Project title: European Federation for Cancer Images

Project acronym: EUCAIM

Grant Agreement: 101100633

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

D5.3. The EUCAIM CDM and Hyper-Ontology for Data Interoperability: final version

Responsible partner: FORTH

Authors: Valia Kalokyri (FORTH), Mirna El Ghosh (LIMICS), Stelios Sfakianakis (FORTH), Christel Daniel (APHP-LIMICS), Catherine Duclos (LIMICS), Xavier Tannier (LIMICS), Eirini Kaldeli (MAG), Laure Saint Aubert (MEDEX), Gianna Tsakou (MAG), Ferdinand Dhombres (LIMICS), Manolis Tsiknakis (FORTH)

Contributors: Maciej Bobowicz (GUMED), Victor Sóñora Pombo (BAHIA), Irene Marín Radoszynski (HULAFE), Celia Martin Vicario (QUIBIM), Olga Giraldo (DKFZ), Carles Hernandez-Ferrer (BSC), Alexandra Kosvyra (AUTH), Ioannis Ladakis (AUTH), Ioanna Chouvarda (AUTH), Eleftherios Pappas (UoA), Ioannis Seimenis (UoA), Maria Gonzalez Lopez (SAS), Sophie Mendjel (APHP), Loick Chardon (APHP), Taeyang Choi (KI), Ignacio Blanquer (UPV)

Reviewers: Alberto Talaya (HULAFE), Teresa García Lezana (CRG)

Final Deadline: 31 December 2025

Date of delivery: 23 December 2025

Version: 1.4

Table of Contents

List of Tables

Table of Figures

1. Executive Summary

1.1 Purpose of the Deliverable

This deliverable presents the final specification of the EUCAIM hyper-ontology and Common Data Model (CDM), and documents their implementation within the EUCAIM ecosystem. It describes the modular structure of the hyper-ontology, the vocabularies it incorporates, and its mappings to external standards, together with the associated governance processes. In parallel, it provides a detailed specification of the final EUCAIM CDM, including its entities, attributes, value sets and constraints covering clinical, biological and imaging data. A central focus of the document is the way in which the CDM and hyper-ontology are integrated and bound together in order to enable semantic interoperability from high-level dataset metadata down to patient-level records. Their role within the Minimum Data Federation and Interoperability Framework (min-FIF) is explained, highlighting their contribution to Tier 1 (catalogue), Tier 2 (federated search) and Tier 3 (federated processing). The deliverable also reports on validation activities, drawing on the experience gained through seven technical pilots supporting the onboarding of federated nodes, and on feedback from clinical and technical experts.

1.2 Scope and Audience

The scope of this deliverable is confined to the semantic and syntactic interoperability artefacts under the responsibility of WP5. Specifically, it covers three main components: the EUCAIM DCAT-AP specification, which provides the dataset-level metadata model adopted in the public catalogue at Tier 1, the EUCAIM hyper-ontology, which acts as a common semantic meta-model for heterogeneous cancer imaging data sources for federated search in Tier 1 and Tier 2 only, and the EUCAIM CDM, which serves as the reference model both for federated search in Tier 2 and for federated processing and analysis in Tier 3.

Legal and organizational aspects of interoperability, including the rules of participation, governance of data access and contractual arrangements, are out of scope of this deliverable. These aspects are handled in D4.4 “Final rules of participation” and related WP4 deliverables and are only considered in this document when they directly influence the CDM and hyper-ontology design.

The primary audience of this deliverable includes data modelers, semantic interoperability experts, technical architects and developers within EUCAIM, as well as clinical leads and data stewards at participating nodes who need to understand the semantic framework underpinning data sharing and federated analysis. It may also be of interest to external initiatives and infrastructures exploring alignment or interoperability with EUCAIM.

1.3 Relation to other EUCAIM deliverables

This deliverable relates to and builds upon:

- **D5.1 – Early release of the Data Federation Framework**, which introduced the initial design of the EUCAIM data federation and the first proposals for metadata and semantic models.
- **D5.2 – The EUCAIM CDM and Hyper-Ontology for Data Interoperability: initial version**, which detailed the first release of the hyper-ontology and CDM, together with preliminary proof-of-concept scenarios based on the AI4HI clinical and imaging knowledge.
- **D5.6 – Minimum Data Federation and Interoperability Framework**, which operationalizes the semantic and technical requirements in the three tiers and provides minimum datasets and guidelines for node setup and data preparation.

The present document should therefore be read as the semantic and data modelling counterpart to D5.6: it specifies *what* needs to be represented and *how* it is semantically organized, while D5.6 specifies the corresponding technical requirements and workflows to implement this in practice.

2. Introduction

2.1 Background and Motivation

EUCAIM brings together a large number of heterogeneous data holders, each with their own clinical information systems, imaging archives, coding practices and local data models. Without a shared semantic framework, it would be extremely difficult to catalogue available datasets in a consistent manner, to locate and select relevant data across nodes, or to execute reproducible analysis workflows on harmonized data. The need for semantic harmonization and common data structures is therefore fundamental to the project's ambition to create a pan-European infrastructure for cancer imaging data.

The hyper-ontology and CDM respond to this need by providing, respectively, a common semantic meta-model and a harmonized data representation that can be used across the federation. Together, they allow EUCAIM to bridge differences between local schemas, terminologies and technical implementations, while preserving data holders' autonomy and data sovereignty. This is particularly important in the context of federated analytics and AI development, where consistent interpretation of variables, measurements and outcomes across sites is critical for the validity and comparability of results.

Furthermore, the approach taken in EUCAIM is designed to be compatible with, and reusable by, other European health data initiatives. By aligning with widely adopted standards and ontologies, the hyper-ontology and CDM contribute to the broader goals of interoperability, FAIR data, and cross-infrastructure collaboration within the European Health Data Space.

2.2 Role of the Hyper-ontology and CDM

Within this overall framework, the hyper-ontology and CDM play complementary roles. The hyper-ontology provides a structured, machine-interpretable representation of the key concepts, relationships and value sets that are relevant for describing cancer patients, diseases, laboratory tests, imaging procedures, findings, treatments and outcomes, etc. It sits

at the semantic mediation layer and is used to define controlled vocabularies, to harmonise the meaning of terms across datasets, and to act as a knowledge basis to support query formulation and federated search (through contextual constraints, expansion, etc).

The CDM, in turn, provides a concrete schema for representing patient-level clinical, biological and imaging data suitable both for federated search, and federated processing, analysis and AI model training. It captures the entities and attributes needed for both cohort building and typical analytic and AI workflows, and it incorporates constraints and value sets that reflect the semantics defined in the hyper-ontology. In practice, local nodes could adopt the EUCAIM CDM directly or map their existing OMOP-, FHIR- or DICOM-based schemas to it through dedicated transformation components. In this case, a node can participate in federated processing/analysis by creating the materialized view of what is requested (in a .csv) following the EUCAIM CDM, without transforming the whole local database into an EUCAIM CDM compliant database.

By linking the two artefacts, using the hyper-ontology to define the semantics of CDM elements and to anchor them to external standards, EUCAIM ensures a coherent interoperability stack. Dataset descriptions in the public metadata catalogue, semantic search across nodes and federated analytic workflows all rely on a single, consistent semantic foundation, thereby reducing ambiguity and improving the robustness and reusability of results across the federation. Within this framework, the CDM and hyper-ontology function as follows:

- At Tier 1 (Dataset metadata), the CDM and hyper-ontology influence which elements become mandatory or recommended in the EUCAIM DCAT-AP profile, and they provide controlled vocabularies for key fields such as cancer type, anatomical site and imaging modality and annotation related information.
- At Tier 2 (Federated search), the hyper-ontology is the central semantic artefact: query criteria defined in D5.6 (e.g. “malignant neoplasm of prostate”, “MRI of breast”) are represented as ontology concepts that can be mapped to heterogeneous local schemas via mediator services. This is applicable for nodes implementing Tier 1 and Tier 2 only. In case a DH implements Tier 3, the node exposes an EUCAIM CDM compliant dataset, and therefore this dataset is used for both federated search and federated processing analysis.
- At Tier 3 (Federated processing), the CDM is the primary reference model for representing patient-level clinical and imaging data in a harmonized way across federated and EUCAIM reference nodes, either by direct adoption (through ETL) or via mapping components from OMOP or FHIR-based local models into the EUCAIM CDM only for the requested sub-cohort relevant for the analysis

This layered positioning ensures a coherent semantic stack: DCAT-AP for discoverability, hyper-ontology for semantic mediation, and CDM for analytic harmonization.

2.3 Main changes and refinements in this final version

The final version reported in this deliverable incorporates several important evolutions driven by:

- The definition of the Minimum Data Federation and Interoperability Framework (min-FIF),
- Feedback from node onboarding and mapping exercises (technical pilots).

- The open-call accepted data holder applications.

The main changes include:

1. **Tighter alignment with min-FIF tiers and minimum datasets**:
    - For Tier 1, the EUCAIM DCAT-AP has been refined to respect the mandatory elements of DCAT-AP v3.0 and the latest version of HealthDCAT-AP release 5, (released on 22 September 2025), while adding and reviewing EUCAIM-specific controlled vocabularies derived from the hyper-ontology.
    - For Tier 2, the hyper-ontology has been structured to support the specific query dimensions ("diagnosis", "image modality", "treatment", "body site", etc.) identified as minimum federated query criteria by D4.4, along with other important cancer related categories ("tumor marker tests", "cancer staging" etc.) and imaging related metadata ("sequence types", "injection metadata/contrast agents" etc.).
    - For Tier 3, the CDM explicitly distinguishes between mandatory and optional/conditional elements required for cohort building, supports data minimization (selecting only patients and corresponding data relevant to the data users' objectives) and federated processing, and aligns these with the minimum information specified in D4.4 and D5.6.
2. **More complete terminology binding and mappings** between CDM attributes and hyper-ontology concepts, including explicit links to OMOP domains, FHIR resources and DICOM attributes where relevant.
3. **Expanded coverage of vocabularies and cancer entities** in the hyper-ontology, including additional ICD-O-3 morphological types, diagnostic/treatment procedures and imaging parameters (e.g. semantic annotations on image segmentation masks) that emerged from new data sources and use cases.
4. **Documentation of governance and maintenance**: the final version clarifies how ontology and CDM changes are proposed, reviewed with clinical experts, and adopted.

Overall, the final version transitions the CDM and hyper-ontology from conceptual artefacts to operational components with federated node pilot validation.

3. EUCAIM DCAT-AP for dataset cataloguing

3.1 Final EUCAIM DCAT-AP profile

Our methodology for extending DCAT-AP, as it was described in D5.1, D5.2, D5.6, 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, the initiatives undertaken by the European Network of Cancer Registries (ENCR), the work of other European initiatives, such as the BBMRI-ERIC biobank metadata catalogue, the AI4HI project metadata catalogues, as well as the required data elements from the AI interoperability in imaging White Paper. 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 used the EUCAIM Hyper-ontology specification (described in section 4).

An updated metadata model that complies with the latest version of HealthDCAT-AP is given below in which the general dataset metadata and the domain-specific EUCAIM dataset metadata are outlined.

It is important to mention that extending DCAT for EUCAIM - and creating the so-called EUCAIM DCAT Application Profile - comes with a specific set of requirements that are met for making sure that the profile complies with both DCAT v3.0.0 and the HealthDCAT-AP. As such:

- The mandatory requirements defined in the DCAT-AP and HealthDCAT-AP are respected.
- The controlled vocabularies of the DCAT-AP and HealthDCAT-AP specification are respected.
- Some recommended or optional properties have become mandatory (have stricter semantics)
- New domain-specific controlled vocabularies have been defined for newly added EUCAIM specific properties.

In order to fine-tune the EUCAIM DCAT-AP and, in particular, to decide on the final property cardinalities and ranges, a structured questionnaire was first circulated among WP5 partners and subsequently shared with the wider consortium, including data holders, data users, software developers and other EUCAIM partners. In the initial EUCAIM DCAT-AP specification, some metadata properties had been defined as mandatory, whereas in the current HealthDCAT-AP these same properties are treated with looser semantics. The questionnaire therefore asked for feedback on which metadata fields could realistically sustain stricter cardinalities and where a more relaxed treatment would be appropriate. The outcomes of this questionnaire were used to adjust the set of mandatory, recommended and optional properties in the EUCAIM DCAT-AP, and to validate the introduction of EUCAIM-specific extensions. A summary of the questionnaire, including response statistics and the resulting cardinality decisions for the properties under discussion, is provided in **Annex C** - EUCAIM DCAT-AP Questionnaire results**.**

Table 1 and Table 2 present the final EUCAIM DCAT-AP dataset and distribution specifications, respectively, outlining the DCAT-AP, the HealthDCAT-AP and the EUCAIM-DCAT-AP specific properties. ***All added properties, differences in cardinalities with the HealthDCAT-AP specification, and EUCAIM defined controlled vocabularies are highlighted in*** *red*. The hyper-ontology specification is publicly available through the URI: https://cancerimage.eu/ontology/EUCAIM#. All newly defined EUCAIM properties and EUCAIM controlled vocabularies described in **Table 1**, **Table 2** can be accessed publicly by substituting the prefix "*eucaim:*" with the ontology URI. **Annex D** - EUCAIM DCAT-AP examples also provides a set of examples for each property.

Table 1: *The EUCAIM DCAT-AP* ***Dataset*** *specification.*

	Property	Description	Property IRI	Range	Cardinality	Usage Note
DCAT-AP v3	Title	A clear and concise name given to the dataset.	dct:title	rdfs:Literal	1..n	A title(s) for your Dataset, which can be repeated in multiple languages. The English version is mandatory.
DCAT-AP v3	Description	A detailed description of the dataset's content, purpose, and scope.	dct:description	rdfs:Literal	1..n	This property can be repeated for multiple language versions of the description. The English version is mandatory.

HealthDCAT-AP	Theme	A category of the dataset.	dcat:theme	skos:Concept	1..1	fixed to: http://publications.europa.eu/resource/authority/data-theme/HEAL
DCAT-AP v3	Provenance	Information about how the data was collected, including methodologies, tools, and protocols used.	dct:provenance	dct:ProvenanceStatement	1..n	Information about how the data was created, or processed, including methodologies, tools, and protocols used.
DCAT-AP v3	Keyword	A keyword or tag describing the Dataset.	dcat:keyword	rdfs:Literal	1..n	Add keywords to increase dataset discoverability. You can include keywords in different languages, submitting each keyword as a separate entry. Example: Prostate Cancer, mpMRI.
HealthDCAT-AP	Purpose	The primary objective for which the dataset was created.	dpv:hasPurpose	dpv:Purpose	0..n	One or many categories or sub-categories of the purposes can be chosen from the taxonomy provided by dpv https://w3c.github.io/dpv/2.1/dpv/#vocab-purposes. Example value could be: dpv:ResearchAndDevelopment. OR a free text statement of the purpose of the processing of data or personal data.
HealthDCAT-AP	Access Rights	Information that indicates whether the Dataset is publicly accessible, has access restrictions or is not public.	dct:accessRights	dct:RightsStatement	1..1	Usage note: This property allows you to indicate whether the dataset is publicly accessible, has restrictions, or is not publicly available. This is a mandatory property in HealthDCAT-AP. Use one of the following values :public, :restricted, :non-public, which means: • NON_PUBLIC – The dataset is not publicly available and access is subject to restrictions. • PUBLIC – The dataset is openly accessible to everyone. • RESTRICTED – The dataset has access restrictions but may be available under specific conditions.
HealthDCAT-AP	Health Category	The health category to which this dataset belongs as described in the Commission Regulation on the European Health Data Space laying down a list of categories of electronic data for	healthdcatap:healthCategory	skos:Concept	1..n	This property indicates the EHDS data category the dataset belongs to, based on Article 51 (https://eur-lex.europa.eu/legal-content/EN/TXT/HTML/?uri=OJ:L_202500327#ntr35-L_202500327EN.000101-E0035) of the EHDS Regulation. Use the controlled

		secondary use, Art.51.				vocabulary http://13.81.34.152:1101/resource/authority/healthcategories/ to ensure standardized classification across all datasets
HealthDCAT-AP	Health Theme	A category of the Dataset or tag describing the Dataset.	healthdcatap:healthTheme	skos:Concept	1..n	Usage note: A Dataset may be associated with multiple health themes. By default the value "Cancer" is registered.
DCAT-AP v3	Geographical Coverage	A geographic region that is covered by the Dataset.	dct:spatial	dct:Location	1..n	The EU Vocabularies Name Authority Lists must be used for continents, countries and places that are in those lists; if a particular location is not in one of the mentioned Named Authority Lists, Geonames URIs must be used.
DCAT-AP v3	Applicable Legislation	The legislation that mandates the creation or management of the Dataset.	dcatap:applicableLegislation	rdfs:Resource	1..n	The ELI of the EHDS was published in March 2025 and can now be included as the applicable legislation, the value must include the ELI of the EHDS Regulation (http://data.europa.eu/eli/reg/2025/327/oj). As multiple legislations may apply to the resource the maximum cardinality is not limited.
DCAT-AP v3	Contact Point	Contact information of the individual/managing organization of the Dataset for sending comments about the Dataset.	dcat:contactPoint	vcard:Kind	1..n	Contact information is limited to the contact email and/or the contact page. At least one of the two MUST be provided.
EUCAIM DCAT-AP	Publisher Name	An entity (organization) responsible for making the Dataset available.	dct:publisher	foaf:Organization	1..1	This property allows you to specify the organization responsible for making the dataset available. The publisher is typically the data holder - the entity accountable for ensuring the dataset can be accessed and used. Provide the name of the organization that manages and maintains access to the dataset.
	Publisher Contact Point	Contact information that can be used to contact the Publisher.				A webpage (URL) that either allows to make contact (i.e. a webform) or the email of the organization to get into contact.
HealthDCAT-AP	Publisher Type	A type of the agent that makes the Dataset available.	healthdcatap:publishertype	skos:Concept	0..1	One of: Research Institute, Hospital or Healthcare System Repository, European project, Cancer screening program, Patient association, Data

						altruism organization, ERIC and EDIC.
HealthDCAT-AP	Publisher Note	A description of the publisher activities.	dct:description	rdfs:Literal	0..1	Provide specific details about the health publisher on the dataset. This field can be repeated for different language versions of the description.
EUCAIM DCAT-AP	Type	A type of the Dataset.	dct:type	eucaim:DatasetType and dpv:Data	1..n	One of the EUCAIM defined properties: Original Dataset, Annotated Dataset, Processed Dataset. and values based on the “DPV Data Taxonomy” such as Personal Data, Pseudonymized Data, Anonymized Data, Synthetic Data etc.
HealthDCAT-AP	Minimum Typical Age	The minimum age of subjects within the dataset.	healthdcatap: minTypicalAge	xsd:nonNegativeInteger	0..1	The approximate minimum age of subjects in the dataset, if applicable. Approximate age is given to protect potentially sensitive information of subjects in the dataset.
HealthDCAT-AP	Maximum Typical Age	The maximum age of subjects within the dataset.	healthdcatap: maxTypicalAge	xsd:nonNegativeInteger	0..1	The approximate maximum age of subjects in the dataset, if applicable. Approximate age is given to protect potentially sensitive information of subjects in the dataset.
EUCAIM DCAT-AP	BirthSex	BirthSex of subjects in the dataset.	eucaim:hasBirthSex	eucaim:COM1001396	1..n	EUCAIM Controlled vocabulary: Subclasses of “Sex assigned at birth”
EUCAIM DCAT-AP	Number of Imaging Studies	Total count of Imaging Studies. (e.g., DICOM Studies)	healthdcatap:numberOfRecords maps to eucaim:nbrOfStudies	xsd:nonNegativeInteger	1..1	Use this property to indicate the total number of imaging studies in the dataset, giving users an idea of the dataset’s size and whether it suits their analysis or research needs. A patient might have undergone multiple imaging studies during different timepoints, e.g. diagnosis, treatment, follow up after treatment etc.
EUCAIM DCAT-AP	Number of Unique Individuals	Total count of unique individuals in the dataset.	healthdcatap:numberOfUniqueIndividuals	xsd:nonNegativeInteger	1..1	Use this property to indicate the number of unique individuals represented in the dataset. It helps users understand the dataset’s scale and assess whether it meets their needs for secondary use. Provide the total or an approximate count of distinct individuals whose data is included. If the exact number is unknown, a well-founded estimate is still valuable.

EUCAIM DCAT-AP	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	eucaim:SPEC1000002	1..n	EUCAIM Controlled vocabulary.
DCAT-AP v3	Image Acquisition Period	A temporal period that the dataset covers. This corresponds to the year range that the actual (DICOM) images were created/acquired.	dct:temporal	dct:PeriodOfTime	0..n	This can be extracted from the DICOM acquisition date (0008,0022), if this has not been changed/removed in the anonymization process. If this is not available, an approximation should be added.
EUCAIM DCAT-AP	Cancer Condition	The primary cancer condition of individuals in the dataset.	eucaim:hasCondition	eucaim:MaligantNeoplasticDisease	1..1	EUCAIM controlled vocabulary based on ICD-10 subclasses of “Malignant neoplastic disease”. In case the datasets contain only metastatic subjects without knowing the primary cancer condition, the proper secondary concept code should be added.
EUCAIM DCAT-AP	Image Modality	The set of modalities for the images in the dataset.	eucaim:hasImageModality	eucaim:ImagingModality	1..n	EUCAIM controlled vocabulary based on Radlex: subclasses of “Imaging Modality”
EUCAIM DCAT-AP	Image Equipment Manufacturer	Manufacturer of the imaging device as it is defined in DICOM tag (0008,0070).	eucaim:hasEquipmentManufacturer	eucaim:Manufacturer	1..n	EUCAIM controlled vocabulary based on Birnlex: subclasses of “Manufacturer”
EUCAIM DCAT-AP	Image Body Part/Structure	Anatomical areas captured in the images.	eucaim:hasImageBodyPart	eucaim:BodyStructure	1..n	EUCAIM controlled vocabulary based on ICD-O3: subclasses of “Body structure”
EUCAIM DCAT-AP	Segmentation Label	User-defined or ontology-defined label identifying the segment.	eucaim:hasAnnotationLabel	eucaim:BodyStructure	0..n	EUCAIM controlled vocabulary. For organ annotations subclasses of “Body structure” should be used.
EUCAIM DCAT-AP	Segmentation Method	Type of algorithm used to generate the segment (manual, automatic, semiautomatic)	eucaim:hasAlgorithmType	eucaim:SegmentationMethod	0..n	EUCAIM controlled vocabulary.
EUCAIM DCAT-AP	Number of Segmentations	Total number of annotated patient DICOM studies	eucaim:nbrOfSegmentations	rdfs:Integer	0..1	Each DICOM segmented study counts as one segmentation.

EUCAIM DCAT-AP	Identifier	A unique identifier for the dataset, i.e. the URI in the context of the EUCAIM Public Catalogue. ((in compliance with the findability aspect of the FAIR principles))	dct:identifier	rdfs:Literal	1..1	For Health DCAT Application Profile the identifier is mandatory and its value must be unique from the origin database of the Dataset. In the process of dataset transmission to the next database (EUCAIM), the 'identifier' property is overwritten by the new database identifier, and the original identifier becomes the value of the 'other identifier' property.
EUCAIM DCAT-AP	Version	The version of the dataset.	dcat:version	rdfs:Literal	1..1	in SemVer or CalVer format
EUCAIM DCAT-AP	Interoperability Tier	The EUCAIM data federation and interoperability tier the specific dataset belongs to.	adms:interoperabilityLevel	eucaim:SPEC1000008	1..1	One of “Tier 1”, “Tier 2”, “Tier 3”, “Tier 1A+”, “Tier 1C+”, “Tier 2A+”,”Tier 2C+”, “Tier 3A+”, “Tier3C+“.
DCAT-AP v3	Language	A language of the Dataset.	dct:language	dct:LinguisticSystem	0..n	For this property the values from the EU Vocabularies Languages Named Authority List must be used. If your Dataset contains multiple languages, this property can be repeated.
HealthDCAT-AP	Population Coverage	A definition of the population within the dataset.	healthdcatap:populationCoverage	rdfs:Literal	0..n	This field is a free text description of the population covered in the dataset. For example, “Patients between 35 and 87 years old with prostate cancer treated with prostatectomy in hospitals in France between 2018 and 2023.”
HealthDCAT-AP	Personal Data	Key elements that represent an individual in the dataset, Indicates association with Personal Data.	dpv:hasPersonalData	dpv:PersonalData	0..n	The different types of personal information that are collected in the dataset (real data not the metadata record) can be indicated with this property. Values can be picked from the dpv taxonomy https://w3c.github.io/dpv/2.0/pd/. For example: dpv-pd:Gender.
DCAT-AP v3	Temporal Resolution	The minimum time period resolvable in the dataset.	dcat:temporalResolution	xsd:duration	0..1	If the dataset is a time-series, this should correspond to the spacing of items in the series. For other kinds of dataset, this property will usually indicate the smallest time difference between items in the dataset. The time period has to be provided in the xsd:duration format.

DCAT-AP v3	Frequency	The frequency at which the Dataset is updated.	dct:accrualPeriodicity	dcterms:Frequency	0..1	The value of this property should be the IRI from the listed controlled vocabulary indicating the frequency at which the dataset is updated. For example: http://publications.europa.eu/resource/authority/frequency/WEEKLY
HealthDCAT-AP	Legal Basis	Legal basis used to justify processing of data or use of technology in accordance with a law.	dpv:hasLegalBasis	dpv:LegalBasis	0..n	The legal basis can be provided as a value from the dpv taxonomy https://w3c.github.io/dpv/2.0/dpv/modules/legal_basis.html#vocab-legal-basis While the applicable legislation indicates which legislation mandates the publication of the dataset, the legal basis property described the legal basis for initial collection and processing of (personal) data. Example value for this property could be: dpv:Consent
HealthDCAT-AP	Retention Period	A temporal period for which the dataset is available for secondary use.	healthdcatap:retentionPeriod	xsd:duration	0..1	This property makes use of the class dct:PeriodOfTime (more details on PeriodOfTime class tab). The start and end of the interval SHOULD be given by using properties dcat:startDate or time:hasBeginning, and dcat:endDate or time:hasEnd, respectively. The interval can also be open - i.e., it can have just a start or just an end
DCAT-AP v3	Conforms To	An established standard to which the described resource conforms.	dct:conformsTo	dct:Standard	0..n	If your data conforms to an established standard or specification, use this property to indicate which one.
HealthDCAT-AP	Coding System	Coding systems in use (ex: ICD-10-CM, DGRs, SNOMED-CT, ...).	healthdcatap:hasCodingSystem	dct:Standard	0..n	This property allows you to specify the coding systems used within your dataset. For example, if a dataset uses ICD-10 for disease classification, this property allows data users to search for datasets with the same coding system. As a machine-actionable property, it also facilitates automated processes, making dataset discovery more efficient. Specify the standardized coding systems used in the

						dataset (e.g. ICD-10-CM, SNOMED CT, DRGs).
HealthDCAT-AP	Code Values	Health classifications, or other classification-like systems and their codes that can be associated with the dataset.	healthdcatap:hasCodeValues	skos:Concept	0..n	A dataset may be associated with multiple health classification and/or classification-like systems such as ontologies, terminologies or thesauri and their codes. Inside this property you can provide the coding system of the dataset.
DCAT-AP v3	Related Resource	A related resource	dct:relation	rdfs:Resource	0..n	This property allows you to indicate a general link between this dataset and another resource (e.g., another dataset, document, or service). Use it when there is a connection, but none of the more specific relation properties below apply. - This is the most generic way to describe a connection. - Use it when you want to indicate a related resource but cannot define the relationship more precisely. When the relationship is known and semantically clear , use one of the more specific properties instead
DCAT-AP v3	Is Referenced By	A related resource, such as a publication, that references, cites, or otherwise points to the dataset.	dct:isReferencedBy	rdfs:Resource	0..n	This property identifies publications, articles, studies, websites that include in their content a reference to the Dataset. For example: If there is a website that has a reference to this Dataset, then the URL to the website can be added in this property "is referenced by".
DCAT-AP v3	Landing Page	A web page that provides access to the Dataset, its Distributions and/or additional information	dcat:landingPage	foaf:Document	0..n	It is intended to point to a landing page at the original data holder, not to a page on a site of a third party, such as an aggregator.
DCAT-AP v3	Documentation	A page or document about this Dataset.	foaf:page	foaf:Document	0..n	Any additional source of information about the Dataset.
HealthDCAT-AP	Sample	A sample distribution of the dataset.	adms:sample	dcat:Distribution	0..n	This property makes use of the dcat:Distribution class to describe a sample distribution of the dataset, which can be anonymized or synthetic data,

						or the data dictionary provided in CSVW format.
HealthDCAT-AP	Analytics	An analytics distribution of the dataset.	healthdcatap:analytics	dcat:Distribution	0..n	Publishers are encouraged to provide URLs pointing to document repositories where users can access or request associated resources such as technical reports of the dataset, quality measurements, usability indicators, etc.
HealthDCAT-AP	Quality Annotation	A statement related to quality of the Dataset, including rating, quality certificate, feedback that can be associated to the dataset.	dpv:hasQualityAnnotation	dqv:QualityCertificate	0..n	This property makes use of the class (dqv:QualityCertificate) from Data Quality Vocabulary, in which it is indicated the quality certificate, linked to the described dataset. In EUCAIM the dpv:inDimension is used for specifying which specific technical dimension of quality has been assessed. For example, for the integrity dimension, when the “DICOM Integrity Checker Tool” is used and has successfully assessed the integrity of the DICOM images, the appropriate certificate will be used.
DCAT-AP v3	Creator Name	An entity responsible for producing the dataset.	dct:creator	foaf:Agent	0..n	This property points to an Agent who plays the creation role of the dataset. This property can be repeated for different versions of the name (e.g. the name in different languages).
	Creator Contact Point	Contact information that can be used to contact the creator. A webpage that either allows to make contact (i.e. a webform) or the email of the agent to get into contact.				This property points to a contact point (Individual, Organization, Location, Group) that can answer questions about the dataset. Details on how to describe these are provided under class vcard:Kind.
	Creator Type	A type of the agent that produced the dataset	dct:type	skos:Concept	0..1	If the agent type exists in the EU corporate bodies NAL the corresponding entry should be used. If the Agent represents a publisher of the Dataset the new publisher type controlled vocabulary MUST be used.
	Creator Note	A description of the creator's activities	dct:description	rdfs:Literal	0..1	A free-text description of the Creator activities

DCAT-AP v3	Was Generated By	An activity that generated, or provides the business context for, the creation of the dataset.	prov:wasGeneratedBy	prov:Activity	0..n	The activity associated with generation of a dataset will typically be an initiative, project, mission, survey, on-going activity etc. Multiple prov:wasGeneratedBy properties can be used to indicate the dataset production context at various levels of granularity.
DCAT-AP v3	Spatial Resolution	The minimum spatial separation resolvable in a dataset, measured in meters.	dcat:spatialResolutionInMeters	xsd:decimal	0..1	If the dataset is an image or grid this should correspond to the spacing of items. For other kinds of spatial datasets, this property will usually indicate the smallest distance between items in the dataset.
DCAT-AP v3	Qualified Attribution Agent Name	A name of the Qualified Attribution Agent. Used to link to an Agent where the nature of the relationship is known but does not match one of the standard [DCTERMS] properties (dct:creator, dct:publisher, healthdcatap:hdab). Use dcat:hadRole on the prov:Attribution to capture the responsibility of the Agent with respect to the Resource.	prov:qualifiedAttribution	prov:Attribution	0..n	This property can be repeated for different versions of the name (e.g. the name in different languages).
	Qualified Attribution Agent Contact Point	Contact information that can be used to contact the Qualified Attribution Agent. A webpage that either allows to make contact (i.e. a webform) or the email of the agent to get into contact.				
	Qualified Attribution Agent Role	The function of an entity or agent with respect to another entity or resource.				Choose one of the roles as listed in the controlled vocabulary. Note that for HealthDCAT-AP, the list of roles might be extended in the future. Example: https://standards.iso.org/iso/19115/resources/Codelists/gml/CI_RoleCode.xml#processor

DCAT-AP v3	Other Identifier	A secondary identifier of the Dataset, such as MAST/ADS17, DataCite18, DOI19, EZID20 or W3ID21.	adms:identifier	adms:Identifier	0..n	Examples for secondary identifiers are MAST/ADS, DataCite, DOI, EZID or W3ID (if not used for the original identifier). This property makes use of another, small class: adms:Identifier, where you provide the identifier and the name of the identifier schema (e.g. DOI). This property will hold the original 'identifier' of the dataset provided during creation.
DCAT-AP v3	Version Notes	A description of the differences between this version and a previous version of the Asset.	adms:versionNotes	rdfs:Literal	0..n	Provide a short description of changes made to the dataset from the previous version.
DCAT-AP v3	Release Date	The date of formal issuance (e.g.: publication) of the Dataset.	dct:issued	rdfs:TemporalLiteral	0..1	The values must be data typed as either xsd:date, xsd:dateTime, xsd:gYear or xsd:gYearMonth Example: 2023-12-10T13:16:10.246Z.
DCAT-AP v3	Modification Date	The most recent date on which the Dataset was changed or modified.	dct:modified	rdfs:TemporalLiteral	0..1	The values must be data typed as either xsd:date, xsd:dateTime, xsd:gYear or xsd:gYearMonth.

Table 2: *The EUCAIM DCAT-AP* ***Distribution*** *specification.*

	Property	Description	Property IRI	Range	Cardinality	Usage Note
DCAT-AP v3	Access URL	A URL of the resource that gives access to a distribution of the dataset. E.g., landing page, feed, SPARQL endpoint.	dcat:accessURL	rdfs:Resource	1..n	Access URL is used to indicate the URL of a service or location that can provide access to this distribution, typically through a Web form, query or API call. The resource at the access URL may contain information about how to get the Dataset. In EUCAIM, this is the URL of the negotiator service for the specific dataset.
DCAT-AP v3	Applicable Legislation	The legislation that mandates the creation or management of the Distribution.	dcatap:applicableLegislation	rdfs:Resource	1..n	The value must include the ELI of the EHDS Regulation. Multiple legislations may apply to the dataset.
DCAT-AP v3	Description	A free-text account of the Distribution.	dct:description	rdfs:Literal	0..n	Provide detailed information about the distribution's content, format characteristics, intended use cases, and any special considerations for health data usage.

DCAT-AP v3	Format	The file format of the Distribution included in the Dataset.	dct:format	dct:MediaTypeOrExtent (IANA Media Types)	0..1	Specify the technical format using standardized media types (e.g., CSV, JSON, XML, FHIR) to indicate data structure and required processing tools. Imaging data: the imaging format of the images in your dataset (e.g. DICOM, Nifti), Annotation data: the format of the annotations (e.g. DICOM-SEG, Nifti), if available. Clinical data: the format of the available clinical data (e.g. CSV, XLS, JSON, parquet). Note for each of these different distributions should be added.
DCAT-AP v3	License	A licence under which the Distribution is made available.	dct:license	dct:LicenseDocument	0..1	Reference the legal terms and conditions governing access, use, and redistribution of this distribution, ensuring compliance with health data regulations. For interoperability, it is recommended to use canonical IRIs of well-known licenses such as those defined by Creative Commons.
DCAT-AP v3	Title	A name given to the Distribution.	dct:title	rdfs:Literal	0..*	Provide a descriptive name that clearly identifies this specific distribution format, version, or subset to distinguish it from other distributions of the same dataset.
DCAT-AP v3	Access Service	A data service that gives access to the distribution of the dataset.	dcat:accessService	dcat:DataService	1..1	Indicate the specific data service through which this distribution can be accessed programmatically.
DCAT-AP v3	Availability	An indication how long it is planned to keep the Distribution of the Dataset available.	dcatap:availability	skos:Concept	0..1	Specify the planned duration of availability using controlled vocabulary terms that indicate retention policies and access guarantees. The NAL planned availability from publication office must be used.
EUCAIM DCAT-AP	Image Size (in GB)	The size of a Distribution in bytes.	dcat:byteSize	xsd:decimal	0..1	Describes the size of the distribution (the actual files) in bytes, and is therefore expressed as a non-negative integer. If the actual size is not know, it can be estimated. Provide the exact file size in bytes to help users understand storage requirements and

						download expectations if applicable.
	Checksum	A mechanism that can be used to verify that the contents of a distribution have not changed.	spdx:checksumValue	xsd:hexBinary	0..1	The checksum is related to the downloadURL. This property makes use of the spdx:Checksum class, which itself has two properties to indicate checksum algorithm and checksum value (see Checksum class for further details). Provide cryptographic hash values (e.g., MD5, SHA-256) to enable verification of file integrity and detect any unauthorized modifications.
	Checksum Algorithm		spdx:algorithm	spdx:ChecksumAlgorithm		
DCAT-AP v3	Compression Format	The format of the file in which the data is contained in a compressed form, e.g. to reduce the size of the downloadable file.	dcat:compressFormat	dct:MediaType	0..1	Specify compression algorithms used (e.g., gzip, zip, bzip2) using IANA media types to inform users about decompression requirements.
DCAT-AP v3	Documentation	A page or document about this Distribution.	foaf:page	foaf:Document	0..n	Reference technical documentation, data dictionaries, schema files, or usage guides that help users understand and properly utilize this distribution.
DCAT-AP v3	Download URL	A URL that is a direct link to a downloadable file in a given format. E.g., CSV file or RDF file. The format is indicated by the distribution's dcterms:format and/or dcat:mediaType.	dcat:downloadURL	rdfs:Resource	0..n	Download URL should be used for the URL at which this distribution is available directly, typically through a HTTP Get request.
DCAT-AP v3	Has Policy	The policy expressing the rights associated with the distribution if using the ODRL vocabulary.	odrl:hasPolicy	odrl:Policy	0..1	Information about rights expressed as an ODRL policy [ODRL-MODEL] using the ODRL vocabulary [ODRL-VOCAB] MAY be provided for the distribution. Define machine-readable usage policies using ODRL vocabulary to specify permissions, prohibitions, and duties related to health data access and use.
DCAT-AP v3	Language	A language used in the Distribution.	dct:language	dct:LinguisticSystem	0..n	The language of the Dataset. For this property, the values from the EU Vocabularies Languages Named Authority List must be used. If your

						Dataset contains multiple languages, this property can be repeated.
DCAT-AP v3	Linked Schemas	An established schema to which the described Distribution conforms.	dct:conformsTo	dct:Standard	0..n	This property SHOULD be used to indicate the model, schema, ontology, view or profile that this representation of a dataset conforms to, in a machine-readable form. Reference technical standards, data models, or schema specifications (e.g., HL7 FHIR, OMOP CDM) that define the structure and semantics of this distribution.
DCAT-AP v3	Media Type	The media type of the Distribution as defined in the official register of media types managed by IANA.	dcat:mediaType	dct:MediaType	0..1	It SHOULD be expressed using a media type as defined in the official register of media types managed by IANA. Example: https://www.iana.org/assignments/media-types/text/csv (for csv) If IANA media types do not sufficiently describe the format, use "format" to describe it in more detail. Specify the precise IANA media type (e.g., application/json, text/csv) to enable proper content handling by consuming applications.
DCAT-AP v3	Modification Date	The most recent date on which the Distribution was changed or modified.	dct:modified	rdfs:TemporalLiteral	0..1	Record the timestamp when the distribution was last updated to help users assess data currency and track version history.
DCAT-AP v3	Packaging Format	The format of the file in which one or more data files are grouped together, e.g. to enable a set of related files to be downloaded together.	dcat:packageFormat	dct:MediaType	0..1	Specify container formats (e.g., tar, zip, 7z) used to bundle multiple files together, using IANA media types for format identification. It SHOULD be expressed using a media type as defined in the official register of media types managed by IANA.
DCAT-AP v3	Release Date	The date of formal issuance (e.g., publication) of the Distribution.	dct:issued	rdfs:TemporalLiteral	0..1	Document when this distribution was first published or made available to establish temporal context and version tracking.
DCAT-AP v3	Rights	A statement that specifies rights associated with the Distribution.	dct:rights	dct:RightsStatement	0..1	Detail intellectual property rights, usage restrictions, and access permissions governing this distribution,

						complementing license information.
DCAT-AP v3	Spatial Resolution	The minimum spatial separation resolvable in a dataset distribution, measured in meters.	dcat:spatialResolutionInMeters	xsd:decimal	0..1	The NAL planned availability from publication office must be used.
DCAT-AP v3	Status	The status of the distribution in the context of the maturity lifecycle.	adms:status	skos:Concept	0..1	It MUST take one of the values Completed, Deprecated, Under Development, Withdrawn from the provided controlled vocabulary.
DCAT-AP v3	Temporal Resolution	The minimum time period resolvable in the dataset distribution.	dct:temporal	xsd:Duration	0..1	Specify the finest temporal granularity of data points using ISO 8601 duration format (e.g., daily, monthly, yearly data collection intervals).
EUCAIM DCAT-AP	accessConditions	A statement about the conditions of access and usage of the dataset.	dct:rights	eucaim: Dataset Access Condition	1..1	fixed to a predefined set of values: “Authorization to download the datasets” “Authorization to access, view and process in-situ the datasets” “Authorization to remotely process the datasets without the ability to access and visualise data, even remotely.”

4. The EUCAIM Hyper-Ontology

4.1 Introduction

The EUCAIM hyper-ontology is a common semantic meta-model developed to support and maintain semantic interoperability among heterogeneous cancer image data models/standards. The ontology model defines a structured, controlled vocabulary that enables disparate, heterogeneous data models/standards to communicate and integrate unambiguously. EUCAIM's hyper-ontology, formalized using OWL (Ontology Web Language), a W3C standard for knowledge representation, serves as a domain ontology that reflects the essentials of oncology across clinical/biological and medical imaging contexts, and as an application ontology that semantically supports various EUCAIM components, including the EUCAIM CDM, HealthDCAT-AP, the ETL process, federated querying and processing and image annotation/segmentation. Primarily, the hyper-ontology is developed from multiple datasets provided by AI4HI, which use heterogeneous healthcare and imaging standards (OMOP/FHIR/DICOM), overlapping terminologies and vocabularies, and non-standardized information. The main cancer types initially covered in the hyper-ontology are: breast, prostate, liver, lung, colon, and rectum.
Furthermore, the ontology model is extended and enriched with new cancer types from datasets acquired during onboarding. These cancer types include thyroid, brain, ovary, bladder, kidney, and pancreas. Developing and extending the hyper-ontology are challenging processes due to the domain complexity, data heterogeneity, and disparity. Additionally, data ambiguity or incompleteness, different interpretations of data, and rapid knowledge evolution constitute the main conceptual barriers throughout the building and enrichment phases, necessitating continuous expert intervention and validation. Thereby, the hyper-ontology must harmonize heterogeneous data representations and interpretations, ensure reusability and extensibility, maintain evolving knowledge, and ensure practical integration with other EUCAIM components, all while preserving accurate semantics.

4.2 Methods

The hyper-ontology is developed using a well-founded, iterative, and hybrid approach that simplifies the building process while addressing the complexity and heterogeneity of the application domain and the diversity and disparity of the clinical/biological and imaging knowledge. In the ontology design, the ontological structure is organized into layers (domain-specific, domain, core, and upper) and modules (clinical & biological, imaging, and common). This structure facilitates the building process and supports the ontology reusability and extension. The development process is composed of two main strategies: top-down and bottom-up.

- **Top-down**: to ground the ontology model on the mCODE specifications by applying high-level conceptual modeling language OntoUML, whose modeling types are based on the ontological distinctions of the Unified Foundational Ontology. This strategy has helped to clarify conceptually and explicitly represent the semantics of the oncology domain, including the main categories or aspects considered in mCODE, such as primary and secondary cancer conditions, cancer staging and grading, tumor marker tests, tumor, health assessment methods, and cancer-related procedures (surgical,

medical administration, radiotherapy), as well as their associated connections. The top-down strategy mainly supports the core and upper layers.

- **Bottom-up**: to represent the domain-specific knowledge, including the clinical/biological and imaging data and metadata provided by AI4HI and the new dataset collections, the semantic relations, as well as the associated semantic mappings across standards. All provided information (standard or non-standard) has been documented in the Ontology Requirements and Specifications Document (ORSD) using domain-specific Competency Questions (CQs) and answers, classified by cancer type. ORSD, which helps organize data and identify incomplete or overlapping information or sources, has supported the hyper-ontology revision and evaluation. To resolve data heterogeneity and ensure harmonization at the domain and domain-specific levels, various semantic or mapping patterns have been designed and incorporated into the ontology under the supervision of clinical experts. These patterns have helped to semantically and explicitly represent non-standard, ambiguous, or incomplete data, and to classify them with high accuracy. Additionally, applying semantic patterns has supported mapping heterogeneous terms that differ only at the syntactic level (e.g., different labels) but are semantically equivalent, thereby aligning multiple terminologies.

The hyper-ontology has been evaluated and validated by experts. Additionally, to ensure ontology sustainability, a versioning strategy has been implemented, publishing multiple ontology versions on Zenodo with persistent universal identifiers (DOIs). The ontology development process, evaluation, and validation are presented in Deliverable D5.2.

4.3 Classes and Properties

The EUCAIM hyper-ontology version 2.0 (December 2025) defines more than 3,900 classes, 170 object properties, 60 data properties, and 55 annotation properties (**Figure 1**). Classes are represented at varying levels of granularity, from the most generic to the most domain-specific. Additionally, they are classified into modules based on thematic categories, covering the following areas: Clinical and Biological, Imaging, and Common (primarily qualifiers). Two additional modules, Generic and Specific, have been defined to ensure the semantic integration of the hyper-ontology with other EUCAIM components, including the Common Data Model and Health DCAT-AP. The hyper-ontology classes and properties are documented and publicly accessible via the ontology URI: https://cancerimage.eu/ontology/EUCAIM#

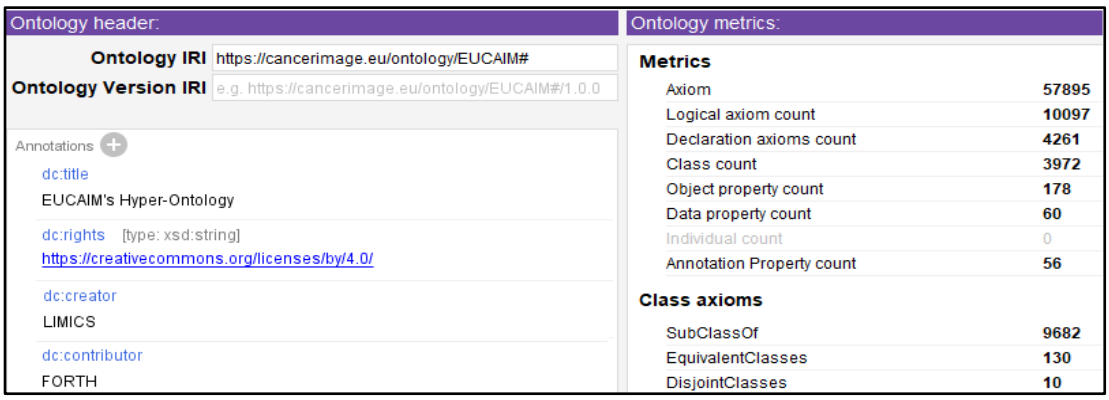


Figure 1: *Excerpt of the hyper-ontology header and metrics.*

4.3.1 Classes

This section presents selected classes and concepts from the various modules, highlighting the essentials of oncology and cancer imaging. In the hyper-ontology, classes are codified using EUCAIM codes, with the ontological module to which the class belongs as a prefix (CLIN, IMG, COM, GEN, SPEC). The hyper-ontology classes are represented using standard concepts from standard terminologies and vocabularies, including ICD-10, ICD-O-3, SNOMED, LOINC, NAACCR, Cancer Modifier, OMOP Genomic, UCUM, RxNorm, NCIT, and RadLex. Classes that are not defined in standard terminological resources are standardized and codified using EUCAIM codes.

Clinical and Biological Module

This module encompasses classes that are significant in the oncology domain, particularly in clinical and biological contexts, such as primary and secondary cancer conditions, histology/morphology, body parts, staging and grading methods/systems and values, surgical and therapeutic procedures, laboratory and tumor marker tests, medication, clinical findings, and comorbidities.

Table 3: *Selected classes from the Clinical and Biological Module, their standard sources, and mappings with external sources.*

Category	Standard source	Mappings (direct)	Examples of classes (source code, EUCAIM ID)
Cancer types	ICD-10	SNOMED, ICD-O-3, NCIT	Malignant neoplasm of breast (C50, CLIN1000060) Malignant neoplasm of rectum (C20, CLIN1000077) Malignant neoplasm of uterus (C55, CLIN1003316) Malignant neoplasm of prostate (C61, CLIN1000075) Malignant neoplasm of brain (C71, CLIN1007976) Malignant neoplasm of gallbladder (C23, CLIN1000089) Malignant neoplasm of liver (C22, CLIN1000071) Malignant tumor of kidney (C64, CLIN1000180) Malignant neoplasm of colon (C18, CLIN1000057) Malignant neoplasm of thyroid gland (C73, CLIN1059392)
Cancer subtypes	SNOMED	NCIT, ICD-O-3	Carcinoma of breast (254838004, CLIN1000169) HER2-Positive (427685000, CLIN1046795) ER positive (416053008, CLIN1046663) Glioblastoma multiforme of brain (276828006, CLIN1007993) Malignant tumor of hepatic flexure (363407001, CLIN1053460) Adenocarcinoma of uterus (309245001, CLIN1003318)
Histology and Morphology	ICD-O-3	SNOMED	Neoplasm, malignant (8000/3, CLIN1049561) Adenocarcinoma, NOS (8140/3, CLIN1047138) Papillary adenocarcinoma, NOS (8260/3, CLIN1049465) Follicular carcinoma, NOS (8330/3, CLIN1049604) Adenosquamous carcinoma (8560/3, CLIN1049457) Lobular carcinoma, NOS (8520/3, CLIN1052212) Large cell carcinoma, NOS (8012/3, CLIN1063535) Malignant lymphoma, NOS (9590/3, CLIN1049661)

Topography	ICD-O-3	SNOMED, NCIT, RadLex	Prostate (C61, BP1000021) Brain (C71, BP1000051) Heart (C38.0, BP1000075) Liver (C22.0, CLIN1063728) Gallbladder (C23, BP1000142) Kidney (C64, BP1000298) Pancreas (C25, BP1000209) Colon (C18, CLIN1063722) Breast (C50, CLIN1063727) Thyroid (C73.9, BP1000286) Vagina (C52, BP1000299) Bladder (C67, BP1000301) Stomach (C16, BP1000302) Esophagus (C15, BP1000303) Rectum (C20, CLIN1063724)
Staging and grading systems	SNOMED		American Urological Association staging system for prostate cancer (254381008, CLIN1044305) FIGO staging system of gynecological malignancy (254383006, CLIN1044306) American Joint Committee on Cancer, Cancer Staging Manual, 8th edition neoplasm staging system (897275008, CLIN1044309) TNM tumor staging classifications (258234001, CLIN1033314)
			Nottingham histologic grading system (449205006, CLIN1037299) Gleason grading system for prostatic cancer (106241006, CLIN1037295) WHO CNS tumor grading system (277460003, CLIN1037293) Histological grading systems (277457005, CLIN1037296) International Society of Urological Pathology prostate cancer staging system (1812481000004109, CLIN1037300) Gleason scoring system for malignant neoplasm of prostate (897202005, CLIN1037294)
	NAACCR		TNM Path M (900, CLIN1033404) TNM Path N (890, CLIN1033068) TNM Path T (880, CLIN1033007) TNM Clin M (960, CLIN1031282) TNM Clin N (950, CLIN1031528) TNM Clin T (940, CLIN1031297) Tumor Regression Grade (rectum@2920, CLIN1049083)
	NCIT		Miller and Payne Classification (C63611, CLIN1049086) Modified Ryan Scheme for Tumor Regression (C155939, CLIN1049085)
Staging and grading values	Cancer Modifier		Stage 1 (Stage-1, CLIN1000422) AJCC/UICC 8th Stage 1 (8th_AJCC/UICC-Stage-1, CLIN1064607) FIGO Stage 1 (FIGO-1, CLIN1000425)
			AJCC/UICC 8th clinical M1a Category (c-8th_AJCC/UICC-M1a, CLIN1071413) AJCC/UICC 8th pathological M1a Category (p-8th_AJCC/UICC-M1a, CLIN1079718) Nottingham Grade 1, (Nottingham-Grade-1, CLIN1022114)

			Primary Gleason Pattern 1, (OMOP4998758, CLIN1031198) Secondary Gleason Pattern 1, (OMOP4997849, CLIN1031205) FIGO grade 1 (FIGO-Grade-1, CLIN1022115)
Performance status assessment	LOINC	SNOMED, NCIT	American society of anesthesiologists morbidity state (97816-3, CLIN1047416) ECOG Performance Status score (89247-1, CLIN1004580) Karnofsky performance status (89243-0, CLIN1047415)
Surgical and therapeutic procedures	SNOMED	NCIT	Radical mastectomy (384723003, CLIN1004693) Simple mastectomy (172043006, CLIN1060362) Total pneumonectomy (49795001, CLIN1060443) Partial lobectomy of brain (67402009, CLIN1004594) Lumpectomy of breast (392021009, CLIN1060316) Rectal polypectomy (274031008, CLIN1060376) Prostatectomy (90470006, CLIN1000248) Radical Perineal prostatectomy (8782006, CLIN1030358) Sigmoid colectomy (84604002, CLIN1060281)
			Radiotherapy (1287742003, CLIN1014933) Thermotherapy (261570006, CLIN1014761) Chemotherapy (367336001, CLIN1024528) Hormone therapy (169413002, CLIN1016082) Combined chemotherapy and radiation therapy (703423002, CLIN1035091)
Tumor marker tests	LOINC	NCIT	Estrogen receptor Ag [Presence] in Breast cancer specimen by Immune stain (85337-4, CLIN1045815) Progesterone receptor Ag [Presence] in Breast cancer specimen by Immune stain (85339-0, CLIN1046085) Prostate specific Ag [Mass/volume] in Serum or Plasma (2857-1, CLIN1033410)
	OMOP Genomic		ATRX (ATRX chromatin remodeler) gene variant measurement (886, CLIN1051954) IDH1 (isocitrate dehydrogenase (NADP(+)) 1) gene variant measurement (5382, CLIN1051952) TERT (telomerase reverse transcriptase) gene variant measurement (11730, CLIN1051956) TP53 (tumor protein p53) gene variant measurement (11998, CLIN1051958)

Imaging Module

This module includes classes that are significant in the context of medical imaging, with a focus on cancer imaging, including modalities, procedures, associated attributes, findings, descriptors, assessment methods and values, and manufacturers.

Table 4: *Selected classes from the Imaging Module, their standard sources, and mappings with external sources.*

Category	**Standard source**	**Mappings (direct)**	**Examples of classes (source code, EUCAIM ID)**
Imaging modalities	RadLex	SNOMED	Magnetic resonance imaging (RID10312, IMG1000038) Magnetic resonance spectroscopy (RID10315, IMG1000083) Computed tomography (RID10321, IMG1000042)

			Ultrasound (RID10326, IMG1000035) Nuclear medicine imaging (RID10330, IMG1000071) PET-CT (RID10341, IMG1004451) PET-MR (RID10342, IMG1004452) Molecular imaging (RID12508, IMG1000073) Tomography (RID28840, IMG1000070) Scintigraphy (RID34428, IMG1000079)
	NCIT		Microscopy (C16853, IMG1004457) Digital Microscopy (C190537, IMG1004458) Slide Microscopy (C190609, IMG1004456)
	EUCAIM		Digitized hematoxylin-eosin (HE) slides (IMG1004463) Photon Counting Computed Tomography (IMG1000085) Low-Dose Computed Tomography (IMG1000086)
Imaging procedures	SNOMED	NCIT	Magnetic resonance imaging (113091000, IMG1000022) Multiparametric magnetic resonance imaging (1144760005, IMG1005606) Magnetic resonance imaging of breast for screening for malignant neoplasm procedure (609223006, IMG1016600) MRI-US fusion guided prostate biopsy (787153004, IMG1016128) Ultrasonography of thorax (25850001, IMG1016135) CT of thorax with contrast (75385009, IMG1000076) Pelvic echography (24848001, IMG1005543) Multiparametric MRI of prostate (719178004, IMG1016133) MRI of breast for screening (711540006, IMG1016199) MRI of brain with arterial spin labeling (720705001, IMG1016598)
Imaging Assessment	RadLex		BI-RADS 1, (RID36028, IMG1005469) BI-RADS 2, (RID36029, IMG1005468) LI-RADS 2, (RID49795, IMG1005505) LI-RADS 3, (RID49796, IMG1005504) LI-RADS Treated, (RID49804, IMG1005508) PI-RADS 3, PI-RADS 3 - Intermediate, (RID50291, IMG1005477) PI-RADS 1 - Very low (Lesion), (RID50296, IMG1005487) PI-RADS 4 - High (Lesion), (RID50299, IMG1005484) PI-RADS 5 - Very high (Lesion), (RID50300, IMG1005483)
Imaging descriptors	RadLex		Enhancement pattern (RID6058, IMG1005498) Microlobulated margin (RID5712, IMG1005421) Nonenhancing (RID6056, IMG1005496) Heterogenously echogenic (RID28488, IMG1016628) No posterior acoustic features (RID34363, IMG1016616) Posterior acoustic enhancement (RID34364, IMG1016615) High density (RID34240, IMG1016691) Circumscribed margin (RID5707, IMG1005419) Non-circumscribed margin (RID34355, IMG1005429)
Laterality	RadLex	SNOMED, LOINC	Left (RID5824, IMG1016670) Right (RID5825, IMG1016682) Bilateral (RID5771, IMG1016659) Unilateral (RID38593, IMG1016695)
Manufacturers	birnlex		Siemens (3066, IMG1000044) Philips (3065, IMG1000046) General Electric (12833, IMG1000047)
	EUCAIM		I.M.S (IMG1000048) MiE (IMG1000058) Mediso (IMG1000056) Marconi (IMG1000053)

			Toshiba (IMG1000045) Agfa (IMG1000051) UIH (IMG1000063) Hologic (IMG1000064)

Common Module

This module defines entities that are common, i.e., reusable in the clinical/biological and imaging contexts, including cancer patient, sex/gender, qualifier values, such as temporal and adjectival modifiers, units of measure and time, answer set values, and disease, absence/presence, and positive/negative qualifiers. Additionally, this module includes staging and grading values, considering their semantic classification as qualifier values.

Table 5: *Selected classes from the Common Module, their standard sources, and mappings with external sources.*

Category	Standard source	Mapping (direct)	Examples of classes (source code, EUCAIM ID)
Sex and Gender	LOINC	NAACCR, Gender	Female (LA3-6, COM1001370) Male (LA2-8, COM1001366)
Qualifiers for Segmentation/Annotation (e.g., annotation type, segmentation method)	SNOMED	LOINC, NCIT	Contouring (367558000, COM1000089) Automatic (8359006, COM1000008) Manual (87982008, COM1000003)
	EUCAIM		Bounding box (COM1000118) Mask (COM1000139) Mesh (COM1000138)
Qualifiers for Answer or Value set (e.g., clinical status, performance status, risk assessment, treatment response)	LOINC	SNOMED	Improved (LA65-8, COM1002075) No evidence of disease (LA32281-0, COM1001896) Pregnant (LA15173-0, COM1001161) Complete response (LA28366-5, COM1001312) ASA1 (no disturbance) (LA32129-1, COM1001197)
	SNOMED	LOINC	Active (55561003, COM1001177) Inactive (73425007, COM1001173) High risk (723509005, COM1000137) Very high (260360000, COM1002064)
Finding Values (e.g., absence/presence) and values specific to medical imaging (e.g., contrast enhancement) or genetic data (e.g., mutation, methylation)	LOINC	SNOMED	Absent (LA9634-2, COM1001886) No (LA32-8, COM1001188) Not assessed (LA9348-9, COM1001188) Not available (LA7338-2, COM1001163)
	NAACCR		Methylated (3889@3, COM1001314) Partial Methylated (3889@1, COM1001316) Not methylated (3889@0, COM1001889)
	SNOMED	LOINC	Positive (10828004, COM1001310) Negative (260385009, COM1001332)
	EUCAIM		No contrast injected (COM1001892)

			No contrast enhancement (COM1001890) No ADC images (COM1001893)
	NCIT		Mutated (C177686, COM1001313) Mutated/c.1-146C>T (C131106, COM1001318) Mutated/c.1-124C>T (C131107, COM1001317)
Units of measure and time, and radiation dose	UCUM	SNOMED	Centimeter (cm, COM1001954) enzyme unit per liter (U/L, COM1001956) Milliliter (mL, COM1001966) Percent (%, COM1000161) Day (d, COM1000153) Month (mo, COM1000154) Year (a, COM1000151) mGy (mGy, COM1001957)
	SNOMED		Gy (229029004, COM1001950) mSv (282250007, COM1001951)
Disease qualifiers	SNOMED	NCIT, RadLex, LOINC	Symptomatic (264931009, COM1000130) Malignant (21594007, COM1000051) Invasive (10179008, COM1000101) Recurrent (255227004, COM1000158)
Cancer grading values (clinical, pathological, and histological grades)	NAACCR	SNOMED, NCIT	cM0 (960@c0, COM1001860) cM1 (960@c1, COM1000246) cM1a (960@c1A, COM1000258) pM1 (900@p1, COM1000531) pM1a (960@p1A, COM1000558)
			G1: Well differentiated (colon_rectum@3844@1, COM1002679) G2: Moderately differentiated (colon_rectum@3844@2, COM1002684)
	SNOMED	NCIT	G1 - American Joint Committee on Cancer grade G1(1228848004,COM1001750) G2 - American Joint Committee on Cancer grade G2 (1228850007, COM1001755) High histologic grade (1155707008, COM1001374) Low histologic grade (1155708003, COM1001373) International Society of Pathology histologic grade group 1 (1525761000004109, COM1001362) Grade group 1 (Gleason score 3 + 3 = 6) (1279715000, COM1001355) International Society of Pathology histologic grade group 2 (1525771000004101, COM1001365) Grade group 2 (Gleason score 3 + 4 = 7) (1279714001, COM1001343)

			ypN0 (1229878000, COM1001080) ypN1 (1229884002, COM1001082) ypN1a (1229887009, COM1001089) ypT0 (1228863001, COM1001607) ypT1 (1228869002, COM1001610) ypTis (1228865008, COM1001724)
	NCIT		Tumor Regression Score 0 (C155941, COM1001776) Tumor Regression Score 1 (C155942, COM1001785)
Diagnostic categories	SNOMED	NCIT	Screening (722021000000100, COM1002624) Control group (718601000000106, COM1002625)
	EUCAIM		Patient with Cancer (COM1001087) Negative screening (COM1002626) Patient with lesion not being a malignant tumor (COM1001086) Subject under discussion with suspicious findings (COM1002628)

Generic & Specific Modules

Both the Generic and Specific modules aim to ensure the semantic integration of the hyper-ontology with other EUCAIM components, including the CDM and HealthDCAT-AP. However, they differ on the specification level. For instance, the *Specific Module* defines specific categories, such as collection methods and interoperability tiers, to support the specification of dataset collection at the metadata level. Additionally, generic categories for classifying observations and procedures (based on mCODE specifications) are explicitly represented in the *Generic Module* to support the EUCAIM CDM semantically.

Table 6: *Selected classes from the Generic and Specific Modules, their standard sources, and EUCAIM IDs.*

Category	**Standard source**	**Classes (source code, EUCAIM ID)**
Collection methods	EUCAIM	Only-Image (SPEC1000003) Longitudinal (SPEC1000006) Disease-specific (SPEC1000007) Patient-based (SPEC1000004) Case-control (SPEC1000005)
	SNOMED	Cohort study (719581000000108, COM1002627)
Interoperability tiers	EUCAIM	Tier 1 (SPEC1000009) Tier 2 (SPEC1000011) Tier 3 (SPEC1000010) Tier 1A+ (SPEC1000012) Tier 1C+ (SPEC10000011) Tier 2A+ (SPEC1000014) Tier 2C+ (SPEC1000013) Tier 3A+ (SPEC1000016) Tier 3C+ (SPEC1000015)
Dataset Access Condition	EUCAIM	Authorisation to access, view and process in-situ the datasets (SPEC1000023)

		Authorisation to download the datasets (SPEC1000024) Authorisation to remotely process the datasets without the ability to access and visualise data, even remotely (SPEC1000022)
Dataset Type	EUCAIM	Annotated Dataset (SPEC1000019) Original Dataset (SPEC1000020) Processed Dataset (SPEC1000018)
Observation Category (mCODE)	EUCAIM	Laboratory (GEN1000022) Social History (GEN1000025) Exam (GEN1000019) Therapy (GEN1000018) Procedure (GEN1000021) Activity (GEN1000017) Imaging (GEN1000023) Vital Signs (GEN1000024) Survey (GEN1000020) Symptom (GEN1000030)
Procedure Category (mCODE)	SNOMED	Counseling (409063005, CLIN1058187) Surgical procedure (387713003, CLIN1004413) Measurement (122869004, CLIN1000003) Imaging (Procedure) (363679005, IMG1005453) Laboratory test (15220000, CLIN1033407) Removal (118292001, CLIN1058186)

4.3.2 Object Properties

Various object properties have been defined in the hyper-ontology to establish meaningful connections and relations among the classes (*Domain*/*Range*) of the same module or across different modules. Two main types of relations are considered: specific and generic. Specific relations establish an association between a specific Domain class as the subject and a specific Range class as the object. Meanwhile, the generic properties allow multiple Domain classes to be associated with a single *Range* class. In this case, the *Domain* class is left empty in the ontology to ensure reusability. Moreover, object properties ensure semantic integration with EUCAIM components, including the CDM and HealthDCAT-AP, thereby enabling data integration and interoperability.

CDM-Aligned: Clinical

***Table 7**: Selected object properties of the clinical context, their descriptions, generic ranges, and examples of range values.*

Property	**Description**	**Range**	**Example of range value**
hasBirthSex	Associates the cancer patient with the sex assigned at birth	Sex assigned at birth (COM1001396)	Female (COM1001370)
hasEthnicity	Associates the cancer patient with the ethnicity	Ethnicity (COM1001058)	White (COM1001311)
hasDiagnosticCategory	Associates the cancer patient with the diagnostic category	Population (COM1001084)	Patient with Cancer (COM1001087)

hasSubject	Links entities, such as primary cancer condition, cancer stage, tumor marker test, with the patient	Patient (COM1001047)	Patient (COM1001047)
hasRelatedCondition	Links entities, such as secondary cancer condition, cancer stage, tumor marker test, with the primary cancer condition	Primary Cancer Condition (CLIN1007988)	Malignant neoplasm of breast (CLIN1000060)
hasRelatedProcedure	Links entities, such as primary and secondary cancer conditions, cancer stage method, and tumor marker test, with specific procedures. For instance, for the primary condition, this property links to a pathology confirmation procedure, such as a biopsy or surgery.	Procedure (CLIN1000005)	Biopsy of breast (CLIN1056766)
hasTopography	Associates primary and secondary cancer conditions with topography from ICD-O-3	ICDO3Topography (BP1000265)	Breast (BP1000136)
hasBodySite	Associates various entities, such as tumor, surgical procedure, and radiotherapy, with the affected body site or structure.	Body structure (BP1000024)	Entire breast (BP1000264)
hasHistologyMorphologyBehavior	Connects cancer conditions (primary and secondary) with their morphologic and behavioral characteristics.	Histology/Morphology (CLIN1000033)	Papillary carcinoma (CLIN1049532)
hasClinicalStatus	Connects cancer conditions (primary and secondary) with their clinical status (e.g., active, inactive).	Condition clinical status value (COM1001176)	Inactive (COM1001173)
hasStageMethod	Specifies the method or system used for cancer staging (e.g., TNM, AJCC, FIGO).	Tumor staging (CLIN1044182)	American Joint Committee on Cancer, Cancer Staging Manual, 8th edition neoplasm staging system (CLIN1044309)
hasStageValue	Associate cancer stage with values (e.g., cM1a, cM1b, pM1)	American Joint Committee on Cancer allowable value (COM1000183)	cM1a (COM1000258)

hasHistologicGradeType	Specifies the type of histologic grade associated with the tumor or neoplasm	Histologic grade of neoplasm (CLIN1049681)	Nottingham grade of primary malignant neoplasm of breast (CLIN1049683)
hasGradeValue	Specifies the value of histologic grading (e.g., G1, G2, G3)	Histological grades (COM1001360)	G2 (COM1001755)
hasAssociatedMorphology	Associates the neoplasm with its morphology (e.g., benign, malignant, metastatic)	Tumor Morphology (CLIN1049522)	Malignant neoplasm (CLIN1049106)
hasTumorSizeMethod	Specifies the method/procedure used to detect the tumor size.	Imaging (Procedure) (IMG1005453)	CT of breast (IMG1000078)
hasLocation	Specifies the specific location of the body site associated with entities, such as surgical procedures, radiotherapy, and neoplasms.	Body structure (BP1000024)	Breast (BP1000136)
hasLaterality	Specifies the laterality qualifier of the body site associated with entities, such as surgical procedures, radiotherapy, and neoplasms.	Laterality (IMG1016305)	Left (IMG1016670)
hasRiskAssessment	Specifies the type of risk assessment.	Risk Assessment (CLIN1051859)	Intergroup Rhabdomyosarcoma Study Group Clinical Staging and Grouping System (CLIN1044304)
hasUnitofTime	Connects various entities, such as secondary cancer condition, surgical procedure, tumor marker test, with units of time (e.g., day, month, year).	Unit of time (COM1000150)	Month (COM1000154)
hasUnitofMeasure	Associates entities, such as the results of tumor marker or lab tests, tumor size, or volume with units of measure (e.g., mL, mm, ng/mL)	Unit of measure (COM1000147)	mm (COM1000152)
hasTreatmentResponse	Associates treatment (surgical procedure, radiotherapy, and medication	RECIST finding (CLIN1037303)	RECIST: partial response (CLIN1037306)

	administration) with response values from the RECIST terminology (e.g., complete response, partial response, stable disease)	Treatment response value (COM1001180)	Partial response (COM1001186)
hasRadiationModality	Specifies the modality associated with radiotherapy (e.g., electrons, protons, High Dose Rate)	Modality radiation value (COM1001181)	Electrons (COM1001178)
hasRadiationTechnique	Specifies the technique associated with radiotherapy (e.g., 3D, 2D, IMRT)	Radiotherapy (CLIN1014933)	Three dimensional conformal radiotherapy (CLIN1005283)
hasRadiationDoseUnit	Specifies the unit of the radiation dose (e.g., mGy, mSv)	'Unit of radiation dose' (COM1000143)	mGy (COM1001957)
hasObservationCategory	Associates the observation with its classification category (e.g., survey, laboratory, exam)	Observation Category (GEN1000014)	Laboratory (GEN1000022)
hasProcedureCategory	Associates the procedure with its classification category (e.g., surgical, imaging)	Procedure Category (GEN1000027)	Surgical procedure (CLIN1004413)

CDM-Aligned: Imaging

Table 8: *Selected object properties of the imaging context, their descriptions, generic ranges, and examples of range values.*

Property	Description	Range	Example of range value
hasImageBodyPart	Associates the dataset collection or the image series with the body part scanned (e.g., brain, prostate, breast)	Body structure (BP1000024)	Prostate (BP1000021)
hasImageModality	Connects the dataset collection or the image series with the image modality (e.g., MR, CT)	Imaging modality (IMG1000009)	Magnetic resonance imaging (IMG1000038)
hasAcquisitionParameter	Associates the image modality with the acquisition parameters or attributes (e.g., slice thickness, echo type)	Imaging procedure property (IMG1016619)	MR Echo Type (IMG1016640)
hasAcquisitionParameter Value	Links the image modality with the associated	Imaging procedure property (IMG1016619)	Spin echo (IMG1016662)

	acquisition parameters values (e.g., spin echo or radiant echo for MR echo type)		
hasLaterality	Associates the image series with the laterality - the side of the body scanned - (e.g, left, right)	Laterality (IMG1016305)	Left (IMG1016670)
hasEquipmentManufacturer	Associates the dataset collection or the image series with the type of equipment manufacturer used (e.g., GE, Fujifilm)	Manufacturer (IMG1000010)	General Electric (IMG1000047)
hasAlgorithmType	Connects the dataset collection or the segmentation series with the type of the annotation/segmentation algorithm (e.g., manual, automatic)	Segmentation Method (COM1001204)	Semiautomatic (COM1000005)
hasAnnotationType	Associates the segmentation series with the annotation type (e.g., mask, bounding box)	Annotation Type (COM1001202)	Bounding box (COM1000118)
hasAnnotatorSpecialty	Links the segmentation series with the annotator specialty (e.g., radiologist, oncologist)	Annotator Specialty (COM1001203)	Radiologist (COM1001104)
hasSegmentLabel	Connects the segment with the associated label	Body structure (BP1000024)	CZ+TZ (BP1000423)
hasAnnotationStatus	Links the segmentation series with the annotation status e.g., final, pending)	Annotation Status (COM1001211)	Final (COM1001213)

HealthDCAT-AP Support

Table 9: *Selected object properties for the HealthDCAT-AP support, their descriptions, generic ranges, and examples of range values.*

Property	Description	Range	Example of range value
hasImageBodyPart	Associates the dataset collection or the image series with the body part scanned (e.g., brain, prostate, breast)	Body structure (BP1000024)	Prostate (BP1000021)
hasImageModality	Connects the dataset collection or the image series with the image modality (e.g., MR, CT)	Imaging modality (IMG1000009)	Magnetic resonance imaging (IMG1000038)
hasEquipmentManufacturer	Associates the dataset collection or the image series with the type of equipment manufacturer used (e.g., GE, Fujifilm)	Manufacturer (IMG1000010)	General Electric (IMG1000047)

hasAlgorithmType	Connects the dataset collection or the segmentation series with the type of the annotation/segmentation algorithm (e.g., manual, automatic)	Segmentation Method (COM1001204)	Semiautomatic (COM1000005)
collectionMethod	Associates the dataset with the collection method (e.g., cohort, longitudinal, patient-based)	Collection method (SPEC1000002)	Cohort study (COM1002627)
hasBirthSex	Specifies the birth sex of the patients associated with the dataset collection	Sex assigned at birth (COM1001396)	Male (COM1001366)
hasCondition	Specifies the cancer condition associated with the dataset collection (e.g., prostate cancer, breast cancer)	Malignant neoplastic disease (CLIN1007977)	Malignant neoplasm of prostate (CLIN1000075)

4.3.3 Data Properties

Various data properties are defined in the hyper-ontology to describe the attributes of entities using data values (e.g., strings, numbers, dates, and booleans).

Table 10: *Selected data properties, their descriptions, generic ranges, and examples of range values.*

Property	**Description**	**Range**	**Example of range value**
PatientIdentifier	A unique identifier associated with the patient class.	xsd:string	“123456”
BirthDate	The date of birth of the patient	xsd:dateTime	1956-04-22
AgeAtDiagnosis	The age at diagnosis of the patient	xsd:decimal	59.2
Deceased	The vital status of patients	xsd:boolean	False
LabTestResultValue	The result value of lab tests	xsd:float	23.4
MaximumDimensionValue	The maximum dimension value of a tumor	xsd:float	10.3
PerformedDate	The occurrence date of procedures, tumor marker tests, and lab tests	xsd:dateTime	2024-10-08
StartDate	The start date of treatment, such as medication administration and radiotherapy	xsd:dateTime	2024-10-22
EndDate	The end date of treatment, such as medication administration and radiotherapy	xsd:dateTime	2024-12-26

4.3.4 Annotation Properties

The hyper-ontology defines several annotation properties that support documentation, clarity, and semantic mapping of ontology classes to external terminological resources in the biomedical domain. They provide information about the classes themselves, including preferred or alternative labels and definitions. Additionally, specific annotations, such as exact matches,

help map ontology terms to biomedical terminology and vocabularies, thereby supporting semantic interoperability. These annotations are reused from *SKOS* (Simple Knowledge Organization System), a data-sharing standard that enables bridging different fields of knowledge by providing a basic vocabulary for associating lexical labels with resources of any type (e.g., skos:prefLabel, skos:*exactMatch*). The following prefix is used to refer to the vocabulary URIs: *skos*: http://www.w3.org/2004/02/skos/core.
Moreover, syntactic alignment with healthcare clinical and imaging standards, such as OMOP/FHIR and DICOM, is recorded using properties defined in the EUCAIM context, including the OMOP CONCEPT ID and the DICOM Tag.
At the metadata level, annotation properties are defined to describe generic information about the hyper-ontology, including the version, date, creator, contributors, language, title, and rights. These properties are reused from DC (Dublin Core), a Metadata Element Set that defines metadata terms or elements identified with URIs, permitting the description of resources (e.g., *dc*:title, *dc*:creator). The following URI is used for the prefix *dc*: http://purl.org/dc/elements/1.1/.

***Table 11**: Selected annotation properties, their descriptions, generic ranges, and examples of range values.*

Property	Source	Description	Example of range value
prefLabel	SKOS	A lexical label that represents the preferred label associated with a concept.	"Malignant neoplasm of breast"
altLabel		A lexical label that represents the alternative label(s), or synonyms, associated with a concept.	"Breast cancer", "Malignant tumor of breast"
exactMatch		A mapping property used to link two concepts, indicating a high degree of confidence that the concepts can be used interchangeably across information retrieval applications.	SNOMEDCT:254837009
OMOP_DOMAIN_ID	EUCAIM (based on OMOP)	A property associated with the *DOMAIN ID* in OMOP	Condition
OMOP_CONCEPT_CODE		A property associated with the *CONCEPT CODE* in OMOP	C50
OMOP_VOCABULARY_ID		A property associated with the *VOCABULARY ID* in OMOP	ICD10
DICOM_Name	EUCAIM (based on DICOM)	A property associated with the *Name attribute* in DICOM	Modality
DICOM_Tag		A property associated with the *Tag attribute* in DICOM	(0008,0060)
creator	DC	An entity primarily responsible for making the resource.	LIMICS
contributor		An entity responsible for making contributions to the resource.	FORTH
title		A name given to the resource.	EUCAIM's Hyper-Ontology

4.4. Mapping Patterns for Data Harmonization

The hyper-ontology defines various semantic patterns to support the harmonization of heterogeneous data and semantic interoperability. Data heterogeneity in EUCAIM affects multiple aspects, including sources, contexts, standards, terminologies, and vocabularies, as well as interpretations and formats.

- **Sources**: data are primarily sourced from AI4HI and supplemented with new datasets collected during onboarding.
- **Contexts**: The EUCAIM hyper-ontology covers clinical and biological information as well as medical imaging.
- **Standards**: clinical and biological data collected from AI4HI are standardized using OHDSI OMOP and HL7 FHIR healthcare standards. Meanwhile, the imaging (meta)data are represented in DICOM format.
- **Terminologies and vocabularies**: Many concurrent coding systems have been used to standardize data. This heterogeneity is exposed even among the data collections that adopted the same healthcare standard.
- **Interpretations**: Data is interpreted heterogeneously across disparate data collections. The same information is represented, labeled, and standardized differently by different data holders (see an example on *Tumor stage* in Table 12).
- **Formats**: Standardized data is provided by AI4HI. Additionally, data holders use non-standardized data due to a lack of standard coding systems for specific health-related terms or imaging values.

Table 12: *Examples of heterogeneous data from AI4HI covering different cancer types.*

Scope	Cancer type	Standard	Data Holder	Variable (source)	Value (source)
Biological sex at birth	Breast	FHIR	EuCanImage	Sex assigned at birth (LOINC)	Male (SNOMED), Female (SNOMED)
		OMOP	ChAImeleon	Gender	MALE (GENDER), FEMALE (GENDER)
	Rectum	FHIR	EuCanImage	Sex assigned at birth (LOINC)	Male (LOINC), Female (LOINC)
Tumor stage	Prostate	OMOP	ProCancer-I	AJCC/UICC 7th pathological M1a Category (Cancer Modifier)	
			ChAImeleon	TNM Path M (NAACCR)	pM1a (NAACCR)
Histological grade	Breast	FHIR	EuCanImage	Tumor histopathological grade status values (SNOMEDCT)	American Joint Committee on Cancer grade GX (SNOMEDCT) G1 - American Joint Committee on Cancer grade G1 (SNOMEDCT) G2 - American Joint Committee on Cancer grade G2 (SNOMEDCT) G3 - American Joint

					Committee on Cancer grade G3 (SNOMEDCT)
	Colon	OMOP	ChAImeleon	Grade Pathological (NAACCR)	G1: Well differentiated (NAACCR) G2: Moderately differentiated (NAACCR) G3: Poorly differentiated (NAACCR) G4: Undifferentiated (NAACCR) Grade cannot be assessed (GX); Unknown (NAACCR)
Tumor regression grade	Rectum	FHIR	EuCanImage	Modified Ryan Scheme for Tumor Regression (NCIT)	Tumor Regression Score 0 (NCIT) Tumor Regression Score 1 (NCIT) Tumor Regression Score 2 (NCIT) Tumor Regression Score 3 (NCIT)
		OMOP	ChAImeleon	Tumor Regression Grade (NAACCR)	Tumor Regression Grade 0|Complete response: No viable cancer cells|No residual tumor (NAACCR) Tumor Regression Grade 1|Moderate response: Single cells or small groups of cancer cells (NAACCR) Tumor Regression Grade 2|Minimal response: Residual cancer outgrown by fibrosis (NAACCR) Tumor Regression Grade 3|Poor response: Minimal or no tumor kill; extensive residual cancer
Pathological response grade	Breast	FHIR	EuCanImage	Miller and Payne Classification (NCIT)	Grade 1 on a scale of 1 to 5 (SNOMED) Grade 2 on a scale of 1 to 5 (SNOMED) Grade 3 on a scale of 1 to 5 (SNOMED) Grade 4 on a scale of 1 to 5 (SNOMED) Grade 5 on a scale of 1 to 5 (SNOMED)
				Residual cancer burden class (SNOMED)	Residual Cancer Burden Class 0 (NCIT) Residual Cancer Burden Class 1 (NCIT) Residual Cancer Burden Class 2 (NCIT) Residual Cancer Burden Class 3 (NCIT)

***Table 13**: Examples of non-standardized terms from AI4HI and new Data Holders.*

Scope	Cancer type	Non-standard term (s)	Source
Cancer subtype	Breast	Luminal B- HER2 positive, HER2-enriched	AI4HI (EuCanImage, Incisive)
Histology/Morphology	Breast	MC: Invasive Mixed Carcinoma, IUC: Invasive Unidentifiable Carcinoma	AI4HI (Incisive)
Image procedure	Colorectal	Colonoscopy Video Recording	New DHs
Body region (Image segmentation)	Prostate	TZ+CZ	AI4HI (ProCancer-I)
Imaging modality	Multiple (breast, lung, prostate, etc.)	Digitized HE slides	New DHs
	Lung	Low-Dose Computed Tomography	New DHs

Several semantic mapping patterns, developed in collaboration with medical experts, have been incorporated into the ontology model to harmonize heterogeneous data and support the classification of non-standard terms. These patterns are represented using the OWL equivalent property *owl:equivalentClass* that associates two classes or concepts that are semantically identical, such as “Tumor Regression Grade 0” and “Tumor Regression Score 0” (**Figure 3**). Also, a single class could correspond to a combination of multiple classes, as in the case of breast cancer subtypes (**Figure 5**) and prostate-specific regions (**Figure 6**). For ambiguous or complex classes, property restrictions are used to define these entities and clarify their meaning. For instance, the cancer staging pattern (**Figure 2**) indicates that the “AJCC/UICC 7th pathological M1a Category” corresponds to the combination of “TNM Path M” and the restriction that this class is associated with the exact staging value “1 pM1a”, as defined by the object property *hasAnswer*. **Figure 4** and **Figure 7** depict additional patterns with restrictions on pathological response grading and imaging modality, respectively. All the figures are illustrated using Protégé[1], an open-source ontology editor.

[1] https://protege.stanford.edu

Cancer Staging: AJCC/UICC and TNM

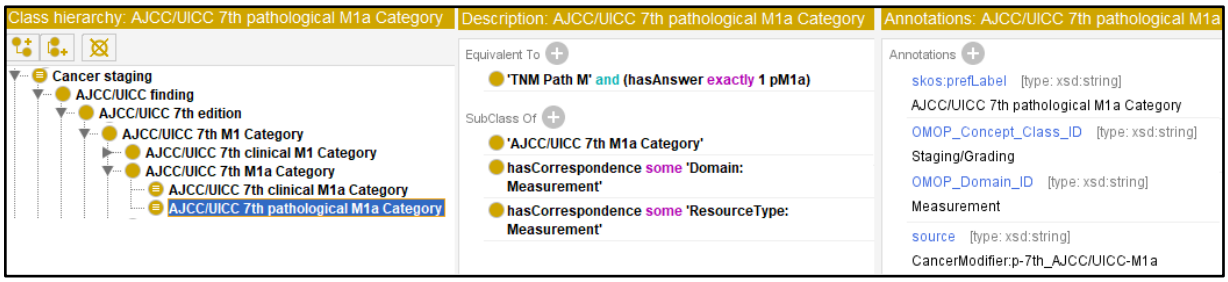


Figure 2: *A mapping pattern that defines the equivalence between the cancer staging structured concept "AJCC/UICC 7th pathological M1a Category" and the TNM classification concept "TNM Path M" associated with the staging value "pM1a" via the object property "hasAnswer".*

Tumor Regression: Grade/Score

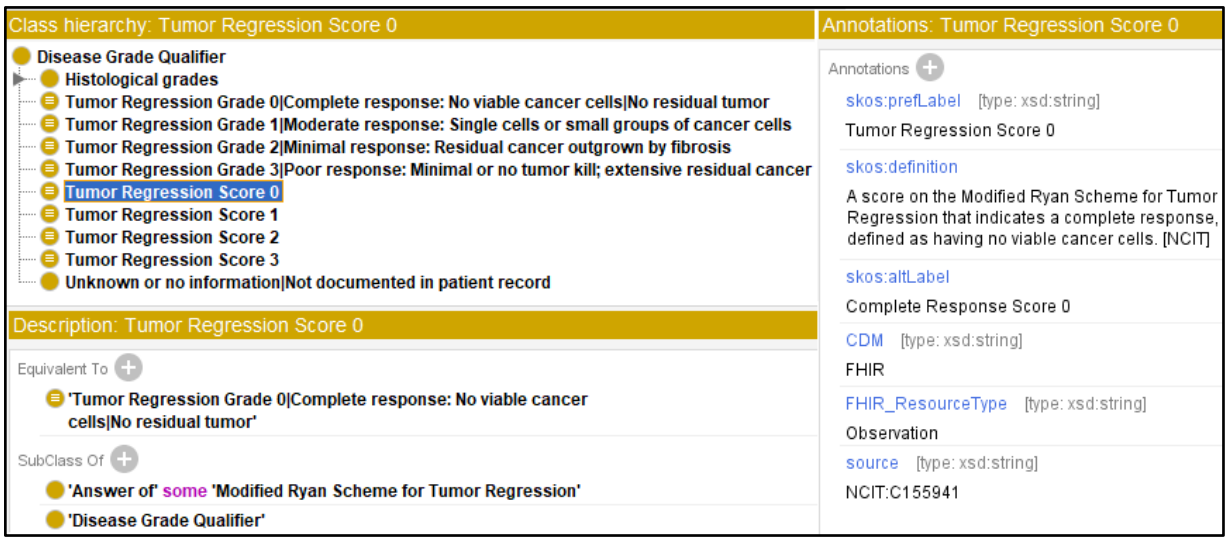


Figure 3: *A mapping pattern that defines the equivalence of the tumor regression score/grade: "Tumor Regression Grade 0|Complete response: No viable cancer cells|No residual tumor" is semantically identical to "Tumor Regression Score 0".*

Pathological Response Grade: Miller and Payne

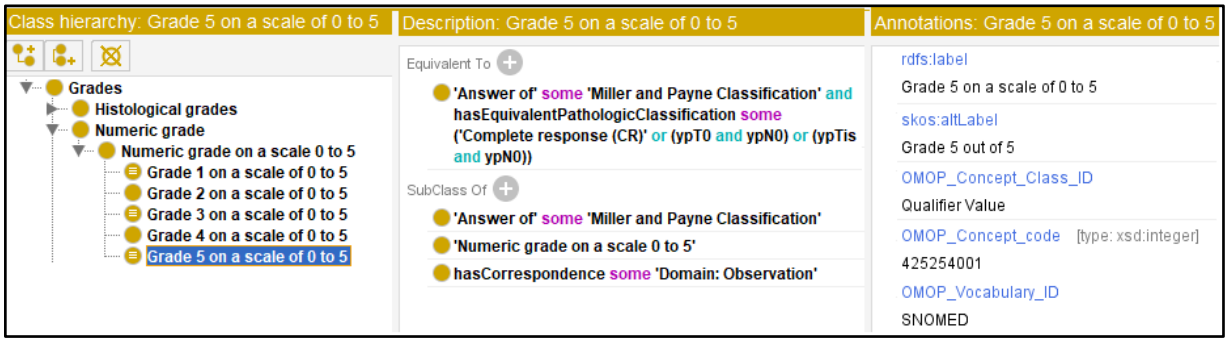


Figure 4: *A mapping pattern that defines the equivalence among the pathological response grading values for breast cancer: The Miller and Payne grading value "Grade 5" has equivalent pathologic classification "Complete Response (CR)", or the combination of AJCC "ypTis" and "ypN0".*

Classification of Non-Standard Terms

Breast Cancer Subtype

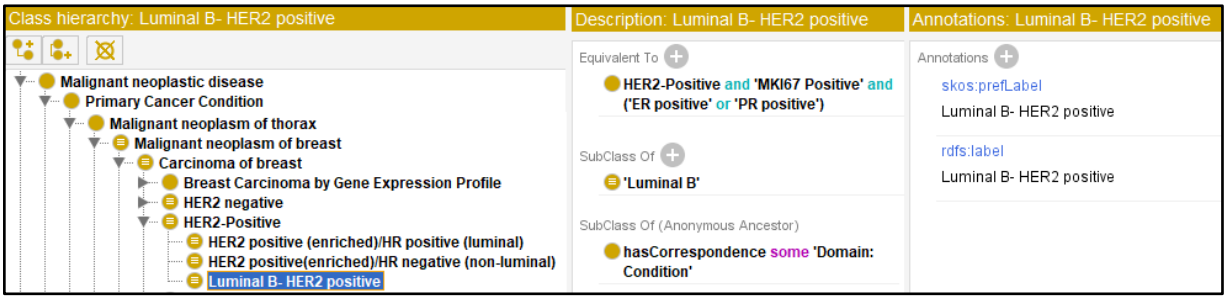


Figure 5: *A mapping pattern that clarifies the meaning of the breast cancer subtype "Luminal B- HER2 positive", which corresponds to the combination of the following tumor marker test results: "HER2-Positive", "MKI67 Positive", and "ER positive" or "PR positive"*

Body Region

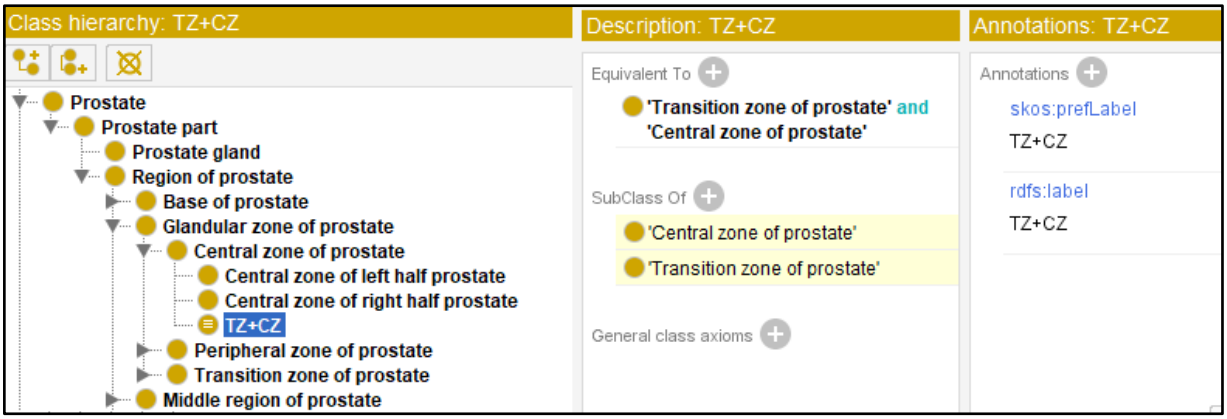


Figure 6: *A mapping pattern that clarifies the meaning of the term "TZ+CZ", which corresponds to the combination of "Transition zone of prostate" and "Central zone of prostate", permitting its semantic classification and integration into the hierarchy as a subclass of these concepts.*

Imaging Modality

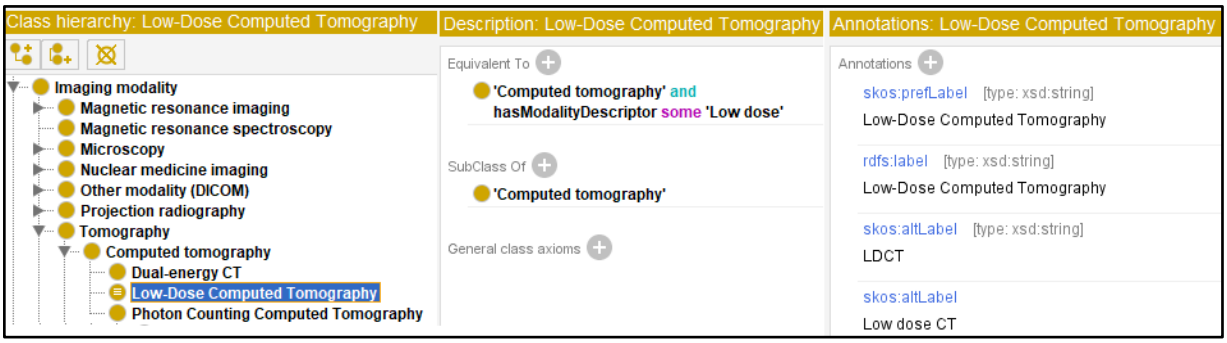


Figure 7: *A mapping pattern that clarifies the semantic meaning of the imaging modality "Low-Dose Computed Tomography" being a "Computed tomography" modality associated with the modality descriptor "Low dose", using the object property hasModalityDescriptor.*

5. The EUCAIM Common Data Model

The EUCAIM Common Data Model (CDM) is the reference schema used in Tier 3 of the federation for representing patient-level clinical, biological, molecular and imaging data in a harmonized way. It has been designed to support the specific needs of cancer imaging research, including multi-centre observational studies, and AI model development, while remaining compatible with widely adopted health data standards and semantically integrated with the EUCAIM hyper-ontology.

The following section presents a quick summary of the business requirements that guided the design of the CDM and explains the rationale behind its selection and design choices, as these were first outlined and presented in D5.2.

5.1 Introduction

Before selecting and defining the common data model for EUCAIM, the consortium conducted an analysis of requirements, expectations and constraints across different stakeholders, with a particular focus on the AI4HI projects that provide a large portion of the initial data. Information was collected from project representatives through the ORSD (Ontology Requirements Specification Document) and WP5 interactions, covering:

- the cancer types and clinical questions addressed in each project,
- the clinical and imaging variables used to answer these questions, distinguishing mandatory from optional data,
- the formats of raw data and the terminologies in use (and their versions),
- the anonymization techniques and profiles applied to meet GDPR and national regulations,
- the modalities and metadata of imaging datasets, including the presence and formats of segmentation masks,
- the data models adopted locally (OMOP-CDM, FHIR or ad-hoc schemas).

The analysis confirmed a high degree of heterogeneity: different cancer types and use cases, differing clinical and imaging variables, not standardized in many cases, multiple terminologies, different levels of granularity, varying anonymization strategies and a mixture of OMOP-, FHIR- and project-specific data models. At the same time, new DHs joining EUCAIM brought additional ad-hoc models and constraints.

From this, a set of key business requirements for the CDM emerged, including:

- support for heterogeneous input formats (standard and ad-hoc),
- terminology-agnostic design, allowing bindings to multiple vocabularies,
- explicit support for imaging metadata alongside clinical data,
- suitability for generating tabular structures required by AI tools, and
- compatibility with the EUCAIM hyper-ontology and existing OMOP/FHIR-based infrastructures.

These requirements framed the subsequent evaluation of candidate models and the design of the EUCAIM CDM.

5.2 Methods

The EUCAIM CDM is conceptually grounded in the FHIR mCODE (Minimal Common Oncology Data Elements) and builds on the OMOP CDM and its oncology extension, as well as on the OSIRIS data framework, as it was originally described in D5.2. In addition, in this final CDM specification the recently published ICGC ARGO data dictionary for standardizing global cancer data with a focus in genomics was explored. The reason why mCODE was selected as the primary conceptual basis is due to the fact that it defines a standardized set of core oncology data elements: patients, primary and secondary cancers, staging, treatments, health assessments and outcomes, specifically for oncology rather than generic clinical care. Although mCODE is expressed in FHIR, its domain model is independent of any particular technical representation and can therefore be reused in other implementations. However, mCODE is limited in capturing longitudinal outcomes. As such the OMOP CDM, in particular its oncology-oriented Episode construct, informed the inclusion of an Episode domain (Episode and Episode Event tables) in the EUCAIM CDM, enabling explicit representation of clinically meaningful periods such as diagnostic work-up and lines of treatment. OMOP also remains an important reference for sites that already maintain OMOP-based repositories. OSIRIS, on the other hand, developed as a minimum data set for clinico-biological oncology data, contributed two key ideas: a dedicated imaging module[2] for integrating imaging data, and a practical approach for exporting cohorts as “pivot tables” (CSV) for AI processing. EUCAIM incorporates this imaging module in its own representation, and adopts the practice of materializing analysis-ready tables in .csv files, for model training and evaluation. In addition, OSIRIS provides a comprehensive list of imaging metadata attributes for extraction, which EUCAIM capitalizes and includes in its imaging metadata model.

The very recently published ICGC-ARGO data dictionary (20 November 2025) also follows a similar approach to EUCAIM, proposing a data model specific for the genomic oncology domain, which is similar in nature with the EUCAIM CDM for the imaging oncology domain. Both are suited for federated multi-institutional precision oncology, offering comprehensive, high granularity clinical data, particularly suited for representing cancer related diagnosis events, treatments, and outcomes over time. In addition, EUCAIM capitalizes on the user-friendly data dictionary ICGC-ARGO published, where data producers and consumers are able to view the details of the data model including each schema and its details such as field descriptions, data tier and attributes, as well as permissible values and additional guidance for submission. EUCAIM has already developed a data dictionary viewer (**Figure 8**) publicly available at: https://eucaim-cdm.ics.forth.gr/ for its own model following other initiatives in the cancer domain, such as the EMBL-EBI’s CancerModels.org which hosts their Metadata Dictionary using ARGO’s Dictionary Viewer.

[2]https://github.com/InstitutNationalduCancer/OSIRIS_Radiomics/blob/main/DataDictionnary_OSIRISCT_2021_MedicalImaging.pdf

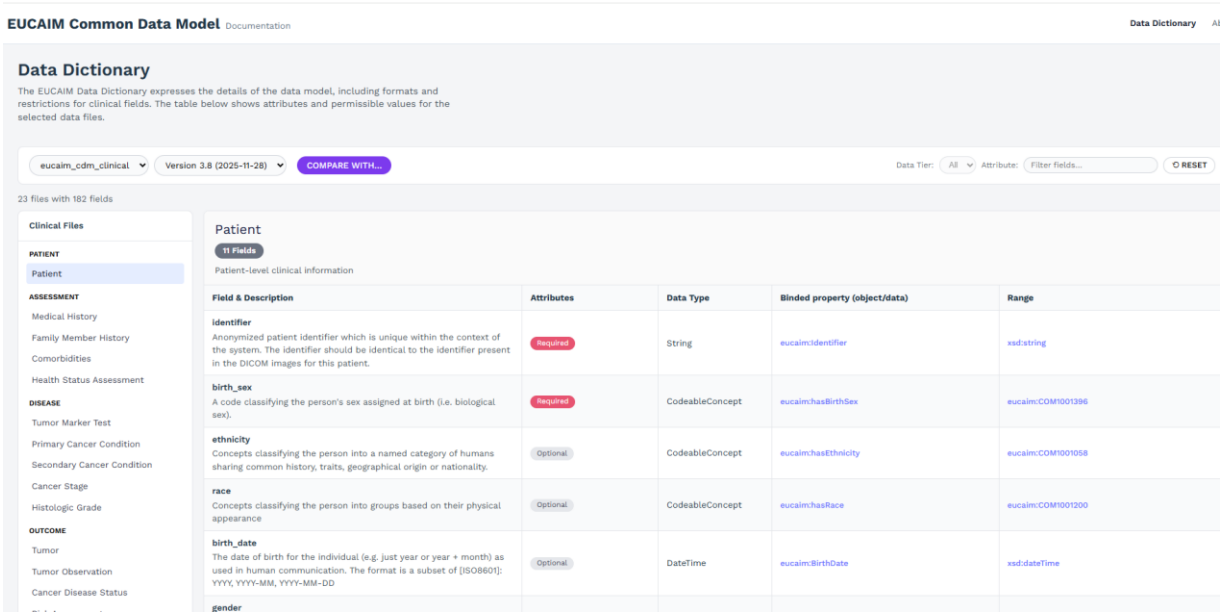


Figure 8: *Interactive EUCAIM CDM Data Dictionary viewer.*

The following section describes the EUCAIM CDM structure and main entities. It also outlines how the CDM handles anonymization constraints, how it is operationalized through a data dictionary, and how it is used in practice for mapping various new datasets, with ad-hoc models.

5.3 Main Entities and Attributes

The EUCAIM CDM follows the six conceptual groups of mCODE—Patient, Health assessment, Disease, Treatment, Outcome and Imaging—while adding Episode constructs and entities for Image related information. The overall conceptual structure of the CDM is summarized in the entity–relationship (ER) diagram shown in **Figure 9**.

The subsections below describe the main entities at a conceptual level, following the mCODE conceptual model and the hyper-ontology, along with their concrete attributes and constraints which are specified in detail in the EUCAIM data dictionary (**Annex A** - EUCAIM CDM Data Dictionary).

5.3.1 Patient

The **Patient** entity represents the individual receiving care and serves as the central anchor of the CDM. It holds a minimal set of demographic attributes required for most analyses and gives the information about the current status of the patient, including:

- a stable pseudonymous/anonymous identifier (unique within the node and the EUCAIM context),
- the patient's managing organization pseudonymous identifier (unique within the node and the EUCAIM context), extracted by the DICOM studies after de-identification
- year or date of birth, depending on anonymization constraints,
- sex assigned at birth or administrative gender,
- stratification variables such as ethnicity, race, where available,

- current patient high level status, including its diagnostic category, deceased or not, date of last contact and cause of death.

The Patient entity links to all other entities in the CDM, such as cancer conditions, episodes, treatments, imaging and outcomes, providing the backbone for integrating multimodal information.

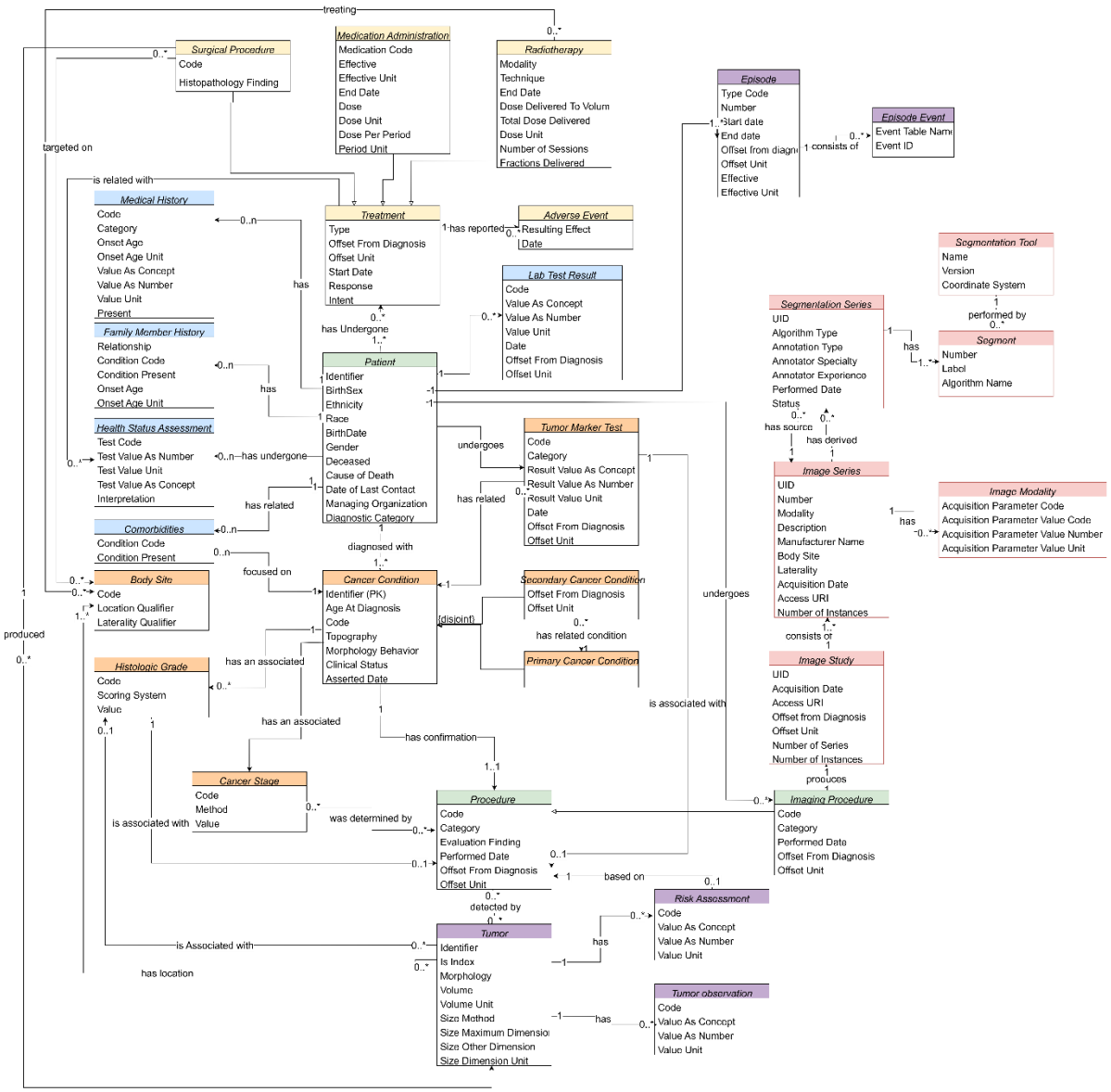


Figure 9: *Entity Relationship diagram (conceptual model) showing relationships between different entities in the EUCAIM CDM.*

5.3.2 Health assessments: comorbidities, laboratory tests, performance status, medical assessments, imaging, pathological assessments

The **health assessment** group includes comorbidities, family history, laboratory tests, vital signs and performance status measures, in line with the mCODE health assessment domain.

It covers:

- comorbidities and medical history as coded conditions or observations
- family history of cancer and other relevant conditions
- laboratory tests coded with LOINC or local codes, with values, units and reference ranges
- vital signs
- performance status scores (e.g. ECOG) recorded as coded assessments.
- medical assessments, including signs or symptoms reported by the medical expert or by the patient.

These variables provide crucial context for interpreting imaging and treatment outcomes and are important covariates in most analytic and AI use cases.

5.3.3 Cancer and Non-Cancer confirmation procedures

To ensure reliable ground-truth information for many AI use cases, the EUCAIM Rules of Participation require that each included case (cancer or healthy individual) has either a pathological or an imaging confirmation procedure. Accordingly, the EUCAIM CDM defines generic Procedure and Imaging Procedure entities to record these confirmation acts together with a corresponding evaluation finding (for example, biopsy result: normal/abnormal; imaging result: normal/abnormal). The admissible findings are defined as subclasses of the Evaluation finding class in the EUCAIM hyper-ontology, ensuring consistent use of controlled values across datasets.

5.3.4 Tumor and related Tumor observations

All the tumor related information extracted from relevant imaging and pathological procedures should be represented through the Tumor entity. This allows for tracking tumor characteristics over time based on different procedures, and specifying:

- the index tumor of the patient,
- tumor morphology, volume, size, body site, location, qualifier
- assessment tests, such as RADs or other risk assessment reports
- tumor observations based on imaging or other procedures, such as VASARI criteria, RADS specific characteristics (e.g. calcifications, breast density, mass margin/orientation/echo pattern etc.).

Note that some assessments can be represented both at a tumor level but also at a patient/imaging study level (e.g. PI-RADS score on each lesion found on MRI vs overall PI-RADS).

5.3.5 Primary and secondary cancer conditions

Cancer diagnoses are represented through **Primary Cancer Condition** and, where applicable, **Secondary Cancer Condition** entities, which represent metastases or other malignant processes causally related to a primary cancer, mirroring mCODE's distinction between primary and secondary malignancies.

Each cancer condition includes:

- a coded diagnosis using terminologies such as ICD-10, SNOMED CT.
- topography, histology morphology behavior
- anatomical site and laterality
- the date of diagnosis (or best available approximation) and/or age at diagnosis
- current clinical status of the condition (active, recurrence, relapse, inactive, remission, resolved)
- associations to staging, grading, confirmation imaging/biopsy related procedures.

5.3.6 Cancer Stage and Histologic Grade

The CDM defines explicit entities for **cancer stage** and **cancer grade**, in line with mCODE with clear semantics.

Stage records include:

- the staging system and edition (TNM, FIGO, etc.).
- the individual T, N and M categories, with indication of whether they are clinical or pathological
- values for each staging category
- link to the corresponding cancer condition.

Grade records capture histological grades or scores (e.g. G1–G3, Nottingham, Gleason), with the grading system explicitly identified. Multiple staging and grading assessments can be stored over time, enabling analyses of baseline stage as well as changes during the disease course.

5.3.7 Cancer-related Treatments

Treatment-related information is structured around a **Treatment** entity, with subclasses that specifically describe **surgical procedures**, **radiotherapy** and **medications administered**, following mCODE's treatment domain.

The general Treatment entity provides a high-level overview of the treatments administered (e.g. Chemotherapy, Immunotherapy, Radiotherapy, Surgical Procedure, and other therapeutic procedures as these are defined in the hyper-ontology), dates/offsets, intent (adjuvant, neoadjuvant, palliative) and treatment response based on RECIST terminology. This entity is always present for all treatments, and only in case more granular information is available for the basic cancer treatments (Surgical Procedure, Medication Administration, Radiotherapy), the respective entities should be used. For example, surgical procedures are captured with procedure codes, body site and laterality, and dates. In case there is a histopathological finding through this surgery it should be registered here. Radiotherapy is represented at course or plan level, start and end dates, total dose, number of fractions and, where available, technique and irradiated volumes. Systemic therapies capture regimens and their component drugs, and record treatment periods and, where possible, drug doses and periods of use. All treatments are linked to the patient, relevant cancer conditions and episodes, enabling reconstruction of complete treatment histories and their alignment with imaging and outcome data.

5.3.8 Tumor biomarkers

The CDM includes a focused representation of **tumor biomarkers**, aligned with the mCODE "Tumor Marker Test", but constrained to elements that are commonly available and relevant for EUCAIM use cases.

It supports:

- gene-level or variant-level findings where available, using standard nomenclatures,
- key driver mutations, recorded with appropriate codes and values.
- immunohistochemistry (IHC) or similar pathology stains

This enables integration of molecular characteristics into cohorts and analysis workflows, including imaging–genomics studies and multimodal AI models.

5.3.9 Imaging studies, series and modality parameters

The imaging component of the CDM reflects the central role of imaging in EUCAIM and builds on FHIR ImagingStudy/ImagingSeries, the MI-CDM extension of OMOP, the ProCAncer-I imaging extension, as well as the OSIRIS imaging concepts.

The model distinguishes:

- **Image Study**, representing a DICOM study with stable identifiers (e.g. StudyInstanceUID), acquisition date/offset, access URI (either on a DICOM web server (e.g. via the WADO-RS DICOMweb REST-API) or on local machine via the path name to the folder containing the study), number of series/images.
- **Image Series**, representing DICOM series with modality, series-specific acquisition parameters, body site information, manufacturer, accessURI.
- **Image Modality** parameters, capturing key acquisition details for modalities such as MR and CT (e.g. sequence name, field strength, kVp, slice thickness), in a flexible structure that allows storage of additional parameters without changing the schema.

The imaging entities standardize important metadata extracted from DICOM headers, which are often heterogeneous across institutions (e.g. free-text "Series Description"). This enables consistent selection of imaging data across nodes and supports modality- and protocol-specific queries. The specific imaging tags extracted from the DICOM images that can be queried in the CDM are included in **Annex B** - Imaging metadata for extraction.

5.3.10 Image segmentations and lesion-level information

The CDM also provides entities for **image annotations** and lesion-level information, extending mCODE's tumor profile concepts into the imaging domain.

These entities capture:

- links to imaging studies/series from which the segmentations were derived, and annotator descriptions (years of experience, specialty)
- types of annotations (e.g. manual segmentation, automated mask, semiautomatic),
- standardized controlled vocabularies to describe the anatomical structures or lesions segmented,

This structured representation makes it possible to trace lesions across time points, associate them with specific imaging protocols and clinical episodes, and feed lesion-level data into AI models.

5.3.11 Episodes

The **Episode** entity, adopted and adapted from OMOP, introduces a temporal layer that groups related events into clinically meaningful periods/timepoints such as diagnostic workup, therapeutic procedures and lines of therapy, or active surveillance.

Each episode is characterized by:

- an episode type (i.e. Diagnosis, Treatment, Progression, Relapse, Remission, Active Surveillance)
- a start and end date (or relative temporal markers)
- associations to the patient
- links to the clinical events, imaging and any other kind of assessment occurring during that period linked through the Episode Event entity.

This structure simplifies episode-based cohort definitions (for example, “lab values during first-line treatment, tumor markers/imaging studies after radiotherapy”), clarifies longitudinal relationships between interventions and outcomes, and supports reproducible temporal logic in federated analysis and observational studies.

5.3.12 Vocabularies: Concept, Concept Ancestor, Concept Relationship, Concept Synonym, Concept External Mapping

Inspired by OMOP, the EUCAIM CDM also has a set of vocabulary tables that map the EUCAIM hyper-ontology codes to multiple standard terminologies. This means that the EUCAIM hyper-ontology’s logic is transformed into a relational-based model (similar to OMOP concept, concept_relationship, and concept_ancestor, enabling semantic queries via relational logic (e.g., ontology subsumption via concept_ancestor equivalent). As such EUCAIM follows the same assumptions with OMOP:

- All the hyper-ontology classes are loaded into the CONCEPT table.
- The key is a newly created EUCAIM concept_id, not the original code of the terminology, because source codes are not unique identifiers across terminologies.
- Records in the CONCEPT_RELATIONSHIP table define semantic relationships between concepts as these are defined in the hyper-ontology. Such relationships can be hierarchical or lateral.
- Chains of hierarchical relationships are recorded in the CONCEPT_ANCESTOR table.
- A new CONCEPT_EXTERNAL_MAPPING is created for all the concepts that are defined in the hyper-ontology as annotations and not individually as classes supporting semantic interoperability with a great number of terminologies implicitly (e.g. source terminology codes, OMOP related concepts etc.)

Figure 10 presents the vocabulary schemas, whereas **Figure 11** presents the logical schema of the EUCAIM CDM.

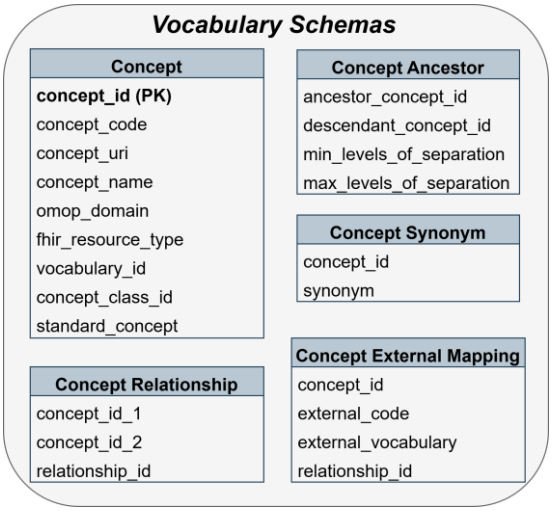


***Figure 10**: The logical schema of the Vocabulary tables.*

5.4 Handling anonymization and temporal information

A key constraint identified in the AI4HI projects and also verified in the pilots during the onboarding process is that many clinical datasets either cannot expose full date information, due to GDPR-compliant anonymization and project-specific de-identification strategies or they have not registered this information in the first place.

While all standards, FHIR, mCODE, OMOP-CDM and OSIRIS define numerous date attributes (e.g. exact diagnosis dates, procedure dates, assessment dates), many of these were not available in the AI4HI repositories or in other contributing datasets. Instead, temporal information is often expressed as relative timing (for example, “X months after diagnosis” or “Y weeks after treatment start”).

The EUCAIM CDM therefore supports two patterns:

- use of absolute dates where available and where local regulations and anonymization policies permit
- use of relative temporal attributes—i.e. intervals from diagnosis—where dates have been shifted or fully anonymized.

This flexibility allows nodes with stricter anonymization to still participate in analyses and federated learning, while preserving patient privacy. The Episode concept further reduces dependence on absolute dates, as episodes can be defined using relative order and duration as well as calendar dates.

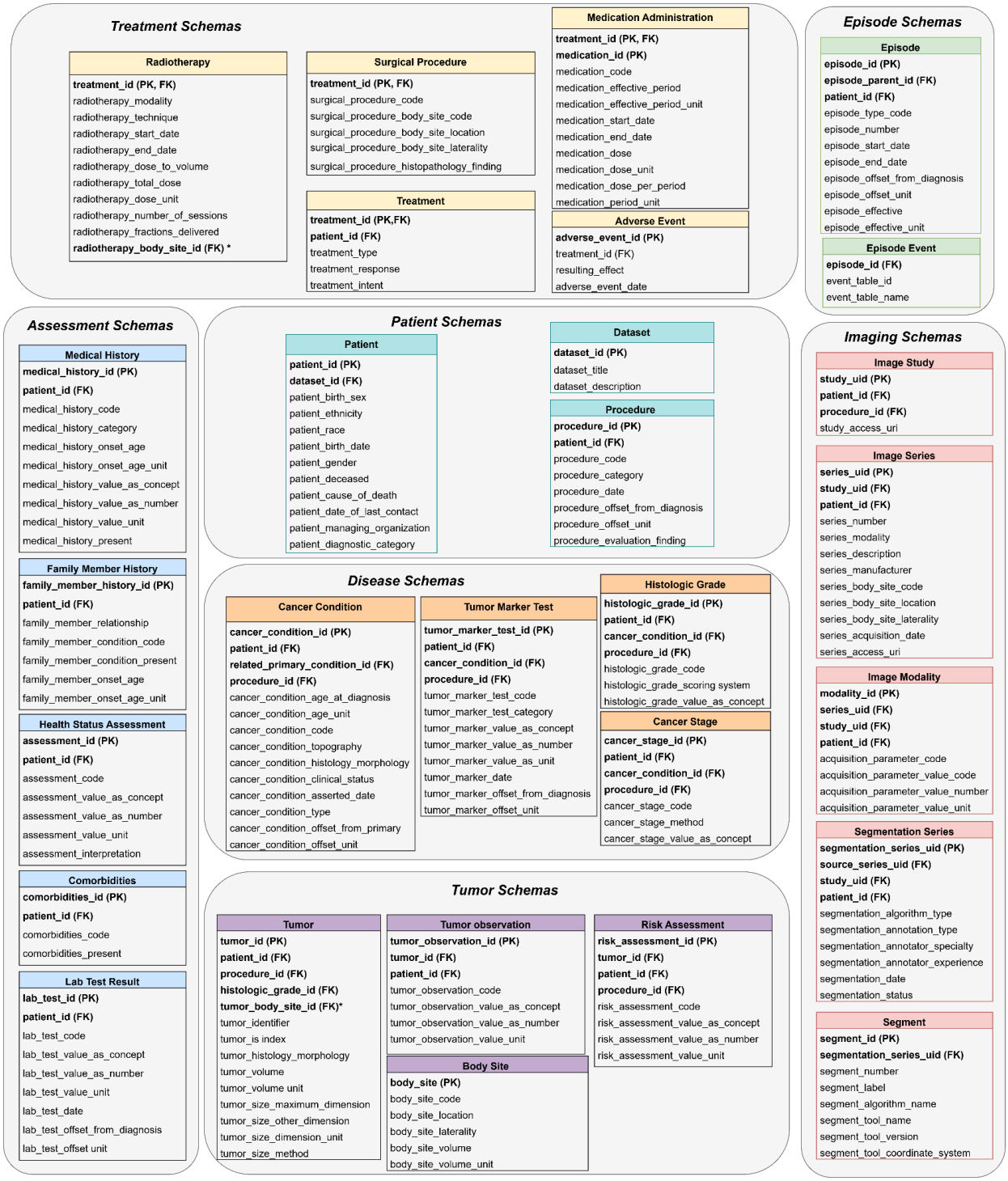


Figure 11: *The logical schema of the EUCAIM CDM.*

5.5 CDM groups and data dictionary

To operationalize the CDM, EUCAIM maintains a data dictionary that classifies clinical and imaging data into six high-level groups derived from mCODE—Patient, Health assessment, Disease, Cancer treatment, Outcome and Imaging—and defines the entities, data elements, cardinalities and datatypes within each.

The initial version of this dictionary was presented in D5.2 as a set of tables (e.g. Tables 7–12), and has since been refined and extended based on experience with node onboarding. The current version is presented in the **Annex A** - EUCAIM CDM Data Dictionary to this deliverable.

For each data element the dictionary specifies:

- a unique name and definition,
- whether it is required or optional, and its allowed occurrences,
- the datatype and, where appropriate, the reference value set in the hyper-ontology,
- any specific constraints (e.g. permissible ranges or conditional requirements).

This dictionary is the main reference for data holders implementing ETL to the CDM and for tool developers building queries and analytics on top of EUCAIM CDM-compliant data.

5.6 Data structure and format for analysis and federated learning

Although the CDM is defined as a structured model with related entities, most analysis and AI workflows operate on tabular representations of cohorts. Building on the OSIRIS experience, EUCAIM routinely transforms CDM-based data into “pivot tables” (and most specifically CSV files) tailored to the variables required by a given study or AI model use case.

The typical pattern is:

1. Define a cohort and variables of interest using CDM entities and attributes based on the requested data and applying the data minimization GDPR principle.
2. Materialize data into one or multiple .CSV files depending on the use case.
3. A materialized view of the mandatory clinical and imaging metadata will always be provided, each as a single CSV file, as these were defined in D4.4 and D5.6 (Tables 20, 21, 7)
4. Feed the clinical data files into local or federated analytic pipelines and AI tools (including those described in D5.9).

This approach preserves the semantic consistency and interoperability of the CDM while providing AI experts with familiar tabular inputs. In federated settings, the same analytic specification can be executed across nodes as long as each node can produce the required CDM-derived tables.

This is the expected data structure for each of the EUCAIM datasets:

```
/DatasetID/
  ├── clinical_mandatory_view.csv <-- View with mandatory clinical data
  ├── imaging_mandatory_view.csv  <-- View with mandatory imaging metadata
  ├── clinical_data               <-- full clinical data
  │   ├── patient.csv
  │   ├── cancer_condition.csv
  │   └── treatment.csv ...
  └── imaging_data                <-- Hierarchical structure for imaging files
      ├── patient-uid1
      │   ├── study-uid1/
      │   │   ├── series-uid1/
      │   │   │   ├── DICOM file 1
      │   │   │   ├── DICOM file 2
      │   │   │   └── DICOM file ...
```

```
│   │   └── ...
│   └── study-uid2/ ...
├── patient-uid2
│   └── ...
└── ...
```

Below there is an example of the `clinical_mandatory_view.csv` and the `imaging_mandatory_view.csv` file in vertical orientation for better readability.

Table 14: *Mandatory clinical variables (first treatment attributes might be empty in case there is no treatment recorded).*

```
dataset_id
patient_id
patient_diagnostic_category
patient_birth_sex
cancer_condition_age_at_diagnosis
cancer_condition_age_unit
cancer_condition_code
cancer_condition_asserted_date
cancer_condition_type
cancer_condition_topography
cancer_condition_histology_morphology
evidence_procedure_code
evidence_procedure_category
evidence_procedure_date
evidence_procedure_offset_from_diagnosis
evidence_procedure_offset_unit
first_treatment_type
first_treatment_date
first_treatment_offset_from_diagnosis
first_treatment_offset_unit
```

Table 15: *Mandatory imaging variables. Each record corresponds to an image series.*

```
dataset_id
patient_id
image_procedure_code
image_procedure_date
image_procedure_offset_from_diagnosis
image_procedure_offset_unit
study_uid
series_uid
series_modality
series_manufacturer
series_body_site_code
series_acquisition_date
```

6. Integration of CDM and Hyper-Ontology

6.1 Mapping Concepts to Data Structures

Mapping ontology concepts to data structures involves binding the CDM entities and their attributes to the hyper-ontology classes and properties. Each entity is mapped to an ontology class, and its attributes are bound to object or data properties, depending on the attribute's data type (*CodeableConcept* or data values (e.g., *Integer*, *String*, *DateTime*, *Boolean*)), thereby establishing explicit semantic integration between the EUCAIM CDM and the hyper-ontology. This integration supports semantic interoperability and consistent data interpretation across applications, including federated querying, segmentation/annotation, and ETL.

- **Entity**: entities in the CDM correspond to core ontology classes in the hyper-ontology. For clinical data elements, both the CDM and hyper-ontology are grounded in the mCODE conceptual model, supporting semantic alignment between the ontology's core concepts and the CDM entities, such as Patient (COM1001047), Primary Cancer Condition (CLIN1007988), Secondary Cancer Condition (CLIN1007989), Cancer Stage (CLIN1044182), Histologic Grade (CLIN1049681), Tumor Marker Test (CLIN1000224), and Treatment (GEN1000015). Similarly, for imaging, entities are semantically aligned with the classes defined in the imaging module, such as Image Study (IMG1016308), Image Series (IMG1016309), and Image modality (IMG1000009).

- **Attribute**: an attribute in the CDM describes an entity's properties. They are mapped to object properties or data properties depending on the associated data type. Data values, such as *String*, *DateTime*, and *Boolean*, are mapped to data properties, and *CodeableConcept* and *Reference* (foreign key) are mapped to object properties associated with a specific range of ontology concepts. For instance, Patient.*Identifier* is assigned a *String* data type, thereby mapped to the data property *PatientIdentifier* with “Patient” (COM1001047) domain and *xsd:string* range. Meanwhile, Patient.*BirthSex* is assigned a *CodeableConcept* data type, which entails a mapping with the object property *hasBirthSex* with specific range values (*Male*, *Female*) defined under “Sex assigned at birth” (COM1001396).

- **Data Type**: data values, such as *String*, *DateTime*, and *Boolean*, are mapped to standard datatypes from the XML Schema Definition (XSD), such as *xsd:string*, *xsd:dateTime*, and *xsd:boolean*. *Reference* and *CodeableConcept* data types are aligned with specific value ranges in the ontology, based on the data element definitions provided in the CDM and the terminology binding specified in mCODE.

Clinical & Biological

Table 16: *A subset of the EUCAIM CDM entities and attributes and their semantic mapping with the Hyper-Ontology properties and classes or data types (Clinical and Biological context).*

Common Data Model			Hyper-Ontology	
Entity	**Attribute**	**Data Type**	**Property (Object/Data)**	**Range (Ontology Class/ Data Type)**

Patient	Identifier	String	*PatientIdentifier*	xsd:string
	BirthSex	CodeableConcept	*hasBirthSex*	Sex assigned at birth (COM1001396)
Medical History	Code	CodeableConcept	*hasMedicalHistory*	
	Subject	Reference: Patient	*hasSubject*	Patient (COM1001047)
	Category	CodeableConcept	*hasCategory*	Generic Category (GEN1000011)
Family Member History	Relationship	CodeableConcept	*hasRelationship*	Person in the family (COM1001043)
	Condition Present	Boolean	*ConditionPresent*	xsd:boolean
Comorbidities	Focus	Reference:PrimaryCancerCondition	*hasFocus*	Primary Cancer Condition (CLIN1007988)
	Condition Code	CodeableConcept	*hasComorbidity*	Disease (CLIN1002855)
Health Status Assessment	Code	CodeableConcept	*hasHealthStatusAssessment*	Health Status Assessment (CLIN1037292)
	Value As Concept	CodeableConcept	*hasAnswer*	Performance Status Assessment Answer Value (COM1001110)
			hasAnswerAsFinding	Performance Status Assessment Answer Finding (CLIN1052860), Clinical finding (CLIN1000004)
Tumor Marker Test	Code	CodeableConcept	*hasTumorMarkerTest*	Tumor marker measurement (CLIN1000224)
	Category	CodeableConcept	*hasCategory*	Observation Category (GEN1000014)
	Result Value As Concept	CodeableConcept	*hasResultValue*	Finding value (COM1000181)
	Related Condition	CodeableConcept	*hasRelatedCondition*	Malignant neoplastic disease (CLIN1007977)
Primary Cancer Condition	Code	CodeableConcept	*hasPrimaryCancerCondition*	Primary Cancer Condition (CLIN1007988)
	Age At Diagnosis	Decimal	*AgeAtDiagnosis*	xsd:decimal
	Topography	CodeableConcept	*hasTopography*	ICDO3Topography (BP1000265)
	Histology Morphology Behavior	CodeableConcept	*hasHistologyMorphologyBehavior*	Histology/Morphology (CLIN1000033
	Clinical Status	CodeableConcept	*hasClinicalStatus*	Condition clinical status value (COM1001176)

SecondaryCancer Condition	Related Primary Cancer Condition	CodeableConcept	*hasRelatedPrimaryCancerCondition*	Primary Cancer Condition (CLIN1007988)
	Offset From Diagnosis	Integer	*OffsetFromDiagnosis*	xsd:int
	Offset Unit	CodeableConcept	*hasUnitofTime*	Unit of time (COM1000150)
Cancer Stage	Code	CodeableConcept	*hasCancerStage*	Tumor staging (CLIN1044182)
	Method	CodeableConcept	*hasStageMethod*	Tumor staging (CLIN1044182)
	Value	CodeableConcept	*hasStageValue*	American Joint Committee on Cancer allowable value (COM1000183), Malignant tumor staging (CLIN1044308)
	Related Procedure	CodeableConcept	*hasRelatedProcedure*	Procedure (CLIN1000005)
Histologic Grade	Code	CodeableConcept	*hasHistologicGrade*	Histologic grade of neoplasm (CLIN1049681)
	Scoring System	CodeableConcept	*hasHistologicalGradingSystem*	Histological grading systems (CLIN1037296)
	Value	CodeableConcept	*hasHistologicGradeValue*	Histological grades (COM1001360)
Tumor	Identifier	String	*TumorIdentifier*	xsd:string
	Morphology	CodeableConcept	*hasAssociatedMorphology*	Tumor Morphology (CLIN1049522)
	Size Method	CodeableConcept	*hasTumorSizeMethod*	Tumor Size Method (GEN1000028)
	Size Maximum Dimension	Float	*MaximumDimensionValue*	xsd:float
	Size Dimension Unit	CodeableConcept	*hasUnitofMeasure*	Unit of measure (COM1000147)
Risk Assessment	Code	CodeableConcept	*hasRiskAssessment*	Risk Assessment (CLIN1051859)
	Value As Concept	CodeableConcept	*hasRiskAssessmentValue*	Risk assessment value (COM1001198), subclasses of Assessment (IMG1005456)
	Value As Number	Float	*RiskAssesmentValue*	xsd:float

Imaging

Table 17: *A subset of the EUCAIM CDM entities and attributes and their semantic mapping with the Hyper-Ontology properties and classes or data types (Imaging context).*

Common Data Model			Hyper-Ontology	
Entity	**Attribute**	**Data Type**	**Property (Object/Data)**	**Range (Ontology Class/ Data Type)**
Image Study	UID	Identifier	*ImageStudyUID*	xsd:string
	Acquisition Date	dateTime	*AcquisitionDate*	xsd:dateTime
	Offset from Diagnosis	Integer	*OffsetFromDiagnosis*	xsd:int
	Offset Unit	CodeableConcept	*hasUnitofTime*	Unit of time (COM1000150)
	Number Of Series	unsignedInt	*NumberOfSeries*	xsd:unsignedInt
Image Series	UID	Identifier	*ImageSeriesUID*	xsd:string
	Number	unsignedInt	*ImageSeriesNumber*	xsd:unsignedInt
	Description	string	*ImageSeriesDescription*	xsd:string
			hasImageSeriesDescription	
	Modality	Reference	*hasImageModality*	Imaging modality (IMG1000009)
	Manufacturer Name	CodeableReference	*hasEquipmentManufacturer*	Manufacturer (IMG1000010)
	Body Site	CodeableReference	*hasImageBodyPart*	Body structure (BP1000024)
	Laterality	CodeableConcept	*hasLaterality*	Laterality (IMG1016305)
Image Modality	Acquisition Parameter Code	CodeableConcept	*hasAcquisitionParameter*	Imaging procedure property (IMG1016619)
	Acquisition Parameter Value Code	CodeableConcept	*hasAcquisitionParameterValue*	Imaging procedure property (IMG1016619)
	Acquisition Parameter Value Number	Float	*AcquisitionParameterValueNumber*	xsd:float
	Acquisition Parameter Value Unit	CodeableConcept	*hasUnitofMeasure*	Unit of measure (COM1000147)
Segmentation Series	Algorithm Type	CodeableConcept	*hasAlgorithmType*	Segmentation Method (COM1001204)

	Annotation Type	CodeableConcept	*hasAnnotationType*	Annotation Type (COM1001202)
	Annotator Specialty	CodeableConcept	*hasAnnotatorSpecialty*	Annotator Specialty (COM1001203)
	Status	CodeableConcept	*hasAnnotationStatus*	Annotation Status (COM1001211)
Segment	Number	unsignedInt	*SegmenNumber*	xsd:unsignedInt
	Label	CodeableConcept	*hasSegmentLabel*	Body structure (BP1000024)
Segmentation Tool	Name	String	*SegmentationToolName*	xsd:string
	Coordinate System	CodeableConcept	*hasImageCoordinateSystem*	Image Coordinate System (COM1001216)

7. Validation and Use Cases

7.1 Pilot Studies and Demonstrators

The technical pilot studies, designed to test the setup of federated nodes, provided an excellent opportunity to evaluate the EUCAIM hyper-ontology and CDM as a common representation for heterogeneous datasets across cancer types, imaging modalities, and formats.

During the pilots, DHs were asked to map their local data models to the EUCAIM CDM. To support them in this process, the WP5 pilot support team held several one-to-one meetings with each DH, guiding them through the challenges and nuances of the mapping activity. As noted in D5.8, mapping local data models to the CDM remains one of the most effort-intensive and time-consuming steps in the technical onboarding workflow. Throughout the pilot activities, considerable effort has been devoted to streamlining and optimizing this step, including:

- the preparation of a comprehensive mapping template (Excel format - **Annex E** - Clinical data template file) with detailed instructions and a clear guiding structure. This template defines how each local dataset concept is associated with the EUCAIM CDM entities and attributes,
- the development of various training materials, including dedicated tutorial videos on the CDM and its use, available at the EUCAIM youtube channel and the materials in moodle developed by WP2,
- ongoing personalized support to DHs.

The mapping template is then provided to the ETL tool implementation team, which translates the declared mappings into concrete transformation rules and applies them to the actual data. The tool outputs a Postgres database that implements the EUCAIM CDM. The work dedicated to the revision and enhancement of the technical tools that materialize the mapping process, namely the ETL tool complemented by the DICOM tag extraction tool for handling the imaging metadata, is described in Deliverable D5.9 - The EUCAIM tools for Data Preprocessing and Data/Metadata Management & Interoperability. It is worth highlighting that, for many DHs, one-to-one guidance from semantic experts familiar with the EUCAIM CDM and the Hyper-ontology

will remain necessary during the onboarding process, especially for accurately mapping custom case-specific concepts and attributes.

The analysis of the heterogeneous data models contributed to several revisions and extensions of the EUCAIM CDM and the Hyper-ontology. These include restructuring entities and attributes, expanding the supported controlled vocabularies, and refining specific modelling choices so that the CDM can adequately represent the diversity of concepts found across DH data models.

As part of the ongoing pilot activities, the mapping of 10 datasets—each with a distinct local ad-hoc data model and different cancer types—across 6 DHs has been completed to date, resulting in validated mapping template files. These datasets include:

- 1 dataset on sarcoma cancer (by UoA)
- 2 datasets on prostate cancer following different models (by UoA and APHP)
- 1 dataset on brain cancer (by UoA)
- 1 dataset on lung cancer (by SAS)
- 3 datasets on breast cancer following different models (by SAS, AUTH, and KI)
- 1 dataset on thyroid cancer (AUTH)
- 1 dataset on brain cancer (HULAFE)

In the following sections we will present the mappings for the following 4 datasets, each representing a new cancer type (which was not considered in the AI4HI projects and in D5.2), and as such various concepts and modeling choices had to be added and/or modified. These datasets are: the sarcoma dataset from UoA, the brain (glioblastoma) dataset from HULAFE, the lung cancer from SAS, and the thyroid cancer from AUTH.

Structure of the mapping

The mapping was documented in the EUCAIM CDM mapping template (Excel format). For each local data element/field, the template records:

- the local field name,
- the local values,
- the corresponding EUCAIM CDM entity and attribute,
- the terminology used, if available.

This format makes the mapping explicit, auditable and reusable by the ETL implementation team and the local data holders.

Semantic annotation of the clinical and imaging data

Semantic annotation of clinical and imaging data from real-world datasets involves mapping variables, labels, or codes to standard concepts defined in biomedical or imaging standards, thereby supporting data integration and enriching the hyper-ontology with new concepts, relations, and annotations. The mapping process, which focuses on new data values not included in previous versions of the hyper-ontology, requires a (manual) data curation phase to address ambiguous or incomplete information and terms not precisely covered in standard coding systems. Furthermore, to address overlapping terminologies, appropriate standard

sources need to be identified or verified. For instance, for cancer conditions, ICD-10 is used as the primary source. ICD-O-3 is used for histology/morphology and topography, and LOINC is used for tumor marker tests. Switching to another source is needed if concepts are not covered in the primary source or if the source lacks the required specificity, necessitating semantic integration across different terminologies. This complexity and data heterogeneity are addressed in the hyper-ontology using various techniques, including semantic alignment with external resources and semantic patterns, supported by a unified coding system based on thematic categorization. The labels and codes provided by data holders (DHs) are mapped to precise standard labels or synonyms via an exact-match similarity approach. To prevent semantic misinterpretation or inconsistency and ensure accurate granularity, the mapping results need to be validated by domain experts (e.g., clinicians, radiologists).

7.1.1 Sarcoma dataset to the EUCAIM CDM

To illustrate the feasibility of applying the EUCAIM CDM to heterogeneous real-world datasets, this subsection summarizes the mapping of the UoA sarcoma dataset to the CDM using the EUCAIM mapping template (**Annex E** - Clinical data template file).

The UoA sarcoma dataset focuses on patients with histologically confirmed sarcoma and includes patient-level clinical data, tumor characteristics, staging, treatments, imaging studies and selected outcome information. In its original form, the dataset has a local ad-hoc schema, and uses a combination of coding systems, and free-text fields. The goal of the mapping was to align this schema with the EUCAIM CDM without losing clinically relevant information.

Table 18: Sarcoma dataset - Diagnosis Episode: mapping of local data elements to EUCAIM CD shows the mappings from the UoA sarcoma dataset to the EUCAIM CDM. They illustrate different patterns: simple one-to-one mappings, composite local fields that map to several CDM attributes, and value standardization into controlled vocabularies through the hyper-ontology.

Table 18: *Sarcoma dataset - Diagnosis Episode: mapping of local data elements to EUCAIM CDM entities, attributes, value sets.*

Local Model		**EUCAIM Common Data Model**		
Local field	**Local Values**	**Entity**	**Attribute**	**Value Set**
Patient ID	<String>	Patient	Identifier	Literal
Gender	female, Male	Patient	Gender	Female (COM1001370), Male (COM1001366)
Sex	F, M (LOINC)	Patient	Birth Sex	FEMALE (COM1000178), MALE (COM1000176)
Status	TRUE, FALSE	Patient	Deceased	FALSE (COM1000035) TRUE (COM1000036)
Population	Patient with Cancer	Patient	Diagnostic Category	Patient with Cancer (COM1001087)
Site	<String>	Cancer Condition	Managing Organization	<String>
Age at diagnosis	<Integer>	Cancer Condition	Age at diagnosis	<Decimal>
Pathology	C49 (ICD10 code)	Cancer Condition	Code	Malignant neoplasm of other connective and soft tissue (CLIN1007996)

			Type	Primary (COM1000017)
Histological subtype according to WHO 2020	Malignant Peripheral Nerve Sheath Tumor, Myxofibrosarcoma (WHO 2020 terms)	Cancer Condition	Histology Morphology Behavior	Malignant peripheral nerve sheath tumor, NOS (CLIN1049536), Myxofibrosarcoma (CLIN1049514)
Clinical status	Active (custom)	Cancer Condition	ClinicalStatus	Active (COM1001177)
Topography: Organ	C49.22, C49.21 (ICD-O-3)	Cancer Condition	Topography	The local code values provided by the DH are not defined in ICD-O-3. Manual data curation was needed to ensure an accurate semantic representation of body organs. This is clarified in the challenges section.
Grade (FNCLCC)	<Integer>	Histologic Grade	Scoring System	French Federation of Cancer Centers Sarcoma Group (FNCLCC) grading system (CLIN1037302)
			Value	<Integer>
Tumor max dimension (cm)	<Decimal>	Tumor	Size Maximum Dimension	<Decimal>
			Size Dimension Unit	centimer (COM1001954)
Imaging procedure (CT, MRI)	CT, MRI	Procedure	Code	Computed Tomography (IMG1000026), Magnetic resonance imaging (IMG1000022)
			Category	Imaging (Procedure) (IMG1005453)
Image acquisition date	<Date>		Performed Date	<Date>

Table 19: *Sarcoma dataset - Treatment Episode: mapping of local data elements to EUCAIM CDM entities, attributes, value sets.*

Treatment: Radiotherapy		Treatment	Category	Radiotherapy (CLIN1014933)
	<Date>		StartDate	<Date>
Radiotherapy modality	TomoTherapy (custom)	Radiotherapy	Modality	TomoTherapy (CLIN1005282)
Radiotherapy technique	VMAT (custom)		Technique	Volumetric modulated arc therapy (CLIN1005272)

Delivered dose	<Integer>		Total Dose Delivered	<Integer>
Number of fractions (Gy)	<Integer>		DoseUnit	Gy (COM1001950)
			FractionsDelivered	<Integer>
Treatment	Wide excision of malignant neoplasm. (SNOMED-CT: 409063005)	Surgical Procedure	Code	The local code provided by the DH doesn't correspond to the documented label. Instead, it maps to the SNOMED-CT concept "Counseling". Manual data curation was needed to ensure an accurate semantic representation. This is clarified in the challenges section.
Surgery date	<Date>		Performed Date	<Date>

Impact of the UoA sarcoma mapping on the CDM

The mapping of the sarcoma dataset did not require any change to the overall structure of the EUCAIM CDM. All clinically relevant concepts could be accommodated within existing entities and attributes. The main work consisted of:

- completing and refining terminology mappings for sarcoma-specific histology and primary site codes,
- extending certain value sets (e.g. body-region codes) to include additional locations frequent in sarcoma (such as specific limb segments or retroperitoneal sites)

These refinements have been incorporated into the current version of the hyper-ontology. The pilot therefore both confirmed the feasibility of applying the CDM to a sarcoma real dataset and helped to evolve the controlled vocabularies and bindings needed for broader multi-cancer coverage.

Challenges in the Sarcoma Dataset Semantic Annotation

The Sarcoma dataset has helped to extend and enrich the hyper-ontology with a new cancer type, "Malignant neoplasm of other connective and soft tissue", and associated histologic types, topography, and procedures. The extension has affected the domain and domain-specific layers of the hyper-ontology without modifying the core content. The main challenge in the semantic annotation of this dataset is data incompleteness and the absence of exact standard concepts for mapping specific labels/codes, necessitating a manual data-curation phase (see

Table 20). For instance, C49.21 and C49.22 are ICD-O-3 codes provided by the data holder for particular body locations ("Left upper limb, including shoulder" and "Right upper limb, including shoulder"). However, ICD-O-3 defines broader concepts (Connective, Subcutaneous and other soft tissues of upper limb and shoulder, C49.1) and (Connective, Subcutaneous and other soft tissues of lower limb and hip, C49.2), which necessitate a semantic mapping across ICD-O-3 and SNOMED to ease data integration and build a taxonomy of concepts with consistent and accurate levels of specificity. Another example is the SNOMED code "409063005," which is labeled "Wide excision of malignant neoplasm". However, the provided code does not correspond to the documented label, which is non-standardized in coding systems. Instead, it maps to the SNOMED concept "Counseling". This inconsistency can lead to misinterpretation of results if data is used without correction. Thereby, a dataset validation check is required to ensure codes are aligned with labels.

Table 20: *Examples of value sets (label, source terminology, code) from the sarcoma dataset and their semantic annotation with standard sources.*

Local Data Element	Local Value Label (DH)	Local Value Source (DH)	Local Value Code (DH)	Standard Label	Standard Source: Code	EUCAIM ID
Pathology	-	ICD-10	C49	Malignant neoplasm of other connective and soft tissue	ICD10:C49	CLIN1007996
Histological subtype according to WHO 2020	Malignant Peripheral Nerve Sheath Tumor	WHO 2020	-	Malignant peripheral nerve sheath tumor, NOS	ICDO3:9540/3	CLIN1049536
	Myxofibrosarcoma	WHO 2020	-	Myxofibrosarcoma	ICDO3:8811/3	CLIN1049514
Topography: Organ	Left upper limb, including shoulder	ICD-O-3	C49.22	Structure of soft tissues of left upper extremity	SNOMEDCT:764402008	BP1000411
	Right upper limb, including shoulder	ICD-O-3	C49.21	Structure of soft tissues of right upper extremity	SNOMEDCT:764401001	BP1000412
The correct ICD-03 codes and labels for "Topography: Organ" local data element				Connective, Subcutaneous and other soft tissues of upper limb and shoulder	ICDO3:C49.1	BP1000410
				Connective, Subcutaneous and other soft tissues of lower limb and hip	ICDO3:C49.2	BP1000309
Grade scoring system (FNCLCC)	-	SNOMED	426757001	French Federation of Cancer Centers Sarcoma Group (FNCLCC) grading system	SNOMEDCT:426757001	CLIN1037302
Grade	3	FNCLCC	-	Grade 3 tumor	CancerModifier:Grade-3	CLIN1022188
Imaging procedure (CT MRI)	-	RadLex	10318	N/A	N/A	

Radiotherapy modality	TomoTherapy	custom	-		Tomotherapy	SNOMEDCT:113310002241 00	CLIN1005282
Radiotherapy technique	VMAT	custom	-		Volumetric modulated arc therapy	SNOMEDCT:1156530009	CLIN1005284
Treatment	Wide excision of malignant neoplasm	SNOMED	409063005		Counseling	SNOMEDCT:409063005	CLIN1058187

7.1.2 Glioblastoma dataset to the EUCAIM CDM

The Glioblastoma dataset focuses on patients with glioblastoma multiforme confirmed by biopsy and MRI. The purpose of the dataset is to detect recurrence and estimation of survival in patients with glioblastoma multiforme after total tumor resection.

Table 21: *Glioblastoma dataset – Diagnosis episode: mapping of local data elements to EUCAIM CDM entities, attributes and value sets.*

Local Model		**EUCAIM Common Data Model**		
Local field	**Local Values**	**Entity**	**Attribute**	**Value Set**
PatientID	<Integer>	Patient	Identifier	<Literal>
Population	Patient with Cancer	Patient	Diagnostic Category	Patient with Cancer (COM1001087)
Gender	Male, Female (LOINC)		BirthSex	FEMALE (COM1000178) MALE (COM1000176)
Diagnosis Age	<Decimal>	Cancer Condition	Age at diagnosis	<Decimal>
Age Unit	Year		Age Unit Concept	Year (COM1000151)
Histological Type	Glioblastoma	Cancer Condition	Code	Glioblastoma Multiforme (CLIN1007990)
			Type	Primary (COM1000017)
D-Pathological subtype	Epithelioid, Gliosarcoma, Giant Cell, Oligodendroglioma component (SNOMED)	Cancer Condition	Histology Morphology Behavior	Epithelioid (CLIN1000069) Gliosarcoma (CLIN1000082) Giant cell (CLIN1000073) Oligodendroglioma (CLIN1000081)
Tumor Grade	Grade-4	Histologic Grade	Value	Grade 4 tumor (CLIN1022227)
ASA Classification System (performance status)	Class I Class II Class III Class IV Class V Class VI	Health Status Assessment	Value as Concept	Class I (CLIN1052872) Class II (CLIN1052871) Class III (CLIN1052870) Class IV (CLIN1052869) Class V (CLIN1052868) Class VI (CLIN1052867)
			Code	American Society of Anesthesiologists physical status classification (CLIN1047414)
D-KPS (Karnofsky)	<Decimal>	Health Status Assessment	Code	Karnofsky performance status (CLIN1047415)

			Value as Number	<Decimal>
Pathology Confirmation	Biopsy	Procedure	Code	Biopsy (CLIN1001712)
			Category	Removal (CLIN1058186)
D-Date Biopsy	<Date>		Performed Date	<Date>
		Cancer Condition	Asserted Date	<Date>
D-MR pre-surgery Date	<Date>	Procedure	Code	Magnetic Resonance imaging (IMG1000022)
			Category	Imaging (Procedure) (IMG1005453)
			Performed Date	<Date>
Organ	Brain	Tumor	Body Site Code	Brain (BP1000051)
Tumor location	Frontal (1) Temporal (2) parietal (4) occipital (5)	Tumor	Body Site Location	Frontal lobe (BP1000351) Temporal lobe (BP1000350) Parietal lobe (BP1000352) Occipital lobe (BP1000354)
Side of tumor epicenter	Left, Right	Tumor	Body Site Laterality	Left (IMG1016670) Right (IMG1016682)
D-IDH1	Mutated Not mutated Unknown	Tumor Marker Test	Code	IDH1 (isocitrate dehydrogenase (NADP(+)) 1) gene variant measurement (CLIN1051952)
			Result Value As Concept	Mutated (COM1001313) Not mutated (COM1001887) Unknown (COM1001289)
D-ATRX	Mutated Not mutated Unknown	Tumor Marker Test	Code	ATRX (ATRX chromatin remodeler) gene variant measurement (CLIN1051954)
			Result Value As Concept	Mutated (COM1001313) Not mutated (COM1001887) Unknown (COM1001289)
D-p53mut	Negative Positive Unknown	Tumor Marker Test	Code	TP53 (tumor protein p53) gene variant measurement (CLIN1051958)
			Result Value As Concept	Negative (COM1001332) Positive: (COM1001310) Unknown (COM1001289)
D-MGMT	Methylated Not methylated Unknown	Tumor Marker Test	Code	MGMT (O-6-methylguanine-DNA methyltransferase) gene variant measurement (CLIN1051955)
			Result Value As Concept	Methylated (COM1001314) Not methylated (COM1001889) Unknown (COM1001289)
D-IDH1-BM	Mutated Not mutated Unknown	Tumor Marker Test	Code	IDH1 and IDH2 genes targeted mutation analysis in Blood or Tissue by Molecular genetics method (CLIN1045804)

			Result Value As Concept	Mutated (COM1001313) Not mutated (COM1001887) Unknown (COM1001289)
D-TERT	M - 124 M - 146 Not mutated Unknown	Tumor Marker Test	Code	TERT (telomerase reverse transcriptase) gene variant measurement (CLIN1051956)
			Result Value As Concept	Mutated/c.1-124C>T (COM1001317) Mutated/c.1-146C>T (COM1001318) Not mutated (COM1001887) Unknown (COM1001289)
D-Proliferation rate KI67	<Integer>	Tumor Marker Test	Code	MKI67 (marker of proliferation Ki-67) gene variant measurement (CLIN1051957)
			Result Value As Number	<Integer>
Eloquent brain	No eloquent brain; speech motor; speech receptive; motor; vision;thalamus /hypothalamus; internal capsule; brainstem; cerebellar peduncles; deep cerebellar nuclei	Tumor Observation	Code	Eloquent brain (IMG1016750)
			Value As Concept	any subclass of “Eloquent Brain Areas” (BP1000438) No eloquent brain (COM1001899)
Enhancement Quality	No contrast enhancement; mild; marked; no contrast injected		Code	Enhancement Quality (IMG1016743)
			Value As Concept	any subclass of “Enhancement Quality value” (COM1016743)
Proportion enhancing	>2/3; 1/3-2/3; <1/3; minimal		Code	Proportion Enhancing (IMG1016744)
			Value As Concept	any subclass of “Proportion Enhancing value” (COM1016744)
Proportion nCET	>2/3; 1/3-2/3; <1/3; minimal; No nCET		Code	Proportion nCET (IMG1016745)
			Value As Concept	any subclass of “Proportion nCET value” (COM1016745)
Proportion Necrosis	>2/3; 1/3-2/3; <1/3; minimal; none		Code	Proportion Necrosis (IMG1016747)
			Value As Concept	any subclass of “Proportion Necrosis value” (COM1016747)
Proportion of Edema	>2/3; 1/3-2/3; <1/3; minimal; none		Code	Proportion of Edema (IMG1016748)
			Value As Concept	any subclass of “Proportion of Edema value” (COM1016748)
Cyst(s)	Absent; Present		Code	Cysts (IMG1016749)
			Value As Concept	any subclass of “Cysts value” (COM1016749)

Multifocal or multicentric	Focal; Multifocal o multicentric; gliomatosis		Code	Multifocal or Multicentric (IMG1016730)
			Value As Concept	Multifocal or Multicentric value (COM1016730)
T1/FLAIR Ratio	No FLAIR images; T1~FLAIR; T1<FLAIR; T1<<FLAIR		Code	T1/FLAIR Ratio (IMG1016733)
			Value As Concept	T1/FLAIR Ratio value (COM1016733)
Thickness of enhancing margin	Minimal; Thick/nodular (>=3 mm); solid		Code	Thickness of enhancing margin (IMG1016735)
			Value As Concept	Thickness of enhancing margin value (COM1016735)
Enhancing margin	Well-defined; poorly defined		Code	Enhancing margin (IMG1016736)
			Value As Concept	Enhancing margin value (COM1016736)
Haemorrhage volume	Yes; No; Cannot determine		Code	Hemorrhage volume (IMG1016739)
			Value As Concept	Hemorrhage volume value (COM1016739)
Diffusion	Facilitated; Restricted; Cannot determine; No ADC images		Code	Diffusion Characteristics (IMG1016720)
			Value As Concept	Diffusion Characteristics value (COM1016720)
Pial Invasion	Absent; Present			Pial Invasion (IMG1016721)
			Value As Concept	Pial Invasion value (COM1016721)
Ependymal invasion	Absent; Present		Code	Ependymal Extension (IMG1016722)
			Value As Concept	Ependymal Extension value (COM1016722)
Cortical involvement	Absent; Present		Code	Cortical Involvement (IMG1016723)
			Value As Concept	Cortical Involvement value (COM1016723)
Deep White Matter invasion	None; Corpus Callosum; Internal Capsule; Brainstem		Code	Deep White Matter Invasion (IMG1016725)
			Value As Concept	Brain Deep White Matter Structure (BP1000441) None (COM1001375)
Enhancing tumor Crosses Midline	Yes; No		Code	Enhancing tumor Crosses Midline (IMG1016727)
			Value As Concept	Enhancing tumor Crosses Midline value (COM1016727)
Satellites	Absent; Present		Code	Satellites (IMG1016710)
			Value As Concept	Satellites value (COM1016710)
Calvarial remodeling	Absent; Present		Code	Calvarial Remodeling (IMG1016711)
			Value As Concept	Calvarial Remodeling value (COM1016711)
Non-enhancing margin	Well-defined; poorly defined		Code	Non-enhancing margin (IMG1016738)
			Value As Concept	Non-enhancing margin value (COM1016738)

nCET tumor Crosses Midline	Yes; no		Code	nCET tumor Crosses Midline (IMG1016738)
			Value As Concept	nCET tumor Crosses Midline value (COM1016738)
Total tumor Volume	milliliter	Tumor	Volume	Total Tumor Volume (IMG1003881)
Enhancing tumor volume	milliliter	Tumor Observation	Code	Enhancing Tumor Volume (IMG1003887)
			Value As Number	<Decimal>
			Value Unit	milliliter (COM1001966)
Non-enhancing tumor volume	milliliter		Code	Non-Enhancing Tumor Volume (IMG1003883)
			Value As Number	<Decimal>
			Value Unit	milliliter (COM1001966)
Necrosis volume	milliliter		Code	Necrosis Volume (IMG1003884)
			Value As Number	<Decimal>
			Value Unit	milliliter (COM1001966)
Edema volume	milliliter		Code	Edema Volume (IMG1003885)
			Value As Number	<Decimal>
			Value Unit	milliliter (COM1001966)
Haemorrhage volume	milliliter		Code	Haemorrhage Volume (IMG1003886)
			Value As Number	<Decimal>
			Value Unit	milliliter (COM1001966)
Primary tumor sample ID	ID number	Tumor	identifier	<String>

Table 22: *Glioblastoma dataset – Treatment episode: mapping of local data elements to EUCAIM CDM entities, attributes and value sets.*

Local Model		**EUCAIM Common Data Model**		
Local field	**Local Values**	**Entity**	**Attribute**	**Value Set**
Chemotherapy	Yes; No	Treatment	Treatment Type	If yes: Chemotherapy (CLIN1024528)
Chemotherapy date	<Date>	Treatment	Start Date	<Date>
Type of chemotherapy	TMZ; Others	Medication Administration	Code	IF TMZ: Temozolomide (CLIN1035872)
Radiotherapy	Yes; No	Treatment	Treatment Type	if yes: Radiotherapy (CLIN1014933)
Radiotherapy date	<Date>	Treatment	Start Date	<Date>
Type of Radiotherapy	Free text	-	-	-

***Table 23**: Glioblastoma dataset – Relapse episode: mapping of local data elements to EUCAIM CDM entities, attributes and value sets.*

Local Model		EUCAIM Common Data Model		
Local field	**Local Values**	**Entity**	**Attribute**	**Value Set**
Progression Date	<Date>	Episode	Start Date	<Date>
		Episode	Type Code	Progression (GEN1000010)
R-MR before-surgery date	<Date>	Procedure	Code	Magnetic Resonance imaging (IMG1000022)
			Category	Imaging (Procedure) (IMG1005453)
			Performed Date	<Date>
R-KPS	0-100	Health Assessment Status	Code	Karnofsky performance status (CLIN1047415)
			Value As Concept	Subclasses of Performance status qualifier (COM1000025)
R-ECOG	0-5	Health Assessment Status	Code	ECOG Performance Status score (CLIN1004580)
		Health Assessment Status	Value As Concept	Subclasses of ECOG performance status finding (CLIN1015980)
R-tumor biopsy ID	ID number	Procedure	Code	Biopsy (CLIN1001712)
			Category	Removal (CLIN1058186)
R-biopsy date	<Date>		Performed Date	<Date>
R-pathological subtype	Giant cells; Oligodendroglioma component; Epithelioid; Gliosarcoma	Tumor	Histology Morphology Behavior	Epithelioid (CLIN1000069) Gliosarcoma (CLIN1000082) Giant cell (CLIN1000073) Oligodendroglioma (CLIN1000081)
R-IDH1	Mutated Not mutated Unknown	Tumor Marker Test	Code	IDH1 (isocitrate dehydrogenase (NADP(+)) 1) gene variant measurement (CLIN1051952)
			Result Value As Concept	Mutated (COM1001313) Not mutated (COM1001887) Unknown (COM1001289)
R-ATRX	Mutated Not mutated Unknown	Tumor Marker Test	Code	ATRX (ATRX chromatin remodeler) gene variant measurement (CLIN1051954)
			Result Value As Concept	Mutated (COM1001313) Not mutated (COM1001887) Unknown (COM1001289)
R-p53	Negative Positive Unknown	Tumor Marker Test	Code	TP53 (tumor protein p53) gene variant measurement (CLIN1051958)
			Result Value As Concept	Negative (COM1001332) Positive: (COM1001310) Unknown (COM1001289)
R-MGMT	Methylated Not methylated	Tumor Marker Test	Code	MGMT (O-6-methylguanine-DNA methyltransferase) gene variant

	Unknown			measurement (CLIN1051955)
			Result Value As Concept	Methylated (COM1001314) Not methylated (COM1001889) Unknown (COM1001289)
R-IDH1-BM	Mutated Not mutated Unknown	Tumor Marker Test	Code	IDH1 and IDH2 genes targeted mutation analysis in Blood or Tissue by Molecular genetics method (CLIN1045804)
			Result Value As Concept	Mutated (COM1001313) Not mutated (COM1001887) Unknown (COM1001289)
R-TERT	M - 124 M - 146 Not mutated Unknown	Tumor Marker Test	Code	TERT (telomerase reverse transcriptase) gene variant measurement (CLIN1051956)
			Result Value As Concept	Mutated/c.1-124C>T (COM1001317) Mutated/c.1-146C>T (COM1001318) Not mutated (COM1001887) Unknown (COM1001289)
R-Proliferation rate KI67	<Integer>	Tumor Marker Test	Code	MKI67 (marker of proliferation Ki-67) gene variant measurement (CLIN1051957)
			Result Value As Number	<Integer>
R-Date of surgery	<Date>	Treatment	Code	Surgical Procedure (CLIN1004413)
		Treatment	Start Date	<Date>
R-MR after-surgery date	<Date>	Procedure	Code	Magnetic Resonance imaging (IMG1000022)
			Category	Imaging (Procedure) (IMG1005453)
			Performed Date	<Date>
R-MR Extent of resection	Gross-total resection; subtotal resection	Surgical Procedure	Code	Gross Total Resection (CLIN1004587); Subtotal Resection (CLIN1004586)
R-MR Volume Extent of resection	milliliter	Tumor	Volume	<Float>
			Volume Unit	milliliter (COM1001966)
R-MR % Extent of resection	percentage			
R-Chemotherapy	Yes; No	Treatment	Treatment Type	If yes: Chemotherapy (CLIN1024528)
R-Chemotherapy date	Date after diagnosis - before present	Treatment	Start Date	<Date>
R-Type of chemotherapy	TMZ Fotemustina Others Unknown N/A	Medication Administration	Code	Temozolomide (CLIN1035872) Bevacizumab (CLIN1035837) Fotemustine (CLIN1035873)

	Bevacizumab			

Table 24: *Glioblastoma dataset – overarching episode: mapping of local data elements to EUCAIM CDM entities, attributes and value sets.*

Local Model		EUCAIM Common Data Model		
Local field	**Local Values**	**Entity**	**Attribute**	**Value Set**
Date of Last visit	<Date>	Patient	Date Of Last Contact	<Date>
Death	Yes, No	Patient	Deceased	<Boolean>
Date of death	<Date>	Patient	Date Of Last Contact	<Date>
Death related to GBM	Yes, No	Patient	Cause Of Death	If yes: Glioblastoma Multiforme (CLIN1007990)
Time to progression	days	These were not directly mapped as they can be calculated through the Episode and Episode Event instances.		
Overall survival	days			
Progression-free survival (PFS)	days			
Time to reoperation	days			
Survival from reoperation	days			

Impact of the Glioblastoma dataset to the EUCAIM CDM

The Glioblastoma dataset has had a significant impact on the refinement of the EUCAIM CDM. Beyond demonstrating that the CDM can represent a detailed neuro-oncology cohort with multiple clinical timepoints (diagnosis, treatment, relapse and follow-up), the mapping exercise led to several concrete improvements in the structure, value sets and modelling guidelines of the CDM.

First, the dataset provided a clear example of one local concept being decomposed into several CDM attributes. For example, “Progression Date” is mapped to the “Episode” entity with episode type “Progression”, and the “start date” of the “Episode” be equal to the “Progression Date”. In addition, tumor location in the source data was recorded through concise fields such as “tumor location” (frontal, temporal, parietal, occipital) and “side of tumor epicenter” (left, right), and a field indicating whether the tumor involved eloquent brain areas. In the CDM these were mapped to a combination of Tumor.BodySite.Code, Tumor.BodySite.Location, Tumor.BodySite.Laterality and Tumor Observation for eloquent brain involvement. This mapping pattern clarified how anatomical location, sub-regional qualifiers and laterality should be represented, and it motivated the extension of the body-site and qualifier value sets in the hyper-ontology to properly cover neuro-anatomical structures.

Second, the Glioblastoma dataset helped to consolidate the representation of performance and health status assessments. The source variables included ASA physical status, Karnofsky Performance Status (KPS) and ECOG performance scores at diagnosis and relapse. These were all mapped to *Health Status Assessment* entities, using coded concepts for the assessment type (ASA, KPS, ECOG) and either categorical or numerical result fields as appropriate. This confirmed that the CDM can uniformly handle different performance scales

and led to clearer documentation of how “value as concept” and “value as number” should be used for such assessments.
Third, the dataset was instrumental in deciding the handling of tumoral specimens and genomics. In this dataset, molecular pathology information included IDH1, ATRX, p53, MGMT, TERT and Ki-67, across different timepoints, each recorded with local categories such as “mutated / not mutated / unknown”, “Methylated / Not methylated / Unknown” or numerical values, and were assessed using different methods such as immunohistochemistry or sequencing. During the mapping process, these variables were modelled as tumor biomarkers with a coded test concept (for example, “IDH1 gene variant measurement”, “Ki-67 marker of proliferation”) and either a categorical result (“mutated”, “not mutated”) or a numerical result (percentage for Ki-67). This exercise led to the formulation of a general modelling rule: Since biomarker tests refer both to the analysis of a patient’s tissue and/or blood (tumoral specimen) and to genomic tests (searching for specific mutations, multiple gene alterations, etc., that are somatic (not inherited) and/or non-genomic biomarkers (such as protein expression), all biomarkers should be represented under the generic “Tumor Marker Entity”. In addition, some genetic tests refer to testing for germinal (inherited mutations or cancer risk): inherited element that increases the risk for cancer (such as BRCA mutation or the Lynch syndrome). This rule has been adopted in the CDM documentation and is now reused when mapping molecular information from other datasets.
Fourth, the glioma dataset validated and refined the modelling of procedures, treatments and timepoints in a highly time-structured context. Biopsy dates, pre-surgical and post-surgical MR studies, extent of resection, radiotherapy and chemotherapy (including specific regimens such as temozolomide and bevacizumab) were consistently mapped to *Procedure*, *Surgical Procedure*, *Radiotherapy*, and *Medication Administration* entities, linked to the appropriate Episodes (Diagnosis episode, Treatment episode, Relapse episode). Each of the episodes groups all the related clinical events present in the dataset. This was a major refinement in the CDM, since the original CDM specification described in D5.2, based on mCode, didn’t support natively the grouping and association of clinical events into high-level abstractions (e.g. progression period) as OMOP-CDM supports with its oncology extension. As such in this CDM version we extended our model with the addition of OMOP Episode and Episode Event entities to support “time to progression”, “survival from reoperation” etc. related clinical questions.

Challenges in the Glioblastoma Dataset Semantic Annotation

The Glioblastoma dataset has enriched the hyper-ontology with a new cancer subtype, “Glioblastoma multiforme", with the associated histological types and subtypes, topography, tumor marker or genomic tests and results, medication, and standard imaging features (VASARI Criteria) and associated values. While the core and upper layers remain intact, the domain and domain-specific layers have been extended to cover the new concepts and relationships. The main challenge in the semantic annotation of this dataset is the non-standard format of variables and values, which introduces ambiguity, incompleteness, and heterogeneity, necessitating a manual data curation phase. For instance, tumor locations (e.g., Frontal (1), Temporal (2), parietal (4), and occipital (5)) could be mapped to generic brain region from SNOMED (e.g., Frontal brain region, Temporal brain region, Parietal region, Occipital region) or specific lobes of brain from ICD-O-3 (e.g., Frontal lobe, Temporal lobe, Parietal lobe, Occipital lobe). Moreover, tumor marker or genomic test acronyms are provided, such as IDH1, IDH1-BM, MGMT, ATRX, and TERT, which creates difficulties for accurate semantic

annotation. For instance, MGMT could be mapped to “MGMT (O-6-methylguanine-DNA methyltransferase) gene variant measurement” (Genetic Variation) from OMOP Genomic or “MGMT gene methylation [Presence] in Tissue by Molecular genetics method” (Lab Test) from LOINC. For data disambiguation, it is necessary to manually curate and revise these data with clinical experts to guide terminology selection and validate mappings and semantic classifications for tumor markers and genomic tests. However, data curation and disambiguation are both time- and effort-intensive tasks that require domain expertise and manual review alongside automated tools to ensure accuracy, consistency, and interoperability. Additional examples of incomplete or ambiguous information are presented in **Table 25**.

Table 25: *Examples of data (variables and associated values) from the Glioblastoma dataset and their semantic annotation with standard sources.*

Variable (DH)	Value (DH)	Standard Label	Standard Source:Code	EUCAIM ID
Histological type	Glioblastoma	Glioblastoma	ICDO3:9440/3	CLIN1000067
D-Pathologycal subtype	Giant cells; Oligodendroglioma component; Epithelioid; Gliosarcoma	Giant cell glioblastoma	ICDO3:9441/3	CLIN1000073
		Oligodendroglioma	ICDO3:9450/3	CLIN1000081
		Epithelioid glioblastoma	SNOMEDCT:733837004	CLIN1000069
		Gliosarcoma	ICDO3:9442/3	CLIN1000082
Tumor location	Frontal (1); Temporal (2); Insular (3); parietal (4); occipital (5); brainstem (6); cerebellum (7); basal ganglia(8); thalamus (9); corpus callosum (10)	Frontal lobe	ICDO3:C71.1	BP1000351
		Temporal lobe	ICDO3:C71.2	BP1000350
		Insular structure	SNOMEDCT:36169008	BP1000364
		Parietal lobe	ICDO3:C71.3	BP1000352
		Occipital lobe	ICDO3:C71.4	BP1000354
		Brain stem	ICDO3:C71.7	BP1000061
		Cerebellum	ICDO3:C71.6	BP1000373
		Basal ganglia	SNOMEDCT:32610002	BP1000375
		Thalamus	SNOMEDCT:42695009	BP1000376
		Corpus callosum	SNOMEDCT:88442005	BP1000378
D-IDH1	Mutated; Not mutated	IDH1 (isocitrate dehydrogenase (NADP(+)) 1) gene variant measurement	OMOPGenomic:5382	CLIN1051952
		IDH1 gene mutations found [Identifier] in Blood or Tissue by Molecular genetics method Nominal	LOINC:100022-3	CLIN1051946
		Mutated	NCIT:C177686	COM1001313
		Not mutated	NCIT:C177687	COM1001887

D-ATRX	Mutated;Not mutated (% expression)	ATRX (ATRX chromatin remodeler) gene variant measurement	OMOPGenomic:886	CLIN1051954
		ATRX gene mutations found [Identifier] in Blood or Tissue by Molecular genetics method Nominal	LOINC:53841-3	CLIN1051949
D-p53mut	Positive; Negative	TP53 (tumor protein p53) gene variant measurement	OMOPGenomic:11998	CLIN1051958
		P53 protein Ag [Presence] in Tissue by Immune stain		CLIN1051945
D-MGMT	Metilated; Not metilated	MGMT (O-6-methylguanine-DNA methyltransferase) gene variant measurement	OMOPGenomic:7059	CLIN1051955
		MGMT gene methylation [Presence] in Tissue by Molecular genetics method	LOINC:60252-4	CLIN1051951
		Methylated	NAACCR:3889@3	COM1001314
		Not Methylated	NAACCR:3889@0	COM1001889
D-IDH1-BM	mutated; not mutated	DH1 and IDH2 genes targeted mutation analysis in Blood or Tissue by Molecular genetics method	LOINC:95772-0	CLIN1045804
D-TERT	M-124;M-146; Not mutated Mutated 124; Mutated 146;	TERT (telomerase reverse transcriptase) gene variant measurement	OMOPGenomic:11730	CLIN1051956
		TERT gene promotor region targeted mutation analysis in Tissue by Molecular genetics method	LOINC:95778-7	CLIN1051950
		Mutated/c.1-124C>T	NCIT:C131107	COM1001317
		Mutated/c.1-146C>T	NCIT:C131106	COM1001318
D-Proliferation rate KI67	0-100 (%)	MKI67 (marker of proliferation Ki-67) gene variant measurement	OMOPGenomic:7107	CLIN1051957
ASA Physical Status Classification System	ASA I; ASA II; ASA III; ASA IV; ASA V; ASA VI Class I, Class II, Class III, Class IV, Class V, Class VI	American Society of Anesthesiologists physical status classification	SNOMEDCT:273270000	CLIN1047414
		ASA physical status class 1	SNOMEDCT:413495001	CLIN1052872
		ASA physical status class 2	SNOMEDCT:413496000	CLIN1052871
		ASA physical status class 3	SNOMEDCT:413497009	CLIN1052870

		ASA physical status class 4	SNOMEDCT:413498004	CLIN1052869
		ASA physical status class 5	SNOMEDCT:413499007	CLIN1052868
		ASA physical status class 6	SNOMEDCT:413500003	CLIN1052867

7.1.3 Thyroid dataset to the EUCAIM CDM

The Thyroid contains a collection of patients with scintigraphies and clinical data who have confirmed thyroid gland dataset cancer. It provides a basis for a) predicting response to radioactive iodine therapy in patients who have undergone thyroid cancer surgery and b) improvement in identifying risk of remaining thyroid cancer cells.

Table 26: *Thyroid dataset – Diagnosis episode: mapping of local data elements to EUCAIM CDM entities, attributes and value sets.*

Local Model		**EUCAIM Common Data Model**		
Local field	**Local Values**	**Entity**	**Attribute**	**Value Set**
Patient ID	<String>	Patient	Identifier	<Literal>
Population	0: Patient with Cancer, 1: Patient with lesion not being a malignant tumor	Patient	Diagnostic Category	Patient with Cancer (COM1001087) Patient with lesion not being a malignant tumor (COM1001086)
Age at Diagnosis	<Integer>	Cancer Condition	Age at diagnosis	<Decimal>
Sex	0: Male 1: Female	Patient	Birth Sex	FEMALE (COM1000178) MALE (COM1000176)
Date of pathology confirmation / diagnosis date *	<Date>	Cancer Condition	Asserted Date	<Date>
Pathology confirmation*	0: histological surgery, 1: biopsy, 2: Imaging	Procedure	Code	if 0: Histologic test (CLIN1034154) if 1: Biopsy(CLIN1001712) if 2: Imaging (Procedure) (IMG1005453)
			Category	If 0: Laboratory test (CLIN1033407) If 1: Removal (CLIN1058186) If 2: Imaging (Procedure) (IMG1005453)
Topography : Organ	Thyroid BED	Cancer Condition	Topography	Thyroid (BP1000286)
Pathology*	Thyroid Cancer	Cancer Condition	Code	Malignant neoplasm of thyroid gland (CLIN1059392)

			Type	Primary (COM1000017)
Histological subtype	0: P (papillary), 1: FVP (follicular variant of papillary thyroid carcinoma), 2: F (follicular), 3: HC(Hurthle Cell), 4: IC (Insular carcinoma), 5: TVP (Trabecular variant papillary)	Cancer Condition	Histology Morphology Behavior	0: Papillary adenocarcinoma, NOS (CLIN1049465) 1: Papillary carcinoma, follicular variant (CLIN1049605) 2: Follicular carcinoma, NOS (CLIN1049604) 3: Oxyphilic adenocarcinoma (CLIN1049463) 4: Insular carcinoma (CLIN1049458) 5: Trabecular adenocarcinoma (CLIN1049606)
Imaging procedure protocol*	WB scintigraphy with I-131	Procedure	Code	Scintigraphy (IMG1000079)
			Category	Imaging (Procedure) (IMG1005453)

Table 27: *Thyroid dataset – Treatment episode #1: mapping of local data elements to EUCAIM CDM entities, attributes and value sets.*

First Treatment	Surgical Procedure	Treatment	Treatment Type	If yes: Surgical Procedure (CLIN1004413)
Date of first treatment*	Months from diagnosis		Offset from Diagnosis	<Date>
			Offset Unit	month (COM1000154)
T	T0, T1a, T1b, T2, T3a, T3b, T4a, T4b	Cancer Stage	Code	pT category (CLIN1033391)
			Method	AJCC/UICC 8th edition
			Value	pT0 (COM1000708), pT1a (COM1000742), pT1b (COM1000743), pT2 (COM1000775), pT3a (COM1000960), pT3b (COM1001000), pT4a (COM1001577), pT4b (COM1001585)
N	N0, N1a, N1b	Cancer Stage	Code	pN category (CLIN1033368)
			Method	AJCC/UICC 8th edition
			Value	pN0 (COM1000708), pN1a (COM1000744),

				pN1b (COM1000745)
M	M0, M1	Cancer Stage	Code	pM category (CLIN1033349)
			Method	AJCC/UICC 8th edition
			Value	pM0 (COM1000708), pM1 (COM1000531)

Table 28: *Thyroid dataset – Treatment episode #2: mapping of local data elements to EUCAIM CDM entities, attributes and value sets.*

Date of Timepoint 1 (Months from surgery)	<Integer>	Treatment	Offset from Diagnosis	Calculated sum ((this)+Treatment_offset when type=surgical_procedure)
			Offset Unit	month (COM1000154)
Treatment TP1	Type of treatment received by the patient in ATC code		Type	Medication Administration (CLIN1034187)
		Medication Administration	MedicationCode	subclasses of CLIN1034110 (Drug or medicament)
Imaging procedure TP1	SCINT, SPECT-CT	Procedure	Code	Scintigraphy (IMG1000079), Single photon emission computed tomography with computed tomography (IMG1004158)
			Category	Imaging (Procedure) (IMG1005453)
Tg TP1	NG/ML	Lab Test Result	Code	Thyroglobulin [Mass/volume] in Serum or Plasma
			Value as number	<Decimal>
			Value Unit	nanogram per milliliter (COM1000156)
TgAb TP1	IU/ml	Lab Test Result	Code	Thyroglobulin Ab [Units/volume] in Serum or Plasma (CLIN1049787)
			Value as number	<Decimal>
			Value Unit	international unit per milliliter (COM1000168)
TSH TP1	mU/l	Lab Test Result	Code	Thyrotropin [Units/volume] in Blood (CLIN1033900)
			Value as number	<Decimal>
			Value Unit	milliunit per liter (COM1000170)
Tg stim TP1	NG/ML	Lab Test Result	Code	Stimulated Thyroglobulin measurement (CLIN1050572)

			Value as number	<Decimal>
			Value Unit	nanogram per milliliter (COM1000156)
TgAb stim TP1	IU/ml		Code	Stimulated Thyroglobulin antibody measurement (CLIN1050466)
			Value as number	<Decimal>
			Value Unit	international unit per milliliter (COM1000168)

Table 29: *Thyroid dataset – Treatment episode #3: mapping of local data elements to EUCAIM CDM entities, attributes and value sets.*

Date of Timepoint 2	Months from diagnosis	Treatment	Offset from Diagnosis	Calculated sum ((this)+Treatment_offset when type=surgical_procedure)
			Offset Unit	month (COM1000154)
Treatment TP2	Type of treatment received by the patient in ATC code		Type	Medication Administration (CLIN1034187)
Imaging procedure TP2	SCINT, SPECT-CT	Procedure	Code	Scintigraphy (IMG1000079), Single photon emission computed tomography with computed tomography (IMG1004158)
			Category	Imaging (Procedure) (IMG1005453)
Tg TP2	NG/ML	Same as above for TP1.		
TgAb TP2	IU/ml			
TSH TP2	mU/l			
Tg stim TP2	NG/ML			
TgAb stim TP2	IU/ml			

Table 30: *Thyroid dataset – Treatment episode #4: mapping of local data elements to EUCAIM CDM entities, attributes and value sets.*

Date of Timepoint 3	Months from diagnosis	Same as above for TP1, TP2.
Treatment TP3	Type of treatment received by the patient in ATC code	

Imaging procedure TP3	SCINT, SPECT-CT	
Tg TP3	NG/ML	
TgAb TP3	IU/ml	
TSH TP3	mU/l	
Tg stim TP3	NG/ML	
TgAb stim TP3	IU/ml	

Impact of the Thyroid dataset to the EUCAIM CDM

The thyroid dataset mainly stressed the terminological and value-set dimension of the EUCAIM CDM rather than its structure. From a structural point of view, all information in the dataset—primary thyroid cancer, surgery, radioactive iodine therapy, follow-up scintigraphies and laboratory results—could be represented using existing CDM entities (Patient, Cancer Condition, Cancer Stage, Procedure / Imaging Procedure, Medication Administration, Lab Test Result). No new tables or attributes were required.

The impact was therefore concentrated on extending controlled vocabularies in the hyper-ontology:

- In the imaging domain, new modality concepts and codes were added for *scintigraphy* and *SPECT-CT* and linked to the Imaging Procedure category, so that nuclear medicine examinations can be represented in the same way as CT, MR or PET.
- In the disease domain, additional thyroid-specific histological subtypes (papillary, follicular variant, follicular, Hurthle cell, insular, trabecular) were incorporated into the histology value set used by the Cancer Condition entity.
- In the laboratory domain, specific endocrine and tumour-marker tests—such as thyroglobulin (Tg), thyroglobulin antibodies (TgAb), stimulated Tg and stimulated TgAb, as well as TSH—were added or aligned to the Lab Test Result code sets, with harmonized units (ng/mL, IU/mL, mU/L).

These additions confirmed that the CDM's existing structure is flexible enough to accommodate nuclear medicine workflows and thyroid-specific oncology practice, while highlighting where new modality, histology and lab-test terms need to be curated at the terminology layer. The thyroid dataset thus primarily acted as a driver for vocabulary enrichment and mapping rules rather than for structural evolution of the CDM.

Challenges in the Thyroid dataset Semantic Annotation

The Thyroid dataset has extended the hyper-ontology with a new cancer type, "Malignant neoplasm of thyroid gland", with the associated histological types, topography, procedures, cancer staging, and lab tests and results. As with previous datasets, the enrichment process primarily affected domain- and domain-specific content, while core aspects remained intact.

The main challenge of this dataset is the non-standard format of the provided variables and values, which introduces ambiguity and incompleteness. A manual data-curation phase was required to ensure accurate semantic annotation and coherent integration into the hyper-ontology at precise specificity levels.

Table 31: *Examples of data (variables and associated values) from the Thyroid dataset and their semantic annotation with standard sources.*

Variable (DH)	Value (DH)	Standard Label	Standard Source:Code	EUCAIM ID
Primary cancer condition	Thyroid cancer	Malignant neoplasm of thyroid gland	ICD10:C73	CLIN1059392
Histology Morphology Behavior	0: P (papillary), 1: FVP (follicular variant of papillary thyroid carcinoma), 2: F (follicular), 3: HC(Hurthle Cell), 4: IC (Insular carcinoma), 5: TVP (Trabecular variant papillary)	0: Papillary adenocarcinoma, NOS	ICDO3:8260/3	CLIN1049465
		1: Papillary carcinoma, follicular variant	ICDO3:8340/3	CLIN1049605
		2: Follicular carcinoma, NOS	ICDO3:8330/3	CLIN1049604
		3: Oxyphilic adenocarcinoma	ICDO3:8290/3	CLIN1049463
		4: Insular carcinoma	ICDO3:8337/3	CLIN1049458
		5: Trabecular adenocarcinoma	ICDO3:8190/3	CLIN1049606
Imaging procedure	SCINT, SPECT-CT	Scintigraphy	RadLex:RID34428	IMG1000079
		Single photon emission computed tomography with computed tomography	SNOMEDCT:450438002	IMG1004158
Laboratory test	Thyroglobulin test	Thyroglobulin [Mass/volume] in Serum or Plasma	LOINC:3013-0	CLIN1050573
	Thyroglobulin Antibody test	Thyroglobulin Ab [Units/volume] in Serum or Plasma	LOINC:8098-6	CLIN1049787
	Thyrotropin test	Thyrotropin [Units/volume] in Blood	LOINC:3015-5	CLIN1033900
	Stimulated Thyroglobulin test	Thyroid stimulating immunoglobulins [Units/volume] in Serum	LOINC:30567-2	CLIN1033781
	Stimulated Thyroglobulin Antibody test	N/A	N/A	CLIN1050466

8. Governance and Sustainability

The EUCAIM hyper-ontology and Common Data Model (CDM) have evolved from initial conceptual artefacts into operational components that support the federation. Their long-term value, however, depends on robust governance and on a sustainable maintenance model that can cope with the evolution of standards, new use cases and new data holders joining the

federation. This section outlines how changes will be managed over time and how continuity beyond the project's lifetime can be ensured.

8.1. Versioning and Change Management

The hyper-ontology and CDM are both living artefacts. New cancer types, modalities and use cases will emerge. External standards such as SNOMED CT, ICD, DICOM, OMOP and FHIR will change, and experience from pilots and production use will reveal areas for refinement. To handle this evolution in a controlled way, EUCAIM adopts a structured approach to versioning and change management.

Versioning scheme

Both artefacts have been following a semantic versioning scheme (*MAJOR.MINOR*) throughout their development and they will continue to use the same scheme. Each release is accompanied by a changelog and updated documentation. During the project, versions were also assigned persistent identifiers (e.g. DOIs) and published on public repositories (such as Zenodo, e.g. https://zenodo.org/records/17237843), so that specific versions can be reliably cited and referenced. This will continue to happen after the end of the project.

Change proposal and review process

Change requests can originate from WP5, from other work packages, from data holders during onboarding, or from external stakeholders. Proposals should be captured in a structured form through the EUCAIM helpdesk, including:

- description of the requested change
- affected entities, attributes, value sets or mappings
- example use cases and datasets

Proposals are then screened by the WP5 technical team, comprising ontology engineers, data modelers and clinical experts. Approved changes are scheduled into upcoming releases with a clear indication of whether they are major or minor.

Relationship to EUCAIM min-FIF and max-FIF

The current versions of the hyper-ontology and CDM are aligned with the minimum data Federation and Interoperability framework (min-FIF), which defines the essential semantic and technical requirements for dataset cataloguing, federated querying and basic federated processing. As EUCAIM progresses towards a maximized framework (max-FIF) with richer clinical, imaging and molecular content and more advanced query criteria, new requirements will feed into the change process described above. This ensures that evolution towards max-FIF is incremental, governed and backwards compatible whenever possible.

8.2. Long-term Sustainability

The long-term sustainability of the hyper-ontology and CDM is critical to the success of the Cancer Image Europe platform and its alignment with other European data spaces. Sustainability has organizational, technical and community dimensions.

Governance framework and roadmap

Interoperability is not purely a technical task, but rather it also requires ongoing governance. Building on the mechanisms described above, the future operational entity for EUCAIM (e.g. within the EDIC framework) is expected to host

- a small central team responsible for maintaining repositories, tooling and releases,
- a technical team with semantic expertise, which oversees the evolution of the models,
- a clear roadmap for planned extensions (e.g. new tumour types, modalities, alignment with emerging EU profiles).

This roadmap will be updated regularly, taking into account new use cases from hospitals and laboratories joining the community, as well as changes in external standards and EU policy (e.g. EHDS, HealthData@EU, XSHARE, QUANTUM).

Furthermore, to remain comprehensive and relevant, the CDM and hyper-ontology must reflect the needs of a broad set of stakeholders. EUCAIM therefore promotes open access to the models themselves (OWL/RDF, schemas, data dictionaries), training materials, tutorial videos and worked examples. Feedback and contributions from the wider community are encouraged through workshops, helpdesk channels, and scientific publications that describe the approach and lessons learned. This will allow other projects and infrastructures to reuse and extend the artefacts, and to feed improvements back into EUCAIM.

Resources and sustainability planning

Maintaining and evolving the hyper-ontology and CDM requires both financial and human resources. These needs are being analyzed in detail in WP8, including options such as:

- dedicated funding for a central interoperability team,
- in-kind contributions of expert time from participating institutions,
- partnerships with other European infrastructures and initiatives.

This analysis will inform the long-term governance model and ensure that upkeep of the semantic and structural framework is adequately resourced.

9. Conclusions and Next Steps

This deliverable presents the final specification of the EUCAIM hyper-ontology and Common Data Model. It marks a transition from prototype designs to an operational semantic and technical framework used in technical pilots and node onboarding.
Despite the major progress achieved with the EUCAIM CDM and hyper-ontology, several challenges remain. Till the end of the project and the definition of the max-FIF, the CDM and

hyper-ontology will be gradually extended to additional cancer types, imaging modalities and biomarker domains, when necessary, guided by concrete use cases and new data holders. Alignment with emerging European frameworks—such as the European Health Data Space, future versions of HealthDCAT-AP and cross-infrastructure initiatives in the cancer and imaging domains—will be pursued to maximize reuse and interoperability beyond EUCAIM.

In addition, a major remaining challenge lies in the sustainability and evolution of the hyper-ontology and its terminology bindings. Oncology and medical imaging knowledge are changing rapidly: new molecular subclasses, biomarkers, therapies (e.g. cancer vaccines) and imaging techniques (e.g. molecular imaging, hybrid and fusion modalities) continually appear and need to be reflected at the semantic layer. In parallel, external standards such as ICD, ICD-O, SNOMED CT and OMOP are themselves evolving (e.g. transitions from ICD-10/ICD-O-3 to ICD-11/ICD-O-4), sometimes introducing changes in the way primary and secondary neoplasms are classified. This can lead to semantic inconsistencies and even logical contradictions if earlier modelling choices are not systematically revised, and it requires continuous monitoring and expert curation to keep the hyper-ontology aligned with current versions of these standards. The imaging domain presents additional difficulties. Many relevant acquisition details and protocol descriptions are encoded in DICOM using non-standardized or locally defined values. Accurately representing such terms in the hyper-ontology—clarifying their meaning, mapping them to standard concepts and integrating them into the imaging taxonomy—demands targeted expert effort and the use of semantic patterns. While these patterns improve precision and interoperability, applying them systematically risks increasing the size and logical complexity of the ontology, which can make it harder to handle for non-expert users and lightweight applications. Managing this trade-off between semantic richness and practical usability, while keeping pace with evolving medical knowledge and standards, will remain a core focus of the future work.

10. Publications

- El Ghosh M., Kalokyri V., Bobowicz M., Sambres M., Vaterkowski M., Giraldo O., Charlet J., Fournier L., Duclos C., Tannier X., Tsakou G., Tsiknakis M., Dhombres F., Daniel C. Integrating Heterogeneous Real-World Cancer Image Data for Semantic Interoperability in Oncology and Medical Imaging: Development and Usability of the Cancer Image Europe Hyper-Ontology. Journal of Medical Internet Research. UNDER_REVIEW

- Daniel C., El Ghosh M., Kalokyri V. Fournier L., Tsakou G., Tsiknakis M. Ontology-based FAIRification of Europe Cancer Image (EUCAIM) data sets applied to prostate and liver cancers. AMIA informatic summit. UNDER_REVIEW

- Kalokyri V., El Ghosh M., Sfakianakis S., Daniel C., Kaldeli E., Radoszynski M.I., Pombo V.S., Saint-Aubert L., Dhombres F., Tsakou, G, Tsiknakis, M. The EUCAIM Common Data Model: a unified representation for pan-European cancer imaging data and its semantic integration with the EUCAIM hyper-ontology. Nature, Scientific Data. UNDER_PREPARATION

- Tsave O., Kalokyri V., El Ghosh M., Sfakianakis S., Mazzetti S., Daniel C., Dhombres F., Tachos N., Marias K., Tsiknakis M., Chouvarda I. “The Necessity of Harmonized Quality Data in Medical Repositories. Challenges and Best Practices in Cancer Imaging Data Pre-validation.” In Trustworthy AI in Cancer Imaging Research. Springer, Cham. 2025.

- El Ghosh, M., Kalokyri, V., Sambres, M., Vaterkowski, M., Duclos, C., Tannier, X., Tsakou, 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., Tsakou, 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.

Annex A - EUCAIM CDM Data Dictionary

Entity: Patient (Required)

Prefix eucaim: https://cancerimage.eu/ontology/EUCAIM#

Element Name	Attribute Name	Definition	Required	Cardinality	Data Type	Binded property (object/data)	Range
Identifier	patient_id	Anonymized patient identifier which is unique within the context of the system. The identifier should be identical to the identifier present in the DICOM images for this patient.	Required	1..1	String	eucaim:PatientIdentifier	xsd:string
Birth Sex	patient_birth_sex	A code classifying the person's sex assigned at birth (i.e. biological sex).	Required	1..1	CodeableConcept	eucaim:hasBirthSex	eucaim:COM1001396
Ethnicity	patient_ethnicity	Concepts classifying the person into a named category of humans sharing common history, traits, geographical origin or nationality.	Optional	0..1	CodeableConcept	eucaim:hasEthnicity	eucaim:COM1001058
Race	patient_race	Concepts classifying the person into groups based on their physical appearance	Optional	0..1	CodeableConcept	eucaim:hasRace	eucaim:COM1001200
Birth Date	patient_birth_date	The date of birth for the individual (e.g. just year or year + month) as used in human communication. The format is ISO8601: YYYY-MM-DD	Conditional (Required if diagnosis age is not available)	0..1	DateTime	eucaim:BirthDate	xsd:dateTime
Gender	patient_gender	Administrative Gender - the gender that the patient is considered to have for administration and record keeping purposes.	Optional	0..1	CodeableConcept	eucaim:hasGender	eucaim:COM1000175
Deceased	patient_deceased	Indicates if the individual is deceased or not.	Optional	0..1	Boolean	eucaim:Deceased	eucaim:COM1000034/ xsd:boolean
Cause Of Death	patient_cause_of_death	Main cause of death of the patient (i.e. Condition)	Optional	0..1	CodeableConcept	eucaim:hasCauseOfDeath	
Date Of Last Contact	patient_date_of_last_contact	Date of last contact if not deceased, or date of death if deceased.	Optional	0..1	Date	eucaim:DateOfLastContact	eucaim:COM1001949/ xsd:dateTime

Managing Organization	patient_managing_organization	Organization that is the custodian of the patient record. Need to know who recognizes this patient record, manages and updates it. This should be anonymized and be identical to the DICOM "Site ID", in case the organization managing the patient record is the same with the organization in which the DICOM studies were performed.	Required	1..1	String	eucaim:ManagingOrganizationIdentifier	xsd:string
Diagnostic Category	patient_diagnostic_category	The categorization of the subject/patient in the dataset based on their status. Possible values for negative cases: "Subject on Screening with a negative result", "Subject on a Control group" or possible values for positive cases: "Patient with Cancer", "Patient with lesion not being a malignant tumor", or indeterminate cases: "Subject under discussion with suspicious findings"	Required	1..1	CodeableConcept	eucaim:hasDiagnosticCategory	eucaim:COM1001084

Entity: Medical History (Optional)

Element Name	Attribute Name	Definition	Required	Cardinality	Data Type	Binded property (object/data)	Range
Code	medical_history_code	Code of the related medical history that is known for this patient. The medical history might correspond to prior conditions unrelated to cancer, symptoms, prior medications, or prior medical procedures. In addition, smoking or alcohol habits should be registered here.	Required	0..1	CodeableConcept	eucaim:hasMedicalHistory	
Patient	medical_history_patient	The patient that the medical history is about.	Required	0..1	Reference: Patient	eucaim:hasSubject	eucaim:COM1001047
Category	medical_history_category	The category refers to the type of medical history being recorded. Possible values are: Observation, Condition, Measurement, Medication, Procedure	Required	0..1	CodeableConcept	eucaim:hasCategory	eucaim:GEN1000011
Onset Age	medical_history_onset_age	When the medical history observation/condition/medication/procedure first manifested on the patient.	Optional	0..1	Integer	eucaim:OnsetAge	xsd:int

Onset Age Unit	medical_history_onset_age_unit	Unit concept of the age (e.g. months, years, days) unit of time (UCUM)	Conditional (required if Onset Age is present)	0..1	CodeableConcept	eucaim:hasUnitofMeasure	eucaim:COM1000147
Value As Concept	medical_history_value_as_concept	The value of the medical history being reported as a concept	Optional	0..1	CodeableConcept	eucaim:hasMedicalHistoryValue	
Value As Number	medical_history_value_as_number	The value of the medical history being reported as a numerical value	Optional	0..1	Float	eucaim:MedicalHistoryValue	xsd:float
Value Unit	medical_history_value_unit	The value unit of the medical history being reported as a numerical value	Conditional (required if Value As Number is present)	0..1	CodeableConcept	eucaim:hasUnitofMeasure	eucaim:COM1000147
Present	medical_history_present	Indicates if this particular medical history being reported is currently present or not.	Optional	0..1	boolean	eucaim:Present	xsd:boolean

Entity: Family Member History (Optional)

Element Name	Attribute Name	Definition	Required	Cardinality	Data Type	Binded property (object/data)	Range
Patient	family_member_history_patient	The patient that the family history is about.	Required	0..1	Reference: Patient	eucaim:hasSubject	eucaim:COM1001047
Relationship	family_member_history_relationship	Relationship to the subject	Required	0..1	CodeableConcept	eucaim:hasRelationship	eucaim:COM1001043
Condition Code	family_member_history_condition_code	Condition that the related person has/doesn't have.	Required	0..1	CodeableConcept	eucaim:hasCondition	eucaim:CLIN1007988
Condition Present	family_member_history_condition_present	Indicates if the condition of the relative is present or not.	Optional	0..1	boolean	eucaim:ConditionPresent	xsd:boolean
Onset Age	family_member_history_onset_age	When condition first manifested on the relative.	Optional	0..1	Integer	eucaim:OnsetAge	xsd:int
Onset Age Unit	family_member_history_onset_age_unit	Unit concept of the age (e.g. months, years, days) unit of time (UCUM)	Conditional (required if Onset Age is present)	0..1	CodeableConcept	eucaim:hasUnitofMeasure	eucaim:COM1000147

Entity: Comorbidities (Optional)

Element Name	Attribute Name	Definition	Required	Cardinality	Data Type	Binded property (object/data)	Range
Condition Code	comorbidities_condition_code	A comorbid condition that is known to be present/absent	Required	0..1	CodeableConcept	eucaim:hasComorbidity	eucaim:CLIN1002855
Condition Present	comorbidities_condition_present	The information determines if the comorbid condition is present	Optional	0..1	boolean	eucaim:ConditionPresent	xsd:boolean
Patient	comorbidities_patient	The patient whose comorbidities are recorded.	Required	0..1	Reference: Patient	eucaim:hasSubject	eucaim:COM1001047

Entity: Health Status Assessment (Optional)

Element Name	Attribute Name	Definition	Required	Cardinality	Data Type	Binded property (object/data)	Range
Code	health_status_assessment_code	The name of the Health Status Assessment test performed on a patient during the whole patient's cancer journey. The assessment can either be a Performance Status Assessment (Measurement) such as ECOG, Karnofsky etc. or a Medical Assessment. A Medical Assessment may include symptoms directly communicated by the patient (e.g. breast pain, fatigue, etc) or signs reported by the medical expert. In case the symptoms are part of adverse events by a treatment, these need to be registered in the respective Adverse Event entity.	Required	0..1	CodeableConcept	eucaim:hasHealthStatusAssessment	eucaim:CLIN1037292
Patient	health_status_assessment_patient	Patient whose health assessment status is recorded.	Required	0..1	Reference: Patient	eucaim:hasSubject	eucaim:COM1001047
Value As Number	health_status_assessment_value_as_number	The information determined as a result of making the observation, if the information has a numerical value.	Optional	0..1	Decimal	eucaim:AssessmentTestValue	xsd:decimal
Value Unit	health_status_assessment_value_unit	If a numeric value, the unit **SHALL** be selected from [UCUM](http://unitsofmeasure.org).	Conditional (required if value as number is	0..1	CodeableConcept	eucaim:hasUnitofMeasure	eucaim:COM1000147

			present)				
Value As Concept	health_status_assessment_value_as_concept	The information determined as a result of making the observation, if the information has a concept value	Optional	0..1	CodeableConcept	eucaim:hasAnswer, eucaim:hasAnswerAsFinding	eucaim:COM1001110, eucaim:CLIN1052860, eucaim:CLIN1000004
Interpretation	health_status_assessment_interpretation	A categorical assessment of an observation value. For example, high, low, normal.	Optional	0,,1	CodeableConcept	eucaim:hasInterpretationValue	eucaim:COM1000181

Entity: Tumor Marker Test (Optional)

Element Name	Attribute Name	Definition	Required	Cardinality	Data Type	Binded property (object/data)	Range
Patient	tumor_marker_test_patient	Patient whose tumor marker test result is recorded.	Required	1..1	Reference: Patient	eucaim:hasSubject	eucaim:COM1001047
Code	tumor_marker_test_code	The test that was performed. A LOINC **SHALL** be used if the concept is present in LOINC.	Required	1..1	CodeableConcept	eucaim:hasTumorMarkerTest	eucaim:CLIN1000224
Category	tumor_marker_test_category	A code that classifies the general type of observation being made.	Required	1..1	CodeableConcept	eucaim:hasCategory	eucaim:GEN1000014
Result Value As Concept	tumor_marker_test_result_value_as_concept	The Laboratory result value. If a coded value, the Result Value as Concept **SHOULD** be selected from [SNOMED CT](http://hl7.org/fhir/ValueSet/uslab-obs-codedresults) if the concept exists.	Optional	1..1	CodeableConcept	eucaim:hasResultValue	eucaim:COM1000181
Result Value As Number	tumor_marker_test_result_value_as_number	The Laboratory result value if numeric	Optional	0..1	Float	eucaim:TumorMarkerTestValue	xsd:float

Result Value Unit	tumor_marker_test_result_value_unit	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.	Conditional (Required if result is a numeric value)	0..1	CodeableConcept	eucaim:hasUnitofMeasure	eucaim:COM1000147
Date	tumor_marker_test_date	The date the tumor marker test happened.	Optional	0..1	Date	eucaim:PerformedDate	xsd:dateTime
Offset From Diagnosis	tumor_marker_test_offset_from_diagnosis	The interval of time elapsed before/after the diagnosis.	Optional	0..1	Integer	eucaim:OffsetFromDiagnosis	xsd:int
Offset Unit	tumor_marker_test_offset_unit	Unit concept of the time interval (e.g. months, years, days) unit of time (UCUM)	Conditional (Required if offset is present)	0..1	CodeableConcept	eucaim:hasUnitofTime	eucaim:COM1000150
Related Condition	tumor_marker_test_related_condition	Associates the tumor marker test with a condition, if one exists. In the case of a screening test such as prostate-specific antigen (PSA), there may be no existing condition to reference.	Optional	0..1	Reference: CancerCondition	eucaim:hasRelatedCondition	eucaim:CLIN1007977
Related Procedure	tumor_marker_test_related_procedure	Associates the tumor marker test with a procedure, if one exists.. In the case of a screening test such as prostate-specific antigen (PSA), there may be no existing procedure to reference.	Optional	0..1	Reference: Procedure	eucaim:hasRelatedProcedure	eucaim:CLIN1000005

Entity: Cancer Condition (Required)

Element Name	Attribute Name	Definition	Required	Cardinality	Data Type	Binded property (object/data)	Range
Age At Diagnosis	cancer_condition_age_at_diagnosis	The patient age on which the existence of the Condition was first asserted or acknowledged.	Required	1..1	Decimal	eucaim:AgeAtDiagnosis	xsd:decimal
Patient	cancer_condition_patient	Indicates the patient who the condition record is associated with.	Required	1..1	Reference: Patient	eucaim:hasSubject	eucaim:COM1001047
Code	cancer_condition_code	Identification of the cancer condition. ICD-10 is the preferred condition terminology	Required	1..1	CodeableConcept	eucaim:hasCondition	eucaim:CLIN1007977

Topography	cancer_condition_topography	The topography where this condition manifests itself. (Values from ICDO3 topography)	Required	1..*	CodeableConcept	eucaim:hasTopography	eucaim:BP1000265
Histology Morphology Behavior	cancer_condition_histology_morphology_behavior	The morphologic and behavioral characteristics of the cancer. (Values from ICDO3 histology)	Required	1..1	CodeableConcept	eucaim:hasHistologyMorphologyBehavior	eucaim:CLIN1000033
Clinical Status	cancer_condition_clinical_status	The clinical status of the condition. Possible values: Active, Recurrence, Relapse, Inactive, Remission, Resolved.	Optional	0..1	CodeableConcept	eucaim:hasClinicalStatus	eucaim:COM1001176
Asserted Date	cancer_condition_asserted_date	Date the condition was first asserted	Required if available	1..1	Date	eucaim:AssertedDate	xsd:dateTime
Evidence Type	cancer_condition_evidence_type	Categorization of the kind of evidence contributing to a clinical judgment on cancer disease. Possible values: https://build.fhir.org/ig/HL7/fhir-mCODE-ig/ValueSet-mcode-cancer-disease-status-evidence-type-vs.html	Optional	0..1	CodeableConcept	eucaim:hasEvidenceType	eucaim:GEN1000016
Related Procedure	cancer_condition_related_procedure	Associates the cancer condition with a pathology confirmation procedure (biopsy, surgery) or by imaging in specific cases such as HCC	Required	1..*	Reference: Procedure	eucaim:hasRelatedProcedure	eucaim:CLIN1000005
Related Primary Cancer Condition	cancer_condition_related_primary_cancer_condition	A reference to the primary cancer condition that provides context for this resource.	Optional	1..1	Reference: CancerCondition	eucaim:hasRelatedPrimaryCancerCondition	eucaim:CLIN1007988
Related Procedure	cancer_condition_related_procedure	Associates the cancer condition with a pathology confirmation procedure (biopsy, surgery) or by imaging in specific cases such as HCC	Required	1..*	Reference: Procedure	eucaim:hasRelatedProcedure	eucaim:CLIN1000005
Type	cancer_condition_type	If this the a primary or a secondary cancer condition	Required	1..1	CodeableConcept	eucaim:hasCancerConditionType	eucaim:COM1001133
Offset From Primary	cancer_condition_offset_from_primary	The interval of time elapsed before/after the primary diagnosis if this is a secondary condition	Optional	0..1	Integer	eucaim:OffsetFromDiagnosis	xsd:int
Offset Unit	cancer_condition_offset_unit	Unit concept of the time interval (e.g. months, years, days) unit of time (UCUM)	Conditional (Required if offset is present)	0..1	CodeableConcept	eucaim:hasUnitofTime	eucaim:COM1000150

Entity: Cancer Stage (Optional)

Element Name	Attribute Name	Definition	Required	Cardinality	Data Type	Binded property (object/data)	Range
Code	cancer_stage_code	The kind of stage reported, e.g., a pathologic TNM stage, a Lugano lymphoma stage, or a Rai stage for leukemia. The distinction between cancer_stage_code and cancer_stage_method is important. The cancer_stage_code identifies the kind of stage being reported while cancer_stage_code_method represents the staging system used to determine the code. Cancer_stage_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 the cancer_stage_code, the cancer_stage_method is not required. However, when cancer_stage_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 cancer_stage_method. The value (cancer_stage_value as CodeableConcept) 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 wholy self-identifying, it is not a reliable indicator of the type of stage being reported or the method of staging. Therefore, the cancer_stage_code must in all cases be reported.	Required	1..1	CodeableConcept	eucaim:hasCancerStage	eucaim:CLIN1044182
Method	cancer_stage_method	The staging system or protocol used to determine the stage, stage group, or category of	Conditional (required if	1..1	CodeableConcept	eucaim:hasStageMethod	eucaim:CLIN1044182

		the cancer based on its extent. When the staging system is implied by cancer_stage_code, the method is not required. However, when the cancer stage 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 the method.	stage_code doesnt imply the staging system)				
Value	cancer_stage_value	The stage, stage group, category, or classification resulting from the staging evaluation.	Conditional (required if stage_code doesnt imply the value)	1..1	CodeableConcept	eucaim:hasStage Value	eucaim:COM1000183 , eucaim:CLIN1044308
Patient	cancer_stage_patient	The patient associated with staging assessment.	Required	1..1	Reference: Patient	eucaim:hasSubje ct	eucaim:COM1001047
Related Procedure	cancer_stage_related _procedure	The procedure from which the cancer stage was determined. It can either be an imaging examination (MRI), biopsy, surgery.	Optional	0..1	Reference: Procedure	eucaim:hasRelate dProcedure	eucaim:CLIN1000005
Related Condition	cancer_stage_related _condition	Staging is associated with a particular cancer condition.	Required	1..1	Reference: CancerCondition	eucaim:hasRelate dCondition	eucaim:CLIN1007977

Entity: Histologic Grade (Optional)

Element Name	Attribute Name	Definition	Required	Cardi nality	Data Type	Binded property (object/data)	Range
Code	histologic_grade_code	A code that specifies the type of the histologic grade being recorded.	Required	1..1	CodeableConcept	eucaim:hasHistologi cGrade	eucaim:CLIN1049681
Patient	histologic_grade_patient	Patient whose test result is recorded.	Required	1..1	Reference: Patient	eucaim:hasSubject	eucaim:COM1001047
Scoring System	histologic_grade_scoring _system	Indicates the histologic grade scoring system.	Optional	0..1	CodeableConcept	eucaim:hasHistologi calGradingSystem	eucaim:CLIN1037296
Value	histologic_grade_value	The grade result value.	Optional	0..1	CodeableConcept	eucaim:hasHistologi cGradeValue	eucaim:COM1001360
Related Condition	histologic_grade_related _condition	Associates the histologic grade test with a condition, if one exists.	Optional	0..1	Reference: CancerCondition	eucaim:hasRelated Condition	eucaim:CLIN1007977

Related Procedure	histologic_grade_related_procedure	Associates the histologic grade test with a procedure from which the specimen was acquired.	Optional	0..1	Reference: Procedure	eucaim:hasRelatedProcedure	eucaim:CLIN1000005

Entity: Tumor (Optional)

Element Name	Attribute Name	Definition	Required	Cardinality	Data Type	Binded property (object/data)	Range
identifier	tumor_identifier	Stable identifier(s) for this specific tumor. The identifiers MUST be unique within the context of the referenced `Patient`.	Required	1..*	String	eucaim:TumorIdentifier	xsd:string
Is Index	tumor_is_index	This denotes if it is believed that the tumor is the primary tumor/index lesion causing the cancer.	Optional	0..1	Boolean	eucaim:IsIndex	xsd:boolean
Morphology	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	eucaim:hasAssociatedMorphology	eucaim:CLIN1049522
Patient	tumor_patient	The patient associated with this tumor.	Required	1..1	Reference: Patient	eucaim:hasSubject	eucaim:COM1001047
Volume	tumor_volume	The volume of the lesion	Optional	0..1	Float	eucaim:TumorVolumeValue	xsd:float
Volume Unit	tumor_volume_unit	The volume unit of the lesion	Conditional (required if volume is present)	0..1	CodeableConcept	eucaim:hasUnitofMeasure	eucaim:COM1000147
Size Method	tumor_size_method	Method for measuring the size or the volume of the tumor	Optional	0..1	CodeableConcept	eucaim:hasTumorSizeMethod	eucaim:GEN1000028
Size Maximum Dimension	tumor_size_maximum_dimension	The longest tumor dimension in cm or mm.	Optional	0..1	Float	eucaim:MaximumDimensionValue	xsd:float
Size Other Dimension	tumor_size_other_dimension	The second or third tumor dimension in cm or mm.	Optional	0..2	Float	eucaim:OtherDimensionValue	xsd:float
Size Dimension Unit	tumor_size_dimension_unit	The unit concept	Optional	0..1	CodeableConcept	eucaim:hasUnitofMeasure	eucaim:COM1000147
Body Site	tumor_body_site	The tumor's body site region, location and laterality. A tumor might be present in multiple distinct locations	Optional	0..*	Reference:BodySite	eucaim:hasBodySite	eucaim:BP1000024

Related Procedure	tumor_related_procedure	Associates this tumor with a related procedure. For example it associates a tumor with an MR examination procedure.	Optional	1..1	Reference: Procedure	eucaim:hasRelatedProcedure	eucaim:CLIN1000005
Risk Assessment	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	eucaim:hasRiskAssessment	eucaim:CLIN1051859
Histologic Grade	tumor_histologic_grade	Reference to the HistologicGrade which specifies the type of the histologic grade of this specific tumor.	Optional	0..1	Reference: HistologicGrade	eucaim:hasHistologicGradeType	eucaim:CLIN1049681
Tumor Observation	tumor_tumor_observation	Reference to the Tumor Observation entity which specifies the observation for this tumor	Optional	0..*	Reference: Tumor Observation	eucaim:hasTumorObservation	eucaim:GEN1000026

Entity: Tumor Observation (Optional)

Element Name	Attribute Name	Definition	Required	Cardinality	Data Type	Binded property (object/data)	Range
Code	tumor_observation_code	The tumor observation as a codeable concept. For example, "Mass Margin" of a tumor, "Extraprostatic invasion of a tumor" or VASARI criteria etc. The tumor observations might be either from the imaging or from the biopsy or any other related procedure	Required	0..1	CodeableConcept	eucaim:hasTumorObservation	eucaim:GEN1000026
Subject	tumor_observation_subject	The patient associated with this tumor.	Required	0..1	Reference: Patient	eucaim:hasSubject	eucaim:COM1001047
Value As Concept	tumor_observation_value_as_concept	The tumor observation value as a codeable concept. For example, "angular", "microlobulated", "spiculated" for the "Mass Margin" codeable concept.	Optional	0..1	CodeableConcept	eucaim:hasTumorObservationValue, eucaim:hasAnswer	eucaim:GEN1000026
Value As Number	tumor_observation_value_as_number	The tumor observation value as a numerical value	Optional	0..1	Float	eucaim:TumorObservationValue	xsd:float

Value Unit	tumor_observation_value_unit	The tumor observation value unit type	Conditional (Required if observation value is numeric)	0..1	CodeableConcept	eucaim:hasUnitofMeasure	eucaim:COM1000147

Entity: Body Site (Optional)

Element Name	Attribute Name	Definition	Required	Cardinality	Data Type	Binded property (object/data)	Range
Code	body_site_code	The anatomical location or region of the lesion. Only SNOMED CT and ICD-O-3 are considered conformant (i.e. SNOMED concept is-a 123037004 (Body structure) and all ICD-O-3 topography codes)	Required	1..1	Reference: BodySite	eucaim:hasBodySite	eucaim:BP1000024
Location	body_site_location	General location qualifier (excluding laterality) for this bodySite	Required	1..1	Reference: BodySite	eucaim:hasLocation	eucaim:BP1000024
Laterality	body_site_laterality	Laterality qualifier for this bodySite	Optional	0..1	Reference: BodySite	eucaim:hasLaterality	eucaim:IMG1016305
Volume	body_site_volume	The volume unit of the body site/organ. For example volume of prostate gland.	Optional	0..1	Float	eucaim:hasVolumeValue	xsd:float
Volume Unit	body_site_volume_unit	The volume unit of the body site/organ.	Conditional (required if volume is present)	0..1	CodeableConcept	eucaim:hasUnitofMeasure	eucaim:COM1000147

Entity: Risk Assessment (Optional)

Element Name	Attribute Name	Definition	Required	Cardinality	Data Type	Binded property (object/data)	Range
Code	risk_assessment_code	The kind of risk assessment reported.	Required	1..1	CodeableConcept	eucaim:hasRiskAssessment	eucaim:CLIN1051859
Patient	risk_assessment_patient	The patient associated with risk assessment.	Required	1..1	Reference: Patient	eucaim:hasSubject	eucaim:COM1001047

Value As Concept	risk_assessment_value_as_concept	The classification resulting from the risk evaluation as a concept	Optional (Conditional)	0..1	CodeableConcept	eucaim:hasRiskAssessmentValue	eucaim:COM1001198, eucaim:IMG1005456
Value As Number	risk_assessment_value_as_number	The classification resulting from the risk evaluation as a number	Optional (Conditional)	0..1	Float	eucaim:RiskAssesmentValue	xsd:float
Value Unit	risk_assessment_value_unit	The classification result unit	Optional (Conditional)	0..1	CodeableConcept	eucaim:hasUnitofMeasure	eucaim:COM1000147
Related Procedure	risk_assessment_related_procedure	The procedure (e.g. imaging, laboratory exam) on which the risk assessment was based on	Optional	0..1	Reference: Procedure	eucaim:hasRelatedProcedure	eucaim:CLIN1000005
Related Condition	risk_assessment_related_condition	The cancer condition associated with risk assessment.	Optional	0..1	Reference:Cancer Condition	eucaim:hasRelatedCondition	eucaim:CLIN1007977

Entity: Treatment (Required if available)

Element Name	Attribute Name	Definition	Required	Cardinality	Data Type	Binded property (object/data)	Range
Type	treatment_type	The generic type of treatment.	Required	1..1	CodeableConcept	eucaim:hasTreatmentType	eucaim:GEN1000015
Response	treatment_response	Response evaluation to an oncology treatment from RECIST terminology.	Optional	0..1	CodeableConcept	eucaim:hasTreatmentResponse	eucaim:CLIN1037303
Intent	treatment_intent	The purpose of a treatment.	Optional	0..1	CodeableConcept	eucaim:hasTreatmentIntent	eucaim:COM1002393
Patient	treatment_patient	The patient on whom the treatment was performed	Required	1..1	Reference: Patient	eucaim:hasSubject	eucaim:COM1001047

Entity: Surgical Procedure (Required if available)

Element Name	Attribute Name	Definition	Required	Cardinality	Data Type	Binded property (object/data)	Range
Code	surgical_procedure_code	The specific procedure that is performed.	Required	1..1	CodeableConcept	eucaim:hasSurgicalProcedure	eucaim:CLIN1004413
Patient	surgical_procedure_patient	The patient on whom the procedure was performed	Required	1..1	Reference: Patient	eucaim:hasSubject	eucaim:COM1001047
Performed Date	surgical_procedure_performed_date	When the procedure was performed	Required (if available)	1..1	Date	eucaim:PerformedDate	xsd:dateTime
Offset From Diagnosis	surgical_procedure_offset_from_diagnosis	Period of time elapsed in months after primary cancer diagnosis	Required (if available)	0..1	Integer	eucaim:OffsetFromDiagnosis	xsd:int

Offset Unit	surgical_procedure_offset_unit_	Unit concept of the time interval (e.g. months, years, days) unit of time (UCUM)	Conditional (on offset)	0..1	CodeableConcept	eucaim:hasUnitofTime	eucaim:COM1000150
Body Site	surgical_procedure_body_site	Target body site	Optional	0..*	Reference:BodySite	eucaim:hasBodySite	eucaim:BP1000024
Response	surgical_procedure_response	Response evaluation to an oncology treatment from RECIST terminology.	Optional	0..1	CodeableConcept	eucaim:hasTreatmentResponse	eucaim:CLIN1037303, eucaim:COM1001180

Entity: Medication Administration (Required if available)

Element Name	Attribute Name	Definition	Required	Cardinality	Data Type	Binded property (object/data)	Range
Medication.Code	medication_administration_medication.code	Codes that identify this medication	Required	1..1	CodeableConcept	eucaim:hasAdministeredMedication	eucaim:CLIN1034109
Patient	medication_administration_patient	The patient receiving the medication.	Required	1..1	Reference: Patient	eucaim:hasSubject	eucaim:COM1001047
Effective	medication_administration_effective	An interval of time during which the administration took place.	Optional	0..1	Integer	eucaim:EffectiveTimeInterval	xsd:int
Effective Unit	medication_administration_effective_unit	A unit of the interval of time during which the administration took place.	Optional	0..1	CodeableConcept	eucaim:hasUnitofTime	eucaim:COM1000150
Start Date	medication_administration_start_date	When the medication started	Required (if available)	1..1	Date	eucaim:StartDate	xsd:dateTime
End Date	medication_administration_end_date	When the medication ended	Required (if available)	1..1	Date	eucaim:EndDate	xsd:dateTime
Offset From Diagnosis	medication_administration_offset_from_diagnosis	Period of time elapsed after primary cancer diagnosis	Required (if available)	0..1	Integer	eucaim:OffsetFromDiagnosis	xsd:int
Offset Unit	medication_administration_offset_unit	Unit concept of the time interval (e.g. months, years, days) unit of time (UCUM)	Conditional (on offset)	0..1	CodeableConcept	eucaim:hasUnitofTime	eucaim:COM1000150
Response	medication_administration_response	Response evaluation to an oncology treatment from RECIST terminology.	Optional	0..1	CodeableConcept	eucaim:hasTreatmentResponse	eucaim:CLIN1037303, eucaim:COM1001180

Entity: Radiotherapy (Required if available)

Element Name	Attribute Name	Definition	Required	Cardinality	Data Type	Binded property (object/data)	Range
Modality	radiotherapy_modality	Capturing a modality of external beam or brachytherapy radiation procedures.	Required	1..1	CodeableConcept	eucaim:hasRadiationModality	eucaim:CLIN1014933; eucaim:COM1001181
Technique	radiotherapy_technique	Capturing a technique of external beam or brachytherapy radiation procedures.	Optional	0..*	CodeableConcept	eucaim:hasRadiationTechnique	eucaim:CLIN1014933
Patient	radiotherapy_patient	The patient on whom the procedure was performed.	Required	1..1	Reference: Patient	eucaim:hasSubject	eucaim:COM1001047
Start Date	radiotherapy_start_date	When the radiotherapy started	Required (if available)	1..1	Date	eucaim:startDate	xsd:dateTime
End Date	radiotherapy_end_date	When the radiotherapy ended	Required (if available)	1..1	Date	eucaim:endDate	xsd:dateTime
Offset From Diagnosis	radiotherapy_offset_from_diagnosis	Period of time elapsed in months after primary cancer diagnosis	Optional	0..1	Integer	eucaim:OffsetFromDiagnosis	xsd:int
Offset Unit	radiotherapy_offset_unit	Unit concept of the time interval (e.g. months, years, days) unit of time (UCUM)	Conditional (on offset)	0..1	CodeableConcept	eucaim:hasUnitofTime	eucaim:COM1000150
Dose Delivered to Volume	radiotherapy_dose_delivered_to_volume	A BodyStructure resource representing volume in the body where radiation was delivered, for example, Chest Wall Lymph Nodes.	Optional	0..1	CodeableConcept	eucaim:hasDoseDeliveredToVolume	eucaim:BP1000024
Total Dose Delivered	radiotherapy_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	Float	eucaim:TotalDoseDelivered	xsd:float
Dose Unit	radiotherapy_dose_unit	The dose unit concept	Conditional (on dose delivered)	0..1	CodeableConcept	eucaim:hasRadiationDoseUnit	eucaim:COM1000143
Number of Sessions	radiotherapy_number_of_sessions	The number of sessions in a course of radiotherapy.	Optional	0..1	unsignedInt	eucaim:NumberOfSessions	xsd:unsignedInt
Fractions Delivered	radiotherapy_fractions_delivered	The number of fractions delivered to this volume.	Optional	0..1	unsignedInt	eucaim:FractionsDelivered	xsd:unsignedInt
Body Site	radiotherapy_body_site	Coded body structure(s) treated in this course of radiotherapy. These codes represent general locations.	Optional	0..*	CodeableConcept	eucaim:hasBodySite	eucaim:BP1000024

Response	radiotherapy_response	Response evaluation to an oncology treatment from RECIST terminology.	Optional	0..1	CodeableConcept	eucaim:hasTreatmentResponse	eucaim:CLIN1037303, eucaim:COM1001180

Entity: Adverse Event (Optional)

Element Name	Attribute Name	Definition	Required	Cardinality	Data Type	Binded property (object/data)	Range
Patient	adverse_event_patient	Subject impacted by event	Required	1..1	Reference: Patient	eucaim:hasSubject	eucaim:COM1001047
Resulting Effect	adverse_event_resulting_effect	Information about the condition that occurred as a result of the adverse event.	Required	1..1	CodeableConcept	eucaim:hasResultingEffect	eucaim:CLIN1002855
Suspect Entity	adverse_event_suspect_entity	A suspected agent causing the adverse event	Optional	0..1	Reference:Treatment	eucaim:causedBy	eucaim:GEN1000015

Entity: Procedure (Required for registering the imaging procedure or pathology confirmation procedure)

Element Name	Attribute Name	Definition	Required	Cardinality	Data Type	Binded property (object/data)	Range
Code	procedure_code	Identification of the procedure	Required	1..1	CodeableConcept	eucaim:hasProcedure	eucaim:CLIN1000005
Category	procedure_category	Classification of the procedure	Optional	0..1	CodeableConcept	eucaim:hasCategory	eucaim:GEN1000027
Patient	procedure_patient	The patient on whom the procedure was performed.	Required	1..1	Reference: Patient	eucaim:hasSubject	eucaim:COM1001047
Performed Date	procedure_performed_date	Estimated or actual date the procedure occurred	Optional	0..1	Date	eucaim:PerformedDate	xsd:dateTime
Offset From Diagnosis	procedure_offset_from_diagnosis	Elapsed interval after diagnosis. If this is the diagnosis related procedure this should be zero.	Optional	0..1	Integer	eucaim:OffsetFromDiagnosis	xsd:int
Offset Unit	procedure_offset_unit	Unit concept of the time interval (e.g. months, years, days) unit of time (UCUM) from the cancer diagnosis	Conditional (on offset)	0..1	CodeableConcept	eucaim:hasUnitofTime	eucaim:COM1000150
Evaluation Finding	procedure_evaluation_finding	Finding as a result of the procedure. For example, biopsy result abnormal/normal, imaging result normal/abnormal	Optional	0..1	CodeableConcept	eucaim:hasEvaluationFinding	eucaim:CLIN1002372

Entity: Lab Test Result (Optional)

Element Name	Attribute Name	Definition	Required	Cardinality	Data Type	Binded property (object/data)	Range
Code	lab_test_result_code	Generic Laboratory Test Name. If the lab test is a tumor marker test it should be recorded in the respective table	Required	1..1	CodeableConcept	eucaim:hasLabTestResult	eucaim:CLIN1033405
Value As Concept	lab_test_result_value_as_concept	Result Value (SHOULD use Snomed CT for coded Results)	Optional	0..1	CodeableConcept	eucaim:hasAnswer	eucaim:COM1000181
Value As Number	lab_test_result_value_as_number	Value as a number	Optional	0..1	float	eucaim:LabTestResultValue	xsd:float
Value Unit	lab_test_result_value_unit	Value unit concept	Conditional (required if result is numeric)	0..1	CodeableConcept	eucaim:hasUnitofMeasure	eucaim:COM1000147
Patient	lab_test_result_patient	The patient on whom the lab test was performed.	Required	1..1	Reference: Patient	eucaim:hasSubject	eucaim:COM1001047
Date	lab_test_result_date	The date when the generic lab test occurred	Optional	0..1	Date	eucaim:PerformedDate	xsd:dateTime
Offset From Diagnosis	lab_test_result_offset_from_diagnosis	Period of time elapsed in months after primary cancer diagnosis	Optional	0..1	Integer	eucaim:OffsetFromDiagnosis	xsd:int
Offset Unit	lab_test_result_offset_unit	Unit concept of the time interval (e.g. months, years, days) unit of time (UCUM)	Conditional (on offset)	0..1	CodeableConcept	eucaim:hasUnitofTime	eucaim:COM1000150

Entity: Episode (Required - for diagnosis episode)

Element Name	Attribute Name	Definition	Required	Cardinality	Data Type	Binded property (object/data)	Range
Patient	episode_patient	The patient for whom the episode is recorded.	Required	1..1	Reference: Patient	eucaim:hasSubject	eucaim:COM1001047
Type Code	episode_type_code	The episode type code represents the kind abstraction related to the disease phase, outcome or treatment. Possible values are: Diagnosis, Treatment Episode, Progression, Relapse, Remission, Active Surveillance	Required	1..1	CodeableConcept		eucaim:GEN1000002

Number	episode_number	For sequences of episodes, this is used to indicate the order the episodes occurred. For example, lines of treatment could be indicated here.	Optional	0..1	Integer (positive)	eucaim:SequenceNumber	xsd:int
Start Date	episode_start_date	The date when the Episode begins.	Optional	0..1	Date	eucaim:StartDate	xsd:dateTime
End Date	episode_end_date	The date when the instance of the Episode is considered to have ended.	Optional	0..1	Date	eucaim:EndDate	xsd:dateTime
Offset From Diagnosis	episode_offset_from_diagnosis	Period of time elapsed in months after primary cancer diagnosis	Optional	0..1	Integer	eucaim:OffsetFromDiagnosis	xsd:int
Offset Unit	episode_offset_unit	Unit concept of the time interval (e.g. months, years, days) unit of time (UCUM)	Conditional	0..1	CodeableConcept	eucaim:hasUnitofTime	eucaim:COM1000150
Parent ID	episode_parent_id	Use this field to find the Episode that subsumes the given Episode record. This is used in the case that an Episode are nested into each other.	Optional	0..1	Reference:Episode	eucaim:partOf	eucaim:GEN1000002

Entity: Episode Event (Required for Diagnosis Episode Events)

Element Name	Attribute Name	Definition	Required	Cardinality	Data Type	Binded property (object/data)	Range
Episode ID	episode_event_episode_id	A foreign key identifier to the Episode that the Episode Event belongs to.	Required	1..1	Reference:Episode		
Related Event Name	episode_event_related_event_name	The name of the Entity corresponding to the table where the underlying event is stored. (CancerCondition, TumorMarketTest, MedicationAdministration)	Required	1..1	CodeableConcept	eucaim:hasEpisodeEventRelatedEntity	
Related Event ID	episode_event_related_event_id	A foreign key identifier to the primary key column of table defined by the concept in the RelatedEventName (cancer_condition_id, procedure_id, tumor_marker_test_id) where the underlying event is stored.	Required	1..1	Integer (Reference to the entity defined by the RelatedEventName)		

Entity: Image Study (Required)

Element Name	Attribute Name	Definition	Required	Cardinality	Data Type	Binded property (object/data)	Range	DICOM Tag Reference IHE radiology TF vol 2 table 4.14-1
Patient	image_study_patient	The patient of the imaging study.	Required	1...1	Reference(Patient)	eucaim:hasSubject	eucaim:COM10001047	(0010/*)
UID	image_study_uid	Identifiers for the ImagingStudy, i.e. as DICOM Study Instance UID.	Required	1…1	String	eucaim:ImageStudyUID	sxd:string	StudyInstanceUID (0020,000D) | study ID (0020,0010)
Acquisition Date	image_study_acquisition_date	The date the study acquisition was obtained.	Optional	0...1	DateTime	eucaim:AcquisitionDate	xsd:dateTime	(0008,0020)+(0008,0030)
Offset from Diagnosis	image_study_offset_from_diagnosis	The interval of time elapsed before/after the primary diagnosis.	Optional	0...1	Integer	eucaim:OffsetFromDiagnosis	xsd:int	-
Offset Unit	image_study_offset_unit	Unit concept of the time interval (e.g. months, years, days) unit of time (UCUM)	Optional	0...1	CodeableConcept	eucaim:hasUnitofTime	eucaim:COM10000150	-
Related Procedure	image_study_related_procedure_	Associates the imaging study with an imaging procedure.	Required	1...1	Reference(Procedure)	eucaim:hasRelatedProcedure	eucaim:IMG1005453	-
Access URI	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 local machine via the path name to the folder containing the study.	Optional	0...*	String			-
Number Of Series	ImageStudy.NumberOfSeries	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	eucaim:NumberOfSeries	xsd:unsignedInt	(0020,1206)
Number Of Instances	ImageStudy.NumberOfInstances	Number of SOP Instances in Study. This value given may be larger than the number of instance elements this resource contains	Optional	0...1	unsignedInt	eucaim:NumberOfInstances	xsd:unsignedInt	(0020,1208)

		due to resource availability, security, or other factors. This element should be present if any instance elements are present.						
Series	ImageStudy.Series	Each study has one or more series of images or other content.	Required	1...*	BackboneElement	eucaim:Composed_Of	Eucaim:IMG1016309	-

Entity: Image Series (Required)

Element Name	Attribute Name	Definition	Required	Cardinality	Data Type	Binded property (object/data)	Range	DICOM Tag
Study UID	image_series_study_uid	The study in which the series belongs to.	Required	1..1	Reference(Image Study)	eucaim:UID	xsd:string	StudyInstanceUID (0020,000D) | study ID (0020,0010)
UID	image_series_uid	The DICOM Series Instance UID for the series.	Required	1...1	String	eucaim:ImageSeriesUID	xsd:string	(0020,000E)
Number	image_series_number	The numeric identifier of this series in the study.	Optional	0...1	unsignedInt	eucaim:ImageSeriesNumber	xsd:unsignedInt	(0020,0011)
Modality	image_series_modality	The distinct modality for this series. This may include both acquisition and non-acquisition modalities.	Required	1...1	Reference(Endpoint)	eucaim:hasImageModality	eucaim:IMG1000009	(0008,0060)
Description	image_series_description	A description of the series.	Optional	0...1	string	eucaim:ImageSeriesDescription, eucaim:hasImageSeriesDescription	xsd:string,	(0008,103E)
Manufacturer Name	image_series_manufacturer_name	Name of the manufacturing company of the imaging equipment	Required	1..*	CodeableReference	eucaim:hasEquipmentManufacturer	eucaim:IMG1000010	(0008,0070)
Body Site	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 SNOMED-CT mappings. The bodySite may indicate the laterality of body part imaged; if so, it shall be consistent with any content of	Required	1...1	CodeableReference(BodyStructure)	eucaim:hasImageBodyPart	eucaim:BP1000024	(0018,0015)

		ImagingStudy.series.laterality.						
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 ImagingStudy.series.bodySite.	Optional	0...1	CodeableConcept	eucaim:hasLaterality	eucaim:IMG1016305	(0020,0060)
Acquisition Date	image_series_acquisition_date	The date the series acquisition was obtained.	Optional	0...1	date	eucaim:AcquisitionDate	xsd:dateTime	(0008,0021) + (0008,0031)
Access URI	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 local machine via the path name to the folder containing the series instances.	Optional	0...*	string	-	-	-
Number Of Instances	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	eucaim:NumberOfInstances	xsd:unsignedInt	(0020,1209)

Entity: Image Modality (Optional)

Element Name	Attribute Name	Definition	Required	Cardinality	Data Type	Binded property (object/data)	Range	DICOM Tag
Study UID	image_series_study_uid	Reference to the series id for which important acquisition parameters are being stored.	Required	1..1	Reference(Image Study)	eucaim:UID	xsd:string	StudyInstanceUID (0020,000D) | study ID (0020,0010)
UID	image_series_uid	The study in which the series	Required	1..1	Reference(Ima	eucaim:UID	xsd:string	(0020,000E)

		belongs to.			ge Series)			
Acquisition Parameter Code	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	eucaim:hasAcquisitionParameter	eucaim:IMG1016619	Any modality specific tag
Acquisition Parameter Value Code	image_modality_acquisition_parameter_value_code	The concept code of the value of the acquisition parameter (e.g. "Spin echo" value of the "MR echo type" concept)	Required(conditional)	1..1	CodeableConcept	eucaim:hasAcquisitionParameterValue	eucaim:IMG1016619	Any modality specific tag
Acquisition Parameter Value Number	image_modality_acquisition_parameter_value_number	The numerical value of the modality acquisition concept (e.g. 0 for the gantry tilt angle in case of a CT)	Optional	0..1	Float	eucaim:AcquisitionParameterValueNumber	xsd:float	Any modality specific tag
Acquisition Parameter Value Unit	image_modality_acquisition_parameter_value_unit	If a numeric value, the units of measure concept code should be used. (http://unitsofmeasure.org).	Conditional (on Acquisition Parameter Value as Number)	0..1	CodeableConcept	eucaim:hasUnitofMeasure	eucaim:COM1000147	Any modality specific tag

Entity: Segmentation Series (Optional)

Element Name	Attribute Name	Definition	Required	Cardinality	Data Type	Binded property (object/data)	Range	DICOM Tag
Source UID	segmentation_series_source_uid	The unique identifier for the imaging series being annotated.	Required	1..1	Reference (Image Series)	eucaim:UID	xsd:string	(0020,000E)
Study UID	segmentation_series_study_uid	The unique identifier for the imaging study that contains the series that is being annotated.	Required	1..1	Reference (Image Study)	eucaim:UID	xsd:string	(0020,000D)
UID	segmentation_series_uid	The unique identifier for the annotated derived imaging series.	Required	1..1	String	eucaim:UID	xsd:string	(0020,000E)

Algorithm Type	segmentation_series_algorithm_type	The method used to create the annotation, such as manual or automatic, or semiautomatic. It gets deduced based on the Segment.AlgorithmType. If all segments were made manually, or automatically then the method for the annotated series should be manual or automatic respectively. If the annotation was made based on a mixture of manual and automatic means, then the method should be semi-automatic.	Required	1..1	CodeableConcept	eucaim:hasAlgorithmType	eucaim:COM10001204	(0062,0008)
Annotation Type	segmentation_series_annotation_type	The annotation type (e.g. bounding box, contouring, etc..)	Optional	0..1	CodeableConcept	eucaim:hasAnnotationType	eucaim:COM10001202	-
Annotator Specialty	segmentation_series_annotator_specialty	Specific role of the annotator (e.g. Radiologist, Imaging technician, etc.).	Optional	0..*	CodeableConcept	eucaim:hasAnnotatorSpecialty	eucaim:COM10001203	-
Annotator Experience	segmentation_series_annotator_experience	Years of experience of the annotator as a whole number (integer) without decimals.	Optional	0..*	UnsignedInt	eucaim:AnnotatorExperience	xsd:unsignedInt	-
Performed Date	segmentation_series_performed_date	The date and time the annotation was made.	Optional	0..1	Date	eucaim:PerformedDate	xsd:dateTime	(0008,0023)
Status	segmentation_series_status	The current status of the annotation, such as final or pending.	Optional	0..1	CodeableConcept	eucaim:hasAnnotationStatus	eucaim:COM10001211	-

Entity: Segment (Optional)

Element Name	Attribute Name	Definition	Required	Cardinality	Data Type	Binded property (object/data)	Range	DICOM Tag
Number	segment_number	Unique identification number of the segment	Required	1..1	UnsignedInt	eucaim:SegmentNumber	xsd:unsignedInt	(0062,0004)

Label	segment_label	Ontology-defined label identifying the segment in the case the segmentations are made in the context of EUCAIM	Required	1..1	CodeableConcept	eucaim:hasSegmentLabel	eucaim:BP1000024	(0062,0005)
Algorithm Name	segment_algorithm_name	The name(s) and version of the algorithm(s) used to generate the segment. If the Segment algorithm type is not Manual, the algorithm name is mandatory.	Conditional (if automatic)	0..1	String	eucaim:SegmentAlgorithmName	xsd:string	(0062,0009)
Tool Name	segment_tool_name	The name of the tool through which the segmentation was performed	Optional	0..1	String	eucaim:SegmentationToolName	xsd:string	-
Tool Version	segment_tool_version	The version of the tool through which the segmentation was performed	Optional	0..1	String	eucaim:SegmentationToolVersion	xsd:string	-
Tool Coordinate System	segment_tool_coordinate_system	Defines if the annotation tool does not use the same coordinate system as the image. Values: Physical, Pixel	Optional	0..1	CodeableConcept	eucaim:hasImageCoordinateSystem	eucaim:COM1001216	-

Annex B - Imaging metadata for extraction

ImageStudy	ModelName	Manufacturer's model name	DICOM	dicomTag (0008,1090)
	Manufacturer	Manufacturer's name	DICOM	dicomTag (0008,0070)
	StudyInstanceUID	Study Instance UID	DICOM	dicomTag (0020,000D)
	StudyDescription	Study description	DICOM	dicomTag (0008,1030)
	AcquisitionDate	Acquisition date	DICOM	dicomTag (0008, 0032), (0008, 0022) or dicomTag (0008, 0021), (0008, 0023) or dicomTag (0008, 0031), (0008, 0033)
	ModalitiesInStudy	Modalities in study	DICOM	dicomTag (0008,0061)
	NbStudyRelatedSeries	Number of Study Related Series	DICOM	dicomtag(0020,1206)
ImageSeries	SeriesNumber	Series Number	DICOM	dicomTag(0020,0011)
	SeriesInstanceUID	Series Instance UID	DICOM	dicomTag (0020,000E)
	Modality	Modality	DICOM	dicomTag (0008, 0060)
	Description	Description	DICOM	dicomTag (00e1, 1040), (0008, 103e), (0008, 1030)
	SoftwareVersion	Software's version	DICOM	dicomTag (0018,1020)
	BodyPartExamined	Body Part Examined	DICOM	dicomTag (0018,0015)
	NbSeriesRelatedInstances	Number of Series Related Instances	DICOM	dicomtag (0020,1209)
	SliceThickness	Slice thickness	DICOM	dicomTag (0018,0050)
	PixelSpacing	Pixel spacing	DICOM	dicomTag (0028,0030)
	FieldOfView	Field of view		To be calculated
	Rows	rows	DICOM	dicomTag (0028,0010)
	Columns	columns	DICOM	dicomTag (0028,0011)
MRImage	ScanningSequence	Scanning sequence	DICOM	dicomTag (0018,0020)
	MagneticFieldStrength	Magnetic field strength	DICOM	dicomTag (0018,0087)
	MRAcquisitionType	MR acquisition type	DICOM	dicomTag (0018,0023)
	RepetitionTime	Repetition time	DICOM	dicomTag (0018,0080)
	EchoTime	EchoTime	DICOM	dicomTag (0018,0081)
	FlipAngle	FlipAngle	DICOM	dicomTag (0018,0024)
	InversionTime	InversionTime	DICOM	dicomTag (0018,0082)
	ReceiveCoilName	ReceiveCoilName	DICOM	dicomTag (0018,1250)
	ImagingFrequency	Imaging Frequency	DICOM	dicomTag (0018,0084)
CTImage	KVp	KVp	DICOM	dicomTag (0018,0060)
	XRayTubeCurrent	XRayTubeCurrent	DICOM	dicomTag (0018,1151)
	CTDIvol	CTDIvol	DICOM	dicomTag (0018,9345)
	ExposureTime	Exposure Time	DICOM	dicomTag (0018,1150)
	SpiralPitchFactor	Spiral Pitch Factor	DICOM	dicomTag (0018,9311)

	FilterType	Filter Type	DICOM	dicomTag (0018,1160)
	ConvolutionKernel	Convolution Kernel	DICOM	dicomTag (0018,1210)
Injection	Radiopharmaceutical	Radiopharmaceutical	DICOM	dicomTag (0018,0031)
	RadionuclideTotalDose	Radionuclide total dose	DICOM	dicomTag (0018,1074)
	ContrastBolusAgent	Contrast/bolus agent	DICOM	dicomTag (0018,0031)
	ContrastBolusStartTime	Contrast/bolus start time	DICOM	dicomTag (0018,0010)
	ContrastBolusStopTime	Contrast/bolus stop time	DICOM	dicomTag (0018,1042)
	RadiopharmaceuticalStartTime	Radiopharmaceutical start time	DICOM	dicomTag (0018,1043)
	RadionuclideTotalDose	Radionuclide total dose	DICOM	dicomTag (0018,1072)
PTImage	AttenuationCorrectionMethod	AttenuationCorrectionMethod	DICOM	dicomTag (0054,1101)
	Reconstruction Method	ReconstructionMethod	DICOM	dicomTag (0054,1103)
	ScatterCorrectionMethod	ScatterCorrectionMethod	DICOM	dicomTag (0054,1105)
NMImage	AttenuationCorrectionMethod	AttenuationCorrectionMethod	DICOM	dicomTag (0054,1101)
	Reconstruction Method	ReconstructionMethod	DICOM	dicomTag (0054,1103)
	ScatterCorrectionMethod	ScatterCorrectionMethod	DICOM	dicomTag (0054,1105)
DXImage	ImageLaterality	ImageLaterality	DICOM	dicomTag (0020,0062)
	PatientOrientation	PatientOrientation	DICOM	dicomTag (0020,0020)
	AnatomicRegionSequenceCodeMeaning	AnatomicRegionSequenceCodeMeaning	DICOM	dicomTag (0008,0104)
	Anatomic Region SequenceCodeValue	Anatomic Region SequenceCodeValue	DICOM	dicomTag (0008,0100)
	KVP	KVP	DICOM	dicomTag (0008,0060)
	Exposure	Exposure	DICOM	dicomTag (0018,1152)
	ExposureTime	ExposureTime	DICOM	dicomTag (0018,1150)
GenericImage	Temporal Position	Temporal Position	DICOM	dicomTag (0020,0100)
	Number of Temporal Positions	Number of Temporal Positions	DICOM	dicomTag (0020,0105)
	Temporal Resolution	Temporal Resolution	DICOM	dicomTag (0020,0110)
	Lossy Image Compression	Lossy Image Compression	DICOM	dicomTag(0028,2110)

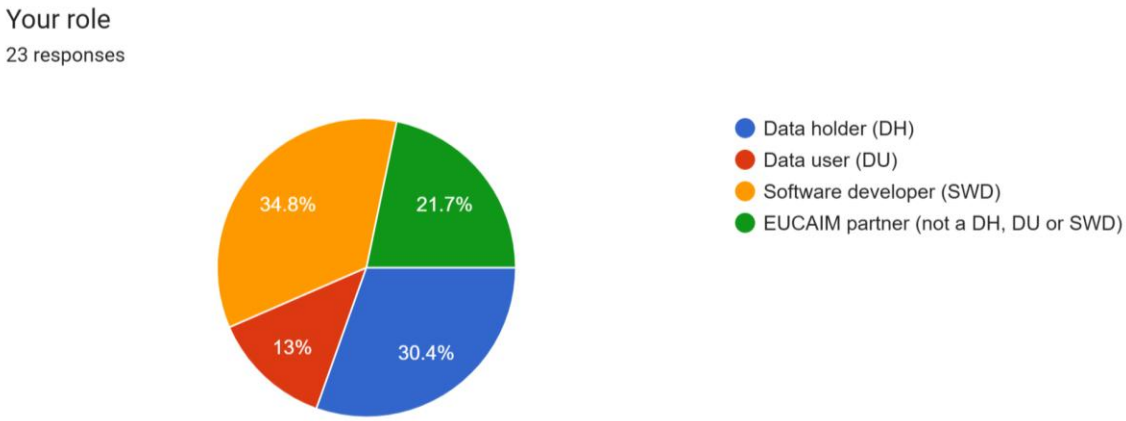
Annex C - EUCAIM DCAT-AP Questionnaire results

A. Questionnaire administration

To support the fine-tuning of the EUCAIM DCAT-AP (in particular, to inform final property cardinalities and ranges), WP5 circulated a structured questionnaire via Google Forms.

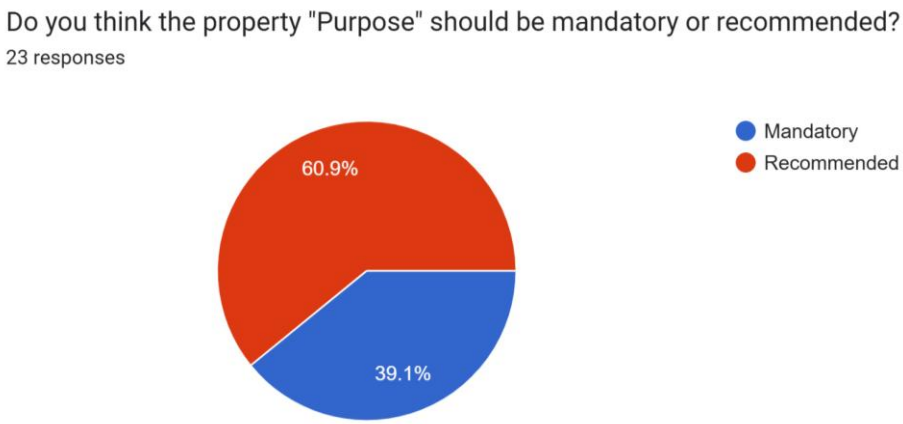
- **Dissemination:** first circulated within WP5, then announced to the full consortium.
- **Response window:** 05 Nov 2025 – 08 Dec 2025
- **Total responses: 23**
- **Note on participation:** although shared consortium-wide, responses were received primarily from WP5 partners / stakeholders.
- **Open ended comments:** each property and question had an open-ended comments section through which participants were asked to provide their opinion on each of the requested questions.

B. Respondent profile (self-reported)



C. Results

Q1. “Purpose”



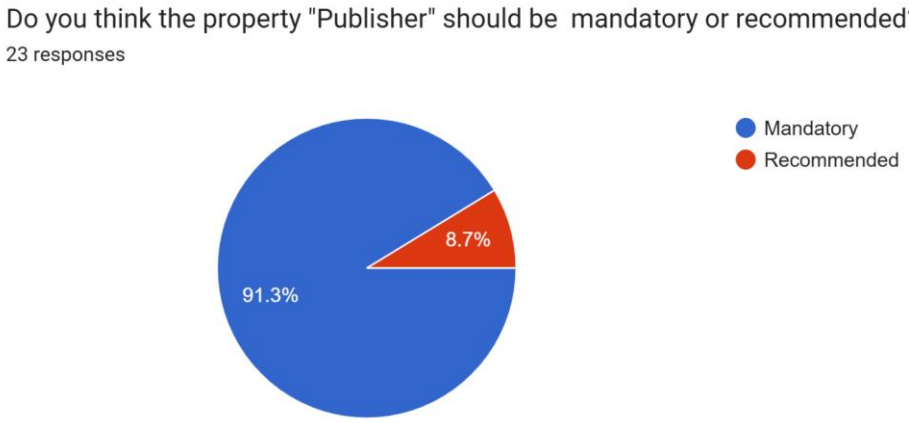
Qualitative feedback

- Strong feeling that *Purpose* is **important**, often leaning towards *mandatory*:
  a. *"It should be considered mandatory because it specifies why data is being collected, processed, or used."*
- It should be possible to specify **more than one purpose** for a dataset.
- The field is seen as helpful for:
  a. Transparency (*"would make everything clearer and more transparent"*)
  b. Discoverability (*"facilitate filtering according to the potential usages of a dataset"*).
- Concern that the **DPV vocabulary is too generic**; some feel the purpose should reflect *the original dataset's creation* rather than just EUCAIM-specific secondary use.
- One comment notes that the original purpose may influence whether data can be used *commercially* later.

Outcome

"Purpose: widely valued for transparency and filtering, but DPV may be too generic"

Q2. "Publisher"



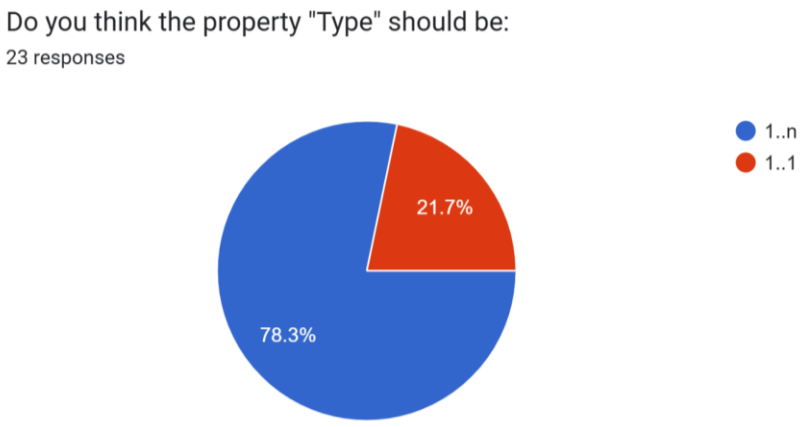
Qualitative feedback

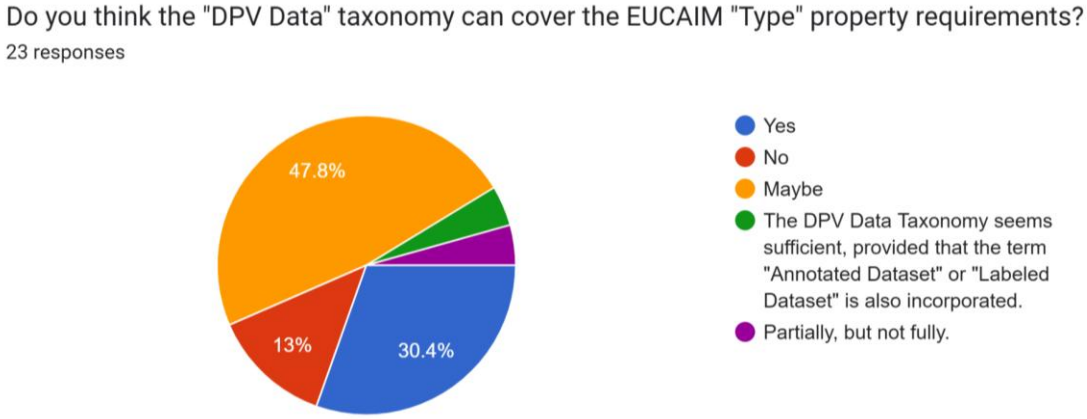
- Generally seen as useful and important for:
  - Identifying the data holder and the party responsible for the dataset.
  - Ensuring stewardship, accountability and FAIR compliance.
- Multiple comments request a stable contact point:
  - *"It should be possible to capture a sustainable/generic contact (mail + tel +/- website) of the Publisher."*
- Also seen as helpful for aggregated statistics (datasets per publisher).
- Needs to coexist with anonymization of site_id:
  - *"As long as it is not incompatible with the anonymization of the site_id."*
- One respondent links Publisher to AI Act "provider" responsibilities, especially when data is stored in a third-party infrastructure.

Outcome

"Publisher: key for accountability and contact".

Q3. "Type"





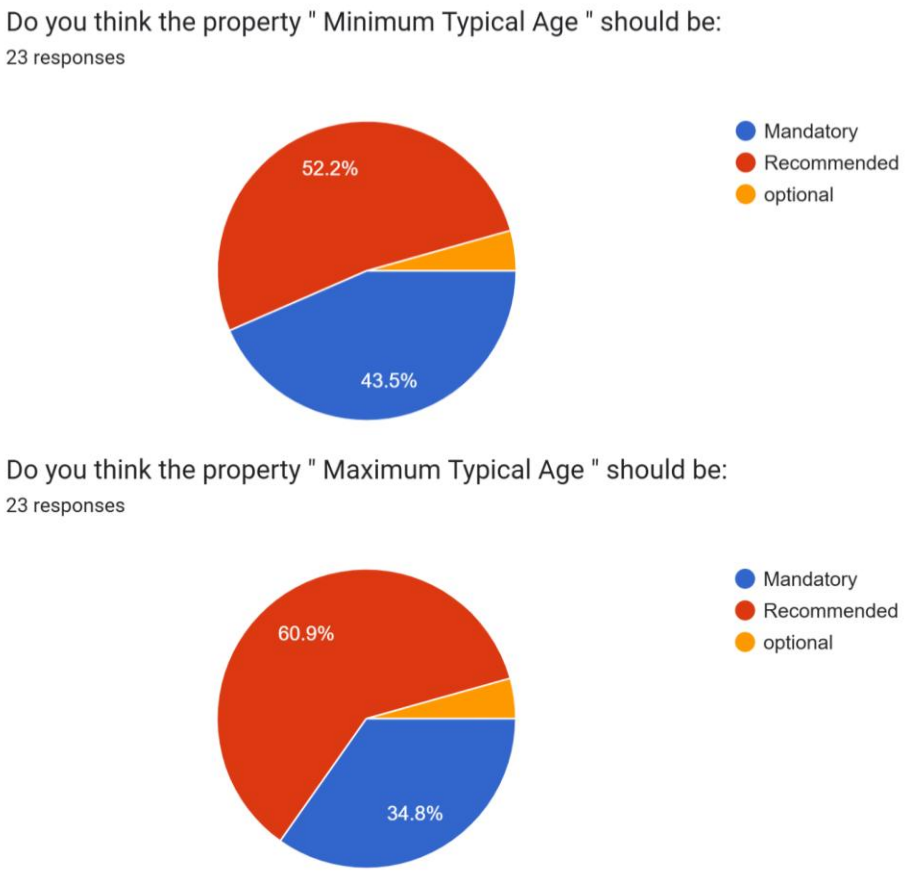
Qualitative feedback

- Strong sense that “Type” is multi-dimensional:
  - Distinguishing imaging vs clinical, annotated vs processed, etc.
  - One participant calls it “a multiaxial concept, each axis/facet being important (e.g. annotations, samples, privacy level, IP confidentiality …)”.
- Cardinality should reflect both EUCAIM type and DPV data type:
  - “The cardinality depends on the value set, e.g. EUCAIM type vs DPV data.”
- Some doubt about DPV’s adequacy:
  - DPV does “not make a clear distinction between ‘annotated’ and ‘processed’”; this might limit expressiveness.
  - One notes: “I am not sure if this ontology brings much to our case.”
- A view that DPV taxonomy is more about privacy / compliance than scientific dataset categorization and should not replace an EUCAIM-specific Type.

Outcome

“Type: needs multiple facets. DPV useful but not sufficient alone”.

Q4, Q5. “Min/Max Typical Age”



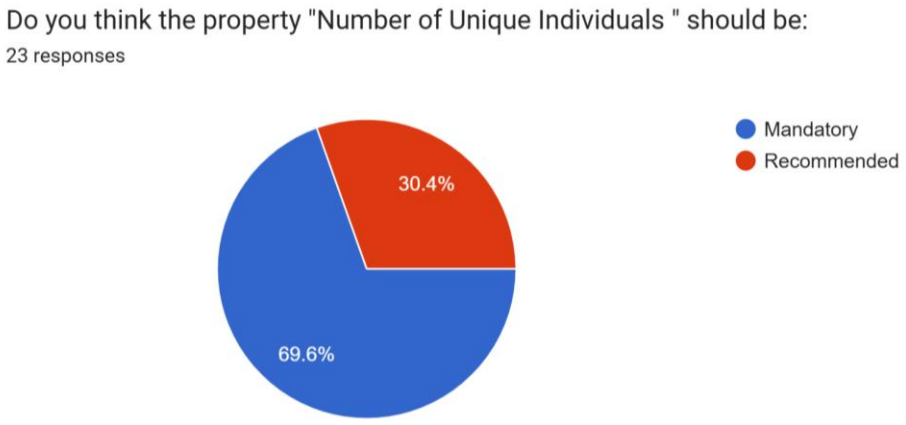
Qualitative feedback

- Seen as useful but not always essential:
    a. Provides “interesting additional information, but not required for processing or filtering data”.
- Important for cohort definition (pediatric, adult, elderly populations).
- For some datasets, this info is only available at restricted tiers and thus it is important
    a. “For tier 1 datasets this information is only available at this level (before access is granted).”
- Several comments suggest these fields are better as recommended rather than strictly mandatory as *“If the goal is to grow the catalogue as much as possible, predominantly by onboarding from/connecting to existing catalogues and (meta)data sources, keep in mind that if 'Minimum typical age' is not present there, these datasets cannot be onboarded automatically in the EUCAIM catalogue if 'Minimum typical age' is mandatory in EUCAIM DCAT-AP.*”

Outcome

“Typical age: useful for understanding cohorts, but availability may be limited”.

Q6. “Number of Unique Individuals”



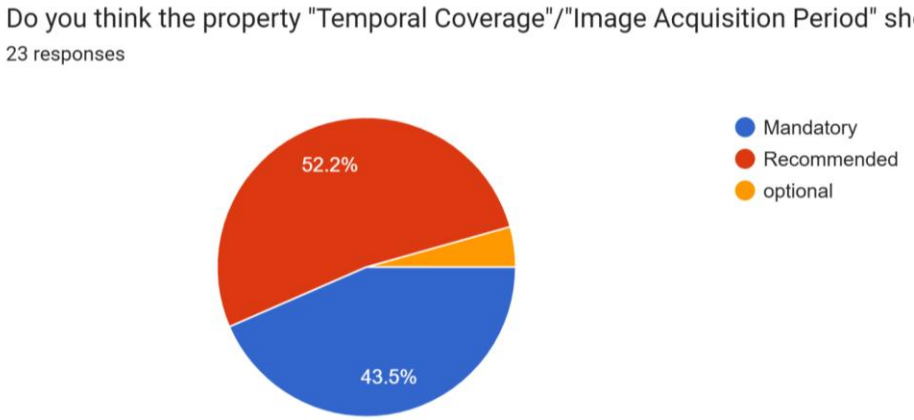
Qualitative feedback

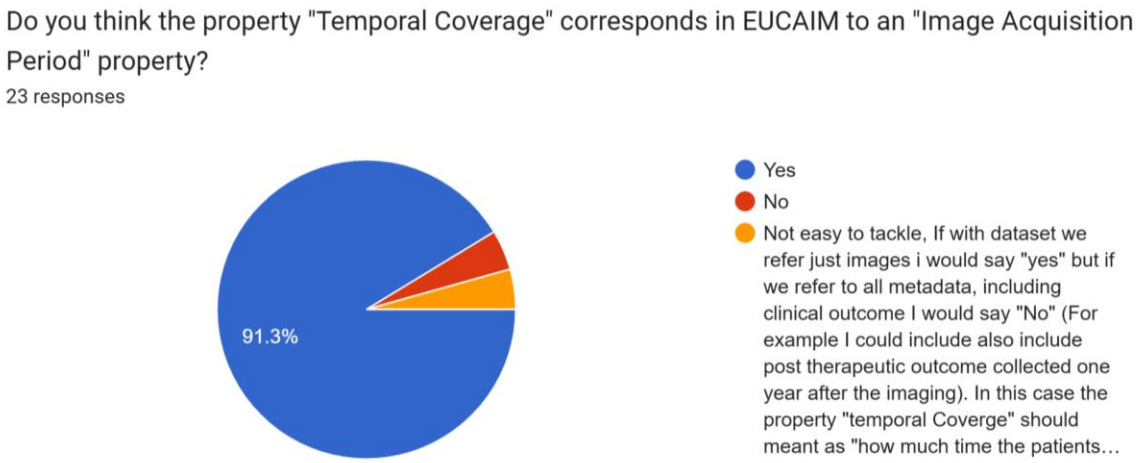
- Often seen as a key indicator of a dataset's population size.
- Important for regulatory submissions and product development, where population size matters more than number of images:
    a. "Regulatory submission for product development typically requires a certain number of individuals. A dataset may be large in number of images or samples but low on population."
- Same constraints as age: for some tiered access models this information is only available at this level.

Outcome

"Unique individuals: crucial for population understanding and regulatory use"

Q7. "Temporal Coverage"





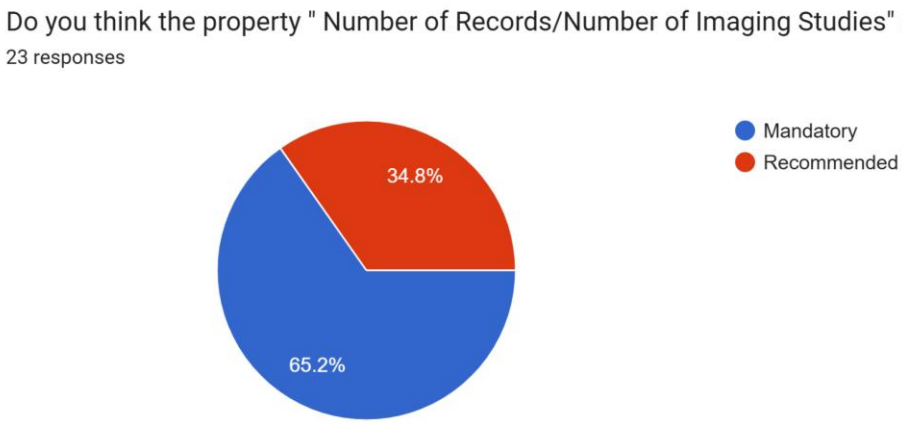
Qualitative feedback

- Strong agreement that for EUCAIM, this should be framed as image acquisition period:
  a. “As it is the Atlas of Cancer Images, I agree on focusing on the imaging period rather than on the diagnosis.”
  b. “In EUCAIM: ‘Temporal Coverage = Image Acquisition Period’.”
- Seen as very useful for:
  a. Publications and grants (“very helpful in publications/grants”).
  b. Understanding whether data reflects an older vs more recent clinical practice.
- Some suggestions to define it more globally:
  a. E.g. an overall dataset period, plus a specific one for medical images.
- Concerns:
  a. Might not exist in legacy datasets, so making it mandatory could be problematic.
  b. Privacy risks for very narrow time windows: “Difficult due to privacy concerns if a small temporal period.”

Outcome

“Temporal coverage: highly informative but constrained by legacy data and privacy”

Q8. “Number of Records”



Do you think the above property could be used in EUCAIM as "Number of Imaging Studies" (e.g. DICOM studies)?

23 responses

Qualitative feedback

- Clearly perceived as important and informative:
  a. Helps users judge the scale and statistical power of a dataset.
  b. Supports filtering (e.g. looking for “large” vs “small” datasets).
- Some technical nuance:
  a. Needs clear alignment with DICOM concept of ‘Study’ (vs series, images, etc.).
  b. Counting can be non-trivial in complex longitudinal datasets.

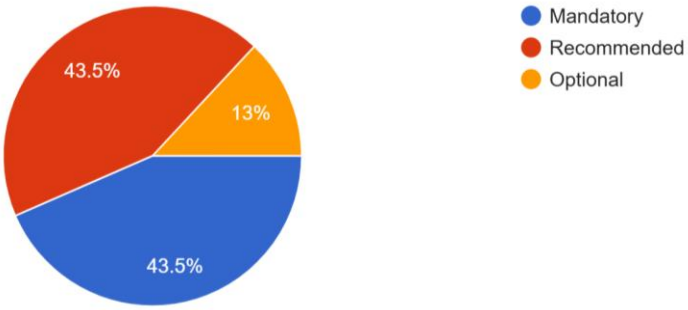
Outcome

“Number of records: can be associated with the number of DICOM studies, although complex to count them”.

Q9. “Version”

Do you think the property "Version" should be:

23 responses



Qualitative feedback

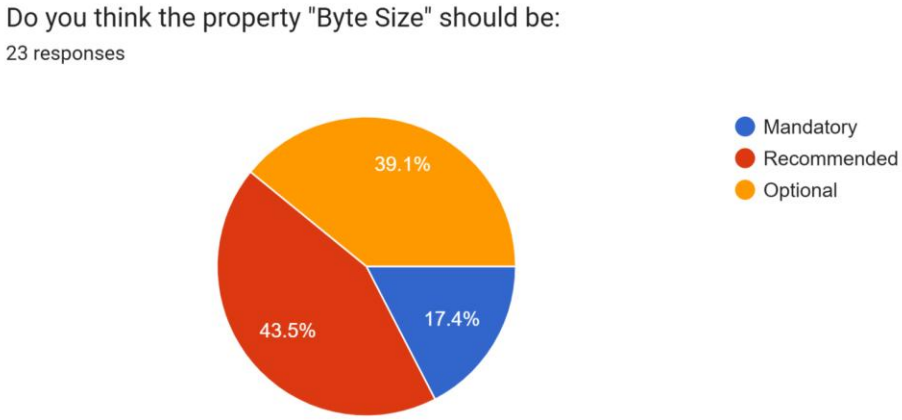
- Strong consensus that versioning is important. Two main reasons:
  a. Traceability and provenance (knowing which data were used in which analysis). “*Important for reproducibility of workflows.*”

b. Regulatory and quality purposes when datasets change over time. "*Mainly for traceability… addition or removal of patients can have a great regulatory importance.*"

- Although important it should not be mandatory, as "*heavily depends on the technical capabilities of the data holder institute or the data storage platform. If the data holder can provide an accurate version, it would be a great property to have, but making it mandatory would likely lead to people entering meaningless or quickly outdated version numbers*."
- Some mention should not be mandatory as "*The unique identifier differentiates datasets and the parent attribute can be used to show the provenance*."

Outcome
"Version: important for traceability and reproducibility but should be recommended".

Q10. "Byte Size"

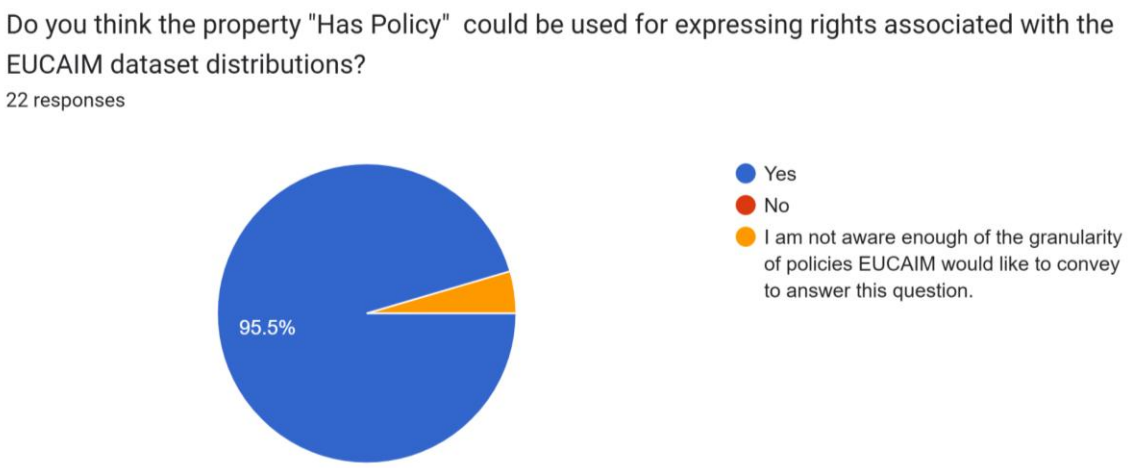


Qualitative feedback

- Not useful since data users are not going to download data.
  a. *"Is it really relevant for data users, given that they are not going to download the datasets?"*
  b. *"Sometimes it is hard to compute and since we do not allow downloading it is irrelevant."*
- Some still see value:
  a. *"It is certainly useful, although it can depend on many factors (formats, compression, etc.)."*
- Helpful for infrastructure planning (storage, transfer time).
- Overall: seen as nice-to-have, but not strong enough to be mandatory.

Outcome
"Byte size: nice-to-have for infrastructure, low value for end-users"

Q11. "Has Policy"



Qualitative feedback

- Generally **positive** towards using an explicit policy property:
  a. *"The Has Policy property is appropriate and recommended."*
- The use of Open Digital Rights Language (ODRL) is appreciated but:
  a. Needs to be clearly explained to Data Holders and Publishers.
  b. Guidance is needed on how policies relate to access and processing rights in EUCAIM.

Outcome
"Has Policy: well-received, but requires clear guidance and communication

Annex D - EUCAIM DCAT-AP examples

Dataset

	Property	Examples
DCAT-AP v3	Title	dct:title "Open Challenge Prostate Cancer V1"@en;
DCAT-AP v3	Description	dct:description "This ProCAncer-I project imaging dataset contains a collection of patients with mpMRI examinations (T2ax, DWI, ADC) who have confirmed PCa at biopsy and/or prostatectomy."@en
HealthDCAT-AP	Theme	dcat:theme <http://publications.europa.eu/resource/authority/data-theme/HEAL>;
DCAT-AP v3	Provenance	dct:provenance [a dct:ProvenanceStatement; rdfs:label "This data is sourced from several existing datasets, including the Duke dataset, ParcTauli and TCGA datasets."@en];
DCAT-AP v3	Keyword	dcat:keyword "Prostate Cancer"@en , "mpMRI"@en.
HealthDCAT-AP	Purpose	dpv:hasPurpose[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;] ;
HealthDCAT-AP	Access Rights	dct:accessRights <http://publications.europa.eu/resource/authority/access-right/NON_PUBLIC> .
HealthDCAT-AP	Health Category	healthdcatap:healthCategory <http://13.81.34.152:1101/resource/authority/healthcategories/EHRS> .
HealthDCAT-AP	Health Theme	healthdcatap:healthTheme <http://13.81.34.152:1101/resource/authority/health-theme/CANCER> .
DCAT-AP v3	Geographical Coverage	dct:spatial <http://publications.europa.eu/resource/authority/country/GRC>;
DCAT-AP v3	Applicable Legislation	dcatap:applicableLegislation <http://data.europa.eu/eli/reg/2022/868/oj>;
DCAT-AP v3	Contact Point	dcat:contactPoint [a vcard:Organization; vcard:fn "FORTH"; vcard:hasEmail <mailto:access-commitee@procancer-i.com>];
EUCAIM DCAT-AP	Publisher Name	dct:publisher [a foaf:Organization; locn:address [a locn:Address; foaf:name "FORTH"; foaf:mbox <mailto:access-commitee@procancer-i.com>; foaf:homepage <https://forth.ics.gr>;];];
	Publisher Contact Point	
HealthDCAT-AP	Publisher Type	healthdcatap:publisherType <.../authority-table/publisher-type/ResearchInstitute>;
HealthDCAT-AP	Publisher Note	dct:publisher [a foaf:Agent; dct:description "Publisher Note"@en;].
EUCAIM DCAT-AP	Type	dct:type a skos:Concept ; skos:prefLabel "Annotated Dataset"@en .
HealthDCAT-AP	Minimum Typical Age	healthdcatap:minTypicalAge "18" ^^xsd:int ;
HealthDCAT-AP	Maximum Typical Age	healthdcatap:maxTypicalAge "90" ^^xsd:int ;
EUCAIM DCAT-AP	BirthSex	eucaim:hasBirthSex eucaim:COM1001370 (Female) .

EUCAIM DCAT-AP	Number of Imaging Studies	eucaim:nbrOfStudies "8789" ^^xsd:int ;
EUCAIM DCAT-AP	Number of Unique Individuals	eucaim:nbrOfSubjects "8237" ^^xsd:int ;
EUCAIM DCAT-AP	Collection Method	eucaim:collectionMethod <https://hyperontology.eucaim.cancerimage.eu/#SPEC1000003>.
DCAT-AP v3	Image Acquisition Period	dct:temporal [a dct:PeriodOfTime; dcat:endDate "2023-12-31"^^<http://www.w3.org/2001/XMLSchema#date>; dcat:startDate "2021-01-01"^^<http://www.w3.org/2001/XMLSchema#date>];
EUCAIM DCAT-AP	Cancer Condition	eucaim:hasCondition <eucaim:CLIN1000075> . (Malignant neoplasm of prostate)
EUCAIM DCAT-AP	Image Modality	eucaim:hasImageModality <eucaim:IMG1000022> . (Magnetic Resonance Imaging)
EUCAIM DCAT-AP	Image Equipment Manufacturer	eucaim:hasEquipmentManufacturer <eucaim:IMG1000047> . (General Electric)
EUCAIM DCAT-AP	Image Body Part/Structure	eucaim:hasImageBodyPart <eucaim:BP1000233> . (Neck and chest)
EUCAIM DCAT-AP	Annotation Label	eucaim:hasAnnotationLabel <eucaim:BP1000075> . (Heart)
EUCAIM DCAT-AP	Segmentation Method	eucaim:hasAlgorithmType <eucaim:COM1000003> . (Manual)
EUCAIM DCAT-AP	Number of Segmentations	eucaim:nbrOfSegmentations "789" ^^xsd:int ;
EUCAIM DCAT-AP	Identifier	dct:identifier "https://catalogue.eucaim.cancerimage.eu/#/collection/1a1a6653-975a-4a0a-a79b-b2bfc7317119"^^<http://www.w3.org/2001/XMLSchema#anyURI>;
EUCAIM DCAT-AP	Version	dcat:version "20231122"
EUCAIM DCAT-AP	Interoperability Tier	adms:interoperabilityLevel a skos:Concept ; skos:prefLabel "Tier 1"@en.
DCAT-AP v3	Language	dct:language <http://publications.europa.eu/resource/authority/language/ENG> .
HealthDCAT-AP	Population Coverage	healthdcatap:populationCoverage “Patients between 35 and 87 years old with prostate cancer treated with prostatectomy in hospitals in France between 2018 and 2023.” @en .
HealthDCAT-AP	Personal Data	dpv:hasPersonalData <https://w3c.github.io/dpv/2.0/pd#BirthDate>, <https://w3c.github.io/dpv/2.0/pd#Age>, <https://w3c.github.io/dpv/2.0/pd#DateOfBirth> .
DCAT-AP v3	Temporal Resolution	ex:temporalResolution "P1M"^^xsd:duration .
DCAT-AP v3	Frequency	dct:accrualPeriodicity <http://publications.europa.eu/resource/authority/frequency/MONTHLY> .
HealthDCAT-AP	Legal Basis	dpv:hasLegalBasis [a dpv:LegalBasis ; 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 health data of.. , as part of the EUCAIM project and the subsequent processing of personal data pseudonymized by…"@en;] ;
HealthDCAT-AP	Retention Period	healthdcatap:retentionPeriod [a dct:PeriodOfTime;

		dcat:endDate "2034-12-31"^^<http://www.w3.org/2001/XMLSchema#date>; dcat:startDate "2020-03-01"^^<http://www.w3.org/2001/XMLSchema#date>; rdfs:comment "Provide complementary information"@en] .
DCAT-AP v3	Conforms To	dct:conformsTo <https://www.wikidata.org/entity/Q19597236>
HealthDCAT-AP	Coding System	healthdcatap:hasCodingSystem <https://www.wikidata.org/entity/Q9006342>, <https://www.wikidata.org/entity/Q5969475> .
HealthDCAT-AP	Code Values	healthdcatap:hasCodeValues [a skos:Concept; skos:inScheme [a skos:ConceptScheme; dct:identifier "http://www.wikidata.org/entity/Q45127"^^<http://www.w3.org/2001/XMLSchema#anyURI>; skos:prefLabel "International Classification of Diseases, 10th Revision (ICD-10)"@en; skos:definition "ICD-10 is a medical classification list by the World Health Organization."@en; skos:notation "ICD-10"; owl:versionInfo "Version:2019"]; dct:identifier "https://icd.who.int/browse10/2019/en#/Y59.0"^^<http://www.w3.org/2001/XMLSchema#anyURI>; skos:notation "Y59.0"; skos:prefLabel "Viral vaccines"@en]; healthdcatap:hasCodeValues [a skos:Concept; skos:inScheme [a skos:ConceptScheme; dct:identifier "http://www.wikidata.org/entity/Q45127"^^<http://www.w3.org/2001/XMLSchema#anyURI>; skos:prefLabel "International Classification of Diseases, 10th Revision (ICD-10)"@en; skos:definition "ICD-10 is a medical classification list by the World Health Organization."@en; skos:notation "ICD-10"; owl:versionInfo "Version:2019"]; dct:identifier "https://icd.who.int/browse10/2019/en#/U07.1>"^^<http://www.w3.org/2001/XMLSchema#anyURI>; skos:notation "U07.1"; skos:prefLabel "COVID-19, virus identified"@en] .
DCAT-AP v3	Related Resource	dct:relation <http://another-resource> . # The resource can be Dataset, Distribution, Dataservice fully described
DCAT-AP v3	Is Referenced By	dct:isReferencedBy <https://doi.org/10.1186/s13690-021-00709-x> .
DCAT-AP v3	Landing Page	dcat:landingPage <https://fair.healthdata.be/dataset/example> .
DCAT-AP v3	Documentation	foaf:page <https://fair.healthdata.be/dataset/example> .
HealthDCAT-AP	Sample	https://healthdataeu.pages.code.europa.eu/healthdcat-ap/releases/release-5/#admssample
HealthDCAT-AP	Analytics	healthdcatap:analytics [a dcat:Distribution;

		dcatap:applicableLegislation <http://data.europa.eu/eli/reg/2022/868/oj>; dct:title "Technical report number of unique study subjects available by environment for project HDBP0250"@en; dcat:accessURL <https://fair.healthdata.be/sites/default/files/distribution/d43a158e-7d13-4660-bbc3-9d3f8d5501e5/Technical_report_number_of_unique_study_subjects_available_by_environment_for_project_HDBP0250.csv> ; dct:format <http://publications.europa.eu/resource/authority/file-type/CSV>; dcat:mediaType <http://www.iana.org/assignments/media-types/text/csv>] .
HealthDCAT-AP	Quality Annotation	dqv:hasQualityAnnotation [a dqv:QualityCertificate ; oa:hasTarget <https://pid.eucaim.cancerimage.eu/822ad0bd-02d1-4932-a8b2-7d5679c3d4f0>; dqv:inDimension <https://example.org/quality/dimension/integrity> ; oa:motivatedBy dqv:qualityAssessment] . <https://example.org/quality/dimension/integrity> a dqv:Dimension ; skos:prefLabel "Integrity"@en ; skos:definition "Degree to which the dataset's DICOM files remain complete, internally consistent and follow the DICOM standard (e.g., no corruption)."@en .
DCAT-AP v3	Creator Name	dct:creator [a foaf:Agent, foaf:Organization; foaf:name "Ministry of Health"; vcard:hasURL <https://ministry-health.com/>;] .
	Creator Contact Point	
DCAT-AP v3	Was Generated By	prov:wasGeneratedBy [a prov:Activity; dct:type <http://13.81.34.152:1101/resource/authority/health-activity/RESEARCH_PROJECT> ; foaf:page <https://www.procancer-i.eu/>; rdfs:label "ProCAncer-I AI4HI project"@en ; rdfs:seeAlso <https://www.ehealth.fgov.be/ehealthplatform/fr/service-codage-anonymization-et-ttp>; prov:startedAtTime "2020-10-01"^^xsd:date ;] .
DCAT-AP v3	Spatial Resolution	dcat:spatialResolutionInMeters "1000.0"^^<http://www.w3.org/2001/XMLSchema#decimal> .
DCAT-AP v3	Qualified Attribution Agent Name	prov:qualifiedAttribution [a prov:Attribution; dcat:hadRole <https://standards.iso.org/iso/19115/resources/Codelists/gml/CI_RoleCode.xml#processor>; prov:agent [a foaf:Organization; locn:address [a locn:Address; locn:adminUnitL1 "Belgium"; locn:adminUnitL2 "Brussels capital"; locn:postCode "1050"; locn:postName "Elsene - Ixelles"; locn:thoroughfare "Rue Juliette Wytsmanstraat 14"]; foaf:homepage <https://healthdata.be>; foaf:mbox <mailto:healthdata@sciensano.be>; foaf:name "healthdata.be (Sciensano)";
	Qualified Attribution Agent Contact Point	

		foaf:phone <tel:+3227930142>]] .
DCAT-AP v3	Other Identifier	adms:identifier [a adms:Identifier; skos:notation "https://www.healthinformationportal.eu/health-information-sources/linking-registers-covid-19-vaccine-surveillance"^^xsd:anyURI ; adms:schemaAgency "Health Information Portal"]; adms:identifier [a adms:Identifier ; skos:notation "HDBP0250"^^xsd:string ; adms:schemaAgency "Health Data" ; dct:issued "2021-01-01"^^xsd:date] .
DCAT-AP v3	Version Notes	adms:versionNotes "Updated data collection methodology and extended temporal coverage to include 2023 data. Fixed data quality issues identified in previous version."@en .
DCAT-AP v3	Release Date	dct:issued "2022-09-10"^^<http://www.w3.org/2001/XMLSchema#date> .
DCAT-AP v3	Modification Date	dct:modified "2023-01-15"^^<http://www.w3.org/2001/XMLSchema#date> .

Distribution

	Property	**Example**
DCAT-AP v3	Access URL	dcat:accessURL <https://negotiator.eucaim.cancerimage.eu/collection/a96b56cd-59d4-444a-8e59-32a7fb0d7dea> ;
DCAT-AP v3	Applicable Legislation	dcatap:applicableLegislation <http://data.europa.eu/eli/reg/2022/868/oj>;
DCAT-AP v3	Description	dct:description "This is the DICOM imaging data distribution"@en
DCAT-AP v3	Format	dct:format <https://www.iana.org/assignments/media-types/application/dicom>;
DCAT-AP v3	License	dct:license <https://creativecommons.org/licenses/by/4.0/>
DCAT-AP v3	Access Service	dcterms:accessRights <http://publications.europa.eu/resource/authority/access-right/NON-PUBLIC> ;
EUCAIM DCAT-AP	Image Size (in GB)	dcat:byteSize "325"^^xsd:decimal
DCAT-AP v3	Compression Format	dct:format <https://www.iana.org/assignments/media-types/application/zip>;
DCAT-AP v3	Has Policy	odrl:hasPolicy [a odrl:Policy ; odrl:permission [a odrl:Permission ; odrl:action (<http://www.w3.org/ns/odrl/2/read> <http://www.w3.org/ns/odrl/2/derive>)]; odrl:prohibition [a odrl:Prohibition ; odrl:action <http://creativecommons.org/ns#CommercialUse>] ; odrl:obligation [a odrl:Duty ; odrl:action <https://schema.org/RegisterAction>];] ;
	accessConditions	dct:rights [a dct:RightsStatement;

		rdfs:label "Authorization to access, view and process in-situ the datasets"@en];

Annex E - Clinical data template file

In order to have interoperable data that can be queried and processed, Tier 3 DHs must provide information on their dataset structure using a tabular template file that comes *in addition to* their source dataset.

Presentation of the tabular template file

The tabular template file is organized by tabs, as follows:

- The "Data elements" tab lists the entities and their corresponding data elements for clinical variables, with definition and data type;
- The other 3 tabs show an example of how to structure your datasets of positive or diagnostic cases (for negative screening and control groups, please refer to the corresponding template file);
    - The "Overarching Episode" corresponds to the entire course of the patient's data collection (i.e, from diagnosis to death or last contact). All diagnosis information should be in there (**Table E1**);
    - Each episode recorded in your dataset must be separated from the first tab in another tab in chronological order (example : “Treatment 1”, “Progression”, “Treatment 2”, “Remission”, “Relapse”, “Treatment 3”, “Active Surveillance").

In each tab :

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

Patient ID	Population	Gender / Sex	Date of radiology detection	Age at diagnosis	Cancer topography : organ	Tumor site	Tumor site : Region	Tumor site : Laterality	Pathology / Histological Type		Histological subtype	Imaging acquisition		Pathology Confirmation	Date of pathology confirmation / Biopsy Date / Diagnosis date	Smoking status	
Patient	Patient	Patient	Cancer Condition	Cancer Condition	Cancer Condition	Tumor	Tumor	Tumor	Cancer Condition		Cancer Condition	Procedure	Procedure	Procedure	Procedure	Medical History	Medical History
Identifier	Diagnostic Category	BirthSex	Asserted Date	Age At Diagnosis	Topography	Body Site.Code	Body Site.Location	Body Site.Laterality	Code	Type	Histology Morphology Behavior	Code	Category	Code	Date	Code	Category
	COM1001087	LOINC			SNOMEDCT	SNOMEDCT	SNOMEDCT:450721000;RadLex:RID6392	custom	SNOMEDCT		SNOMEDCT	SNOMEDCT	SNOMEDCT	RadLex		custom	
AB56Xi89r	Patient with Cancer	Male	5/10/2015	66.99	Brain tumor	Brain part	2	Right	Glioblastoma	Primary	Epithelioid	MRI	imaging	Biopsy	15/6/2015	smoker	Observation
					Brain tumor, brain part		1,frontal brain region	Left	Glioblastoma multiforme		Epithelioid,SNOMEDCT:733837004						
							2,temporal brain region	Right			Multiforme,SNOMEDCT:393563007						
							4,parietal brain region				Giant cell,SNOMEDCT:44529004						
							5,occipital brain region										

Table E1: Example of how to complete the “Overarching episode” tab of the tabular template file

Completion of the template file

1. We recommend that the name of the template file also contains the dataset_ID as the first column.

2. The file must contain the *exact* variables' names on the first row (matching the variable's names from the source dataset), and the PatientID as the first variable. Note: The template file provided only contains the mandatory variables; DHs must provide the full list of variables available in their dataset.

3. All episodes must be separated into different tabs as described above.
   Note: episodes may correspond to the following: Diagnosis, Treatment, Progression, Relapse, Remission, Active Surveillance.

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

   Example 1 with "Imaging acquisition" as Procedure: the sequence (CT, MRI) has to be specified as "Code", and "imaging" must be assigned to it as Category (see the example dataset **Table E1**). Note that the name of the variable is then merged on both columns.

   Example 2 with "Smoking Status" as Medical History: the status value itself (smoker, non-smoker, etc) must be specified as "Code", and "Observation" must be assigned to it as Category (see the example dataset **Table E1**). Again, the name of the variable must be merged on both columns.

5. For each variable of the dataset, an example value must be provided on the 5th line (the value must be given exactly as it is spelled in the dataset)

6. For each variable of the dataset:

- if the variable strictly follows a specific standard, the name of the standard must be provided on line 4.

  Example: in the example dataset in **Table E1**, the "Histological type" variable strictly follows the SNOMEDCT standard; line 4 specifies "SNOMEDCT", and an example value is provided on line 5.
  Important: both information must be separated by a comma, without space

- If the variable follows a specific standard with in-house coding or remaining, DH must provide the name of the standard on line 4, and provide the correspondence between all possible values from the dataset and the standard values (lines 6 and onwards).

  Example 1 in the example dataset in **Table E1**, the "Tumor site: Region" variable follows the SNOMEDCT standard using an in-house coding; line 4 specifies "SNOMEDCT", an

example value is provided on line 5, and correspondence for all possible values present in the dataset to the SNOMEDCT codes is listed on lines 6-9, separated by a comma.

Example 2 in the example dataset in **Table E1**, the "Histological subtype" variable follows the SNOMEDCT standard using an in-house naming; line 4 specifies "SNOMEDCT", an example value is provided on line 5, and correspondence for all possible values present in the dataset to the SNOMEDCT codes is listed in lines 6-9, separated by a comma.

- if the variable does not follow a specific standard, "custom" must be stated on line 4, and the list of all possible values from the dataset for that variable must be provided on lines 6 and onwards.

  Example: in the example dataset in **Table E1**, the "Tumor Site: Laterality" variable does not follow a standard, but only uses the label "Left" or "Right"; in that case line 4 specifies "custom", an example value is provided on line 5, and all possible values present in the dataset (here "Left" and "Right" is listed on lines 6-7.