty*. For instance, *Grade group 3 (Gleason score 4 + 3 = 7)* (qualifier value) (SNOMEDCT:1279716004) is equivalent to the union of the following Gleason findings: '*Gleason Primary Pattern Grade 4*' and '*Gleason Secondary Pattern Grade 3*' (see Figure 17). Therefore, *Grade group 3 (Gleason score 4 + 3 = 7)* (qualifier value), which semantically belongs to the *Common Module*, will be automatically classified using the HermiT Reasoner as *subClassOf Gleason Primary Pattern Grade 4* and *Gleason Secondary Pattern Grade 3* in the *Clinical Module*.

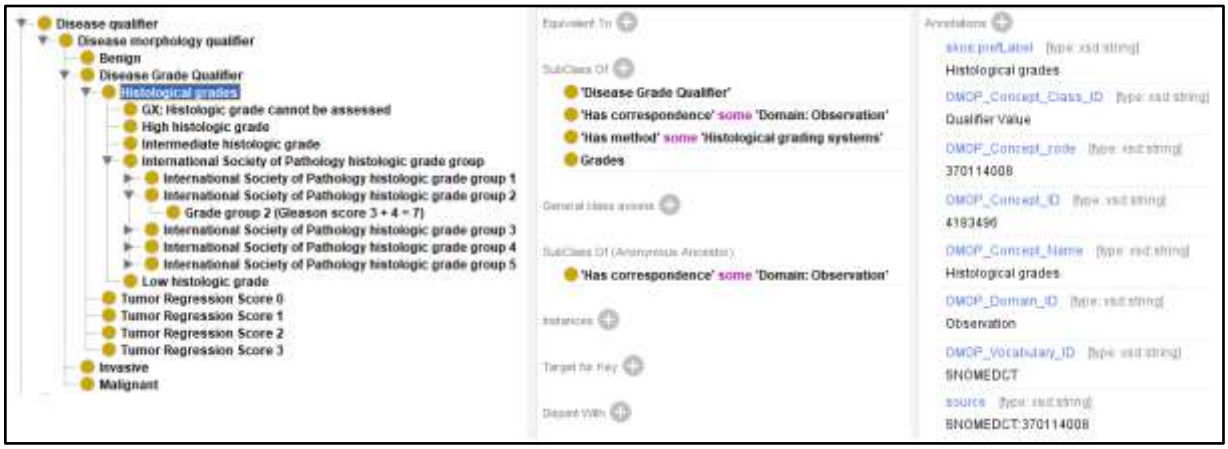


***Figure 15.*** *Part of the hyper-ontology around the concept of "Histological grades" represented using Protege.*


***Figure 16.*** *Part of the hyper-ontology around the concept of "International Society of Pathology histologic grade group" represented using Protege.*


***Figure 17.*** *Part of the hyper-ontology around the concept of "Grade group 3 (Gleason score 4 + 3 = 7)" represented using Protege*

### 4.4.5 Evaluation and Validation

The hyper-ontology is validated as an RDF/OWL formal ontology, and its consistency is verified using Pellet[15], an OWL2 inference engine. To revise the medical and imaging content of the hyper-ontology, workshops are organized with clinical/pathologic and radiologic experts from EUCAIM's community, considering the specified requirements formulated as Competency Questions (CQs) in the ORSD. Also, a *term verification process* has been performed with the help of EUCAIM (WP5) experts to verify that all terms and associated vocabularies are well considered in the ORSD and hyper-ontology as provided by the projects. Besides, meetings with a group of ontology experts are fixed to revise the semantic content of the hyper-ontology, mainly the semantic patterns applied to define specific concepts and the coherence of the hierarchy and modules. Moreover, the hyper-ontology will be evaluated according to its performance in data collection exploration through the Public Catalog, federated querying, and cancer image segmentation/annotation tasks. For the hyper-ontology validation process, we considered real-world use cases around prostate and breast cancers collected from the AI4HI projects (see Section 7, Demonstration Scenarios). Two main validation tasks are applied to verify the pertinence of the hyper-ontology in representing the acquired use cases: 1) we demonstrate hyper-ontology's completeness in representing knowledge from real-world scenarios; and 2) we show the usability of the hyper-ontology for the instantiation of the EUCAIM-CDM based on the provided use cases

[15] https://github.com/stardog-union/pellet

Also, for hyper-ontology validation, we verify the ontology's correctness in answering SPARQL queries (see Annex1) based on the scenarios provided in Section 7.

### 4.4.6 Ontology Enrichment and Maintenance

The process of the hyper-ontology enrichment is continuous throughout the iterative development process. Also, we enrich the hyper-ontology model by considering experts' feedback on each delivered version or any additional requirements and specifications defined by the EUCAIM community, mainly regarding the federated querying or image annotation/segmentation tasks. Moreover, meetings with clinical experts have helped to enrich the *medical-oriented* semantic content of the hyper-ontology by maintaining the semantic patterns connecting various concepts. For instance, in the hyper-ontology, the results of tumor marker tests are represented in two different ways: 1) *conditions* (e.g., *Oestrogen receptor positive tumour* (SNOMEDCT:416053008), *Progesterone receptor negative tumour* (SNOMEDCT:441118006)) and 2) *observations* (e.g., *Estrogen receptor Ag [Presence] in Breast cancer specimen by Immune stain* (LOINC:85337-4), *Progesterone receptor Ag [Presence] in Breast cancer specimen by Immune stain* (LOINC:85339-0)) associated with qualifier values (e.g., *Positive* (SNOMEDCT:10828004), *Negative* (SNOMEDCT:260385009)), indicating the positive or negative detection of tumor markers. Considering the expertise of clinical experts, which states that both aspects reflect similar contexts, we can semantically associate them using an equivalence property (*owl:equivalentProperty)* (see Figure 18 for an example).

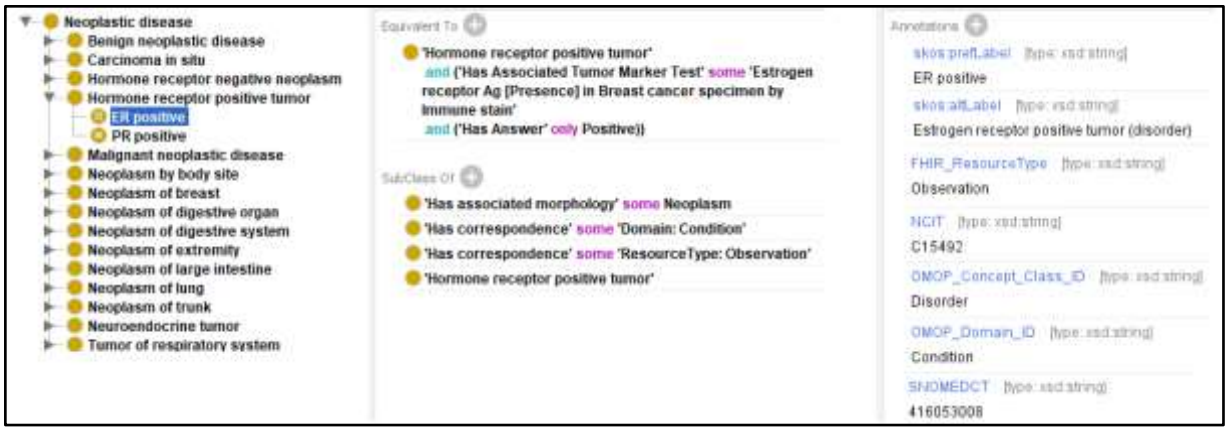


***Figure 18.*** *Part of the hyper-ontology around representing tumor marker test results (Protege).*

Another example concerns the existence of secondary and primary cancers. From a clinical perspective, the term *secondary cancer* may refer to either *metastasis* from primary cancer or *a second cancer* unrelated to the original cancer. Thereby, the existence of a secondary cancer condition (either a metastasis or second cancer) is related to an existing primary (or original) condition. Accordingly, a semantic relationship *(Has Associated Primary Condition*) is defined to link the secondary cancer to primary (see Figure 19). This semantic pattern will help to logically *deduce* the existence of a primary cancer condition for a cancer patient who is suffering from a *clinically identified* secondary cancer condition (see the example of prostate cancer use case - ProCAncer-I, Section 7). The existence relationship is not applicable in the opposite direction; a primary cancer condition does not necessarily entail a secondary cancer condition.

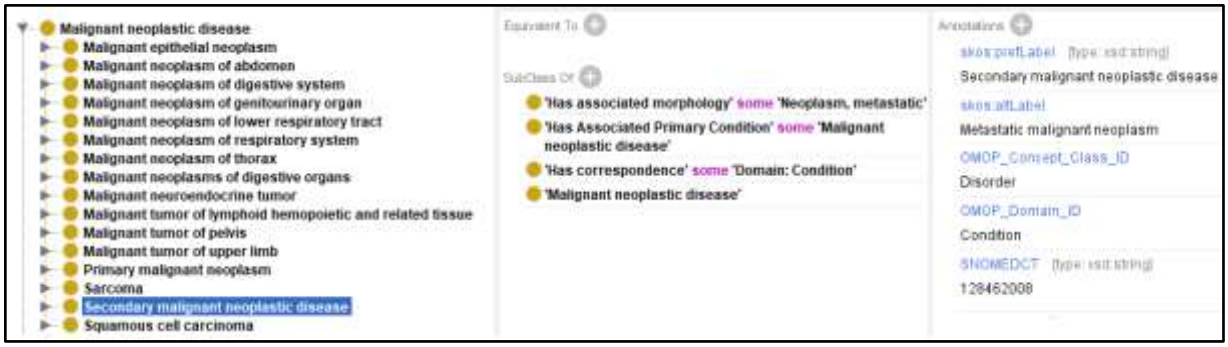


***Figure 19.*** *Part of the hyper-ontology around primary and secondary cancer relationship (Protege).*

Regarding the continuous updates and changes of the hyper-ontology content, there is a need to address the expansion of the semantic content while ensuring that consistency is maintained. Regular evaluation and validation processes on the syntactic and semantic levels (see Section 7) are required to assess the impact of evolution on the consistency and correctness of the hyper-ontology.

# 5. Interoperability framework for federated processing

For enabling federated processing, data holders should implement a semantic and syntactic interoperability layer across their datasets. Semantic as how data meaning is consistent across datasets (this layer should also be implemented in tier 2), and syntactic as how data is structurally persisted within a database.
Syntactic interoperability at this tier is important so that any tool or AI/ML model processing the data is aware of the format and the structure of the local dataset, and these aspects are not addressed by the conceptual specifications (entities, relationships, terminologies) of the hyper-ontology.

## 5.1 CDM business requirements

Prior to selecting a CDM, we conducted an initial analysis of the main requirements, expectations, and constraints from various stakeholders. Our approach involved engaging with representatives from the AI4HI projects and requesting specific information, as follows:

- The specific cancer types that each project focused on.
- The clinical questions/use cases addressed by each project.
- The clinical and imaging data used to answer these questions, including mandatory and optional information.
- The format of the raw data available and whether standardized terminologies were used for different data types, along with the versions of these terminologies.
- The anonymization techniques/profiles employed by each project to ensure compliance with GDPR and national data privacy laws.
- Details about the modalities of radiological images collected and the imaging metadata associated with them, or extracted, if applicable.
- Information regarding the format of segmentation masks, if they exist.
- The chosen common data model and whether it covers all data types, with a straightforward mapping from the raw data.

This information was collected and documented in the ORSD document described in the previous section. The outcome of the analysis was outlined in D5.1 (section 3). It is evident that there are many challenges to be addressed, as the AI4HI projects are dealing with different cancer types, with only three out of five projects to deal with a common type of cancer, i.e. breast and prostate cancer, different use cases, and therefore different clinical and imaging data to support these use cases, different terminologies, different anonymization profiles, different formats for the segmentations, and although all of them have standardized data models, the OMOP-CDM and the FHIR resources as a data model, these are also different. Most importantly, as some of the AI4HI projects are getting finalized, they have no plan of transforming their datasets to a specific standard, as they have all selected and adopted the data model that serves the needs of the respective project. In addition to the AI4HI projects, we need to take into consideration constraints that might arise from new data holders willing to join the EUCAIM federation, which might have either standardized data models or totally ad-hoc models and might also have different capabilities, in terms of technical facilities and resources in general.

Following the collection of information from the AI4HI projects, several group meetings were conducted with different domain experts within the consortium, including AI experts, data

holders, software engineers, and legal teams, to define the data model business requirements for the project. The most critical requirements are presented below:

- EUCAIM should support as many input formats as possible for raw clinical and imaging data, which may or may not comply with interoperability standards.
- The data model should be terminology-agnostic, accommodating different terminologies seamlessly.
- Minimization of the effort required from clinical data managers to prepare data for federated processing and analysis through the platform.
- The data model must fully comply with GDPR and national privacy laws.
- The data model should comprehensively represent all target data types at their intended level of detail, including clinical, demographic, radiomic, and laboratory data.
- The data model should be extensible to allow for additional/new data to be represented.
- The data model must provide an interface for accessing and querying data for the purpose of training federated AI models.
- Data transformations from the raw source to the AI training dataset should be as straightforward as possible.
- The data model should be structured in a way (usually in a tabular format) that simplifies the retrieval of records in the training dataset, regardless of the training plan of an AI algorithm.

Within EUCAIM, two potential frameworks for data harmonization and standardization are being explored, as mentioned in the TEHDAS recommendations on a Data Quality Framework document[16]. One approach involves transforming all datasets held by a data holder to comply with a specific internationally adopted standard (e.g., OMOP-CDM). The other approach entails preparing the dataset for delivery based on a specific data schema that includes the necessary harmonization rules, controlled vocabularies, and standards.

In the first approach, harmonization is driven by a standard design, resulting in a dataset that is comprehensible to the community and can be used for federated analysis and to support interoperability with other research infrastructures and networks (e.g., OHDSI, Darwin EU, EHDEN). However, this method requires significant upfront effort (although only done once per dataset) and is only accessible after extracting, semantically mapping, and transforming all data sources to the standard data model. This ties the research question specification to the semantic constraints of the standard model specification.

In the second approach, harmonization is driven by the materialization of specific information in a bespoke data model, where each transformation is limited to specific entities and variables of interest. This, however, limits the reuse of the data in other contexts and introduces an additional data model for specific purposes. It is important to note that preparing datasets for secondary use should not be limited to mapping concepts. It also requires developing data models that provide a logical harmonized schema, integrating different health data sources among data holders.

In the context of EUCAIM, we explored different approaches to be considered for Tier 3 (federated processing/analysis and AI model development), which is the maximum level of interoperability to be achieved in EUCAIM, based on the two aforementioned harmonization

[16] https://tehdas.eu/app/uploads/2023/09/tehdas-recommendations-on-a-data-quality-framework.pdf

frameworks. These approaches are analyzed in the following section, and which guided many decisions regarding the CDM (e.g. structure, format).

## 5.2 Data harmonization approaches for the federated processing/analysis.

### 5.2.1 Scenario 1: EUCAIM Hyper-Ontology Based CDM for Analysis

The architecture for this scenario is shown in Figure 20. This case outlines two distinct pathways for integrating data from AI4HI repositories or already established repositories adopting standards (OMOP, FHIR) and new data holders with ad-hoc models.

1. ***Established repositories (e.g. AI4HI projects)****: implement a mediator/data access service that dynamically transforms and structures data according to the hyper-ontology and CDM specification.*
2. ***Other data holders (e.g. hospitals):*** *undergo an Extract Transform Load (ETL) process, directly converting their local data into an EUCAIM hyper-ontology based CDM.*

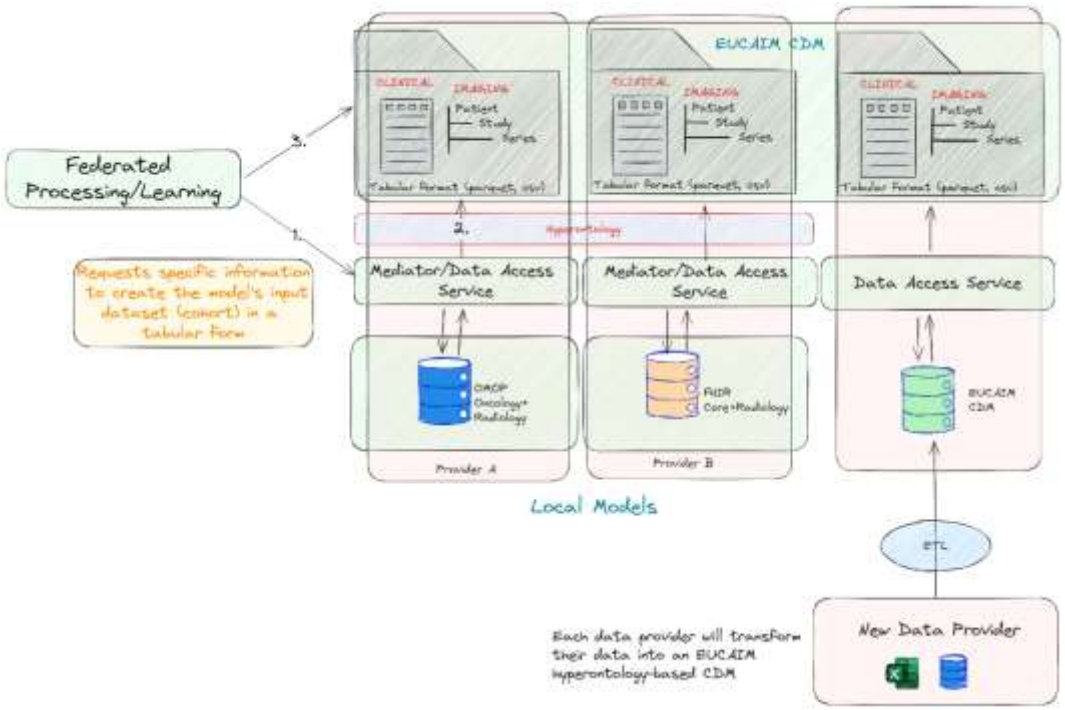


***Figure 20:*** *EUCAIM CDM for analysis & OMOP, FHIR, EUCAIM local data models. For OMOP and FHIR a mediator and mapping component is necessary.*

In this examined scenario, researchers access a Data Access Service in order to request specific information to create their model's input dataset (cohort) in a tabular form (e.g. csv). Established repositories (e.g. AI4HI repositories) utilize a mediator service and a mapping component to transform queries based on the hyper-ontology concepts (e.g., age at diagnosis, modality) to the local CDM query language and the local CDM concepts. It is in a way the same mapping component/service as in the mediator in Tier 2, but in this case, the mediator doesn't return aggregated information, but rather specific hyper-ontology based attributes (e.g. age at

diagnosis, modality, PSA etc.). This required information can be subsequently stored in a tabular form (e.g. csv, parquet) file along with the corresponding images in a POSIX path, that the federated processing service is able to access. For new data holders, an ETL process aligns datasets directly with the EUCAIM hyper-ontology based CDM specification.
The advantages of this approach are:

- The researchers are able to slice and dice the information available according to the needs of their analysis/use case and the inputs of their respective models in an easy and user-friendly way through the data access service.
- Federated Learning scenarios are easier for the researchers since they can specify what type of data (and format) want to be available on each federated node.
- Eliminates the need for AI4HI repositories to go through an ETL process for transforming their data, but rather create a mapping component that transforms only the requested information on the fly and on demand.
- Streamlines data transformation for new data holders through an ETL process, without implementing any mediator/mapping component.

The disadvantages of this approach are:

- A model registry or a UI is required so that researchers are able to specify what's the "granularity" their models/tools want to have their input to (e.g. which variables)
- A data access service is needed to accept specifications of the needed dataset and create (materialize) dynamic cohorts based on these, which increases complexity.
- The mediator component's on-the-fly data transformation (materialization) is technically challenging.
- Adopts a bespoke data model for new providers (based on the hyper-ontology), limiting its utility outside EUCAIM.

### 5.2.2 Scenario 2: Integration with OMOP-FHIR for Wider Compatibility

In this scenario, new data holders can opt to convert their data into either OMOP-CDM or FHIR based standards. This facilitates easier integration with EUCAIM, in a similar way to the AI4HI projects and enhances data utility beyond the EUCAIM ecosystem. Therefore:

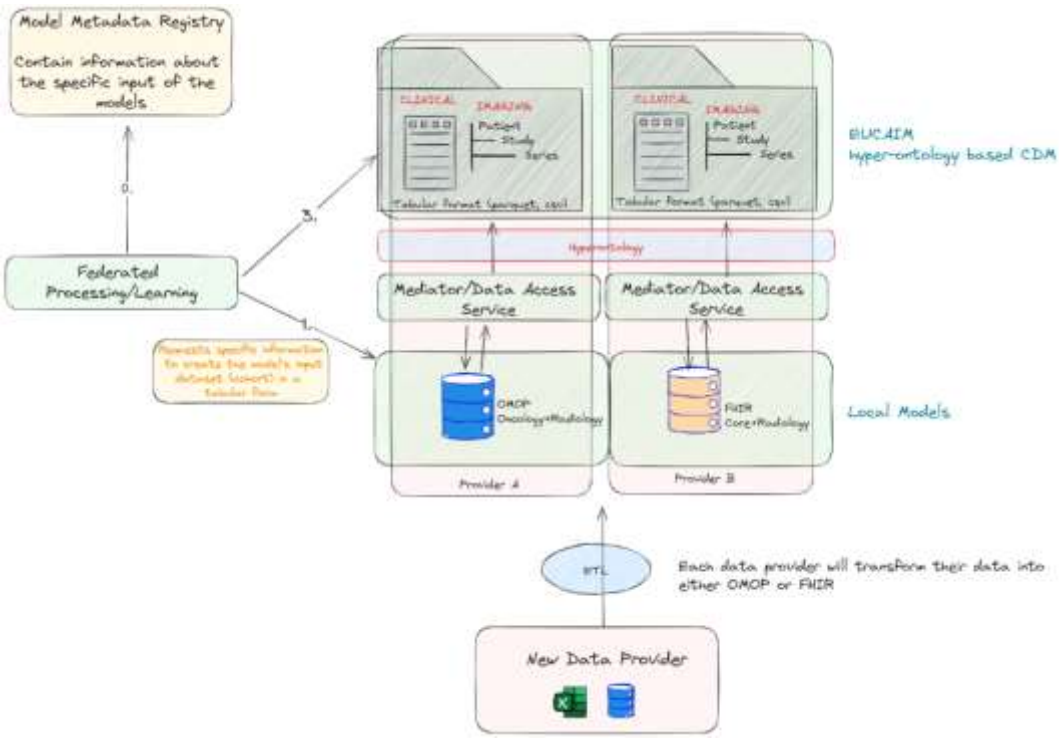

***Figure 21**: OMOP-FHIR local adopted standards– EUCAIM based CDM for analysis with mediator and mapping components necessary for all nodes in the federation.*

1. ***Established (AI4HI) repositories and compliant data holders to OMOP/FHIR** standards use a mediator service as in Scenario 1. (EUCAIM will need to provide mediator components (OMOP/FHIR) to the new data holders (i.e. customized versions of them, as even the same CDM has differences in the way the information is structured as we described in section 4.)*
2. ***Non-compliant data holders to OMOP/FHIR** standards undergo an ETL process to comply with either OMOP or FHIR standards.*

Figure 21 shows the architectural design of this approach. The advantages of this approach compared to Scenario 1 is that new data holders align with well-established standard generic data models, enhancing interoperability and impact beyond EUCAIM. However, the disadvantage of this approach is that a mediator service and a mapping component should be implemented for this case as well, so that all OMOP and FHIR based repositories are harmonized for data analysis, with all the disadvantages this mediator service entails, as described in scenario 1.

### 5.2.3 Scenario 3: Simplifying Integration Through ETL process

This approach mandates **all participating repositories** to undergo a one-time ETL process, conforming to the EUCAIM hyper-ontology based CDM, thereby reducing technical complexities associated with mediator services. In this case all federated nodes can use the same (simpler) Data Access Service implementation that exports data from the CDM into a common format. Figure 22 shows the architectural design of this approach.

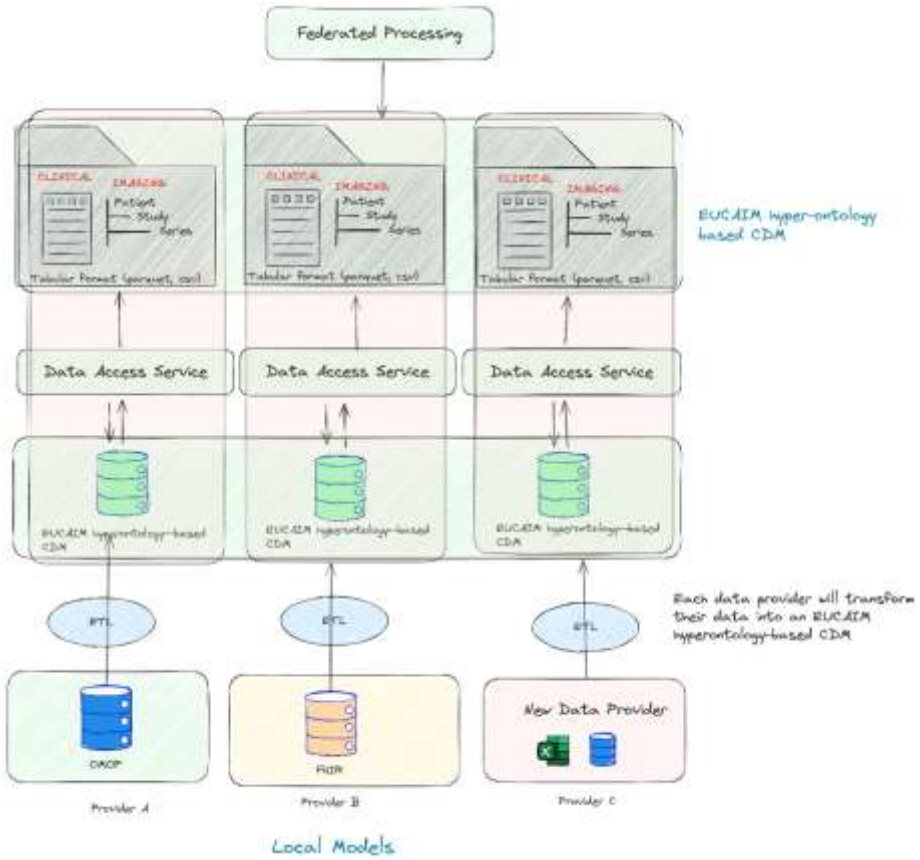

***Figure 22:*** *EUCAIM based CDM for all nodes participating in the federation. This would require a one-time transformation and no mediator/mapping component is necessary.*

### 5.2.4 Scenario 4. EUCAIM hyper-ontology only for federated query purposes, OMOP-CDM for analysis

In this scenario, the EUCAIM hyper-ontology is only applicable for Tier 2 for the federated query purposes and is not used for federated processing. The architectural design of this approach is outlined in Figure 23.

**All participating repositories** should conform to the OMOP-CDM standard data model and go through an ETL process (apart from the OMOP-CDM ones – although some adaptation will be needed to address specific issues as described in section 4.1). The federated processing service could directly access an SQlite[17] file (for example) with the whole OMOP-CDM relational schema available, perform any desired query and transform it to any tabular format for input to the AI model or for analysis.

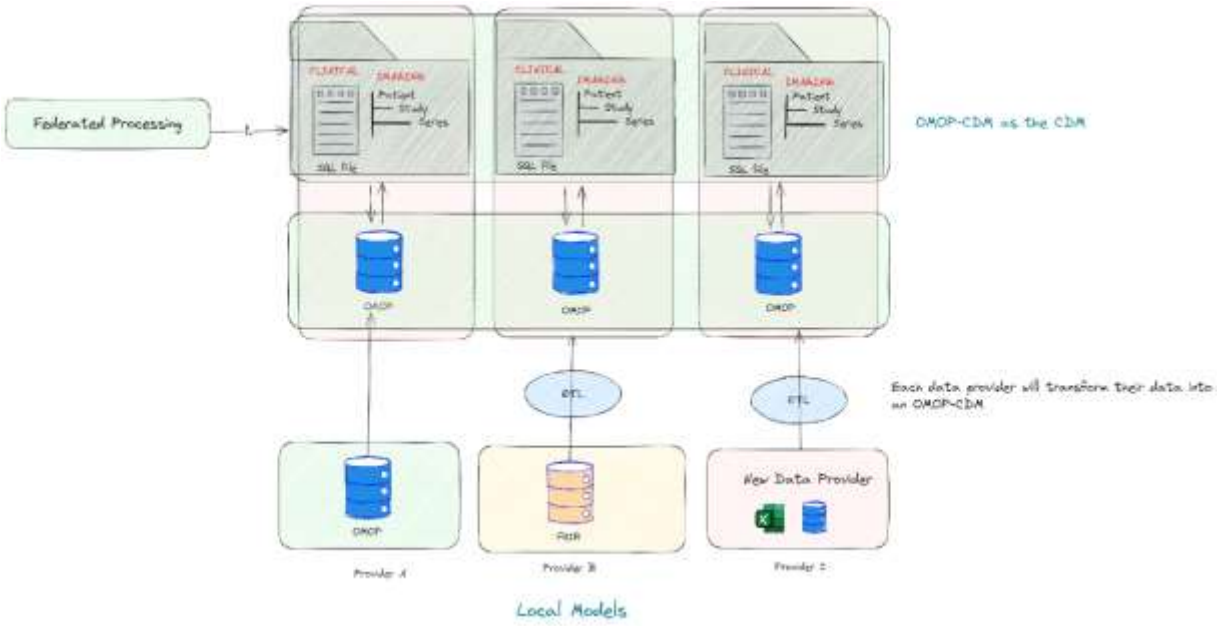


***Figure 23:*** OMOP-CDM as the EUCAIM CDM for federated processing and analysis. Hyper-ontology only for federated queries.

The approach of not having a data access service in this case, but rather providing the whole dataset for researchers to use and slice and dice information, could also be applied to the

previous scenarios as well, regardless of the chosen CDM for analysis. However, the disadvantage of this approach is that all nodes need to both go through an ETL process, but also have a mediator for Tier 2, as this conforms to the hyper-ontology concepts and terms (for bridging the gaps between OMOP and FHIR standards). This approach could also be used with a FHIR-based standard, however, as we described and analyzed in D5.1, OMOP-CDM is more appropriate as a CDM for analysis and AI related operations. In addition, another

[17] https://www.sqlite.org/

drawback of this approach is that researchers are given an SQLite file/relational database to deal with, which requires knowledge of both OMOP-CDM and SQL query language, and not a tabular format that AI experts are usually engaged and accustomed with, which can be dynamically formed for their purposes. In this case, another access service could be added on top of the OMOP-CDM databases for a more user-friendly access to the underlying data.

## 5.3. The EUCAIM Common Data Model

### 5.3.1. CDM Selection Rationale

Based on the aforementioned analysis and the requirements from various stakeholders, i.e., AI experts, data model experts and AI4HI project representatives, Scenario 1 and Scenario 3 were deemed the most appropriate for supporting all the necessary processes for querying and transforming information required by the AI model algorithms and frameworks. Consequently, the EUCAIM CDM for analysis and federated processing/learning will be based on the hyper-ontology specification, which underpins the EUCAIM logical data model.

It is important to note that EUCAIM will not mandate the adoption of Scenario 1 or Scenario 3, which involves either a mediator implementation or a one-time ETL process, respectively. However, the EUCAIM partners agreed that a one-time transformation to the EUCAIM CDM is more straightforward and easier to implement, therefore this will be the recommended approach.

As we initially described in Section 4.4.3, the mCODE conceptual model was identified as the most appropriate basis for grounding the hyper-ontology in the oncology domain, especially to build the core layer of the hyper-ontology model by ontologically analyzing and explicitly and semantically representing the mCODE basic specifications. The rationale behind this decision is multifold.

Although the OMOP-CDM and FHIR standards are widely used for standardizing and exchanging healthcare data, they have limitations when it comes to AI-related tasks, especially those requiring tabular data for model training and analysis. OMOP-CDM excels in transforming and standardizing data from diverse healthcare sources into a common format, which is beneficial for interoperability and large-scale observational studies. However, due to its generic nature, and the fact that it is an observational-based model, it makes it unsuitable and not much straightforward for querying oncology related information by AI experts. For example, through its oncology extension most of the cancer modifiers, as these are defined in the OMOP-CDM specification, are represented as “Measurements”, limiting the semantics of cancer stages, cancer grades, extensions, invasions etc. Similarly, the basic FHIR (Fast Healthcare Interoperability Resources) specification is designed to facilitate real-time data exchange between healthcare systems, with its primary focus being on ensuring that different systems can communicate effectively. However, FHIR’s hierarchical and often complex data structures are not inherently suited for the tabular data formats required by many AI algorithms and frameworks. As a reference, all tools currently available in EUCAIM, which are thoroughly described and analyzed in D5.4 require clinical and imaging metadata in a tabular format.

Due to the aforementioned reasons, EUCAIM explored the two most prominent data models in oncology: mCODE (Minimal Common Oncology Data Elements)[18] and OSIRIS[19] (Interoperability and data sharing of clinical and biological data in oncology) which are both event-based models. mCODE, introduced by the ASCO and a group of collaborators, provides a standardized set of essential oncology data elements, ensuring interoperability and data consistency, which is critical for building reliable AI models. Although mCODE is based on FHIR, it narrows down the scope to oncology-specific data elements, making it easier to extract and query relevant information for cancer research and AI applications. On the other hand, OSIRIS, developed by INCa, offers a minimum data set for the sharing of clinico-biological data in oncology. Its relational model makes it easier to represent and manipulate as tabular data, which is ideal for AI model training. This structure allows for efficient querying, aggregation, and analysis of large datasets.

All options considered, the EUCAIM CDM will leverage and build upon the conceptual model of the mCODE specification and the OSIRIS data framework, leveraging the strengths of each framework, as well as accounting for the specific constraints underpinned by the secondary use of data and the AI4HI projects. For example, both models contain mandatory attributes, which cannot be supported by the available knowledge of the AI4HI projects, and that is due to GDPR and anonymization strategies followed by each project for reducing risks of re-identification of patients, and the fact that the clinical information collected by the projects *accompany the imaging data.* As an example, all date related attributes included in both the OSIRIS and mCODE specifications are not part of the knowledge collected from the AI4HI projects due to the anonymization of the clinical information. Instead, relative relations based on events such as diagnosis or treatment (e.g., events that happened X months after baseline/diagnosis/treatment) are included.

Summarizing, in the context of EUCAIM, mCODE will be the basis conceptual model for representing various cancer types, cancer stages, performance status metrics and scales, as well as assessments (e.g. radiological assessments (ACR Reporting and Data Systems (RADS)[20]), and it is also generally more advantageous due to the fact that it is built on the FHIR based standard, which can be exploited, if necessary, in other contexts, for exchanging purposes. In addition, OSIRIS' relational model nature, and its approach of creating pivot tables (.csv files) for use in AI related processes supports efficient data selection for data preprocessing, feature extraction, and model training, ultimately enhancing the development of AI applications in oncology, and therefore EUCAIM will follow the same approach for facilitating AI experts in selecting specific cohorts as input to their models, by the use of pivot tables.

A first version of the EUCAIM Data Dictionary is described in the following section. A more detailed version is also available at: EUCAIM_CDM_mCODE_based_v1.0.xlsx

[18] https://build.fhir.org/ig/HL7/fhir-mCODE-ig/

[19] Guérin, J., Laizet, Y., Le Texier, V., Chanas, L., Rance, B., Koeppel, F., Lion, F., Gourgou, S., Martin, A. L., Tejeda, M., Toulmonde, M., Cox, S., Hess, E., Rousseau-Tsangaris, M., Jouhet, V., & Saintigny, P. (2021). OSIRIS: A Minimum Data Set for Data Sharing and Interoperability in Oncology. *JCO clinical cancer informatics*, *5*, 256–265. https://doi.org/10.1200/CCI.20.00094

[20] https://www.acr.org/Clinical-Resources/Reporting-and-Data-Systems

## 5.3.2. EUCAIM Data Dictionary

The EUCAIM CDM classifies all the clinical patient data into 6 different domains according to the mCODE specification:

### 5.3.2.1 Patient

The patient information group allows for general information about the patient including demographics, and the patient's managing organization.

***Table 7 7**: The EUCAIM CDM: Patient group*

| Group | Entity | Data Element | Definition | EUCAIM Required | Occurrences Allowed | Data Type |
|---|---|---|---|---|---|---|
| Patient | Patient | Identifier | Anonymized patient identifier which is unique within the context of the system. | Required | 1..1 | string |
| | Patient | Gender | Administrative Gender - the gender that the patient is considered to have for administration and record keeping purposes. | Optional | 0..1 | CodeableConcept |
| | Patient | Ethnicity | Concepts classifying the person into a named category of humans sharing common history, traits, geographical origin or nationality. | Optional | 0..1 | CodeableConcept |
| | Patient | Race | Concepts classifying the person into groups based on their physical appearance | Optional | 0..1 | CodeableConcept |
| | Patient | Birth Year | The year of birth for the individual. | Optional (required if diagnosis age is not available) | 0..1 | Integer (>1900, <current year) |
| | Patient | Managing Organization | Organization that is the custodian of the patient record. Need to know who recognizes this patient record, manages and updates it. | Required | 0..1 | Organization |
| | Patient | Care Provider | Patient's primary care provider or ganization. | Optional | 0..1 | Organization |

| | | | | | | |
|---|---|---|---|---|---|---|
| | Patient | Birth Sex | A code classifying the person's sex assigned at birth. | Required | 1..1 | CodeableConcept |
| | Cancer Patient | Deceased | Indicates if the individual is deceased or not. | Optional | 0..1 | boolean |
| | Cancer Patient | Cause of death | Main cause of death of the patient | Optional (conditional on deceased) | 0..1 | CodeableConcept |
| | Cancer Patient | Date of last contact | Date of last contact if not deceased, or date of death if deceased. | Optional (conditional on deceased) | 0..1 | Date |
| | Organizatio n | Identifier | Identifies this organization across multiple systems | Optional | 1..1 | String |
| | Organizatio n | Name | Name used for the organization | Optional | 1..1 | String |

### 5.3.2.2 Health Assessment

The health assessment group contains information related to the patient's general health before and after treatment. This includes Comorbidities, Laboratory Tests, Performance Assessments (ECOG), Vital Signs, Family Member History, and Patient History of Metastatic Cancer.

***Table 8 8**: The EUCAIM CDM: Health assessment group*

| Group | Entity | Data Element Name | Definition | EUCAIM Required | Occurrences Allowed | Data Type |
|---|---|---|---|---|---|---|
| Health Assessment | Family Member History | Subject | The patient that the family history is about | Required | 1..1 | Reference: Patient |
| | Family Member History | Relationship | Relationship to the subject | Required | 1..1 | CodeableConcept |
| | Family Member History | Condition Code | Condition that the related person had | Required | 1..1 | CodeableConcept |
| | Family Member History | Onset Age | When condition first manifested on the relative. | Optional | 0..1 | Age |

| | | | | | | |
|---|---|---|---|---|---|---|
| | History of Metastatic Cancer | Code | Type of observation | Optional | 0..1 | CodeableConcept |
| | History of Metastatic Cancer | Value | The information determined as a result of making the observation, if the information has a simple value. | Optional | 0..1 | boolean |
| | Comorbidities | Focus | Comorbid conditions are typically defined with respect to a specific 'index' condition. For example, comorbid condition categories would be those specified by CDC, namely obesity, renal disease, respiratory disease, etc. | Optional | 0..* | Reference: PrimaryCancerCondition |
| | Comorbidities | Comorbid Condition Present | A comorbid condition that is known to be present Required (conditional) | Required (conditional) | 0..* | CodeableConcept |
| | Comorbidities | Comorbid Condition Absent | A condition that is NOT present, related to the patient. Required (conditional) | Required (conditional) | 0..* | CodeableConcept |
| | Comorbidities | Code | Describes what was observed. Sometimes this is called the | Required | 1..1 | CodeableConcept |

| | | | | | | |
|---|---|---|---|---|---|---|
| | | | observation "name". | | | |
| | Comorbidities | Subject | The patient whose comorbidities are recorded. | Optional | 0..1 | Reference: CancerPatient |
| | ECOG Performance Status | Category | A code that classifies the general type of observation being made. | Required | 1..1 | CodeableConcept |
| | ECOG Performance Status | Code | The name of the non-imaging or non-laboratory test performed on a patient. A LOINC **SHALL** be used if the concept is present in LOINC. | Required | 1..1 | CodeableConcept |
| | ECOG Performance Status | Subject | Patient whose performance status is recorded. | Required | 1..1 | Reference: CancerPatient |
| | ECOG Performance Status | Value | The information determined as a result of making the observation, if the information has a simple value. | Optional | 0..1 | integer |
| | ECOG Performance Status | Interpretation | A categorical assessment of an observation value. For example, high, low, normal. | Optional | 0,,* | CodeableConcept |

### 5.3.2.3 Disease

The disease group includes information specific to the tumor markers, the cancer diagnosis, the histological classification, grade, morphology, and behavior of tumors, the staging of cancer, as well as any cancer risk assessment metrics.

***Table 9 9**: The EUCAIM CDM: Disease group*

| Group | Entity | Data Element Name | Definition | EUCAIM Required | Occurrences Allowed | Data Type |
|---|---|---|---|---|---|---|
| Disease | Tumor Marker Test | Related Condition | Associates the tumor marker test with a condition, if one exists. Condition can be given by a reference or a code. In the case of a screening test such as prostate-specific antigen (PSA), there may be no existing condition to reference. | Optional | 0..* | Reference(PrimaryCancerCondition) |
| | Tumor Marker Test | Code | The tumor marker test that was performed. A LOINC concept shall be used if the concept is present. | Required | 1..1 | CodeableConcept |
| | Tumor Marker Test | Subject | Patient whose test result is recorded. | Required | 1..1 | Reference: CancerPatient |
| | Tumor Marker Test | Value As Concept | The Laboratory result value if it is a coded value. The value CodeableConcept.code shall be selected from SNOMED CT. | Required (conditional) | 1..1 | CodeableConcept, |
| | Tumor Marker Test | Value As Number | The Laboratory result value, if numeric. | Required (conditional) | 1..1 | Float |
| | Tumor Marker Test | Value Unit Concept | If a numeric value, valueQuantity.code **SHALL** be selected from [UCUM](http://unitsofmeasure.org). A FHIR [UCUM Codes value set](http://hl7.org/fhir/STU3/valueset-ucum-units.html) that defines all UCUM codes is in the FHIR specification. | Required (conditional) | 1..1 | CodeableConcept |
| | Tumor Marker Test | Performed | The elapsed time from the baseline (time 0). | Optional | 1..1 | Integer |
| | Tumor Marker Test | Performed Unit Concept | The unit concept of the time interval | Optional | 1..1 | Integer |
| | Primary Cancer Condition | Age of diagnosis/condition | The patient age on which the existence of the Condition was first asserted or acknowledged. | Required | 1..1 | Age |

| | | | | | | |
|---|---|---|---|---|---|---|
| | Primary Cancer Condition | Subject | Indicates the patient or group who the condition record is associated with. | Required | 1..1 | Reference: CancerPatient |
| | Primary Cancer Condition | Code | Identification of the condition, problem or diagnosis. | Required | 1..1 | CodeableConcept |
| | Primary Cancer Condition | Histology Morphology Behavior | A codeable concept describing the morphologic and behavioral characteristics of the cancer.(It takes values from: http://hl7.org/fhir/us/mcode/ValueSet/mcode-histology-morphology-behavior-vs) | Required | 1..1 | CodeableConcept |
| | Primary Cancer Condition | Body Site | The anatomical location where this condition manifests itself. | Required | 1..* | CodeableConcept |
| | Primary Cancer Condition | Body Site > Location Qualifier | General location qualifier (excluding laterality) for this bodySite | Optional | 0..* | CodeableConcept |
| | Primary Cancer Condition | Body Site > Laterality Qualifier | Laterality qualifier for this bodySite | Optional | 0..1 | CodeableConcept |
| | Primary Cancer Condition | Onset Age | Estimated or actual age the condition began, in the opinion of the clinician. | Optional | 0..1 | Age |
| | Primary Cancer Condition | Abatement Age | The date or estimated date that the condition resolved or went into remission. This is called "abatement" because of the many overloaded connotations associated with "remission" or "resolution" - Conditions are never really resolved, but they can abate. | Optional | 0..1 | Age |
| | Secondary Cancer Condition | Histology Morphology Behavior | Describes the morphologic and behavioral characteristics of the cancer. | Optional | 1..1 | CodeableConcept |
| | Secondary Cancer Condition | Related Primary Cancer Condition | A reference to the primary cancer condition that provides context for this resource. | Optional | 1..1 | Reference: Primary Cancer Condition |
| | Secondary Cancer Condition | Code | Identification of the condition, problem or diagnosis. | Required | 1..1 | CodeableConcept |

| | | | | | | |
|---|---|---|---|---|---|---|
| | Secondary Cancer Condition | Body Site | The anatomical location where this condition manifests itself. | Optional | 0..* | CodeableConcept |
| | Secondary Cancer Condition | Body Site > Location Qualifier | General location qualifier (excluding laterality) for this bodySite | Optional | 0..* | CodeableConcept |
| | Secondary Cancer Condition | Body Site > Laterality Qualifier | Laterality qualifier for this bodySite | Optional | 0..1 | CodeableConcept |
| | Secondary Cancer Condition | Subject | Indicates the patient or group who the condition record is associated with. | Required | 1..1 | Reference: CancerPatient |
| | Secondary Cancer Condition | Condition appearance | The number of time elapsed after the primary cancer condition on which the existence of this Condition was first asserted or acknowledged. | Required (conditional on Onset Age) | 1..1 | Integer |
| | Secondary Cancer Condition | Appearance Unit Concept | The unit of time for the time elapsed after the primary cancer condition | Required (conditional condition appearance) | 1..1 | Integer |
| | Secondary Cancer Condition | Onset Age | Estimated or actual age the condition began, in the opinion of the clinician. | Required (conditional on condition appearance) | 1..1 | Age |
| | Secondary Cancer Condition | Abatement Age | The date or estimated date that the condition resolved or went into remission. This is called "abatement" because of the many overloaded connotations associated with "remission" or "resolution" - Conditions are never really resolved, but they can abate. | Optional | 1..1 | Age |

| | | | | | | |
|---|---|---|---|---|---|---|
| | Cancer Stage | Code | The kind of stage reported, e.g., a pathologic TNM stage, a Lugano lymphoma stage, or a Rai stage for leukemia. This element identifies the type of value that is reported in Observation.value and is necessary for the correct interpretation of that value.<br><br>The distinction between Observation.code and Observation.method is important. Observation.code identifies the kind of stage being reported while Observation.method represents the staging system used to determine the code. Observation.code may imply the staging system. For example, the SNOMED CT 103420007 says the reported value is a modified Dukes stage, implying the Modified Dukes staging system (SNOMED CT 385359000) was used to determine the stage. When the staging system is implied by Observation.code, Observation.method is not required. However, when Observation.code does not imply a staging system (for example, if the code is SNOMED CT 385388004 Lymphoma stage), then the staging system must be specified in Observation.method.<br><br>The value (Observation.valueCodeableConcept) may also imply certain things about the kind of stage being reported. For example, the value cN0 implies the value is a clinical stage. However, even if the value is partly or wholly self-identifying, it is not a reliable indicator of the type of stage being reported or the method of staging. Therefore, Observation.code must in all cases be reported. | Required | 1..1 | CodeableConcept |
| | Cancer Stage | Method | The staging system or protocol used to determine the stage, stage group, or category of the cancer based on its extent. When the staging system is implied by Observation.code, Observation.method is not required. However, when Observation.code does not imply a staging system (for example, if the code is SNOMED CT 385388004 Lymphoma stage), then the staging system must be specified in Observation.method. | Optional | 0..1 | CodeableConcept |

| | | | | | | |
|---|---|---|---|---|---|---|
| | Cancer Stage | Value | The stage, stage group, category, or classification resulting from the staging evaluation. | Required | 1..1 | CodeableConcept |
| | Cancer Stage | Subject | The patient associated with staging assessment. | Required | 1..1 | Reference: CancerPatient |
| | Cancer Stage | Related Procedure | The procedure from which the cancer stage was determined. It can either be an imaging examination (MRI), biopsy, surgery. | Required | 1..* | Reference: Procedure |
| | Cancer Stage | Focus | Staging is associated with a particular cancer condition. Observation.focus is used to point back to that condition. | Optional | 0..* | Reference: CancerCondition |
| | Histologic Grade | Related Condition | Associates the histologic grade test with a condition, if one exists. Condition can be given by a reference. | Optional | 0..* | Reference: Condition |
| | Histologic Grade | Category | A code that classifies the general type of observation being made. | Required | 1..1 | CodeableConcept |
| | Histologic Grade | Subject | Patient whose test result is recorded. | Required | 1..1 | Reference: CancerPatient |
| | Histologic Grade | ValueAsConcept | The Laboratory result value. If a coded value, the value CodeableConcept.code should be selected from SNOMED CT, if the concept exists. | Required | 1..1 | CodeableConcept |
| | Histologic Grade | ValueAsNumber | The Laboratory result value. If a numeric value, value Quantity.code shall be selected from [UCUM](http://unitsofmeasure.org). | Required | 1..1 | Quantity |
| | Histologic Grade | Method | Indicates the mechanism used to perform the observation. | Optional | 0..1 | CodeableConcept |

### 5.3.2.4 Cancer Treatments

The cancer treatment group includes treatment techniques used to treat cancer patients, categorized as: medications, surgery, and radiotherapy.

***Table 1010:*** *The EUCAIM CDM: Cancer treatment group*

| Group | Entity | Data Element Name | Definition | EUCAIM Required | Occurrences Allowed | Data Type |
|---|---|---|---|---|---|---|
| Treatment | Cancer-Related | Code | The specific procedure that is performed. | Required | 1..1 | CodeableConcept |

| | | | | | | |
|---|---|---|---|---|---|---|
| | Surgical Procedure | | | | | |
| | Cancer-Related Surgical Procedure | Subject | The patient on whom the procedure was performed. | Required | 1..1 | Reference: Patient |
| | Cancer-Related Surgical Procedure | Performed | Period of time elapsed after baseline | Optional | 0..1 | Integer |
| | Cancer-Related Surgical Procedure | Performed Unit Concept | | | | |
| | Cancer-Related Surgical Procedure | Body Site | Detailed and structured anatomical location information. Multiple locations are allowed - e.g. multiple punch biopsies of a lesion. | Optional | 0..* | CodeableConcept |
| | Cancer-Related Surgical Procedure | Body Site > Location Qualifier | General location qualifier (excluding laterality) for this bodySite | Optional | 0..* | CodeableConcept |
| | Cancer-Related Surgical Procedure | Body Site > Laterality Qualifier | Laterality qualifier for this bodySite | Optional | 0..* | CodeableConcept |
| | Cancer-Related Surgical Procedure | Response | Response evaluation to an oncology treatment from RECIST terminology. | Optional | 0..1 | CodeableConcept |
| | Cancer-Related Medication Administration | Code | Code that identifies this medication | Required | 1..1 | CodeableConcept |
| | Cancer-Related Medication Administration | Subject | The patient receiving the medication. | Required | 1..1 | Reference: CancerPatient |
| | Cancer-Related Medication Administration | Effective | An interval of time during which the administration took place. | Optional | 0..1 | Period |

| | | | | | | |
|---|---|---|---|---|---|---|
| | Cancer-Related Medication Administration | Effective Unit Concept | An interval of time during which the administration took place. | Optional | 0..1 | Period |
| | Cancer-Related Medication Administration | Administered | The time elapsed | Optional | 0..1 | CodeableConcept |
| | Cancer-Related Medication Administration | Administered Unit Concept | Period of time elapsed unit concept. | Optional | 0..1 | CodeableConcept |
| | Cancer-Related Medication Administration | Response | Response evaluation to an oncology treatment from RECIST terminology. | Optional | 1..1 | CodeableConcept |
| | Radiotherapy Course Summary | Modality | Capturing a modality of external beam or brachytherapy radiation procedures. | Required | 1..1 | CodeableConcept |
| | Radiotherapy Course Summary | Technique | Capturing a technique of external beam or brachytherapy radiation procedures. | Optional | 0..* | CodeableConcept |
| | Radiotherapy Course Summary | Actual Number of Sessions | The number of sessions in a course of radiotherapy. | Optional | 0..1 | unsignedInt |
| | Radiotherapy Course Summary | Dose Delivered to Volume | Dose delivered to a given radiotherapy volume. | Optional | 0..* | Radiotherapy Dose Delivered To Volume Extension |
| | Radiotherapy Course Summary | Dose Delivered to Volume > Volume | A BodyStructure resource representing volume in the body where radiation was delivered, for example, Chest Wall Lymph Nodes. | Optional | 0..1 | Reference: RadiotherapyVolume |
| | Radiotherapy Course Summary | Dose Delivered to Volume > Total Dose Delivered | The total amount of physical radiation delivered to this volume within the scope of this dose delivery, i.e., dose delivered from the Procedure in which this extension is used. | Optional | 0..1 | Quantity |

| | | | | | | |
|---|---|---|---|---|---|---|
| | Radiotherapy Course Summary | Dose Delivered to Volume > Fractions Delivered | The number of fractions delivered to this volume. | Optional | 0..1 | unsignedInt |
| | Radiotherapy Course Summary | Code | The specific procedure that is performed. Use text if the exact nature of the procedure cannot be coded (e.g. "Laparoscopic Appendectomy"). | Required | 0..1 | CodeableConcept |
| | Radiotherapy Course Summary | Subject | The patient on whom the procedure was performed. | Required | 1..1 | Reference: CancerPatient |
| | Radiotherapy Course Summary | Performed | Period of time elapsed in months after primary cancer diagnosis | Optional | 0..1 | Period |
| | Radiotherapy Course Summary | Body Site | Coded body structure(s) treated in this course of radiotherapy. These codes represent general locations. For additional detail, refer to the BodyStructures references in the doseDeliveredToVolume extension. | Optional | 0..* | CodeableConcept |
| | Radiotherapy Course Summary | Response | Response evaluation to an oncology treatment from RECIST terminology. | Optional | 1..1 | CodeableConcept |
| | Radiotherapy Volume | Identifier | Unique identifier to reliably identify the same target volume in different requests and procedures, for example, the Conceptual Volume UID used in DICOM. | Optional | 0..* | Identifier |
| | Radiotherapy Volume | Morphology | The kind of structure being represented by the body structure at `BodyStructure.location`. This can define both normal and abnormal morphologies. | Optional | 0..1 | CodeableConcept |
| | Radiotherapy Volume | Location | The location and locationQualifier codes specify a TG263 body | Optional | 0..1 | CodeableConcept |

| | | | | | | |
|---|---|---|---|---|---|---|
| | | | structure comprising the irradiated volume. | | | |
| | Radiotherapy Volume | Location Qualifier | Qualifiers that together with the associated location code specify the TG263 body structure comprising the irradiated volume. | Optional | 0..* | CodeableConcept |
| | Radiotherapy Volume | Description | A text description of the radiotherapy volume, which SHOULD contain any additional information above and beyond the location and locationQualifier that describe the volume. | Optional | 0..* | string |
| | Radiotherapy Volume | Patient | The patient for which a radiotherapy procedure is planned or performed. | Required | 1..1 | Reference: CancerPatient |

### 5.3.2.5 Outcome

The outcome group involves the cancer disease status, e.g., whether it is stable, worsening (progressing), or improving (responding) based on different kinds of evidence (imaging data, tumor markers etc.).

***Table 11 11:*** *The EUCAIM CDM: Outcome group*

| Group | Entity | Data Element Name | Definition | EUCAIM Required | Occurrences Allowed | Data Type |
|---|---|---|---|---|---|---|
| Outcome | Tumor | Body Structure Identifier | Stable identifier(s) for this specific tumor. The identifiers MUST be unique within the context of the referenced `CancerPatient`. This id is used to track the tumor over time, through the related procedures. | Required | 1..* | Identifier |
| | Tumor | Related Condition | Associates this tumor with a cancer condition. This could be a causal association (e.g., this is believed to be the primary tumor causing the cancer) or a different type of | Optional | 0..1 | CodeableConcept or Reference: Condition |

| | | | | | | |
|---|---|---|---|---|---|---|
| | | | relationship (e.g., this tumor is a metastasis) | | | |
| | Tumor | Related Procedure | Associates this tumor with a related procedure. For example it associates a tumor with an MR examination procedure. | Required (conditional on Condition) | 1..1 | Reference: Procedure |
| | Tumor | Risk Assessment | Associates this tumor with a risk assessment report. In case the tumor is identified in an imaging report, this could be used for storing RADS related information. | Optional | 0..1 | Reference: RiskAssessment |
| | Tumor | Morphology | The kind of structure being represented by the body structure at `BodyStructure.location`. This can define both normal and abnormal morphologies. | Optional | 0..* | CodeableConcept |
| | Tumor | Location | The anatomical location or region of the specimen, lesion, or body structure. | Required | 1..* | CodeableConcept |
| | Tumor | Location Qualifier | Qualifier to refine the anatomical location. These include qualifiers for laterality, relative location, directionality, number, and plane. | Optional | 0..* | CodeableConcept |
| | Tumor | Patient | The patient associated with this tumor. | Required | 1..1 | Reference: CancerPatient |
| | Tumor Size | Code | Describes what was observed. Sometimes this is called the observation "name". | Required | 1..1 | CodeableConcept |
| | Tumor Size | Subject | The patient whose tumor was measured. SHALL be a `Patient` resource conforming to `CancerPatient`. | Required | 1..1 | Reference: CancerPatient |

| | | | | | | |
|---|---|---|---|---|---|---|
| | Tumor Size | Focus | Reference to a BodyStructure resource conforming to Tumor. | Optional | 0..1 | Reference: Tumor |
| | Tumor Size | Volume | The volume of the lesion | Optional | 0..1 | Quantity |
| | Tumor Size | Method | Method for measuring the size or the volume of the tumor | Optional | 0..1 | CodeableConcept |
| | Tumor Size | Tumor Longest Dimension | The longest tumor dimension in cm or mm. | Required | 1..1 | Quantity |
| | Tumor Size | Tumor Longest Dimension > Code | Describes what was observed. Sometimes this is called the observation "code". | Optional | 0..1 | CodeableConcept |
| | Tumor Size | Tumor Longest Dimension > Value | The information determined as a result of making the observation, if the information has a simple value. | Optional | 0..1 | Quantity |
| | Tumor Size | Tumor Other Dimension | The second or third tumor dimension in cm or mm. | Optional | 0..2 | Quantity |
| | Tumor Size | Tumor Other Dimension > Code | Describes what was observed. Sometimes this is called the observation "code". | Required | 1..1 | CodeableConcept |
| | Tumor Size | Tumor Other Dimension > Value | The information determined as a result of making the observation, if the information has a simple value. | Optional | 0..1 | Quantity |
| | Cancer Disease Status | Evidence Type | Categorization of the kind of evidence contributing to a clinical judgment on cancer disease progression. | Optional | 0..* | CodeableConcept |
| | Cancer Disease Status | Code | Describes what was observed. Sometimes this is called the observation "name". | Required | 1..* | CodeableConcept |
| | Cancer Disease Status | Subject | Patient whose disease status is recorded. | Required | 1..* | Reference: CancerPatient |

#### 5.3.2.6 Imaging

As the focus of the EUCAIM project is the federation of cancer *imaging* datasets, it is imperative that important imaging metadata are standardized to facilitate the unambiguous representation of the stored information and support federated queries. Although the DICOM standard for collecting, storing, and transferring medical imaging data can be used to access critical image acquisition parameters (such as acquisition method, field of view, and slice thickness) for cohort discovery and quality checking, it lacks essential information needed to query efficiently relevant images. This is due to the fact that certain information is not standardized in the DICOM metadata. For instance, the classification of a series as a T2-weighted axial series is typically recorded in the "Series Description" (0008,103E) DICOM tag, which is free text and highly variable across clinical institutions.

The EUCAIM Imaging component corresponds to important metadata extracted from the DICOM header-related tags, which get standardized to allow for efficient querying and analysis. Although mCODE does not explicitly represent imaging-related procedures and their corresponding metadata, the EUCAIM CDM builds upon the FHIR Resources ImagingStudy and ImagingSeries, the MI-CDM extension of the OMOP-CDM[21] - the ProCAncer-I imaging extension, and the OSIRIS imaging component. The following section presents a first version of the imaging related entities and their associated information:

- **Image Study:** Representation of the content produced in a DICOM imaging study. A study comprises a set of series, each of which includes a set of Service-Object Pair Instances (SOP Instances - images or other data) acquired or produced in a common context. A series is of only one modality (e.g. X-ray, CT, MR, ultrasound), but a study may have multiple series of different modalities.
- **Image Series:** Representation of the content produced in a DICOM imaging series, by representing important metadata across all image modalities. Some of the most important parameters, include the modality, the body region, the patient position, the patient orientation, the laterality etc.
- **Image Modality:** Representation of the distinct modality-related acquisition parameters, in order to enable tailored queries for each modality (e.g. echo time, magnetic field strength for MR modality etc.). It is important to note that the modeling choice of the image modality entity is to allow for storing any modality related acquisition parameter, without the need to change/add new attributes in the model. However, some important acquisition parameters of the two most important modalities (MR, CT) as these are defined in OSIRIS, but also included in the ProCAncer-I collected MR imaging metadata are:
    - MR image: sequence name, magnetic field strength, MR acquisition type, repetition time, echo time, imaging frequency, flip angle, inversion time, receive coil name, diffusion b-value (for DWI).
    - CT image: kVp, xRay tube current, exposure time, spiral pitch factor, filter type, convolution kernel.
- **Image Annotation:** Representation of the most important metadata concerning imaging annotation processes.

[21] Varvara Kalokyri et al., MI-Common Data Model: Extending Observational Medical Outcomes Partnership-Common Data Model (OMOP-CDM) for Registering Medical Imaging Metadata and Subsequent Curation Processes. *JCO Clin Cancer Inform* 7, e2300101(2023). DOI:10.1200/CCI.23.00101

***Table 1212**: The EUCAIM CDM: Imaging group*

| Group | Entity | Data Element Name | Definition | EUCAIM Required? | Occurrences Allowed | Data Type | Mapping: DICOM Tag Mapping |
|---|---|---|---|---|---|---|---|
| Imaging | Image Study | Identifier | The logical id of the resource, as used in the URL for the resource. Once assigned, this value never changes. | Required | 1...1 | id | |
| | Image Study | Subject | The patient of the imaging study. | Required | 1...1 | Reference(Patient) | (0010/*) |
| | Image Study | Study UID | Identifiers for the ImagingStudy, i.e. as DICOM Study Instance UID. | Required | 1…1 | String | StudyInstanceUID (0020,000D) \| study ID (0020,0010) |
| | Image Study | Acquisition Date | The date the study acquisition was obtained. | Optional | 0...1 | dateTime | (0008,0020) +(0008,0030) |
| | Image Study | Part Of | A larger event of which this particular ImagingStudy is a component or step. For example, an ImagingStudy as part of a procedure. | Optional | 0...* | Reference(Procedure) | |
| | Image Study | Access URI | The accessURI of the study, either on a DICOM web server (e.g. via the WADO-RS DICOMweb REST-API) or on a local machine via the path name to the folder containing the study. | Optional | 0...* | String | |
| | Image Study | Number Of Series | Number of Series in the Study. This value given may be larger than the number of series elements this Resource contains due to resource availability, security, or other factors. This element should be present if any series elements are present. | Optional | 0...1 | unsignedInt | (0020,1206) |
| | Image Study | Number Of Instances | Number of SOP Instances in Study. This value given may be larger than the number of instance | Optional | 0...1 | unsignedInt | (0020,1208) |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| | | | elements this resource contains due to resource availability, security, or other factors. This element should be present if any instance elements are present. | | | | |
| | Image Study | Manufacturer Name | Name of the manufacturing company of the imaging equipment. | Required | 1..1 | CodeableConcept | (0008,0070) |
| | Image Study | Manufacturer Model Name | Name of the model of the manufacturing company of the imaging equipment. | Optional | 0..1 | String | (0008,1090) |
| | Image Series | Study identifier | The study in which the series belongs to. | Required | 1..1 | Reference(Image Study) | |
| | Image Series | Identifier | Unique id for the element within a resource (for internal references). This may be any string value that does not contain spaces. | Required | 1...1 | string | |
| | Image Series | Series UID | The DICOM Series Instance UID for the series. | Required | 1...1 | String | (0020,000E) |
| | Image Series | Number | The numeric identifier of this series in the study. | Optional | 0...1 | unsignedInt | (0020,0011) |
| | Image Series | Modality | The distinct modality for this series. This may include both acquisition and non-acquisition modalities. | Required | 1...1 | CodeableConcept | (0008,0060) |
| | Image Series | Description | A description of the series. | Optional | 0...1 | string | (0008,103E) |
| | Image Series | Number Of Instances | Number of SOP Instances in the Study. The value given may be larger than the number of instance elements this resource contains due to resource availability, security, or other factors. This element should be present if any instance elements are present. | Optional | 0...1 | unsignedInt | (0020,1209) |
| | Image Series | Access URI | The accessURI of the series, either on a DICOM web server (e.g. via the WADO-RS DICOMweb REST-API) or on a local machine via the path name to the folder containing the series instances. | Optional | 0...* | String | |
| | Image Series | Body Site | The anatomic structures examined. See DICOM Part 16 Annex L (http://dicom.nema.org/medical/dicom/current/output/chtml/part16/chapter_L.html) for DICOM to | Required | 1...1 | CodeableConcept | (0018,0015) |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| | | | SNOMED-CT mappings. The bodySite may indicate the laterality of body part imaged; if so, it shall be consistent with any content of ImageSeries.laterality. | | | | |
| | Image Series | Laterality | The laterality of the (possibly paired) anatomic structures examined. E.g., the left knee, both lungs, or unpaired abdomen. If present, shall be consistent with any laterality information indicated in ImageSeries.bodySite. | Optional | 0...1 | CodeableConcept | (0020,0060) |
| | Image Series | Specimen | The specimen imaged, e.g., for whole slide imaging of a biopsy. | Optional | 0...* | Reference(Specimen) | (0040,0551) + (0040,0562) |
| | Image Series | Acquisition Date | The date the series acquisition was obtained. | Optional | 0...1 | date | (0008,0021) + (0008,0031) |
| | Image Modality | Identifier | Unique id for the element within a resource (for internal references). This may be any string value that does not contain spaces. | Required | 1..1 | string | |
| | Image Modality | Series identifier | Reference to the series id for which important acquisition parameters are being stored. | Required | 1..1 | Reference(Image Series) | |
| | Image Modality | Acquisition Parameter Code | The concept code of the acquisition parameters relevant to the modality of the series. (e.g. slice thickness for MR modality) | Required | 1..1 | CodeableConcept | |
| | Image Modality | Acquisition Parameter Value As Concept | The concept code of the value of the acquisition parameter (e.g. "Spin echo" value of the "MR echo type" concept) | Optional(conditional on ParamCode) | 0..1 | CodeableConcept | |
| | Image Modality | Acquisition Parameter Value As Number | The numerical value of the modality acquisition concept (e.g. 0 for the gantry tilt angle in case of a CT) | Optional (conditional on ParamCode) | 0..1 | Float | |
| | Image Modality | Acquisition Parameter Value Unit Concept | If a numeric value, the units of measure concept code should be used. (http://unitsofmeasure.org). | Required (conditional on Acquisition Parameter Value as Number) | 0..1 | CodeableConcept | |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| | Image Annotation | Id | A unique identifier for the annotation. | Required | 1..1 | string | |
| | Image Annotation | series.id | The unique identifier for the imaging series being annotated. | Required | 1..1 | Reference(Image Series) | |
| | Image Annotation | study.id | The unique identifier for the imaging study that contains the series that is being annotated. | Required | 1..1 | Reference(Image Study) | |
| | Image Annotation | derived.series.id | The unique identifier for the annotated derived imaging series. | Required | 1..1 | Reference(Image Series) | |
| | Image Annotation | performed | The date and time the annotation was made. | Optional | 0..1 | datetime | |
| | Image Annotation | status | The current status of the annotation, such as final or pending. | Optional | 0..1 | CodeableConcept | |
| | Image Annotation | anatomic location | The anatomic location being annotated (e.g. peripheral zone of the prostate gland) | Optional | 0..1 | CodeableConcept | |
| | Image Annotation | observation | The imaging observation that is reported. (e.g. lesion of the prostate) | Optional | 0..1 | CodeableConcept | |
| | Image Annotation | type | The annotation type (e.g. bounding box, contouring, etc..) | Optional | 0..1 | CodeableConcept | |
| | Image Annotation | method | The method used to create the annotation, such as manual or automatic, or semiautomatic. | Optional | 0..1 | CodeableConcept | |

# 6. Integration of CDM and Hyper-Ontology

The hyper-ontology is developed using a hybrid approach composed of top-down and bottom-up strategies. While the bottom-up considers the clinical and imaging knowledge provided by the AI4HI projects, the top-down grounds the hyper-ontology in the mCODE conceptual model. Therefore, the mCODE profiles and data elements are analyzed and semantically represented in the ontological model using a high-level conceptual modeling language, OntoUML. By applying this strategy, the hyper-ontology ensures seamless integration with the EUCAM CDM, which is based on the mCODE specifications. The mCODE specifications are syntactic representations of entities, their key elements, and the associated value sets. Thus, there is a need for an ontological analysis that helps to unpack the ontological content of the oncology domain based on mCODE generic specifications.

In the following, we give an example of an ontological analysis and formalization of the *Primary Cancer Condition* profile[22] and the associated elements. Table 13 presents basic data elements required for describing a primary cancer condition: *Code*, *HistologyMorphologyBehavior*, *BodySite*, and *Stage*. The value sets of these elements are specified in mCODE, such as *Malignant tumor of prostate* (ICD10:C61) and *Malignant Neoplasm* (SNOMED:1240414004) value sets for the *Code* and *HistologyMorphologyBehavior* data elements.

Section 4.4.3 (Core Layer) outlines the ontological analysis of Primary Cancer Condition and the associated semantic relations (see Figure 6). For instance, the data element *HistologyMorphologyBehavior* is explicitly and semantically represented in the hyper-ontology using the semantic property/relation “Has associated morphology” and *BodySite* is represented using “Has finding site” association, which links the cancer condition to the morphology/histology (e.g., Malignant Neoplasm (SNOMED:1240414004)) and affected body structure (e.g., Prostate (SNOMED:41216001)), respectively. The formalization of this profile using OWL is illustrated in Figure 24 .

***Table 13 13**: Data elements required to describe a primary cancer condition in mCODE*

| Data Element | **Example of Value Set (Standard concepts)** |
|---|---|
| Code | **Malignant tumor of prostate (ICD10:C61)** |
| HistologyMorphologyBehavior | **Malignant Neoplasm (SNOMED:1240414004)** |
| BodySite | **Prostate (SNOMED:41216001)** |
| Stage | **TNM staging classifications (SNOMED:258234001)** |

[22] https://build.fhir.org/ig/HL7/fhir-mCODE-ig/StructureDefinition-mcode-primary-cancer-condition.html

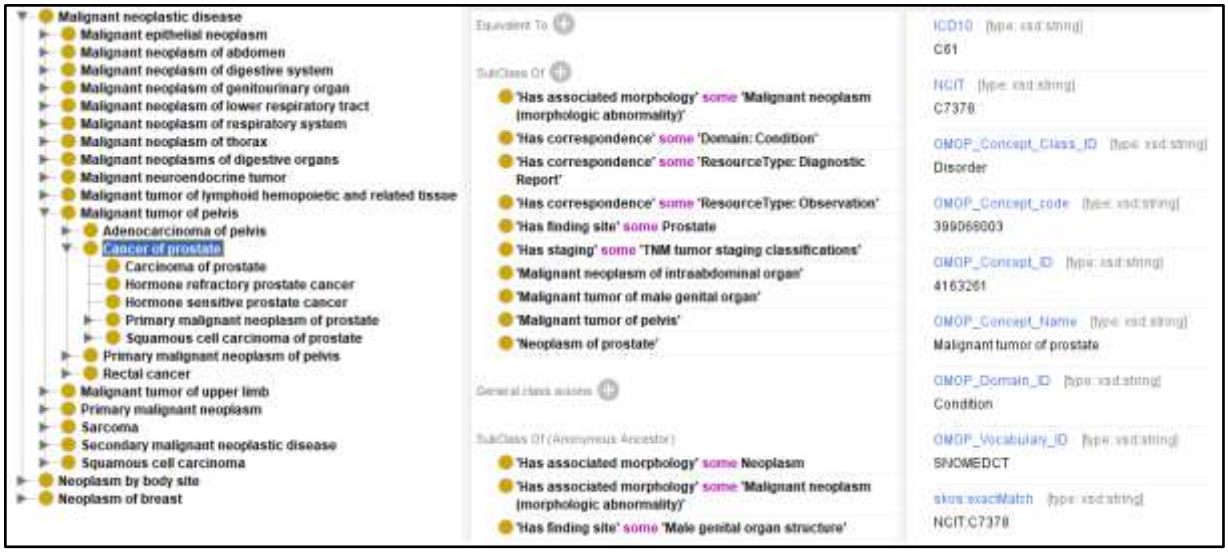


***Figure 24**. An excerpt of the hyper-ontology around "Cancer of prostate" represented in Protege*

As part of our upcoming activities, we will explicitly state each attribute and its associated value set as defined in the EUCAIM CDM, ensuring precise terminology binding using the semantics and terminologies from the EUCAIM hyper-ontology. For example, while SNOMED is the standard terminology for coding conditions in OMOP, the oncology domain uses different reference terminologies: ICD-10 for clinical diagnosis of cancers and ICD-O for histological diagnosis, with ICD-O-3 being the global standard for cancer registries. Given that various terminologies have been used across underlying repositories to represent conditions, the integration of the EUCAIM Hyper-ontology with the CDM will specify the terminologies to be used for specific properties. Additionally, there are multiple ways to represent properties such as tumor marker test results (as discussed in section 4.6), either as a finding (e.g. triple negative) or as an observation ( with an attribute-value ) (e.g. ER negative, PR negative, HER2 negative). The hyper-ontology will clarify the representation and usage of these properties when populating the CDM. These topics will be discussed with experts in the WP5 CDM and hyper-ontology working group and incorporated into the next version of the EUCAIM CDM and Hyper-ontology.

# 7. Demonstration scenarios

To evaluate the efficacy of the EUCAIM CDM and Hyper-ontology, four proof-of-concept scenarios are provided for mapping and structuring clinical and imaging metadata related to prostate and breast cancer information. This information is provided by four AI4HI projects: ProCAncer-I and INCISIVE for the prostate cancer scenario, which adopt the OMOP-CDM and FHIR standards respectively, as well as the CHAIMELEON and EuCanImage projects for the breast cancer related scenarios, that they adopt an OMOP-like CDM and FHIR standards respectively. Two main demonstration strategies are introduced per cancer type: 1) *semantic-based* and 2) *syntactic-based*. The semantic-based strategy aims to demonstrate hyper-ontology's completeness in representing knowledge from real-world scenarios by populating the ontology semantic content (concepts and relations) using individuals extracted from the provided use cases. For the syntactic-based, the objective is to ascertain the usability of the hyper-ontology in instantiating the EUCAIM-CDM.

## 7.1 Prostate Cancer Use Cases

### ProCAncer-I Scenario

The following case is a real case scenario for a patient registered into the ProCAncer-I platform:

#### Patient's journey

The patient is a **59-year-old male**, with a **PSA** value equal to **7.16 (ng/mL)** and **free PSA** equal to **5.04 (ng/mL)**. The patient had a **positive digital rectal examination**, and he was sent by the urologist to perform a **multiparametric MRI**. The **MRI** that was performed **22 days after the PSA lab test** was deemed **positive**, and revealed a **PI-RADS 5** lesion, with a **max diameter** of **10mm**, in the **right peripheral zone basal posterolateral**, with a **clinical stage** of **cT2b**, **cN0**. The patient underwent a **fusion biopsy**, which revealed a **cT2** cancer stage. Because of the **positive** findings the patient was referred to perform a **prostatectomy**. The results of the prostatectomy also confirmed the positive findings, revealing a **4+3 Gleason score lesion** of an **overall volume** of **0.7cc** of **17mm maximum diameter**, with stage **pT3b**, **pN0**, and **intraductal carcinoma**. **After 6 months**, **MRI** and **PET** examinations were performed, where a **liver metastasis** was identified with reported stage **cNX**, **cM1c**.

## Hyper-Ontology Population

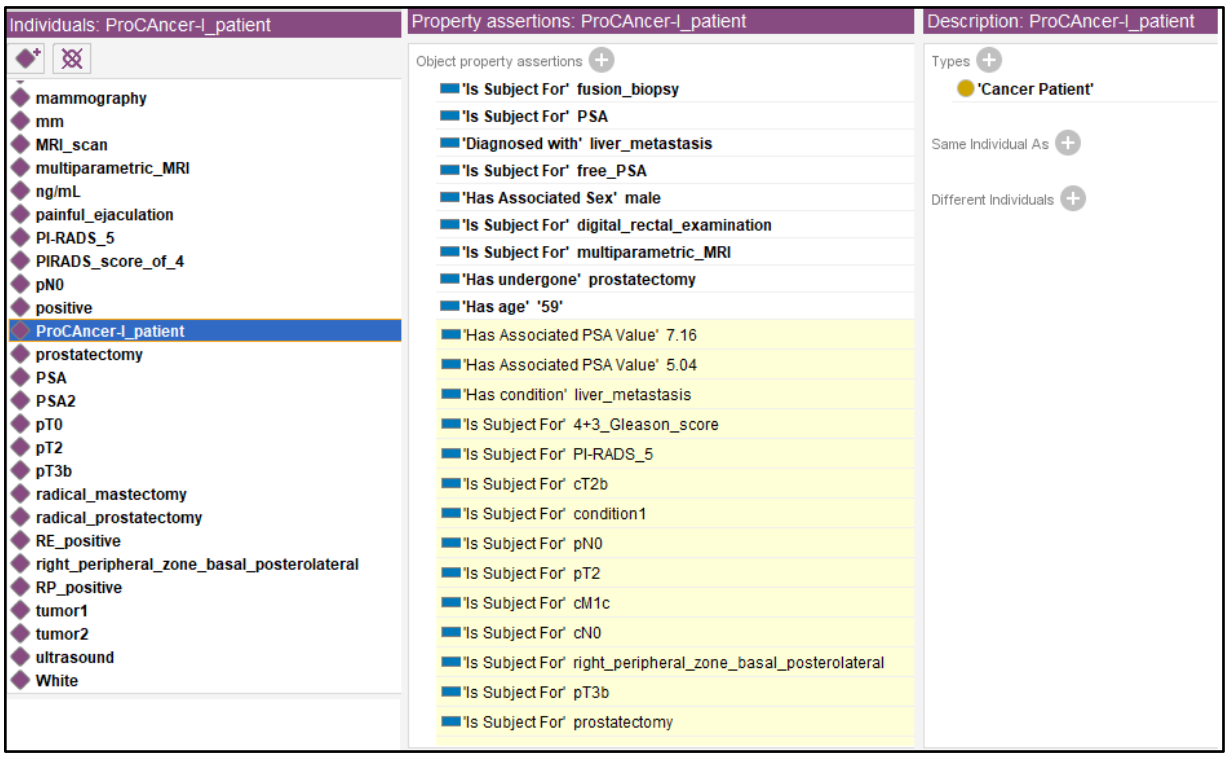


***Figure 25**. A semantic representation and inference of the ProCAncer-I prostate cancer use case (Protege)*

The real-world scenario provided by ProCAncer-I around prostate cancer is considered to (manually) extract a set of instances (individuals) and associate them with the hyper-ontology classes/concepts. Semantic relationships are maintained among the individuals considering the use case presented scenario and the object properties specified in the hyper-ontology. Figure 25 depicts the population results. In this scenario, a diagnosis has been performed on a patient, including imaging (e.g., *multiparametric MRI* and *fusion biopsy*) and clinical/surgical procedures (e.g., *digital rectal examination* and *prostatectomy*). Different imaging and pathologic results have been interpreted based on the performed procedures, such as imaging assessment observations (e.g., *PI-RADS 5*), histological grading (e.g., *4+3 Gleason score*), clinical staging (e.g., *cT2b*, *cN0*), and pathologic staging (e.g., *pT2*, *pN0*). The tumor's maximum dimension and volume have been considered throughout the diagnosis. Also, the PSA labLab test was performed on the patient.

By assigning the various information to their semantic reference, the hyper-ontology is populated with real-world details with which the logic reasoner (Pellet in this example) has deduced the complete diagnosis, including the imaging and clinical results.

## Model Instantiation

The EUCAIM CDM instantiation of the clinical and imaging related information is provided below along with a graphical representation of the events and timepoints in the patient's journey.

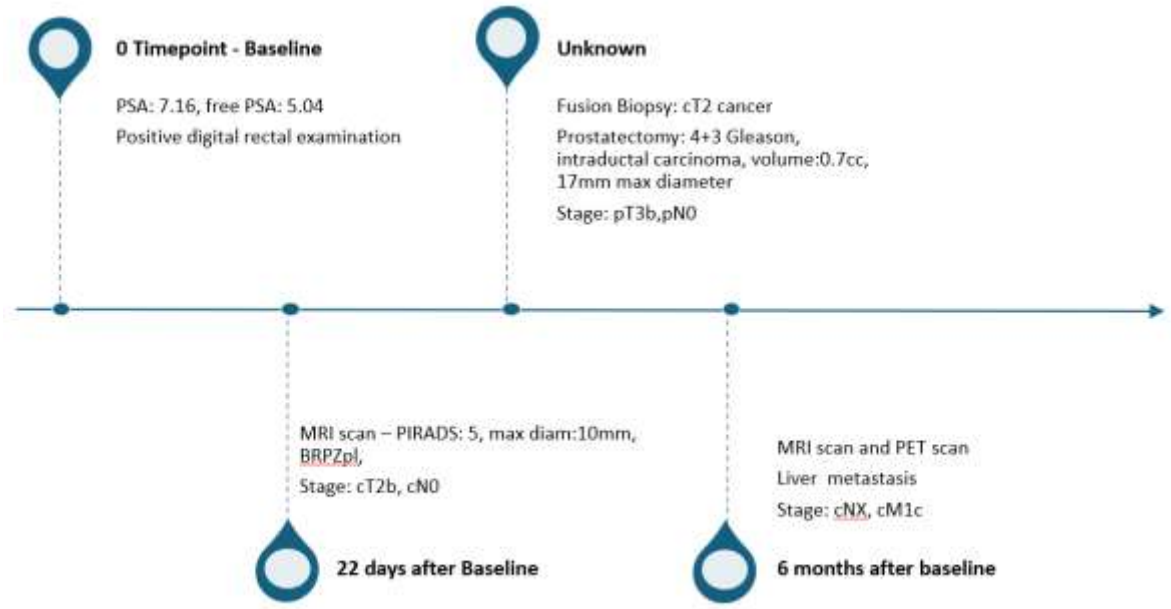


***Figure 26**: The ProCAncer-I prostate cancer patient journey*.

**Cancer Patient**

| | |
|---|---|
| Identifier | ECI-123456 |
| Managing Organization | 1 |
| BirthSex | COM1001366 (**Male**(finding)) |
| Race | - |

**Tumor Marker Test**

| | | |
|---|---|---|
| Identifier | 12 | 23 |
| Related Condition | 1111 | 1111 |
| Subject | ECI-123456 | ECI-123456 |
| Category | laboratory | laboratory |
| Code | CLIN1033410 (**Prostate specific Ag [Mass/volume] in Serum or Plasma**) | CLIN1022987 (**Free prostate specific antigen level**) |
| Value As Number | **7.16** | **5.04** |
| Value Unit Concept | COM1000156 (ng/mL) | COM1000156 (ng/mL) |
| Performed | 0 | 0 |
| Performed Unit Concept | - | - |

**Primary Cancer Condition**

| | |
|---|---|
| Identifier | 1111 |
| Age of diagnosis | **59** |
| Subject | ECI-123456 |
| Code | CLIN1008742 (**Primary malignant neoplasm of prostate**) |
| Histology Morphology Behavior | CLIN1023932 (**Intraductal carcinoma, noninfiltrating, NOS, of prostate gland**) |
| Body Site | BP1000080 (**Right posterolateral basal peripheral zone of prostate**) |

**Secondary Cancer Condition**

| | |
|---|---|
| Related Primary Cancer Condition | 1111 |
| Subject | ECI-123456 |
| Code | CLIN1000192 (**Secondary malignant neoplasm of liver**) |
| Body Site | BP1000105 (**Liver**) |
| Condition appearance | 6 |
| Appearance Unit Concept | COM1000154 (month) |

**Procedure**

| | | | | | |
|---|---|---|---|---|---|
| Identifier | 1000 | 2000 | 3000 | 77 | 4000 | 5000 |
| Code | IMG1016133 (**Multiparametric MRI of prostate**) | CLIN1028950 (**Digital examination of rectum**) | IMG1016128 (**MRI-US fusion guided prostate biopsy**) | CLIN1000248 (**Prostatectomy**) | IMG1016133 (**Multiparametric MRI of prostate**) | IMG1000037 (**Positron emission tomography**) |
| Category | IMG1005453 (Imaging (Procedure)) | CLIN1004055 (Evaluation procedure) | CLIN1001712 (Biopsy) | CLIN1000228 (Procedure on organ) | IMG1005453 (Imaging (Procedure)) | IMG1005453 (Imaging (Procedure)) |
| Subject | ECI-123456 | ECI-123456 | ECI-123456 | ECI-123456 | ECI-123456 | ECI-123456 |
| Performed | 22 | 0 | - | - | 6 | 6 |
| Performed Unit Concept | COM1000153 (day) | | | | COM1000154 (month) | COM1000154 (month) |
| Diagnostic Value | COM1001310 (**Positive**) | COM1001310 (**Positive**) | COM1001310 (**Positive**) | COM1001310 (**Positive**) | | |

**Cancer Stage**

| | | | | | |
|---|---|---|---|---|---|
| Identifier | 9876 | 8765 | 7654 | 6543 | 6543 |
| Subject | ECI-123456 | ECI-123456 | ECI-123456 | ECI-123456 | ECI-123456 |
| Code | CLIN1033379 (**cT category**) | CLIN1033358 (**cN category**) | CLIN1033379 (**cT category**) | CLIN1033391 (**pT category**) | CLIN1033368 (**pN category**) |
| Method | CLIN1000417 (AJCC/UICC 7th edition) | CLIN1000417 (AJCC/UICC 7th edition) | CLIN1000417 (AJCC/UICC 7th edition) | CLIN1000417 (AJCC/UICC 7th edition) | CLIN1000417 (AJCC/UICC 7th edition) |
| Value | COM1000390 (**cT2b**) | COM1000285 (**cN0**) | COM1000351 (**cT2**) | COM1001000 (**pT3b**) | COM1000645 (**pN0**) |
| Procedure | 1000 | 1000 | 3000 | 77 | 77 |
| Rel. Condition | 1111 | 1111 | 1111 | 1111 | 1111 |

**Histologic Grade**

| | | |
|---|---|---|
| Related Condition | 1111 | 1111 |
| Subject | **ECI-123456** | **ECI-123456** |
| Category | laboratory | laboratory |
| Value | CLIN1022393(**Gleason Primary Pattern Grade 4**) | CLIN1022267 (**Gleason Secondary Pattern Grade 3**) |
| Method | CLIN1037295 (Gleason grading system for prostatic cancer) | CLIN1037295 (Gleason grading system for prostatic cancer) |
| Related Procedure | 77 | 77 |

**Cancer Related Surgical Procedure**

| | |
|---|---|
| Identifier | 77 |
| Subject | ECI-123456 |
| Code | CLIN1000248 (**Prostatectomy**) |

**Tumor**

| | | |
|---|---|---|
| Identifier | 7 | 8 |
| Body Structure Identifier | 9999 | 9999 |
| Subject | ECI-123456 | ECI-123456 |
| Related Condition | 1111 | 1111 |
| Related Procedure | 1000 | 77 |
| Risk Assessment Method | IMG1005481 (**PI-RADS lesion assessment**) | |
| Risk Assessment Value | IMG1005475 (**PI-RADS 5 - Very high (lesion)**) | |
| Morphology | - | CLIN1023932 (**Intraductal carcinoma, noninfiltrating, NOS, of prostate gland**) |
| Location | BP1000080 (**Right posterolateral basal peripheral zone of prostate**) | |

**Tumor Size**

| | | |
|---|---|---|
| Identifier | 345678 | 456789 |
| Body Structure Identifier | 9999 | 9999 |
| Subject | ECI-123456 | ECI-123456 |
| Focus | 7 | 8 |
| Volume | | **0.7** |
| Volume Unit Code | | COM1000148(cc) |
| Tumor Longest Dimension | **10** | **17** |
| Tumor Longest Dimension Unit Code | COM1000152 (mm) | COM1000152 (mm) |

***Figure 27**: The EUCAIM CDM instantiation with the ProCAncer-I prostate cancer clinical information.*

**Procedure**

| | | |
|---|---|---|
| Identifier | 1000 | 4000 |
| Subject | **ECI-123456** | **ECI-123456** |
| Code | IMG1016133 (Multiparametric MRI of prostate) | IMG1016133 (Multiparametric MRI of prostate) |
| Category | IMG1005453 (Imaging (Procedure)) | IMG1005453 (Imaging (Procedure)) |
| Elapsed Days | 22 | 180 |

**Image Study**

| | | |
|---|---|---|
| Identifier | **1001** | 4001 |
| Subject | **ECI-123456** | **ECI-123456** |
| Study UID | 1.3.6.1.4.1.58108.1.3...12345678... | 1.3.6.1.4.1.58108.1.3...9876543... |
| Manufacturer Name | IMG1000044 (Siemens) | IMG1000047 (General Electric) |
| Acquisition Date | - | - |
| Part Of (FK) | 1000 | 4000 |
| Access URI | https://procancer-i.eu/studies/1.3.6.1.4.1.... | https://procancer-i.eu/studies/1.3.6.1.4.1.... |
| Number Of Series | 3 | 3 |
| Number Of Instances | 96 | 84 |

**Image Series**

| | | | |
|---|---|---|---|
| Identifier | **10011** | 10012 | 10012 |
| Study identifier | **1001** | **1001** | **1001** |
| Series UID | 1.3.6.1.4.1.58108.1.2...12345678... | 1.3.6.1.4.1.58108.1.2...9876543... | 1.3.6.1.4.1.58108.1.2...4578433... |
| Number | 6 | 8 | 850 |
| Modality | IMG1000022 (MRI) | IMG1000022 (MRI) | IMG1000022 (MRI) |
| Description | Ax T2 | DWI b1000 | ADC (10^-6 mm²/s) |
| Body Site | BP1000021 (Prostate) | BP1000021 (Prostate) | BP1000021 (Prostate) |
| Laterality | - | - | - |
| Access URI | https://procancer-i.eu/studies/..../series/...1.3.6..... | https://procancer-i.eu/studies/..../series/...1.3.6..... | https://procancer-i.eu/studies/..../series/...1.3.6..... |
| Acquisition Date | - | - | - |
| Number Of Instances | 24 | 48 | 24 |

**Image Modality**

| | | | | | |
|---|---|---|---|---|---|
| Identifier | 100111 | 100112 | 100113 | 100113 | ... |
| Series identifier | **10011** | **10011** | **10011** | **10011** | **...** |
| Parameter Code | IMG1016306 (slice thickness) | IMG1016649 (tissue contrast) | IMG1016641 (echo time) | IMG1016648 (MR imaging coil) | ... |
| Parameter Value As Concept | | IMG1016647 (T2 weighted) | | IMG1016643 (endorectal coil) | ... |
| Parameter Value As Number | 3.0 | | 109.48 | | ... |
| Parameter Value Unit Concept | COM1000152 (milliliter) | | COM1001955 (millisecond) | | ... |

***Figure 28**: The EUCAIM CDM instantiation with the ProCAncer-I prostate cancer imaging information.*

## INCISIVE Scenario

The following case is a real case scenario for a patient registered into the INCISIVE platform:

### Patient's journey

The patient is a **74-year-old white male** with a history of **Dyslipidemia**, who initially presented with **painful ejaculation**. An **MRI scan** revealed a tumor with a **PIRADS score of 4**. His **PSA** level was measured at **5.6**, and staging was determined as **T1, N0, M0**. **One month later**, a **targeted biopsy** was performed, resulting in a **Gleason score of 6** and **ISUP grade 5**. **Two months post-diagnosis**, the patient underwent a **radical prostatectomy**. Follow-up screenings began **one month after surgery**, showing a **complete response** with a **PSA level of 0.04**. Subsequent PSA tests were conducted **2, 5, 9, and 12 months after surgery**, with values of **0.07, 0.04, 0.04, and 0.04** respectively.

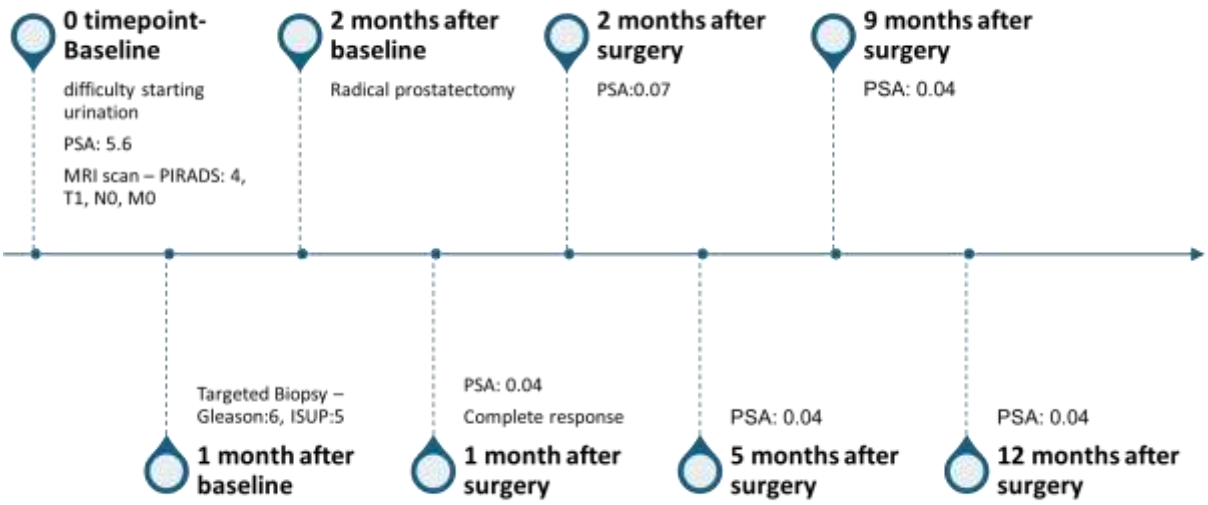


***Figure 29**: The INCISIVE prostate cancer patient journey.*

## Hyper-Ontology Population

Similarly to the ProCAncer-I scenario, we assigned the individuals extracted fro the INCISIVE use case to the hyper-ontology concepts and relations. Figure 30 depicts the population results. In this scenario, a diagnosis has been performed on a patient who is initially suffering from *Dyslipidemia*, including imaging (e.g., *MRI scan* and *biopsy*) and clinical/surgical procedures (e.g., *radical prostatectomy*). Different results have been interpreted based on the performed procedures, such as imaging assessment observations (e.g., *PI-RADS 4*), histological grading (e.g., *Gleason score 6, ISUP grade 5*), and cancer staging (e.g., *T1*, *N0, M0*). PSA lab tests were also performed throughout the diagnostic process. By running the Pellet reasoner, the complete cancer patient diagnosis, including the imaging and clinical interpretation results, is deduced (see Figure 30 ).

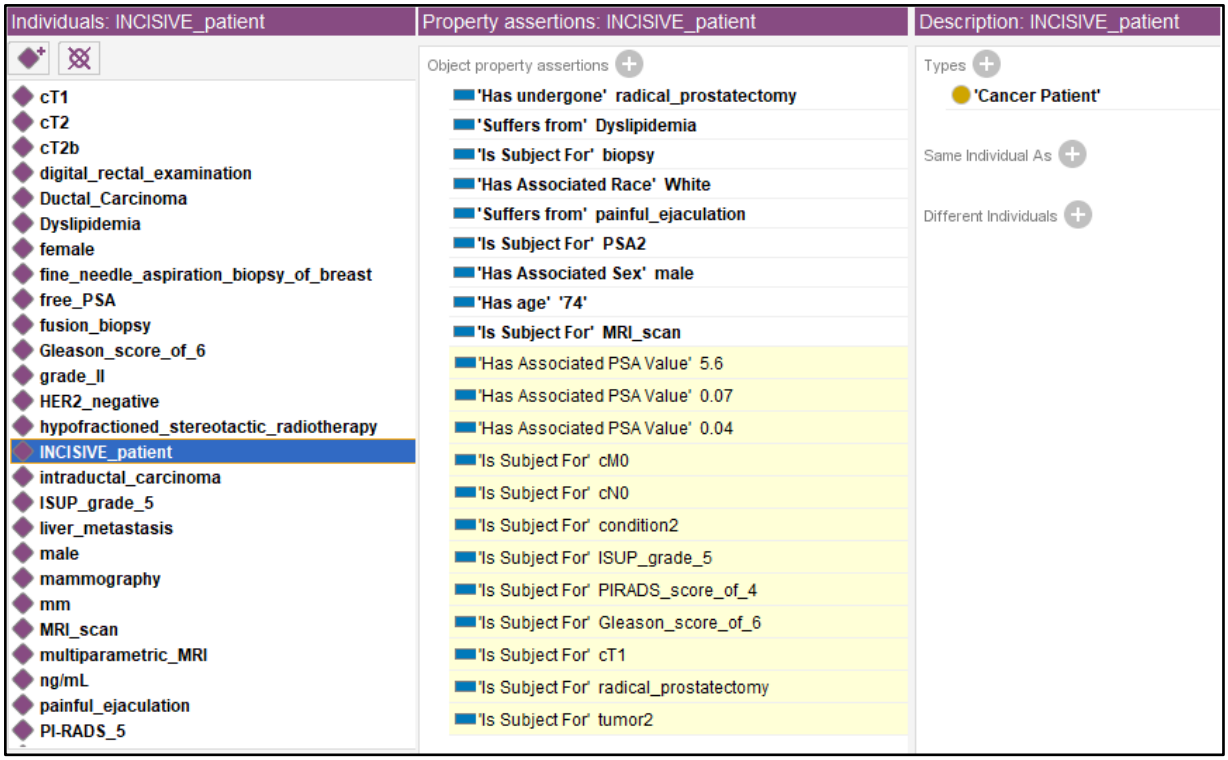


***Figure 30.*** *A semantic representation and inference of the INCISIVE prostate cancer use case (Protege)*

## Model Instantiation

**Cancer Patient**

| | |
|---|---|
| Identifier | ECI-98765 |
| Managing Organization | 2 |
| BirthSex | COM1001366 (**Male**(finding )) |
| Race | COM1001311 (**White**) |

**Tumor Marker Test**

| | | | | | | |
|---|---|---|---|---|---|---|
| Identifier | 34 | 45 | 56 | 67 | 78 | 89 |
| Related Condition | 2222 | 2222 | 2222 | 2222 | 2222 | 2222 |
| Subject | ECI-98765 | ECI-98765 | ECI-98765 | ECI-98765 | ECI-98765 | ECI-98765 |
| Category | laboratory | laboratory | laboratory | laboratory | laboratory | laboratory |
| Code | CLIN1033410 (**Prostate specific Ag [Mass/volume] in Serum or Plasma**) | CLIN1033410 (**Prostate specific Ag [Mass/volume] in Serum or Plasma**) | CLIN1033410 (**Prostate specific Ag [Mass/volume] in Serum or Plasma**) | CLIN1033410 (**Prostate specific Ag [Mass/volume] in Serum or Plasma**) | CLIN1033410 (**Prostate specific Ag [Mass/volume] in Serum or Plasma**) | CLIN1033410 (**Prostate specific Ag [Mass/volume] in Serum or Plasma**) |
| Value As Number | 5.6 | 0.04 | 0.07 | 0.04 | 0.04 | 0.04 |
| Value Unit Concept | COM1000156 (ng/mL) | COM1000156 (ng/mL) | COM1000156 (ng/mL) | COM1000156 (ng/mL) | COM1000156 (ng/mL) | COM1000156 (ng/mL) |
| Performed | 0 | 3 | 4 | 7 | 11 | 13 |
| Performed Unit Concept | | COM1000154 (month) | COM1000154 (month) | COM1000154 (month) | COM1000154 (month) | COM1000154 (month) |

**Procedure**

| | | | |
|---|---|---|---|
| Identifier | 6000 | 7000 | 88 |
| Code | IMG1005507 (**MRI of prostate**) | IMG1005488 (**MRI guided biopsy**) | CLIN1016442 (**Radical Prostatectomy**) |
| Category | IMG1005453 (Imaging (Procedure)) | CLIN1001712 (Biopsy) | CLIN1000228 (Procedure on organ) |
| Subject | ECI-98765 | ECI-98765 | ECI-98765 |
| Performed | 0 | 1 | 2 |
| Performed Unit | | COM1000154 (month) | COM1000154 (month) |

**Primary Cancer Condition**

| | |
|---|---|
| Identifier | 2222 |
| Age of diagnosis | 74 |
| Subject | ECI-98765 |
| Code | CLIN1008742 (**Primary malignant neoplasm of prostate**) |
| Histology Morphology Behavior | - |
| Body Site | (BP1000021) **Prostate** |

**Cancer Stage**

| | | | |
|---|---|---|---|
| Identifier | 5432 | 4321 | 3210 |
| Subject | ECI-98765 | ECI-98765 | ECI-98765 |
| Code | CLIN1033379 (**cT category**) | CLIN1033358 (**cN category**) | CLIN1033341 (**cM category**) |
| Method | CLIN1000417 (AJCC/UICC 7th edition) | CLIN1000417 (AJCC/UICC 7th edition) | CLIN1000417 (AJCC/UICC 7th edition) |
| Value | COM1000333 (**cT1**) | COM1000285 (**cN0**) | COM1001860 (**cM0**) |
| Procedure | 6000 | 6000 | 6000 |
| Rel. Condition | 2222 | 2222 | 2222 |

**Histologic Grade**

| | | |
|---|---|---|
| Related Condition | 2222 | 2222 |
| Subject | ECI-98765 | ECI-98765 |
| Category | laboratory | laboratory |
| Value | CLIN1026200(**Gleason grade score 6 out of 10**) | CLIN1047595 (**International Society of Pathology histologic grade group 5**) |
| Method | CLIN1037294 (**Gleason scoring system for malignant neoplasm of prostate**) | CLIN1037300 (**ISUP (International Society of Urologic Pathology) prostate cancer staging system**) |
| Related Procedure | 7000 | 7000 |

**Cancer Related Surgical Procedure**

| | |
|---|---|
| Identifier | 88 |
| Subject | ECI-98765 |
| Code | CLIN1016442 (**Radical Prostatectomy**) |
| Performed | 2 |
| Performed Unit | COM1000154 (month) |
| Response | COM1001312 (**Complete Response**) |

**Tumor**

| | |
|---|---|
| Identifier | 7777 |
| Subject | ECI-123456 |
| Related Condition | 1111 |
| Related Procedure | 6000 |
| Risk Assessment Method | IMG1005457 (**PI-RADS assessment**) |
| Risk Assessment Value | IMG1005476 (**PI-RADS 4 - High**) |
| Morphology | |

**Comorbidities**

| | | |
|---|---|---|
| Identifier | 0123 | 0234 |
| Focus | 2222 | 2222 |
| Subject | ECI-98765 | ECI-98765 |
| Comorbid Condition Present | CLIN1036520 (**Dyslipidemia**) | CLIN1018937 (**Painful ejaculation**) |
| Comorbid Condition Absent | - | - |

***Figure 31**: The EUCAIM CDM instantiation with the INCISIVE prostate cancer clinical information.*

## 7.2 Breast Cancer Use Cases

### CHAIMELEON Scenario

The following case is a real case scenario for a patient registered into the CHAIMELEON platform:

#### Patient's journey

A **Mammography** and **Ultrasound** were performed on a **59-year-old female** patient born in February 1957 that detected a lump in her breast. The **Ultrasound** indicated suspicious cancer (**BI-RADS 5**). For that reason, a **fine needle aspiration biopsy of breast** was performed **6 weeks later**, confirming the suspicion, diagnosing her with **Ductal Carcinoma grade II**, **cT2N0**, **RE positive**, **RP positive**, **HER2 negative**, and **Ki67** at **12%**. A **thorax, abdomen, and pelvis CT scan 2 weeks after the biopsy** showed no evidence of metastatic disease, confirming the **clinical stage** of the patient to **cT2N0M0** (**stage IIA**) .
The patient received **neoadjuvant chemotherapy**, starting **one month after the CT scan**. A **radical mastectomy** was performed **six months after** the chemotherapy, and **no tumor was found** (**pT0N0**). Another **thorax, abdomen, and pelvis CT scan** was performed **3 weeks after surgery**, showing no evidence of metastatic disease (**M0**). **Six weeks after surgery**, the patient began a **hypofractioned stereotactic radiotherapy** and has achieved a **complete response**.

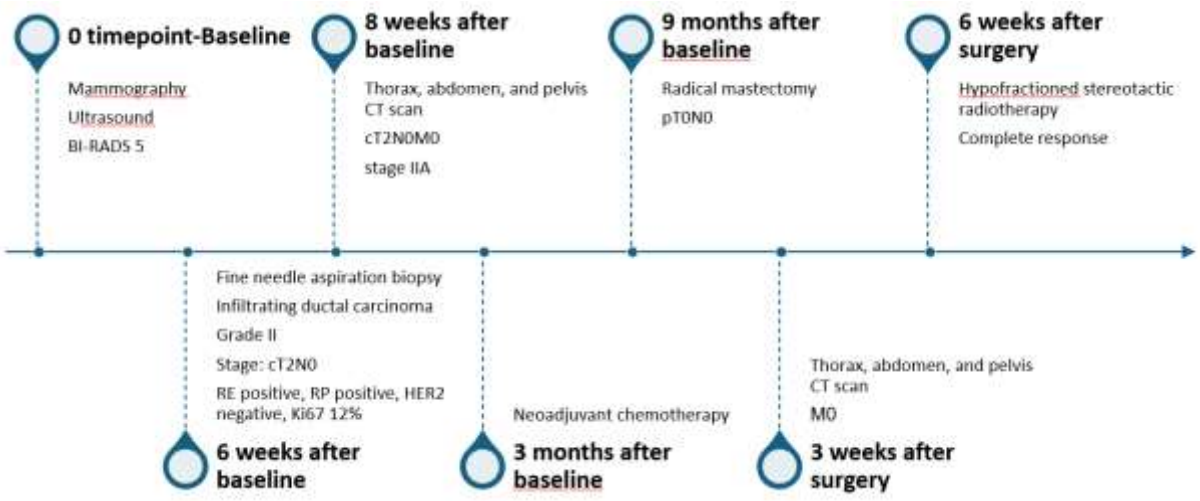


***Figure 32**. The CHAIMELEON breast cancer patient journey.*

### Hyper-Ontology Population

As for prostate cancer use cases, we populate the hyper-ontology with real-world breast cancer individuals (Figure 33). In this scenario, different procedures, including *mammography, ultrasound*, and *fine needle aspiration biopsy of breast*, have been performed on a female patient. Different pathologic and imaging results have been interpreted based on the performed procedures, such as tumor diagnosis (Ductal Carcinoma grade II), imaging assessment observations (e.g., *BI-RADS 5*), clinical staging (e.g., *cT2*, *cN0*), and tumor marker test results (e.g., *ER positive*, *PR positive*). Also, radiotherapy procedure (hypofractioned stereotactic radiotherapy) has been performed with a *complete response* associated result. The complete

diagnosis, including the imaging and pathologic interpretation results, has been inferred and generated by the logic reasoner as depicted in Figure 33.

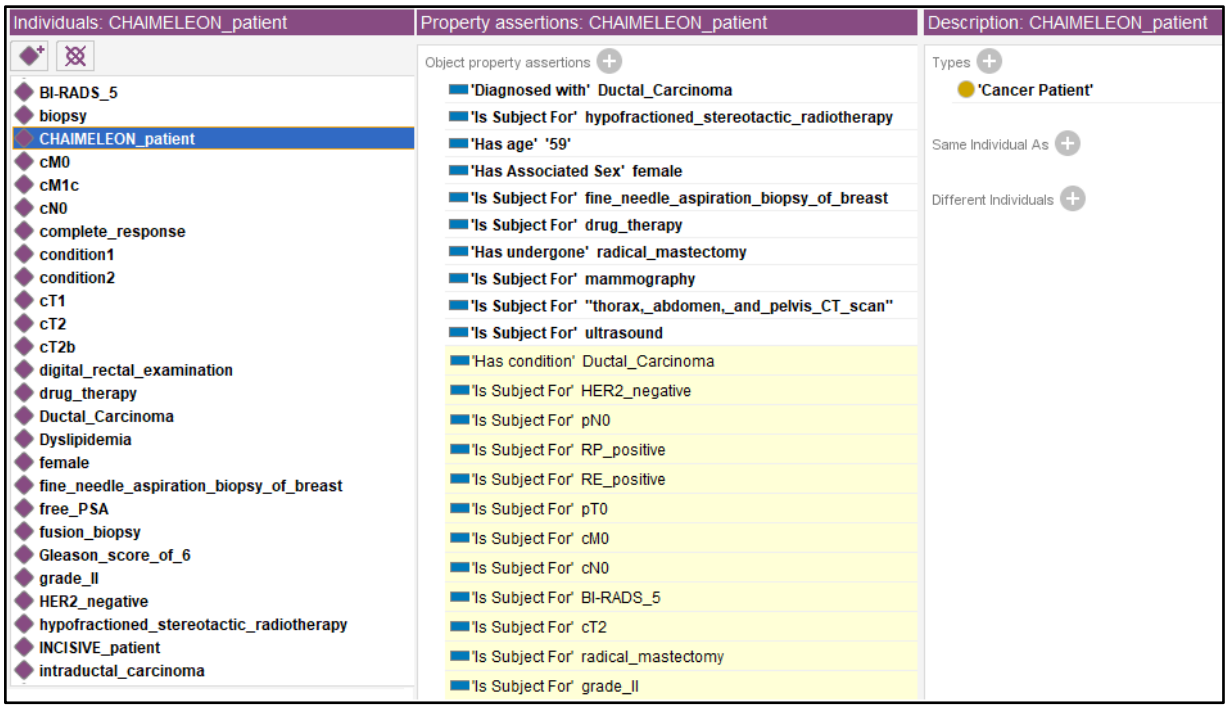


***Figure 33**. A semantic representation and inference of the CHAIMELEON breast cancer use case (Protege)*

## Model Instantiation

**Cancer Patient**

| | |
|---|---|
| Identifier | ECI-45678 |
| Managing Organization | 4 |
| BirthSex | COM1001370 (**Female**(finding)) |
| BirthDate | 1957-02 |

**Procedure**

| Identifier | 6000 | 7000 | 8000 | 9000 | 10000 | 11000 |
|---|---|---|---|---|---|---|
| Subject | ECI-45678 | ECI-45678 | ECI-45678 | ECI-45678 | ECI-45678 | ECI-45678 |
| Code | IMG1000030 (**Mammography**) | IMG1016137 (**Ultrasonography of breast**) | CLIN1001715 (**Fine needle aspiration of breast**) | IMG1000027 (**Computerized Tomography (CT Scan) of Chest, Abdomen and Pelvis**) | CLIN1014762 (**Drug therapy**) | IMG1000027 (**Computerized Tomography (CT Scan) of Chest, Abdomen and Pelvis**) |
| Category | IMG1005453 (Imaging (Procedure)) | IMG1005453 (Imaging (Procedure)) | CLIN1001712 (Biopsy) | IMG1005453 (Imaging (Procedure)) | CLIN1010523 (Therapeutic procedure) | IMG1005453 (Imaging (Procedure)) |
| Performed | 0 | 0 | 6 | 8 | 12 | 19 |
| Performed Unit Concept | | | COM1000155 (week) | COM1000155 (week) | COM1000155 (week) | COM1000155 (week) |

**Primary Cancer Condition**

| | |
|---|---|
| Identifier | 6666 |
| Age of diagnosis | **59** |
| Subject | ECI-45678 |
| Code | CLIN1047300 (**Primary malignant neoplasm of female breast**) |
| Histology Morphology Behavior | CLIN1021458 (**Ductal carcinoma in situ, solid type of breast, NOS**) |
| Body Site | BP1000136 (**Breast**) |
| Body Site Location Qualifier | |
| Body Site Laterality Qualifier | |

**Tumor Marker Test**

| Identifier | 1 | 2 | 3 | 4 |
|---|---|---|---|---|
| Related Condition | 6666 | 6666 | 6666 | 6666 |
| Related Procedure | 8000 | 8000 | 8000 | 8000 |
| Subject | ECI-45678 | ECI-45678 | ECI-45678 | ECI-45678 |
| Category | Laboratory | Laboratory | Laboratory | Laboratory |
| Code | CLIN1045815 (**Estrogen receptor Ag [Presence] in Breast cancer specimen by Immune stain**) | CLIN1046085 (**Progesterone receptor Ag [Presence] in Breast cancer specimen by Immune stain**) | CLIN1045851 (**HER2 [Presence] in Breast cancer specimen by Immune stain**) | CLIN1049762 (**Percent of cell nuclei positive for proliferation marker protein Ki-67 in primary malignant neoplasm by immunohistochemistry**) |
| Value As Concept | COM1001310 (**Positive**) | COM1001310 (**Positive**) | COM1001332 (**Negative**) | |
| Value As Number | | | | 12 |
| Value Unit Concept | | | | COM1000161 (percent) |
| Performed | 6 | 6 | 6 | 6 |
| Performed Unit Concept | COM1000155 (week) | COM1000155 (week) | COM1000155 (week) | COM1000155 (week) |

**Histologic Grade**

| | |
|---|---|
| Related Condition | 6666 |
| Subject | ECI-45678 |
| Category | laboratory |
| Value | CLIN1022150 (**Grade 2 tumor**) |
| Method | CLIN1037299 (**Nottingham histologic grading system**) |
| Related Procedure | 8000 |

**Tumor**

| | |
|---|---|
| Identifier | 7777 |
| Subject | ECI-45678 |
| Related Condition | 6666 |
| Related Procedure | 7000 |
| Risk Assessment Method | IMG1005459 (**BI-RADS assessment**) |
| Risk Assessment Value | IMG1005464 (**BI-RADS 5**) |

**Cancer Stage**

| Identifier | 5432 | 4321 | 3210 | 2109 | 1098 | 8765 | 7654 | 6543 |
|---|---|---|---|---|---|---|---|---|
| Subject | ECI-45678 | ECI-45678 | ECI-45678 | ECI-45678 | ECI-45678 | ECI-45678 | ECI-45678 | ECI-45678 |
| Code | CLIN1033379 (**cT category**) | CLIN1033358 (**cN category**) | CLIN1033358 (**cT category**) | CLIN1033358 (**cN category**) | CLIN1033341 (**cM category**) | CLIN1033391 (**pT category**) | CLIN1033368 (**pN category**) | CLIN1033341 (**cM category**) |
| Method | CLIN1000398 (AJCC/UICC finding) | CLIN1000398 (AJCC/UICC finding) | CLIN1000398 (AJCC/UICC finding) | CLIN1000398 (AJCC/UICC finding) | CLIN1000398 (AJCC/UICC finding) | CLIN1000398 (AJCC/UICC finding) | CLIN1000398 (AJCC/UICC finding) | CLIN1000398 (AJCC/UICC finding) |
| Value | COM1000351 (**cT2**) | COM1000285 (**cN0**) | COM1000351 (**cT2**) | COM1000285 (**cN0**) | COM1001860 (**cM0**) | COM1000708 (**pT0**) | COM1000645 (**pN0**) | COM1001860 (**cM0**) |
| Procedure | 8000 | 8000 | 9000 | 9000 | 9000 | 88 | 88 | 11000 |
| Rel. Condition | 6666 | 6666 | 6666 | 6666 | 6666 | 6666 | 6666 | 6666 |

**Cancer Related Surgical Procedure**

| Identifier | 77 | 88 |
|---|---|---|
| Subject | ECI-45678 | ECI-45678 |
| Code | CLIN1014762 (**Drug therapy**) | CLIN1004693 (**Radical mastectomy**) |
| Performed | 12 | 16 |
| Performed Unit | COM1000155 (week) | COM1000155 (week) |
| Response | | |

**Radiotherapy Course Summary**

| Identifier | 99 |
|---|---|
| Subject | ECI-45678 |
| Code | CLIN1029448 (**Hypofractionated stereotactic radiotherapy**) |
| Performed | 22 |
| Performed Unit | COM1000155 (week) |
| Response | COM1001312 (**Complete Response**) |

***Figure 34**. The EUCAIM CDM instantiation with the CHAIMELEON breast cancer clinical information.*

## EuCanImage scenario

The following case is a real case scenario for a patient registered into the EuCanImage platform:

### Patient's journey

The patient is a **50-year-old postmenopausal female** individual who has **never breastfed** and has **never been pregnant**. There is a **history of breast cancer** in a **second degree relative**. There is **no family history of ovarian cancer**. The patient has **never used hormone replacement therapy** or **hormonal contraception**. Based on a **mammography** followed by a **needle biopsy one month later**, the patient was **diagnosed** with **triple-negative cancer** of the **right breast**, **Grade I LCIS** histological type and clinical stage of **cT1N1M0.** In addition, the following tumor characteristics were assessed: **ER 0% ,PR 0%, HER2 IHC negative**, and **Ki67 0%**. **Neoadjuvant chemotherapy (NAC)** with **Doxorubicin** was started **1 month after the** results of the pathological report from the biopsy, **lasting about 6 months**. After NAC treatment, the patient underwent **breast surgery** where the pathology report revealed: **ypT0N0M0**.

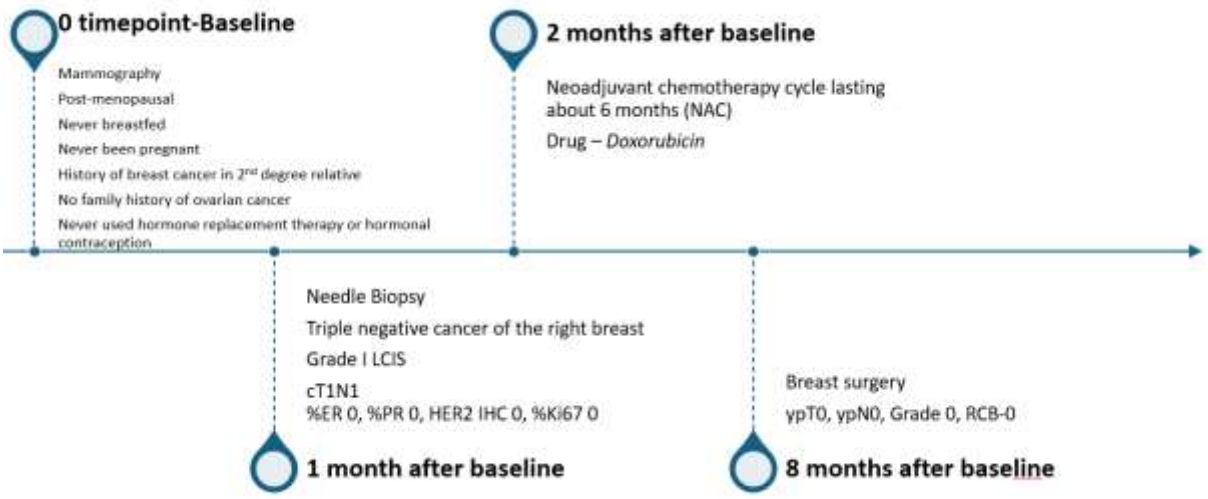


***Figure 35.*** *The EUCANIMAGE breast cancer patient journey.*

## Hyper-Ontology Population

Based on the EuCanImage breast cancer scenario, we have populated the hyper-ontology concepts and relations with real-world instances (Figure 36). In this scenario, procedures, such as *mammography, needle biopsy*, and *breast surgery*, have been performed on a *postmenopausal* female patient who has never used hormone therapies. Different interpretation results have been identified, such as tumor diagnosis (t*riple-negative cancer Grade I LCIS*), clinical staging (e.g., *cT1*, *cN1, ypT0*), and tumor marker test results (e.g., *ER 0%*, *PR 0%, HER2 negative*). The complete diagnosis, including interpretation results, has been inferred and generated by the logic reasoner as depicted in Figure 36.

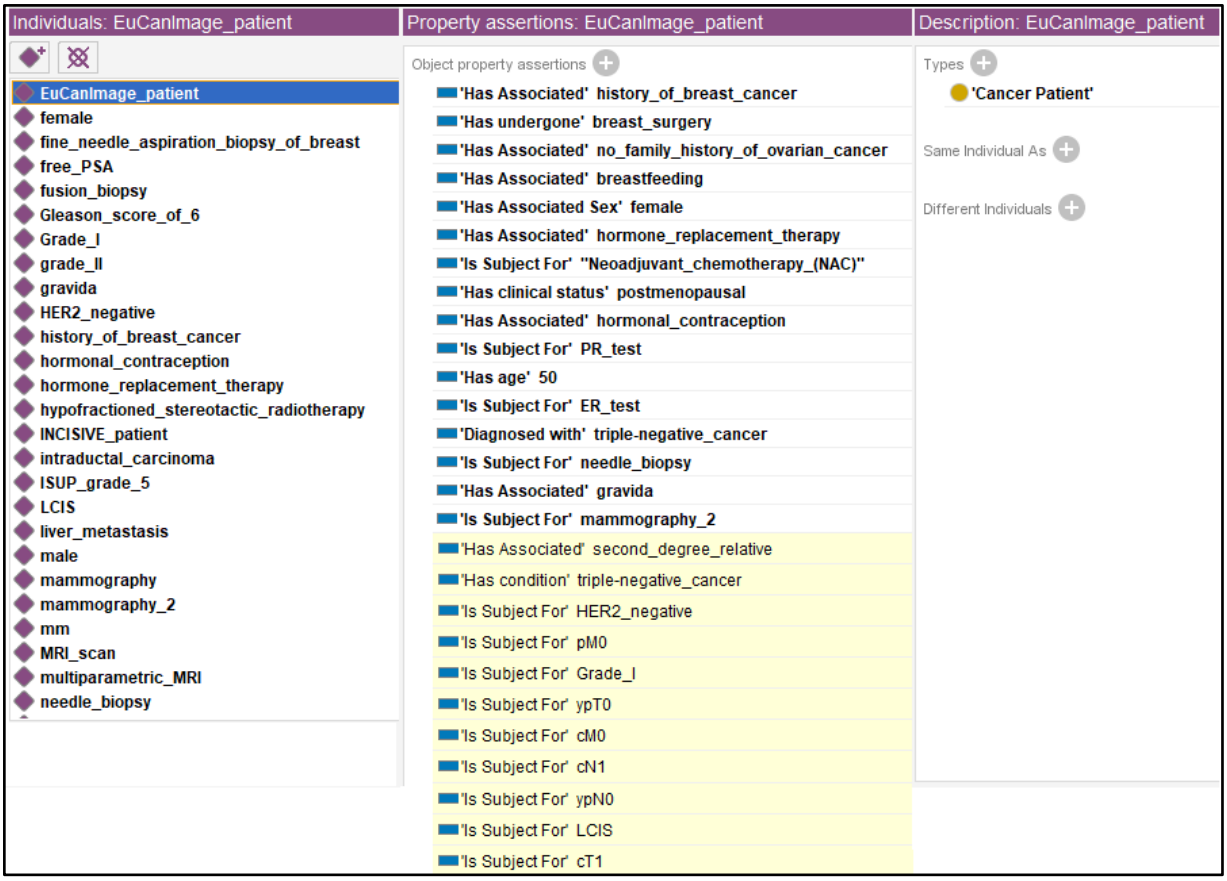


***Figure 36**. A semantic representation and inference of the EuCanImage breast cancer use case (Protege)*

## Model Instantiation

**Cancer Patient**

| | |
|---|---|
| Identifier | ECI-23456 |
| Managing Organization | 3 |
| BirthSex | COM1001370 (**Female**(finding)) |
| Race | - |

**Observation**

| | | | | | |
|---|---|---|---|---|---|
| Identifier | 1000 | 2000 | 3000 | 4000 | 5000 |
| Subject | ECI-23456 | ECI-23456 | ECI-23456 | ECI-23456 | ECI-23456 |
| Code | CLIN1045770 (**Menopause, function**) | CLIN1048535 (**Gravida**) | CLIN1043690 (**Breastfeeding**) | CLIN1035175 (**Hormone replacement therapy**) | CLIN1035208 (**Uses hormonal contraception**) |
| Value As Concept | CLIN1048610 (**Postmenopause - Menopause present**) | | COM1001078 (**No**) | COM1001078 (**No**) | COM1001078 (**No**) |
| Value As Number | | 0 | | | |

**Procedure**

| | | | |
|---|---|---|---|
| Identifier | 7777 | 8888 | 9000 |
| Subject | ECI-23456 | ECI-23456 | ECI-23456 |
| Code | IMG1000030 (**Mammography**) | CLIN1001713 (**Needle biopsy**) | CLIN1034672 (**Excision of breast**) |
| Category | IMG1005453 (Imaging (Procedure)) | CLIN1001712 (Biopsy) | CLIN1000228 (Procedure on organ) |
| Performed | 0 | 1 | 8 |
| Performed Unit Concept | | COM1000154 (month) | COM1000154 (month) |

**Cancer Stage**

| | | | | | |
|---|---|---|---|---|---|
| Identifier | 590432 | 432187 | 25693 | 87365 | 72567 |
| Subject | ECI-23456 | ECI-23456 | ECI-23456 | ECI-23456 | ECI-23456 |
| Code | CLIN1033379 (**cT category**) | CLIN1033358 (**cN category**) | CLIN1033391 (**pT category**) | CLIN1033368 (**pN category**) | CLIN1033349 (**pM category**) |
| Method | CLIN1000398 (**AJCC/UICC finding**) | CLIN1000398 (**AJCC/UICC finding**) | CLIN1000398 (**AJCC/UICC finding**) | CLIN1000398 (**AJCC/UICC finding**) | CLIN1000398 (**AJCC/UICC finding**) |
| Value | COM1000333 (**cT1**) | COM1000285 (**cN1**) | COM1001607 (**ypT0**) | COM1001080 (**ypN0**) | COM1001845 (**pM0**) |
| Procedure | 8888 | 8888 | 9000 | 9000 | 9000 |
| Rel. Condition | 6666 | 6666 | 6666 | 6666 | 6666 |

**Cancer Related Medication Administration**

| | |
|---|---|
| Identifier | |
| Subject | ECI-23456 |
| Code | CLIN1035838 (**Doxorubicin**) |
| Administered | 2 |
| Administered Unit Concept | COM1000154 (month) |
| Effective | 6 |
| Effective Unit Concept | COM1000154 (month) |

**Family Member History**

| | | |
|---|---|---|
| Identifier | 0123 | 0234 |
| Subject | ECI-23456 | ECI-23456 |
| Relationship | COM1001050 (**Second degree blood relative (person)**) | - |
| Condition Code | CLIN1048389 (**Family history of breast cancer**) | (CLIN1029965) **No family history of ovarian cancer** |

**Tumor Marker Test**

| | | | | |
|---|---|---|---|---|
| Identifier | 34 | 45 | 56 | |
| Related Condition | 6666 | 6666 | 6666 | 6666 |
| Subject | ECI-23456 | ECI-23456 | ECI-23456 | ECI-23456 |
| Code | CLIN1045815 (**Estrogen receptor Ag [Presence] in Breast cancer specimen by Immune stain**) | CLIN1046085 (**Progesterone receptor Ag [Presence] in Breast cancer specimen by Immune stain**) | CLIN1045851 (**HER2 [Presence] in Breast cancer specimen by Immune stain**) | CLIN1049762 (**Percent of cell nuclei positive for proliferation marker protein Ki-67 in primary malignant neoplasm by immunohistochemistry**) |
| Value As Number | 0 | 0 | 0 | 0 |
| Value Unit Concept | COM1000161 (%) | COM1000161 (%) | COM1000161 (%) | COM1000161 (percent) |
| Performed | 1 | 1 | 1 | 1 |
| Performed Unit Conc. | COM1000154 (month) | COM1000154 (month) | COM1000154 (month) | COM1000154 (month) |

**Primary Cancer Condition**

| | |
|---|---|
| Identifier | 6666 |
| Age of diagnosis | **50** |
| Subject | ECI-23456 |
| Code | CLIN1046256 (**Triple negative malignant neoplasm of breast**) |
| Histology Morphology Behavior | CLIN1052223 (**Lobular carcinoma in situ, NOS**) |
| Body Site | BP1000136 (**Breast**) |
| Body Site Location Qualifier | - |
| Body Site Laterality Qualifier | IMG1016682 (**Right**) |

**Histologic Grade**

| | | |
|---|---|---|
| Related Condition | 6666 | 6666 |
| Subject | ECI-23456 | ECI-23456 |
| Category | laboratory | laboratory |
| Value | COM1001750 (**G1 - American Joint Committee on Cancer grade G1**) | COM1001750 (**G1 - American Joint Committee on Cancer grade G1**) |
| Method | CLIN1037296 (Histological grading systems) | CLIN1037296 (Histological grading systems) |
| Related Procedure | 8888 | 9000 |

***Figure 37*** *The EUCAIM CDM instantiation with the EuCanImage breast cancer clinical information*

From the populating and instantiating validation tasks, we assume that the hyper-ontology has successfully represented domain-specific knowledge in oncology acquired from real-world prostate and breast cancer scenarios, and fulfilled the requirement of seamless integration with EUCAIM-CDM for the instantiation process.

# 8. Future work and perspective

In further works, we are interested in extending the hyper-ontology cancer types to include new types with the support of clinical experts. The extension process will consider the new use cases expected to be provided by the hospitals or laboratories that will join the EUCAIM community. Besides, the imaging and clinical metadata required for federated querying will be specified explicitly in the hyper-ontology model to permit seamless integration with heterogeneous local datasets and efficient access to these data.

However, one of the main challenges we need to address is the sustainability and evolution of the hyper-ontology facing the continuous syntactic and semantic updates of standard terminologies/ontologies and data models or standards (OMOP/FHIR), especially after the project completion.

The long-term sustainability of the Common Data Model (CDM) and Hyper-Ontology are critical to the success of EUCAIM. We recognize that interoperability is not just a technical challenge but also an organizational one that requires ongoing commitment. To this end, we plan to explore various strategies, including:

- Having a clear data governance framework, to oversee the evolution of the hyper-ontology, ensures that changes are managed in a controlled manner.
- Developing a clear roadmap for the development of the CDM and the hyper-ontology, including regular updates, and adaptation to new technologies and standards. Currently, as we develop the hyper-ontology, we are creating distinct versions, each with unique identifiers and appropriate metadata and documentation in order to track its evolution. All versions are released periodically and published on Zenodo.
- Encouraging contributions and feedback from the wider community to ensure that the CDM and hyper-ontology remain comprehensive, up-to-date, and reflective of the needs of all stakeholders. Towards this end, we have also committed to submitting research papers to workshops and conferences outlining our approach.
- Identifying the resources, both financial and human, required to support the ongoing maintenance and development of the hyper-ontology and the CDM. This might include seeking funding, establishing partnerships, or generating revenue through specific services. This strategy will be further explored in collaboration with WP8.

Finally, we plan to make an impact assessment for becoming compliant to the CDM and the Hyper-Ontology on new data holders providing data or joining EUCAIM. This impact assessment will be two-fold: a) identify the challenges to be faced for complying to the CDM and hyper-ontology but also b) identify the benefits that result from successfully complying with such a framework.

Regarding the challenges, we plan to identify and assess the effort required by new data holders to manage and structure their data according to the hyper-ontology and CDM specifications. This might entail training sessions with clinical and technical staff, therefore assessing the total time and the resources required for new data holders to achieve compliance, and possibly identifying any obstacles they might face in the process.

Regarding the benefits, data holders complying to the CDM might potentially achieve increased interoperability as they will gain the ability to share and integrate data with other entities within the EUCAIM network. Compliance will also ensure that their data meets high standards of quality, aligning with international best practices and standards, and enhancing the credibility of their data contributions.

However, to do such an impact assessment, we need an evaluation process that could include:

- Conducting surveys and interviews with potential new data holders to understand their current data management practices, capabilities, and readiness for compliance with the CDM and Hyper-Ontology.
- Documenting the specific changes and adaptations new data holders would need to make.
- Identifying common issues and problems through the EUCAIM helpdesk and the support groups, which can help us gather feedback on the compliance process and thus make necessary adjustments to our approach if necessary, specifically in cases where data holders consistently experience certain issues.
- Split the onboarding process into stages/tiers, which we have already defined, so that we distribute the required effort across multiple stages till the final adoption of the EUCAIM CDM in order to minimize but also track possible issues/problems.

To achieve this, we will closely collaborate with WP2 and WP4 respective teams.

# 9. Conclusion

This deliverable presents the initial version of the EUCAIM CDM and hyper-ontology for data interoperability. In relation to the first deliverable (D5.1), this document provides a well-established analysis of the strategy developed to achieve the initial goals of the EUCAIM CDM and hyper-ontology. Publications submitted and accepted during the hyper-ontology development support the work accomplished.

Regarding the hyper-ontology development process, we encountered challenges in the knowledge acquisition phase (Section 4.3) to collect the standard clinical/biological and imaging data/metadata provided by the AI4HI projects. For the clinical knowledge, some data/metadata were customized depending on the projects' resources, or standard code/vocabulary was lacking, which required an effort to associate this information with standard ontological/terminological resources. For imaging knowledge, the provided data/metadata was mainly DICOM tags and names used for image querying or segmentation, which is insufficient for a semantic representation of imaging knowledge in the hyper-ontology. Interestingly, the proposed approach (Section 4.4) has helped to overcome these challenges. First, the ORSD document was produced, which helped to organize all the collected data and metadata and classify them by cancer type and project, facilitating the detection of inconsistencies and lack of information. Second, the grounding of the hyper-ontology in mCODE has supported covering the essentials of the oncology domain, mainly for clinical aspects. For the imaging model, we relied on FHIR specifications around *Imaging study* and *Series*, and their relationships with *Modality*, *Laterality*, and other imaging aspects. Although the bottom-up strategy, which relies on the projects' clinical and imaging knowledge, is crucial for developing the hyper-ontology as a domain and application-oriented ontology, the top-down has maintained the ontological model by grounding the hyper-ontology in the oncology domain. Also, the intervention of experts in revising and enriching the semantic content hyper-ontology has enhanced the generic content and expanded it by including *clinically verified* semantic patterns. Finally, the hyper-ontology is validated by:

1- *efficiently* and *explicitly representing* the provided use cases by populating the hyper-ontology semantic content, including the concepts and relations, based on the individuals (instances) harvested from these use cases (Section 7);
2- *instantiating* the EUCAIM-CDM to represent real-world use cases around prostate and breast cancers using the hyper-ontology concepts (Section 7);
3- applying *SPARQL queries* to request cancer patient information, such as lab tests, procedures, imaging and clinical results (Annex 1).

Interestingly, EUCAIM's hyper-ontology, a FAIR-compliant ontology model that effectively reflects oncology's real-world entities, has supported a seamless integration with the EUCAIM CDM, a significant fulfillment for maintaining semantic interoperability in the context of the EUCAIM project.

# 10. Publications

- El Ghosh, M., Kalokyri, V., Sambres, M., Vaterkowski, M., Duclos, C., Tannier, X., Taskou, G., Tsiknakis, M., Daniel, C., and Dhombres, F. (2024). *Towards semantic interoperability among heterogeneous cancer image data models using a layered modular hyper-ontology*. In FOIS 2024.

- El Ghosh, M., Kalokyri, V., Sambres, M., Vaterkowski, M., Duclos, C., Tannier, X., Taskou, G., Tsiknakis, M., Daniel, C., and Dhombres, F. (2024). *From syntactic to semantic interoperability using a hyper-ontology in the oncology domain*. In MIE 2024.

- El Ghosh, M., Daniel, C., Duclos, C., Kalokyri, V., Charlet, J., Sambres, M., Tsakou, G., Tsiknakis, M., and Dhombres, F. (2024). *Grounding a hyper-ontology on mCODE ontological conceptual model and foundational ontologies for semantic interoperability in the oncology domain*. In FOAM@FOIS 2024.

# 11. ANNEX

***Annex 1***: SPARQL Queries

Based on the information acquired from the prostate cancer use cases (Section 7), SPARQL queries are applied to request the hyper-ontology regarding diagnosis details. In the following, we give some examples of SPARQL queries to question the cancer patients (COM1001051) who:

- had a PSA (CLIN1000227) lab test and to return the PSA levels (**Query1**);
- underwent a prostatectomy (CLIN1000248) and to return the associated pathological interpretation results (**Query2**);
- were subject to imaging procedures and to return the associated imaging interpretation results (**Query3**).

**PREFIX ho: <https://cancerimage.eu/ontology/EUCAIM#>**

**Query1: SELECT ?p ?r WHERE {**
**?p rdf:type ho:COM1001051 .**
**?p ho:Is_Subject_For ?a .**
**?a rdf:type ho:CLIN1000227 .**
**?a ho:Has_Value ?r . }**

For Query1, both patients of the ProCAncer-i (uc1) and INCISIVE (uc2) use cases have done the PSA lab test. Thus, by executing Query1, we obtain the following results:

***ProCAncer-I_patient : PSA level = 7.16***
***INCISIVE_patient : PSA level = 0.04***
***INCISIVE_patient : PSA level = 0.07***
***INCISIVE_patient : PSA level = 5.6***

**Query2**: **SELECT ?p ?a ?r WHERE {**
**?p rdf:type ho:COM1001051 .**
**?p ho:HasUndergone ?a .**
**?a rdf:type ho:CLIN1000248 .**
**?a ho:Has_pathologic_interpretation_result ?r . }**

For Query2, only the ProCancer-i patient (uc1) underwent a prostatectomy with different pathologic interpretation results. Thus, the response of this query is obtained as follows:

***ProCAncer-I_patient : prostatectomy -> Result: intraductal_carcinoma***
***ProCAncer-I_patient : prostatectomy -> Result: pN0***
***ProCAncer-I_patient : prostatectomy -> Result: pT3b***
***ProCAncer-I_patient : prostatectomy -> Result: 4+3_Gleason_score***

**Query3**: **SELECT ?p ?a ?r WHERE {**
**?p rdf:type ho:COM1001051 .**
**?p ho:Is_Subject_For ?a .**
**?a ho:Has_imaging_interpretation_result ?r . }**

For Query3, the ProCancer-i patient (uc1) was subject to *multiparametric MRI* and *fusion biopsy* with the following interpretation results: *PI-RADS 5*, *cT2b*, *cN0,* and *pT2*. Meanwhile, the INCISIVE patient (uc2) was subject to *MRI scan* and *biopsy* with the following results: *PI-RADS score 4*, *Gleason score 6*, and *ISUP grade 5*. By executing Query3, we obtain the following results:

***INCISIVE_patient : MRI_scan -> Result: PIRADS_score_of_4***

***INCISIVE_patient : biopsy -> Result: Gleason_score_of_6***
***INCISIVE_patient : biopsy -> Result: ISUP_grade_5***
***ProCAncer-I_patient : fusion_biopsy -> Result: pT2***
***ProCAncer-I_patient : multiparametric_MRI -> Result: PI-RADS_5***
***ProCAncer-I_patient : multiparametric_MRI -> Result: cN0***
***ProCAncer-I_patient : multiparametric_MRI -> Result: cT2b***