Intelligent ecosystem to improve the governance, the sharing, and the re-use of health data for rare cancers

Deliverable 2.5

Metadata Taxonomy

31 August 2023

This project has received funding from the European Union's Horizon Europe research and innovation programme under grant agreement no. 101057048

This project has received funding from the European Union's Horizon Europe research and innovation programme under grant agreement no. 101057048

Distribution List

Organization	Name of recipients
1 - Coord INT	A. Trama, P. Casali, L. Buratti, P. Baili, J. Fleming, L. Licitra, E. Martinelli, G. Scoazec
2 - UDEU	A. Almeida, U. Zulaika Zurimendi, N. Kalocsay
3 - MME	F. Mercalli, S. Copelli, M. Vitali
4 - UPM	E. Gaeta, G. Fico, L. Lopez, I. Alonso, C. Vera, A. Estevan, V. G. Dominguez, I. Alonso, L. Hernandez, C. Vera
5 - HL7	G. Cangioli, C. Chronaki
6 - ECCP	S. Ziegler, S. Miteva. A. Quesada, S. Schiffner, V. Tsiopoulou
7 - ENG	P. Zampognaro, A. Sperlea, E. Mancuso, M. Melideo, F. Saccà, V. Falanga, M. Rosa
8 - CERTH	K. Votis, A. Triantafyllidis, N. Laloumis
9 - UU	S. van Hees, Wouter Boon, E. Moors, M. Kahn-Parker, C. Egher
10 - DICOR	C. Lombardo, G. Pesce, G Ciliberto, A. Tonon,
10° - ACC (Affiliated)	D. De Persis, P. De Paoli, G. Piaggio, M. Pallocca. A. De Nicolo
11 - FBK	A. Lavelli, S. Poggianella, O. Mayora, A.M. Dallaserra
12 – IKNL	E. Bosma. G. Geleijnse, A. Van Gestel, E. Mezei, C. Attanasio, E. Polk, M. Van Swieten
13 - CLB	M. Rogasik, J-Y Blay, H. Crochet, J. Olaz, J. Bollard, C. Chemin-Airiau, C. Bouvier
14 - APHP	B. Baujat, E. Koffi
15 - FJD	J Martin-Broto, N. Hindi, M. Martin Ruiz, A. Montero Manso, C. Roldàn Mogìo, D. Da Silva, A. Herrero, B. Barrios
16 - VGR	Magnus Kjellberg, L. De Verier, A. Muth
17 - MSCI	I. Lugowska, D. Kielczewska, M.Rosinska, A KAwecki , A., P. Rutkowski
18 - MUH	R. Knopp, A. Sediva, K. Kopeckova, A. Nohejlova Medkova, M. Vorisek
19 - OUS	S. Larønningen, J. Nygård, M. Sending, O. Zaikova
20 - MMCI	J. Halamkova, I. Mladenkova, l. Tomastik, V. Novacek, T. Kazda, I. Mladenkova, O. Sapožnikov
21 - CLN	R. Szmuc, J. Poleszczuk, R. Lugowski
22 - FPNS	M. Barbeito Gomez, P. Parente, L. Carrajo Garcia, P. Ramos Vieiro
23 - TNO	E. Lazovik, L. Zilverberg, S. Dalmolen
24 - INF	ML Clementi, C. Sabelli
25 - UKE	S. Bauer, S. Lang, S. Mattheis, N. Midtank

This project has received funding from the European Union's Horizon Europe research and innovation programme under grant agreement no. 101057048

Revision History

Revision	Date of Issue	Author(s)	Brief Description of Change
1	21.07.2023	A. Almeida, U. Zulaika, A. Bilbao (UDEU)	First release
2	29.08.2023	G. Geleijnse, C. Attanasio (IKNL)	Peer review
3	05.09.2023	A. Almeida, U. Zulaika, A. Bilbao (UDEU)	V1 – approved by Coordinator

Addressees of this document

This document is addressed to the whole IDEA4RC Consortium. It is an official deliverable for the project and shall be delivered to the European Commission and appointed experts.

This project has received funding from the European Union's Horizon Europe research and innovation programme under grant agreement no. 101057048

TABLE OF CONTENTS

Table of Contents 4
1. Introduction 6
2. Executive Summary 7
3. State of the art 8
3.1 Data models: Description of the existing data models to describe the oncology domain. *8*
3.1.1 OSIRIS, OMOP CDM and FHIR 17
OSIRIS 17
OMOP 18
OMOP and the oncological domain 19
FHIR 22
3.2 Metadata: Description of the existing metadata taxonomies to describe quality, governance, findability, and provenance *24*
3.2.1 Data quality 24
3.2.2 Governance 34
3.2.3 Findability 41
4. IDEA4RC adopted common data & semantic models 44
5. Idea4rc adopted CDM v1 46
5.1 Data model definition methodology *47*
5.2 Head & Neck Cancer: Description of the model with tables for the selected core variables, draft of the FHIR implementation guide, entity-relationship diagram. *48*
6. Idea4rc ADOPTED Metadata model v1 55
6.1 Data quality: Data quality taxonomy at variable, data source, cohort and federated levels. *56*
6.1.1 Reliability quality checks 57
6.1.2 Relevance quality checks 61
7. Conclusions and future work 63
ANNEX A 65

This project has received funding from the European Union's Horizon Europe research and innovation programme under grant agreement no. 101057048

Abbreviations and definitions

Abbreviation	Definition
CDM	Common Data Model
EHR	Electronic Health Record
ETL	Extract Transform Load
FHIR	Fast Healthcare Interoperability Resources
IG	Implementation Guide
HNC	Head and Neck Cancer
SWRL	Semantic Web Rule Language
OMOP	Observational Medical Outcomes Partnership
CoE	Center of Excellence
OWL	Web Ontology Language
FAIR	Findability, Accessibility, Interoperability, and Reusability
WHO	World Health Organization
OHDSI	Observational Health Data Science and Informatics

This project has received funding from the European Union's Horizon Europe research and innovation programme under grant agreement no. 101057048

1. INTRODUCTION

This document introduces the work done so far on the metadata and CDMs aspects of the IDEA4RC project. It includes the first versions of both models, which describe the data quality metadata and CDMs adopted in the project for head and neck cancers. The following versions will include not only the governance, findability, and provenance metadata but also the structural and semantic data model adopted for rare cancers in general within the project.

The document is organized as follows. Section 1 introduces the topic, while section 2 provides an executive summary of the document. Section 3 analyses the state of the art on the existing CDMs, international ontologies and metadata taxonomies in the healthcare domain. The goal is to review the literature and the current tools, to describe not only the oncology data structurally and semantically, but also the related metadata regarding data quality, governance, findability, and provenance. Section 4 introduces the importance of the FAIR principles in relation to the adoption of one or more CDMs within the IDEA4RC project. Section 5 describes the first version of the conversion of the IDEA4RC core dataset into CDMs, including the variables used to model the head and neck cancers. Section 6 describes the first version of the metadata model adopted, focusing on the data quality metadata. Finally, Section 7 offers some conclusions and discusses the next steps that will be taken for the consequent versions of the CDMs and metadata taxonomies for rare cancers adopted in the IDEA4RC project. To improve the readability of the document, we have included some of the information as annexes, specifically the tables that describe each variable in the adopted models so far.

This project has received funding from the European Union's Horizon Europe research and innovation programme under grant agreement no. 101057048

2. EXECUTIVE SUMMARY

During the first version of this deliverable, the following main decisions regarding the CDM have been taken:

- Use the work done in OSIRIS, FHIR, OMOP and OHDSI as the base from which to build our CDM.
- Reuse standard vocabularies (SNOMED_CT, ICD-O) wherever possible.
- Ensure that the CDM can be mapped to both OMOP and FHIR.
- Model the CDM as an Entity-Relationship Diagram to formalize it.

Regarding the data quality metadata, the following decisions have been taken:

- We have focused our work in three essential categories of metadata: data quality, data governance and data findability. In this first version of the metadata model, we have worked specifically in the data quality metadata.
- Provide a data quality taxonomy that allows to model it at variable, data source, cohort and federated levels.
- Provide specific subsections for relevance and reliability.
- Align with and follow the FAIR principles (see 6.1)

The next versions of the deliverable will encompass:

- The CDM adoption for sarcomas. This extension will include the incorporation of relevant variables, making the model even more comprehensive and applicable.
- Data governance and data findability metadata, completing the trio of essential metadata categories.

This project has received funding from the European Union's Horizon Europe research and innovation programme under grant agreement no. 101057048

3. STATE OF THE ART

This section provides an overview of the existing landscape in CDMs and metadata taxonomies, within the oncology domain, which are relevant to the IDEA4RC project. This section aims to establish a solid foundation of knowledge and understanding of the current practices and standards in the field. By exploring CDMs and metadata taxonomies, we can first identify existing frameworks that have been developed to describe the oncology or the rare cancers domain, and second assess their suitability considering the IDEA4RC's objectives.
Also, through a comparative analysis, we will gain valuable insights into the characteristics and functionalities of these metadata taxonomies, enabling us to make informed decisions about the most suitable approach for implementing metadata standards within the IDEA4RC project.

3.1 Data models: Description of the existing data models to describe the oncology domain.

This subsection presents the data models that can be used to describe the oncology domain, with a focus on two prominent CDMs: the **OMOP** (Observational Medical Outcomes Partnership) CDM and **FHIR** (Fast Healthcare Interoperability Resources). Additionally, the section also introduces the **OSIRIS**[1] project and its CDM, since it has been a foundational data model for the IDEA4RC project. The use of appropriate data models is essential for structuring and organizing clinical data, enabling interoperability, and facilitating data exchange and analysis within the IDEA4RC project. By exploring these CDMs, we can assess their relevance and suitability for achieving the project's objectives.
Efficient and standardized data management is crucial in oncology to facilitate research, improve patient care, and advance our understanding of cancer. This requires robust CDMs and, or taxonomies that can accurately represent and organize complex oncology-related information. In this context, Tables 1.A and 1.B provide an overview of existing data models within the oncology domain, presenting a comprehensive snapshot of the landscape in this field.

[1] https://www.e-cancer.fr/Professionnels-de-la-recherche/Recherche-translationnelle/OSIRIS-projet-national-sur-le-partage-des-donnees

This overview's objective is to establish a foundation of knowledge and understanding of current practices and standards in structural and semantic data modelling within oncology. The structural modelling, called a CDM, gives us direction on the structure of the data, i.e., 'where' to put the information in the data model. The semantic modelling, instead, identified using one or more standard ontologies or vocabularies, guides us on 'how' to encode the information in a standardized way. By exploring these existing frameworks, we can assess their relevance and suitability for addressing the specific requirements of the IDEA4RC project. This comparative analysis will enable informed decision-making regarding the most suitable approach for implementing metadata standards within the project, contributing to improved governance, data sharing, and reusability for rare cancers.

This project has received funding from the European Union's Horizon Europe research and innovation programme under grant agreement no. 101057048

Table 1.A - Overview of Existing Structural CDMs and Metadata Taxonomies in the Oncology Domain.

Data Model	Owner	Function	Area	Metadata	Vocabulary Standard	Database	Data Sharing	Software Tools	Open Community	Flexibility
FHIR	HL7	Data exchange	Healthcare data	Yes	Flexible	Not necessary	Yes	No	No	High: does not constrain on the semantic approach and has flexible data structure
OMOP CDM	OHDSI	Research/analytical purposes	Observational healthcare data	Yes	OHDSI standard vocabularies (including widely adopted international ontologies and vocabularies)	Necessary (flexible DBs)	Yes	Yes	Yes	Intermediate: it constrains on the semantic approach (though facilitating the conversion to it) and it has a fixed data structure
ODM-XML	CDISC	Data exchange	Research data	Yes	CDISC	Not necessary	Yes	Yes	Yes	Intermediate

OpenEHR	openEHR	Data management and storage, retrieval, and exchange	Healthcare data	Yes	openEHR	Necessary (NoSQL)	Yes	Yes	Yes	High

Table 1.B - Overview of Existing Semantic Data Models in the Oncology Domain.

Data Model	Owner	Function	Area	Metadata	Vocabulary Standard	Data Sharing	Software Tools	Open Community	Flexibility
OHDSI standard vocabularies	OHDSI	Health data, both structural and semantic modelling for research purposes.	Observational healthcare data	Yes	OHDSI standard vocabularies (including widely adopted international ontologies and vocabularies)	Yes	Yes	Yes	Intermediate: it provides mappings between different widely adopted vocabularies, but a standard concept is identified at the community level. Also, it does not allow post-coordination.
ICD-O	WHO	Used principally in tumour or cancer registries for coding the site (topography) and the histology	Oncology data	No	ICD-O	Yes	Yes	Yes	High: oncology specific, comprehensive vocabulary capturing the heterogeneity of the cancer diagnosis.

This project has received funding from the European Union's Horizon Europe research and innovation programme under grant agreement no. 101057048

Data Model	Owner	Function	Area	Metadata	Vocabulary Standard	Data Sharing	Software Tools	Open Community	Flexibility
		(morphology) of neoplasms, usually obtained from a pathology report.							
SNOMED-CT	SNOMED International	Most comprehensive clinical terminology in use around the world.	Healthcare data	No	SNOMED-CT	Yes	Yes	Yes	High: comprehensive ontology for coding general health data, allows for controlled post-coordination.

Each of these models and standards offers unique features and functionalities, catering to various aspects of data representation, exchange, and interoperability. By understanding the characteristics and capabilities of these models, we can make informed decisions regarding their suitability for the IDEA4RC project. This preface provides a brief introduction to the key data models discussed in Tables 1.A and 1.B, highlighting their main attributes and contributions to the field of oncology data management.

Data structure modelling (CDMs)

- **FHIR** (Fast Healthcare Interoperability Resources): Developed by HL7, FHIR is a data exchange standard for healthcare data. It offers high flexibility in representing and exchanging healthcare information. While it provides metadata capabilities, it does not enforce a specific vocabulary standard. FHIR has a large and active open community, making it a highly flexible option for data exchange. The FHIR standard needs to be integrated into the IDEA4RC project as the data will be shared by *FHIR Capsules*.
- **OMOP** (Observational Medical Outcomes Partnership) v5.6: Owned by OHDSI (Observational Health Data Sciences and Informatics), OMOP is a data modelling framework for observational healthcare data research. OMOP is supported by the OHDSI vocabularies, a collection of well-known international ontologies and vocabularies, on top of which is provided a hierarchical relationship system to identify standard concepts among overlapping semantic concepts from different vocabularies. This ontological collection can be browsed via the Athena platform and ensures that people among the OHDSI community use the same semantic approach. OMOP requires a database for implementation, though they make one of OMOP's design principles the so-called technology neutrality, for which the CDM does not require a specific technology and it can be realized in any relational database, such as Oracle, SQL Server etc., or as SAS analytical datasets. The OHDSI community also provides analytical software tools on top of the OMOP CDM and supports data sharing, but its flexibility is lower compared to other models. The IDEA4RC decided to additionally integrate the OMOP CDM because many CoEs are currently using it, and it was identified as crucial for the progress of the project. Recently, the Oncology extension, developed to target specifically the oncological domain, was added to the OMOP CDM v5.6, making the

OMOP CDM an emerging framework for observational data studies in oncology. It has an active international community working worldwide on model improvements and extensions, based on experience in large-scale projects run across the data partners in the OHDSI community.

- The inter-SIRIC **OSIRIS** consortium is a multidisciplinary group dedicated to the standardization and sharing of clinical and biological data in oncology. Over five years, the consortium developed a minimum dataset called the "Set OSIRIS," which includes essential clinical and omics items structured using existing reference frameworks. Supported by the French National Cancer Institute, the OSIRIS group aims to enhance data interoperability and facilitate data sharing among different research sites, both academically and industrially. The structured data proposed by the OSIRIS consortium can promote standardized data exchange, ultimately benefiting cancer research and real-world data utilization. Consortium partners CLB participated in the project OSIRIS. The project is similar to IDEA4RC as it proposes to model the oncological domain (not related to rare cancers) and have a mapping to both FHIR and OMOP, in this context, it was decided to use the OSIRIS project as a baseline to start modelling the IDEA4RC data.
- **ODM-XML** (Operational Data Model - Extensible Markup Language): Maintained by CDISC (Clinical Data Interchange Standards Consortium), ODM-XML is a data exchange standard for research data. It utilizes CDISC standards and does not require a specific database type. ODM-XML supports data sharing and has software tools available. IDEA4RC was envisaged to use the FHIR data exchange standard as it fulfilled all the requirements of the project, so the ODM-XML standard is out of scope.
- **OpenEHR**: Maintained by openEHR, OpenEHR is a comprehensive data management and storage framework for healthcare data. It provides metadata capabilities and uses openEHR as its vocabulary standard. OpenEHR requires a NoSQL database for implementation and supports data sharing and software tools. It has an active open community and offers high flexibility. The IDEA4RC project aims to create a platform for research, thus, it will only store data relevant to the research context, it does not pretend to be a full data management, storage, and registration framework for healthcare data.

Semantic modelling (ontologies and vocabularies)

- The **OHDSI standard vocabularies** are an integral part of the OMOP (Observational Medical Outcomes Partnership) CDM. They constitute of a comprehensive and standardized terminology system used to represent medical concepts and clinical data in a consistent and structured manner. The OHDSI standard vocabularies incorporate various international and widely adopted vocabularies and ontologies, such as SNOMED CT, LOINC, RxNorm, and ICD, among others. These vocabularies provide a wide range of codes and concepts for clinical conditions, procedures, medications, laboratory tests, and more. On top of these distinct vocabularies, where semantic overlap happens across them, the OHDSI standard vocabularies on the one side provide relationships between the concepts, and on the other identify one of them as standard, while making all the others non-standard. The use of the OHDSI standard vocabularies within the OMOP CDM enables researchers and healthcare professionals to harmonize and analyse healthcare data from various sources, facilitating robust observational research and real-world evidence generation for medical studies and decision-making. For the IDEA4RC project the OHDSI standard vocabularies will be used, to ensure consistency across the OMOP and FHIR implementations.
- **ICD** (International Classification of Diseases): Owned by the World Health Organization (WHO), ICD is a standard for classifying diseases and medical conditions. While ICD does not provide specific metadata capabilities, it is widely used in healthcare data coding and sharing. ICD-O, its oncology extension, is a relevant coding system for clinicians in the project, therefore, it will be considered when selecting codes in the IDEA4RC project (within the OHDSI standard vocabularies).
- **SNOMED-CT**: Maintained by SNOMED International, SNOMED-CT is a comprehensive clinical terminology for healthcare data. Like ICD, it does not provide metadata capabilities but is widely used for coding and sharing healthcare information. SNOMED is a well-known and used standard in the healthcare domain. Also, the OHDSI standard vocabularies from OMOP contain many codes based on SNOMED-CT, most of which are recognized as standard concepts by the community. Note that OHDSI does not support post-coordination unlike SNOMED does. Thus, the IDEA4RC project will also include the SNOMED-CT system.

By assessing the characteristics of these CDMs and metadata taxonomies, the IDEA4RC project can make informed decisions regarding the most suitable approach for implementing metadata standards within its ecosystem. This evaluation will contribute to the overall success of the project in improving the governance, sharing, and re-use of health data for rare cancers.

Focusing on OMOP (Observational Medical Outcomes Partnership) and FHIR (Fast Healthcare Interoperability Resources) for data modelling and data interchange offers several advantages in the context of IDEA4RC. These widely recognized and adopted CDMs provide a standardized and structured approach to representing, exchanging, and researching healthcare data, ensuring interoperability between different systems and organizations. By adhering to these standards, healthcare data can be easily shared, understood, and utilized across various applications and platforms. The broad community support surrounding OMOP and FHIR contributes to their ongoing development and refinement, leveraging the collective expertise of developers, researchers, and healthcare professionals. This ensures that the CDMs and interchange methodologies stay up-to-date, relevant, and aligned with evolving healthcare needs.

OMOP and FHIR offer flexibility and extensibility to accommodate diverse healthcare data requirements. OMOP's data modelling framework supports comprehensive analysis and health related research by enabling the representation of complex observational healthcare data, such as clinical and national registry data, claims data, EHR data, patient reported outcomes and others. On the other hand, FHIR's resource-oriented approach and support for profiles and extensions allow for the capture and exchange of a wide range of healthcare information, tailored to specific use cases or domain-specific requirements. Both standards facilitate seamless interoperability and integration between different healthcare systems and applications. OMOP CDM allows for the harmonization and integration of data from various sources with a research perspective, enabling comprehensive analysis and research across datasets. FHIR's robust data interchange capabilities enable efficient and secure exchange of healthcare data between systems, promoting interoperability in healthcare ecosystems. With their industry adoption and maturity, OMOP and FHIR provide a solid foundation for implementing data modelling and interchange solutions, minimizing risks, and ensuring compatibility with existing healthcare infrastructures. Especially seen the latest development on bridging the gap between the two undertaken by the joint communities. Leveraging these standards empowers stakeholders in the healthcare domain to achieve standardized data

representation, tap into community expertise, and drive innovation for improved patient care and research outcomes. In IDEA4RC, the use of both OMOP and FHIR allows each CoE to choose the approach they want to follow, whether it is using both to benefit from their advantages or choose one due to the CoE's preference.

3.1.1 OSIRIS, OMOP CDM and FHIR

In this section, we will delve into a detailed description of OSIRIS, OMOP and FHIR, examining their key features, capabilities, and potential applications within the oncology domain. This comparative analysis will provide valuable insights for the project team to make informed decisions regarding the adoption of suitable CDMs within the IDEA4RC ecosystem. By leveraging the strengths of these established models, the project aims to enhance the governance, sharing, and re-use of health data for rare cancers, advancing knowledge and improving outcomes for patients in Europe and beyond.

OSIRIS

The multidisciplinary inter-SIRIC OSIRIS consortium, dedicated to the standardization and sharing of clinical and biological data in oncology, has published its first results in the JCO Clinical Cancer Informatics journal. The consortium aimed to address the challenge of data sharing, which is hindered by the heterogeneity of data and information systems used in cancer research.

Over five years, the OSIRIS group worked towards structuring both clinical and biological data, resulting in a minimum dataset called the "Set OSIRIS." This dataset consists of 67 clinical items and 65 omics items and is structured using terminology based on existing reference frameworks. The Set OSIRIS is intended to be modifiable and expandable, with the possibility of incorporating imaging or immunology data in the future.

The OSIRIS group's efforts in structuring clinical and biological data aim to promote interoperability among different datasets collected in academic and industrial research projects, including real-world data. The widespread adoption of the Set OSIRIS would significantly facilitate data sharing, as it provides a standardized framework for data exchange. The latest developments of the OSIRIS model focus on aligning it with the HL7 Fast Healthcare

Interoperability Resources (FHIR) standard, an international standard for electronic health information exchange.

Overall, the structured clinical and biological data proposed by the OSIRIS consortium have the potential to enhance the interoperability of diverse clinical and biological datasets, promoting data sharing in both academic and industrial research settings, including real-world data.

Utilizing OSIRIS as a baseline for the IDEA4RC project brings numerous benefits, including leveraging the established standardization and data sharing efforts in oncology, saving development time, and capitalizing on the specific focus of OSIRIS on rare cancers. This approach ensures consistency and harmonization across participating centres, facilitating seamless integration and interoperability of data. By building upon the solid foundation provided by OSIRIS, the IDEA4RC project can efficiently adopt a CDM tailored to the unique requirements of rare cancers, enhancing collaboration, and advancing research in the field.

OMOP

The OMOP (Observational Medical Outcomes Partnership) CDM is an open community data standard designed to standardize the structure and content of observational data, enabling reliable evidence generation through efficient analyses. A key component of the OMOP CDM is the standardized vocabularies provided by OHDSI, which allow for the organization and standardization of medical terms across various clinical domains (including SNOMED-CT, ICD-O, TNM, etc.). These vocabularies facilitate standardized analytics, supporting the construction of exposure and outcome phenotypes, characterization studies, population-level effect estimation, and patient-level prediction studies.

Data standardization is crucial for collaborative research, large-scale analytics, and the sharing of tools and methodologies. Healthcare data can vary significantly between organizations, stored in different formats and database systems, and represented using different terminologies. The OMOP CDM addresses these challenges by providing a common format and representation, enabling systematic analysis of disparate observational databases. By transforming data into the OMOP CDM format, researchers can leverage a library of standard analytic routines and tools to analyse and generate evidence from diverse data sources.

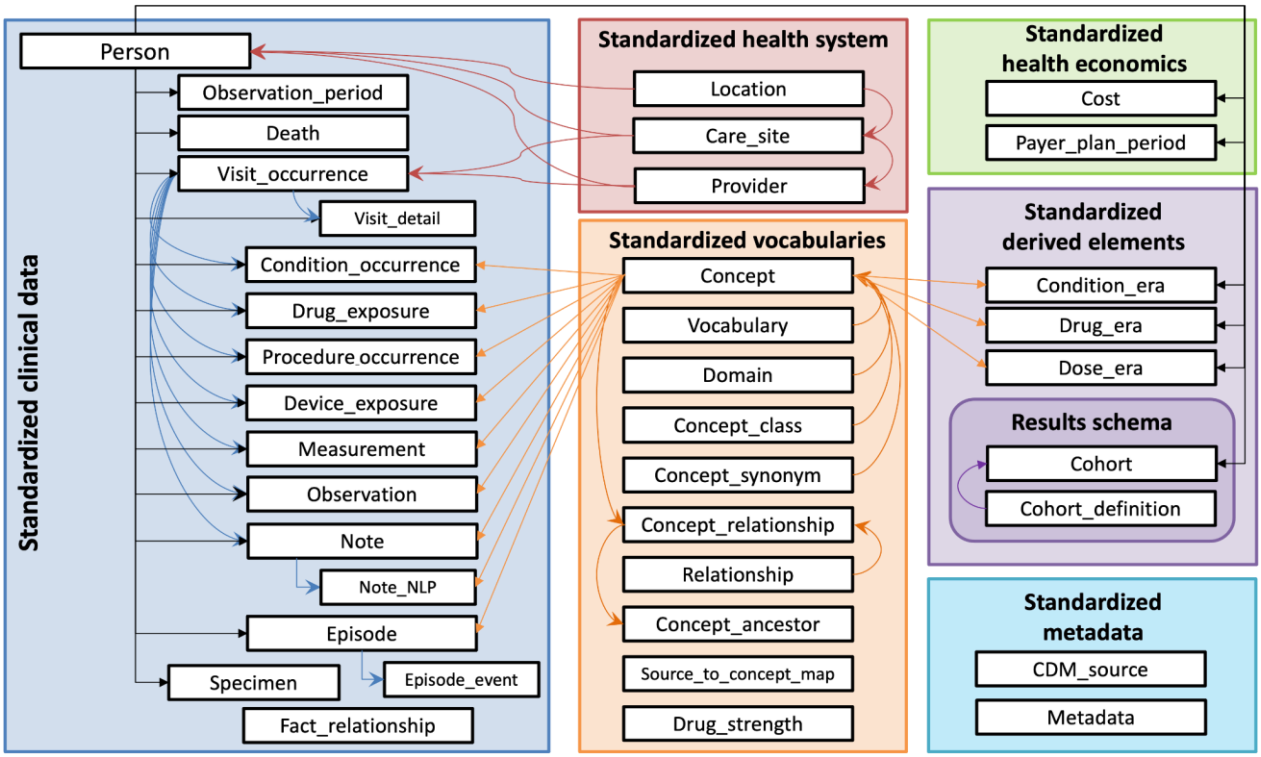
The OMOP CDM accommodates diverse types of observational health data, including electronic medical records (EMR), registries, and administrative claims data. It supports collaborative

research across different data sources and facilitates international collaboration. By utilizing the OMOP CDM, data owners can manage their data more effectively, while data users can leverage standardized analytics tools and methodologies to generate valuable insights. OHDSI, as an active global community, offers resources, expertise, and open-source tools for data conversion, maintenance, data quality assessment, medical product safety surveillance, comparative effectiveness studies, quality of care evaluation, and patient-level predictive modelling. Additionally, there are commercial sources of tools available to further support analysis within the OMOP CDM Framework.

OMOP and the oncological domain

The OMOP CDM v5.6 incorporates specific components for representing cancer diagnoses and treatments. In the context of cancer diagnosis, the model includes the Cancer Diagnostic Model, which consists of cancer diagnoses, diagnostic schemas, and diagnostic modifiers. A cancer diagnosis is defined as a combination of histology (morphology) and topography (anatomic site), while a diagnostic schema represents a group of cancer diagnoses with similar diagnostic features. Diagnostic modifiers encompass various attributes, including stage, grade, laterality, genomic biomarkers, and other relevant factors related to the diagnosis.

Within the OMOP CDM, cancer diagnoses are stored in the CONDITION_OCCURRENCE table. The MEASUREMENT table is utilized to store diagnostic modifiers, which are explicitly linked to the corresponding cancer diagnosis records in CONDITION_OCCURRENCE through specific columns (MEASUREMENT.modifier_of_event_id and MEASUREMENT.modifier_of_field_concept_id). MEASUREMENT.modifier_of_event_id contains the value of the respective condition_occurrence_id, while MEASUREMENT.modifier_of_field_concept_id contains the concept for the condition_occurrence_id field.



Regarding cancer treatments, the OMOP CDM includes the Cancer Treatment Model, comprising cancer treatments and treatment modifiers. Cancer treatments represent higher-level concepts that encompass therapeutic or diagnostic interventions, transcending transactional and observational treatment details. Treatment modifiers refine the description of cancer treatments and encompass attributes such as the number of fractions, radiation primary treatment volume, total dose, lymph nodes examined, surgical margins, and other relevant features.

Cancer treatments are stored in the PROCEDURE_OCCURRENCE and EPISODE tables. Similar to diagnostic modifiers, treatment modifiers are stored in the MEASUREMENT table and explicitly linked to the corresponding cancer treatment records in PROCEDURE_OCCURRENCE through designated columns (MEASUREMENT.modifier_of_event_id and MEASUREMENT.modifier_of_field_concept_id). MEASUREMENT.modifier_of_event_id contains the value of the respective procedure_occurrence_id, while MEASUREMENT.modifier_of_field_concept_id contains the concept for the procedure_occurrence_id field.

To facilitate clinically and analytically relevant representations of cancer diagnoses, treatments, and outcomes, data abstraction in the form of disease and treatment episodes is employed. Episodes capture specific events such as disease first occurrence, disease recurrence, disease remission, disease progression, treatment regimens, and treatment cycles. These episodes are represented in the EPISODE table, with disease episodes corresponding to concepts from the Condition domain and treatment episodes linked to concepts from the Procedure or Regimen domain. The idea is to have a hierarchical and temporal representation of the patient's disease and treatment journey to capture distinct levels of granularity in the data and the disease's progression. The relationship between disease episodes and treatment episodes is established using the self-referencing foreign key column EPISODE.episode_parent_id, enabling the association of cancer treatment with a cancer diagnosis for calculating time from diagnosis to treatment. Additionally, episode modifiers function similarly to condition and procedure modifiers, providing additional attributes for disease episodes and treatment episodes.

Overall, the OMOP CDM and its extensions provide a structured framework for capturing and organizing cancer diagnoses, treatments, and associated attributes, enabling comprehensive analyses and research in oncology.

The utilization of the OMOP (Observational Medical Outcomes Partnership) CDM in the IDEA4RC project offers significant benefits for rare cancer research and analysis. By adopting the standardized structure and vocabularies provided by OHDSI, the project ensures interoperability and harmonization of diverse healthcare datasets, enabling seamless integration and comparison of rare cancer data from multiple sources. The OMOP CDM's flexibility and extensive support for observational data allow for efficient data modelling and analysis, facilitating valuable insights into rare cancer epidemiology, treatment outcomes, and research outcomes. Furthermore, leveraging the established OMOP community and tools empowers IDEA4RC to collaborate, share knowledge, and leverage standardized analytics methodologies, contributing to the advancement of rare cancer research and improving patient care and outcomes.

FHIR

FHIR (Fast Healthcare Interoperability Resources) is a standards-based framework developed by Health Level Seven (HL7) for the exchange and management of healthcare information. It utilizes a resource-oriented approach to represent and organize healthcare data.

- *FHIR Resources*:
  FHIR resources are the building blocks of the FHIR framework, representing several types of healthcare concepts, such as patients, medications, observations, conditions, procedures, and more. Each resource is identified by a unique URL and is designed to be self-contained and shareable. Resources have a consistent structure and include elements to capture specific data related to the concept they represent.
- *Elements and Data Types*:
  FHIR resources consist of elements, which are the individual data components within a resource. Elements can have different data types, such as strings, numbers, booleans, dates, or complex types. Complex types are composed of multiple elements and can represent more structured data, such as addresses or codes. Elements can also have modifiers, such as extensions, which allow for the addition of custom or domain-specific data.
- *References and Relationships*:
  FHIR resources can reference other resources to establish relationships and provide contextual information. These references enable the representation of complex relationships between healthcare concepts. For example, a medication resource may reference a patient resource to indicate the individual for whom the medication is prescribed. References can be used to link resources within the same FHIR instance or across different systems.
- *Search and Query*:
  FHIR provides a powerful search mechanism that allows users to query and retrieve specific data from FHIR servers. It supports a variety of search parameters that can be combined to filter and narrow down the search results. These parameters can include patient demographics, clinical codes, dates, and more. The search capability of FHIR enhances interoperability by enabling targeted data retrieval based on specific criteria.

- *Profiles and Extensions*:

  FHIR allows for the creation of profiles and extensions to further customize and extend the standard resources. Profiles define constraints or additional requirements on resources to meet specific use cases or local requirements. Extensions provide a mechanism to add custom data elements or attributes to resources, enabling the representation of domain-specific or localized information.

- *Versioning and Lifecycle*:

  FHIR resources have built-in support for versioning and tracking changes. Each resource instance can have multiple versions, allowing for historical tracking and auditability. FHIR also supports the concept of transactional bundles, which enable the grouping of multiple resource interactions into a single atomic operation, ensuring data integrity and consistency.

Overall, FHIR's resource-oriented approach, structured elements, references, and search capabilities provide a flexible and standardized framework for representing, exchanging, and managing healthcare information. Its emphasis on interoperability, customization, and extensibility makes it well-suited for a wide range of healthcare use cases and enables seamless integration with existing systems and technologies.

The utilization of FHIR (Fast Healthcare Interoperability Resources) in the IDEA4RC project offers significant advantages. FHIR is a widely adopted data exchange standard that promotes interoperability and seamless data sharing across various healthcare systems. By leveraging FHIR, the IDEA4RC project can ensure compatibility and integration with existing healthcare infrastructures and systems, enabling efficient data interchange and collaboration. FHIR's flexible and extensible nature accommodates the diverse data requirements of rare cancers, allowing for the representation and exchange of complex clinical, genomic, and research data. Its active and supportive community, along with its robust tooling and resources, provide a solid foundation for the successful implementation and adoption of FHIR within the project, enhancing the governance, sharing, and re-use of health data for rare cancers.

3.2 Metadata: Description of the existing metadata taxonomies to describe quality, governance, findability, and provenance

In this subsection, we will analyse and compare metadata taxonomies that play a crucial role in describing various aspects of data quality, governance, findability, and provenance within the healthcare data domain.

3.2.1 Data quality

Data quality plays a pivotal role in ensuring the reliability and validity of research findings generated from health data. When data is not collected systematically for research purposes, its quality may be compromised, leading to potential negative impacts on the outcomes and insights derived from such data. Therefore, it becomes essential to assess and enhance data quality to ensure the accuracy and integrity of the information being utilized.

Electronic Health Records (EHRs), although primarily designed for efficient patient care and nonclinical administrative tasks, exhibit considerable variations in clinical documentation practices, even among users of the same system. These variations can contribute to challenges in data quality when using different data sources, making it imperative to implement robust frameworks focused on data quality. Moreover, the potential value of the secondary use of health data for research and development is widely acknowledged. However, substantial efforts are needed to enhance the quality and usability of such data.

By reviewing existing data quality taxonomies, we aim to identify the most relevant approaches and frameworks that align with the goals and requirements of the IDEA4RC project. Therefore, the state-of-the-art review we are going to carry out will focus on the analysis of 4 different data quality aspects we have considered to be essential for IDEA4RC: variable, dataset, hierarchical and scoring characteristics (see Table 2).

Table 2 - Comparison of different Data Quality taxonomies.

Title	Authors	Variable specific	Dataset Specific	Hierarchical aspect	Scoring
A Harmonized Data Quality Assessment Terminology and Framework for the Secondary Use of Electronic Health Record Data	Kahn et al.	Yes	No	No	No
Dimensions of Data Quality (DDQ)	DAMA NL Foundation	Yes	Yes	Yes	No
Data Quality Framework for EU medicines regulation	EMA	Yes	Yes	No	No
Health data metrics (HDM)	Institut Curie	Yes	No	Yes	No
Development of a data utility framework to support effective health data curation.	Health Data Research UK	Yes	Yes	No	Yes

Kahn's data quality framework for second use of EHR data

One of the most relevant research works regarding Data Quality is the work carried out by Kahn et al.[2] where authors developed a common Data Quality Assessment (DQA) taxonomy after unifying existing terminologies from the biomedical informatics field. Several steps were

[2] Kahn, M. G., Callahan, T. J., Barnard, J., Bauck, A. E., Brown, J., Davidson, B. N., Estiri, H., Goerg, C., Holve, E., Johnson, S. G., Liaw, S. T., Hamilton-Lopez, M., Meeker, D., Ong, T. C., Ryan, P., Shang, N., Weiskopf, N. G., Weng, C., Zozus, M. N., & Schilling, L. (2016). A Harmonized Data Quality Assessment Terminology and Framework for the Secondary Use of Electronic Health Record Data. EGEMS (Washington, DC), 4(1), 1244. https://doi.org/10.13063/2327-9214.1244

carried out for the creation of this taxonomy. First, they vet an initial set of Data Quality terms and definitions through several stakeholders' meetings. Then, feedback from data producers and users was gathered to later build a first draft set of harmonized DQ terms and categories. Finally, multiple refinement iterations were performed to achieve a harmonized terminology which was evaluated comparing it with ten other DQ terminologies. Moreover, the created DQ terminology was validated after aligning it with other published DQ terminologies.
This DQ taxonomy revolves around three distinct categories (conformance, completeness, and plausibility) and two different assessment contexts for each of those categories (verification and validation), as presented in Table 3.

The two assessment contexts, verification, and validation, were first introduced by Weiskopf and Weng[3] and be applied to most of the categories and subcategories of the proposed taxonomy.
The first assessment context, **Verification**, focuses on examining model and metadata data constraints, system assumptions, and local knowledge. Unlike Validation, Verification does not rely on an external reference. Instead, it emphasizes utilizing resources within the local environment to determine expected values and distributions. This context allows us to assess the consistency and compliance of data with internal specifications, ensuring that data aligns with predefined expectations and guidelines.
The second assessment context, **Validation**, emphasizes the alignment of data values with relevant external benchmarks. It involves comparing data values to external references or benchmarks to evaluate their accuracy and reliability. One approach to obtaining external benchmarks is by combining results across multiple data sites, leveraging the collective knowledge and expertise from various sources. Validation provides an essential external validation mechanism, enabling us to assess data quality by comparing it against established standards or measurements.

Regarding the main data quality categories, Kahn et al. proposed three distinct types:

[3] Weiskopf, N. G., & Weng, C. (2013). Methods and dimensions of electronic health record data quality assessment: enabling reuse for clinical research. Journal of the American Medical Informatics Association, 20(1), 144-151.

1. **Conformance** is crucial in verifying if data values adhere to specified standards and formats, encompassing internal or external formatting, relational or computational definitions. This category ensures that data values meet syntactic or structural constraints and can be described through data dictionaries that specify the intended format and allowed values for each data element.
2. **Completeness** focuses on assessing whether all expected data values are present. It ensures that the data collection process captures comprehensive coverage of the required data elements.
3. **Plausibility** evaluation involves assessing the believability or reasonableness of data values. It includes unique plausibility, which examines the presence of unexpected duplications within a database (verification) or when compared to external references (validation). Atemporal plausibility checks if observed data values and distributions align with local or common knowledge, considering factors such as age, gender, or socioeconomic values. Temporal plausibility evaluates if time-varying variables change values as expected based on known temporal properties or in comparison to external comparators or gold standards. It also considers temporal stability, continuity, state transitions, and dependencies between time-varying variables.

Table 3 - Summary of the possible combination of categories and assessments contexts with examples extracted from Kahn et al.

Category	Subcategory	Description	Examples
Conformance	Value	Data values conform to internal formatting constraints **(verification)**, allowable values **(verification)** or ranges and representation constraints based on external standard **(validation)**.	• Sex is only one ASCII character. **(verification)** • Sex only has values "M", "F" or "U". **(verification)** • Values for primary language conform to ISO standards. **(validation)**
	Relational	Unique data values are not duplicated **(verification),** relational constraints are respected **(verification)** and relational constraints based on external standards are respected **(validation)**.	• Patient medical record number links to other tables as required. **(verification)** • A medical record is assigned to a single patient. **(verification)** • Data values conform to all not NULL requirements in a common multi-institutional data exchange format. **(Validation)**

Category	Subcategory	Description	Examples
	Computational	Computed values conform to computation or programming specifications **(verification)** and also, results based on published algorithms yield values that match validation values provided by external sources **(validation)**.	• Database and hard-calculated Body mass Index values are identical. **(Verification)** • Computed BMI percentiles yield identical values compared to test results and values provided by the CDC. **(Validation)**
Completeness	-	The absence of data values at a single moment in time agrees with local or common expectations. The absence of data values measured over time agrees with local or common expectations. The absence of data values at a single moment in time agrees with trusted reference standards or external knowledge. The absence of data values measured over time agrees with trusted reference standards or external knowledge.	• The encounter ID variable has missing values. **(Verification)** • Gender should not be null. **(Verification)** • The number and percent of records with a NULL value in the care_site_id of the PERSON. (Threshold=100%). **(Verification)** • The number and percent of records with a value of 0 in the standard concept field condition_status_concept_id in the CONDITION_OCCURRENCE table. (Threshold=100%). **(Verification)** • Medical discharge time is missing for three consecutive days. **(Verification)** • The current encounter ID variable is missing twice as many values as the insitutionally validated database. **(Validation)** • A drop in ICD-9CM codes matches implementation of ICD-10CM. **(Validation)**
Plausability	Unique	Seeks to determine if objects appear multiple times in settings where they should not be duplicated or cannot be distinguished within a database (**verification**) or when compared with external reference (**validation**).	• Patients from a single institution do not have multiple medical record numbers. (**Verification**)
	Atemporal	Checks if observed data values and distributions agree with local or common knowledge (**verification**). Those values and distributions may vary depending on	• Height and weight values are positive. **(Verification)** • Counts of unique patients as diagnosis are as expected. **(Verification)** • Distribution of encounters per patient or medications per

Category	Subcategory	Description	Examples
		the context (stratified by age, gender or socioeconomic values). Where logic or knowledge do not provide clear guidance, external gold standards created by organizations could be used.	encounter distributions are as expected. **(Verification)** • Serum glucose measurement is similar to finger stick glucose measurement. **(Verification)** • Oral and axillary temperatures are similar. **(Verification)** • Sex value agree with sex specific contexts (pregnancy, prostate cancer). **(Verification)** • For a CONCEPT_ID 30969 (Testicular hyperfunction), the number and percent of records associated with patients with an implausible gender (correct gender = Male). (Threshold=5%). **(Validation)**
	Temporal	Determines if time-varying variables change values as expected based on known temporal properties or across one or more external comparators or gold standards.	• Admission data occurs before discharge date. **(Verification)** • Date of initial immunization precedes the date of a booster immunization. **(Verification)** • Counts of emergency room visits by month show an expected spike during flu season. **(Verification)** • If yes, the number and percent of records with a date value in the observation_period_end_date field of the OBSERVATION_PERIOD table that occurs after death. (Threshold=1%). **(Verification)** • The number and percent of records with a value in the measurement_date field of the MEASUREMENT that occurs prior to the date in the BIRTH_DATETIME field of the PERSON table. (Threshold=1%). **(Verification)** • Immunization sequences match the CDC recommendations. **(Validation)** • Counts of emergency room visits by month shows spike during flu season that are similar to local health

Category	Subcategory	Description	Examples
			department reports.**(Validation)**

Apart from the evaluation performed in Khan et al. for their taxonomy, other research works such as Callahan et al.[4] mapped the data quality checks of six organizations to the harmonized DQA terminology introduced in Khan et al. To ensure consistency in the mapping process of the data quality checks, authors established some mapping conventions and after four iterations, those DQ checks that were difficult to map were discussed among the research team members until a consensus was reached. Among their main findings, they conclude that using the DQA terminology were able to map 99.97% of the DQ checks (49.6% Atemporal Plausibility, 17.84% to Value Conformance and 12.98 to Atemporal Completeness) in the six organizations, claiming that through a common DQA taxonomy or data quality checks could help in the connection between different clinical data networks. However, the distribution of mapped checks varies depending on the organization.

EMA - Data Quality Framework for EU medicines regulation

Following a similar taxonomy, the European Medicines Agency (EMA) provided its own recommendations and guidelines[5] about how data quality should be handled in the health domain. In this new quality framework introduced by the EMA, they identified two broad aspects of data quality: **reliability** and **relevance**. The former one is tightly related to the data quality taxonomy introduced by Kahn et al, even though categories such as conformance are renamed to coherence. Moreover, they also propose other quality/reliability aspects such as the qualification received by other actors, the description of the ETL process and its status, if it has been mapped to a CDM, etc.

The second aspect of data quality, relevance, is defined as "to the extent to which a dataset presents data elements useful to answer a research question". This aspect of data quality is not

[4] Callahan, T. J., Bauck, A. E., Bertoch, D., Brown, J., Khare, R., Ryan, P. B., Staab, J., Zozus, M. N., & Kahn, M. G. (2017). A Comparison of Data Quality Assessment Checks in Six Data Sharing Networks. EGEMS (Washington, DC), 5(1), 8. https://doi.org/10.5334/egems.223

[5] Data Quality Framework for EU medicines regulation. EMA. 2022-09-30.

covered in the work carried out by Khan et al., even though, in our opinion, the quality characteristic related to relevance can be extremely helpful when it comes to analysing how good a data source is or if it has some prerequisites.

Some examples of different types of relevance quality checks suggested by the EMA are:

- Setting.
  - Data source countries.
  - Data source regions.
  - Type of data source:
    - Administrative details
      - Name of the data source.
      - Acronym.
      - Data holder.
      - Contact name.
      - Contact mail.
      - Languages.
      - Establishment of the data source.
      - First collection date.
      - Last collection date.
      - Website.
      - Data source type
        - Administrative
        - Secondary Care
        - Registries (Cancer registry)
      - Care setting
        - Primary Care - GP, community pharmacist level
        - Primary Care- Specialist level
        - Secondary care - specialist level
        - Hospital inpatient care
        - Hospital outpatient care
      - ...
- Population.
  - Population size.
  - Population size by age.
  - Active population size.
  - Active population size by age.
  - Population covered by the data source.
  - Population not covered by the data source.
  - Population age groups.
  - Population covered by the data source.

- Population not covered by the data source.
- Sociodemographic information.
- Lifestyle factors.
- ...

- Exposure.
  - Procedures.
  - Biomarker data.
  - Prescriptions
  - ATMP.
  - ...
- Outcomes.
  - Specific diseases.
  - Hospital admission discharge.
  - Cause of death.
  - Clinical measurements.
  - Diagnostic codes.
- Time elements:
  - First collection date.
  - Last collection date.
  - Median time of the first and last available records.
  - ...

Health Data Research UK - Development of a data utility framework to support effective health data curation

In an analogous manner to the work carried out by Khan et al. and the EMA, the dimensions proposed by Black and Van Nederpelt[6] also mention quality reliability dimensions such as conformance, completeness, plausibility, or timeliness with their respective subcategories. However, in this work, Black and Van Nederpelt listed 60 different dimensions of data quality at different hierarchical levels, such as data values, attributes, data, or datasets. Nevertheless, the analysis performed by the authors was not exclusively centred on clinical data, unlike the rest of the works we are reviewing.

[6] A. Black and P. Van Nederpelt. Dimensions of Data Quality (DDQ). DAMA NL Foundation, 2020.

DAMA - Dimensions of Data Quality (DDQ)

Gordon et al.[7] developed a user-centred data utility framework to evaluate the utility of specific healthcare datasets. The main difficulty they encountered was to detect which metrics or characteristics provide the most value when it comes to evaluating the utility of clinical data. As the authors affirm, the most used metric in this context is data quality, which involves the analysis of various dimensions and some subjective assessment factors depending on the domain or use case, as proposed by Khan et al. However, the authors aim to extend the number of metrics we can use for evaluating the usefulness of a health dataset. To do so, they developed a framework for characterizing datasets through a series of interviews and surveys with data users.

In terms of strict data quality, they defined two main dimensions, each of them ranked with 4 qualitative categories: bronze, silver, gold, and platinum. Those dimensions are:

- Data quality management process: the level of maturity of the data quality management process.
- Data Management Association (DAMA) Quality Dimensions: completeness, uniqueness, accuracy, validity, timeliness, and consistency. As previously stated, these dimensions closely align with those outlined by Khan et al.

However, they also proposed other categories that, even though not related to data quality, are helpful when evaluating the quality and usability of a clinical dataset. The proposed categories are:

- Data documentation: documentation completeness, availability of additional documentation and support; data model, data dictionary; provenance.
- Coverage: pathway coverage and length of the follow-ups.
- Access & provision: allowable uses, time lag and timelines.
- Value & Interest: linkages (ability to link with other datasets) and data enrichments (data sources enriched with annotation, labels, etc.).

[7] Gordon, B., Barrett, J., Fennessy, C., Cake, C., Milward, A., Irwin, C., Jones, M., & Sebire, N. (2021). Development of a data utility framework to support effective health data curation. BMJ health & care informatics, 28(1), e100303. https://doi.org/10.1136/bmjhci-2020-100303

Institute Curie - Health data metrics (HDM)

Finally, it is worth mentioning the work carried out by the Institute *Curie-Data Factory*[8] who developed a Data Quality assessment application named *Health Data Metrics*[9] which to the best of our knowledge, is one of the few systems where different hierarchical levels for data quality metrics have been defined. For them, it is crucial to establish the calculation scopes of the metrics to accurately evaluate the data quality. In this regard, they delineated six hierarchical levels of metrics.

- Level 0: this level pertains to computing metrics across all versions of a database.
- Level 1: metrics are calculated at the level of a specific version of a database.
- Level 2: metrics are computed for tables, specific versions of a database, or the entire database.
- Level 3: metrics are calculated for individual columns, irrespective of their data type (e.g., counting missing or NULL values).
- Level 4: metrics are computed for columns, considering their data types such as numeric, textual, categorical, continuous, date, or identification, among others.
- Level 5: this level involves calculating metrics specifically for categorical variables, including frequency and distinct values.

Unfortunately, these hierarchical levels cannot be reused for IDEA4RC due to the federated nature of our project. However, it will serve as a basis for developing our own levels.

3.2.2 Governance

In the case of the governance-related metadata for health data, several papers have been analysed. To systematically analyze the state of the art, we searched in bibliographic databases (i.e., Scopus and scholar) to find related papers. Three queries have been used:

[8] https://curie-data-factory.github.io/

[9] https://github.com/curie-data-factory/health-data-metrics

Table 4 - Queries used for the analysis.

Query	Number of results
`TITLE ( "governance" ) AND TITLE ( "ontology" )`	64
`TITLE-ABS-KEY ( "governance" ) AND TITLE-ABS-KEY ( "data" ) AND TITLE-ABS-KEY ( "health" )`	7,554
`TITLE-ABS-KEY ( "governance" ) AND TITLE-ABS-KEY ( "data" ) AND TITLE-ABS-KEY ( "health" ) AND TITLE-ABS-KEY ( "ontology" )`	71

Based on these results, we have filtered the papers based on the title and abstract. From the initial list, 115 papers were selected. After performing an analysis of the 115 papers, 9 were finally selected for the state-of-the-art comparison. In the following Table 5, a summary of the conclusions can be found. The table is organized as follows:

- Title: Title of the paper.
- Provenance: If the proposed system provides a mechanism to annotate the data's provenance.
- Control mechanism: How access to the data is managed.
- Used technology: The underlying technology.
- Federated: If the proposed system supports a federated approach to data sharing.
- Granularity: Granularity of the data access control.

This project has received funding from the European Union's Horizon Europe research and innovation programme under grant agreement no. 101057048

Table 5 - Summary of the papers.

Title	Provenance	Control mechanism	Used technology	Federated	Granularity
A Consent Model for Blockchain-Based Health Data Sharing Platforms	Blockchain	Smart contracts	Ethereum	Blockchain	Whole dataset
Ontology-based governance of data-aware processes_	No	Rules	Not implemented	No	user, access, object
Leveraging Algorithms to Improve Decision-Making Workflows for Genomic Data Access and Management	No	Hibrid: human/machine	Not implemented	No	Not specified
DiiS: A biomedical data access framework for aiding data driven research supporting FAIR principles	No	Rules	Not implemented	Not implemented	Not implemented
A collaborative, realism-based, electronic healthcare graph: Public data, CDMs, and practical instantiation	No	No	OMOP	No	No
Semantic security for E-health: A case study in enhanced access	No	Role-Based Access Control	Security Ontology	Yes	Action type and entities

This project has received funding from the European Union's Horizon Europe research and innovation programme under grant agreement no. 101057048

Title	Provenance	Control mechanism	Used technology	Federated	Granularity
control					
Semantic-based privacy protection of electronic health records for collaborative research	No	Role-Based Access Control	Security Ontology, XACML	No	Attribute based access control: policies can be defined, which contain rules for accessing different attributes.
Semantic generation of clouds privacy policies	No	Rules	Security Ontology, SWRL and XACML	No	Action type and entities
IdSM-O: an IoT data Sharing Management Ontology for Data Governance	PROV-O	Rules	Security Ontology, SWRL	No	Two levels: abstraction of traditional access control entities (subject, action, object) into meta entities (role, activity, view

As can be seen in the table several technologies are shared between some of the papers when it comes to implementing the proposed systems. For those systems that are based in security ontologies, SWRL and XACML are popular choices.

SWRL (Semantic Web Rule Language) is an ontology language that extends the Semantic Web languages OWL (Web Ontology Language) and RDF (Resource Description Framework) with rules. It provides a way to represent and reason about knowledge using rules within the framework of the Semantic Web. SWRL allows users to define rules that express relationships and constraints among classes and individuals in an ontology. These rules are typically expressed in the form of logical axioms, which consist of an antecedent (also known as the body) and a consequent (also known as the head). The antecedent specifies the conditions that must be satisfied for the rule to be applicable, and the consequent specifies the action or inference that should be performed when the rule is triggered. The language itself is based on a combination of the OWL and RuleML languages. It uses OWL's ontological constructs to represent concepts, properties, and relationships, while also incorporating RuleML's syntax and semantics for expressing rules. SWRL provides a powerful mechanism for adding inferential capabilities to ontologies. By defining rules, it becomes possible to derive new knowledge from existing knowledge in the ontology. This enables more advanced reasoning and automated inference, allowing systems to make logical deductions and draw conclusions based on the defined rules. SWRL helps enhance the expressivity and reasoning capabilities of Semantic Web applications by allowing the specification of additional logical rules to augment the information contained in ontologies.

On the other hand, XACML (extensible Access Control Markup Language) is a standard for specifying and enforcing access control policies in information systems. It provides a framework for defining and managing fine-grained access control decisions, allowing organizations to control and manage access to resources based on a set of rules and policies. At its core, XACML enables the separation of access control policies from the application logic. It defines a policy language that allows administrators to specify access control rules in a declarative manner, independent of specific applications or systems. These policies can be written to handle complex scenarios involving multiple conditions, user attributes, resource attributes, and environmental factors.

The key components of XACML are as follows:

1. Policy Decision Point (PDP): The PDP is responsible for evaluating access requests against the defined access control policies. It takes the request, along with relevant attributes, and determines whether access should be granted or denied based on the policies.
2. Policy Enforcement Point (PEP): The PEP is the component that intercepts access requests and enforces the access control decisions made by the PDP. It communicates with the PDP, sending access requests and receiving access decisions.
3. Policy Information Point (PIP): The PIP provides additional attribute information to the PDP during the access control evaluation process. It serves as a source of external attribute values required for policy evaluation.
4. Policy Administration Point (PAP): The PAP is responsible for managing the access control policies. It allows administrators to define, update, and delete policies and manage the attributes and other relevant information.

XACML supports attribute-based access control (ABAC), which means that access control decisions can be made based on various attributes such as user roles, resource properties, environmental factors, and more. It offers a flexible and extensible architecture that can be integrated with different systems and applications. By using XACML, organizations can achieve centralized and consistent access control policies across multiple applications and resources. It promotes policy reusability, simplifies policy management, and enables dynamic and adaptive access control decisions based on real-time attributes and conditions. XACML provides a standardized approach to access control policy definition and enforcement, facilitating the implementation of fine-grained access control mechanisms in diverse information systems.

Blockchain-based approaches are also found in the literature. Blockchain is a decentralized and distributed digital ledger technology that allows multiple parties to record and verify transactions in a secure and transparent manner. It was originally introduced as the underlying technology behind the cryptocurrency Bitcoin but has since found applications in various industries beyond finance. At its core, a blockchain is a chain of blocks, where each block contains a list of transactions or other data. These blocks are linked together using cryptographic hashes, creating an immutable and tamper-resistant record of all transactions that have occurred on the network.

Regarding the provenance, for those systems that consider the modelling of data provenance, two approaches are found: blockchain technologies and PROV-O. PROV-O, also known as the PROV Ontology, is a standard ontology for representing and expressing provenance information. Provenance refers to the origin, derivation, and history of a piece of data or artifact, including how it was created, modified, or accessed. The PROV-O ontology is part of the larger W3C PROV (Provenance) standard, which provides a framework and set of specifications for representing provenance in a machine-readable format. PROV-O is expressed using the Web Ontology Language (OWL) and provides a vocabulary for describing entities, activities, and their relationships in the context of provenance.

Key concepts in PROV-O include:

1. Entities: Entities are things of interest that are generated, manipulated, or used. They can represent physical or digital objects, such as documents, datasets, or software artifacts.
2. Activities: Activities represent actions or processes that manipulate or generate entities. They can be computational processes, transformations, or any activity that affects the state of an entity.
3. Agents: Agents are entities that are responsible for activities. They can represent people, organizations, software systems, or any entity capable of performing an action.
4. Relationships: PROV-O defines various relationships to capture the connections between entities, activities, and agents. For example, the "wasGeneratedBy" relationship indicates that an entity was created by an activity, while the "wasAssociatedWith" relationship indicates the association of an agent with an activity.
5. Derivations: PROV-O allows for representing the derivation of entities from other entities. This includes tracking how entities are derived or transformed through activities and other entities, forming a lineage of transformations.

By representing provenance information using PROV-O, applications can capture and communicate the history and context of data or artifacts. This helps with understanding the reliability, quality, and trustworthiness of information, as well as supporting reproducibility, accountability, and auditing in various domains such as scientific research, data integration, cybersecurity, and data governance.

3.2.3 Findability

In this section, we are going to delve into the world of data findability metadata. The purpose of this review is to explore the dimensions and possible values related to data findability (DF) within the IDEA4RC project, with a focus on harmonizing these aspects across all Centers of Excellence (CoE).

As part of the IDEA4RC project, we aim to adhere to the FAIR principles, whenever applicable. In this context, defining Findability aspects within these principles becomes essential. However, it is worth noting that due to the specific setting and inherent limitations associated with healthcare data, not all FAIR principles may be fully applicable to the IDEA4RC data/metadata.

To facilitate data discoverability, we are particularly interested in examining the DCAT Application Profile for Data Portals in Europe (DCAT-AP[10]), following TEHDAS[11] recommendations. The DCAT-AP, based on the Data Catalogue Vocabulary (DCAT)[12] developed by W3C, serves as a standard specification for describing public sector datasets in Europe. Its primary purpose is to facilitate the exchange of dataset descriptions among data portals, enabling data catalogues to describe their dataset collections in a standardized manner while maintaining their respective systems for documentation and storage.
Moreover, DCAT-AP allows content aggregators, like the European Data Portal, to aggregate these standardized dataset descriptions into a centralized point of access, making it easier for data consumers to discover and access datasets from a single unified location.
Taking into consideration TEHDAS[13] recommendations, where they found DCAT-AP and INSPIRE to be best suited for data discoverability, based on Common Assessment Method for

[10] https://joinup.ec.europa.eu/collection/semic-support-centre/solution/dcat-application-profile-data-portals-europe/release/300
[11] https://tehdas.eu/
[12] https://www.w3.org/TR/vocab-dcat-3/
[13] Recommendations to enhance interoperability within HealthData@EU. TEHDAS. 2022-12-21. https://tehdas.eu/app/uploads/2022/12/tehdas-recommendations-to-enhance-interoperability-within-healthdata-at-eu.pdf

Standards and Specifications (CAMSS)[14] evaluation, obtaining the maximum score on most of the analysed criteria:

- Core interoperability principles: openness, transparency, reusability, technological neutrality, and data portability.
- Users' needs and expectations: user-centricity, inclusion and accessibility, security, privacy, and multilingualism.
- Principles for cooperation between institutions: administrative simplification, preservation of information, assessment of effectiveness and efficiency.
- Interoperability layers: interoperability governance, organizational interoperability, semantic interoperability, and technical interoperability.

Despite being evaluated as a powerful metadata taxonomy, the widespread adoption of DCAT-AP, a prominent data interoperability standard, has been a subject of interest. To analyse its level of adoption in various countries, TEDHAS conducted a survey, analysing its implementation in different contexts. Among the countries mentioned in the survey, DCAT-AP has been acknowledged and utilized in France, Finland (with an ad hoc extension), and Norway. Regarding the use of DCAT-AP for health findability, the Norwegian Directorate of eHealth has emerged as the pioneers in adopting data interoperability standards within the healthcare domain. They have taken significant strides by developing a metadata specification that draws inspiration from DCAT-AP properties.

To promote the widespread use of DCAT-AP as the standard for health data discoverability across Europe, the HealthData@EU Pilot[15] project has taken a significant step forward. This two-year-long European project, co-financed by the EU4Health program, is dedicated to building a pilot version of the European Health Data Space (EHDS) infrastructure, facilitating the secondary use of health data for various purposes.

The HealthData@EU Pilot project aims to extend the DCAT-AP standard by developing a health-specific extension. This extension will entail the inclusion of new properties that are particularly relevant for health-related datasets or data registries. By introducing these

[14]https://joinup.ec.europa.eu/collection/common-assessment-method-standards-and-specifications-camss/about

[15] https://ehds2pilot.eu/

specialized properties, the project seeks to enhance the metadata standard's capabilities, ensuring it aligns seamlessly with the unique requirements and complexities of health data. To do so, a dedicated working group has been actively involved in designing the *Health* extension to the DCAT-AP metadata standard. This collaborative approach includes valuable input from project stakeholders and data providers involved in various health-related use cases. Moreover, they are accepting suggestions for new DCAT properties at http://search.healthdataportal.eu/.

Within the IDEA4RC project, we aim to DCAT-AP as the standard for health data discoverability. However, we will closely follow the progress of the HealthData@EU Pilot project in this matter, and we will try to actively participate in the addition of new properties to the standard as new needs are found during the project.

4. IDEA4RC ADOPTED COMMON DATA & SEMANTIC MODELS

The design of the IDEA4RC CDM has been done following the FAIR principles[16]. The FAIR principles, standing for Findable, Accessible, Interoperable, and Reusable, are a set of guidelines that aim to improve the ability to find, access, share, and reuse digital research outputs, such as data, software, and publications. These principles were first introduced in 2016 as a response to the growing concerns around the ability to access and reuse scientific data, which often remains hidden, inaccessible, or not easily reusable by researchers outside the original context in which it was generated.

The **Findable** principle requires that digital research outputs are assigned a unique and persistent identifier, such as a DOI, and described with rich and accurate metadata that allows users to easily discover them. This includes providing contextual information, such as the author, the date, and the license under which the data can be used. The **Accessible** principle mandates that digital research outputs are openly accessible, free of charge or with reasonable costs, and that any necessary authentication and authorization processes are clearly documented and easily navigable. The **Interoperable** principle requires that digital research outputs are represented using a standardized format and language, such that they can be integrated and combined with other digital research outputs, enabling new insights and discoveries. The **Reusable** principle requires that digital research outputs are made available with a clear and open license, allowing others to use, modify, and distribute them with minimum legal or technical restrictions. Additionally, the principle calls for ensuring that the data is documented and structured in a way that allows other researchers to understand and reproduce the results.

The FAIR principles have gained significant importance in the scientific community due to their potential to promote open science, enhance the visibility and impact of research, and facilitate data-driven innovation. By ensuring that research data is findable, accessible, interoperable, and reusable, the FAIR principles enable researchers to build upon each other's work, which accelerates scientific discovery and improves the quality of research outcomes. Additionally,

[16] FAIR Principles. Online. URL: https://www.go-fair.org/fair-principles/

the FAIR principles promote the reproducibility and transparency of research, which ensures that scientific findings are trustworthy and can be verified by others. Furthermore, the FAIR principles contribute to fostering a culture of responsible data sharing and ethical considerations in research. The FAIR principles serve as a framework for achieving these goals by providing a set of guidelines and recommendations for making research data more discoverable, accessible, reusable, and interoperable.

The CDMs adopted within the IDEA4RC project follow the FAIR principles because the consortium believes in the importance of ensuring that rare cancer data is Findable, Accessible, Interoperable, and Reusable. By following these principles, we are creating a framework that makes it easier for researchers to access and use Rare Cancer data, which leads to more efficient and effective research outcomes.

5. IDEA4RC ADOPTED CDM V1

The IDEA4RC project recognizes the importance of establishing a CDM specification for rare cancers to enable seamless data integration and interoperability across different healthcare centres and research institutions. To achieve this goal, the project leverages the baseline provided by the OSIRIS initiative, which focuses on the standardization and sharing of clinical and biological data in oncology.

Building upon the OSIRIS framework, the IDEA4RC project endeavors to ensure compatibility with two widely adopted and recognized CDMs: the OMOP (Observational Medical Outcomes Partnership) CDM and FHIR (Fast Healthcare Interoperability Resources). By aligning the adopted CDMs within IDEA4RC with these standards, the project aims to establish a flexible and agnostic CDM that can be utilized with both OMOP and FHIR, depending on the specific requirements and preferences of each participating centre.

The integration of OMOP and FHIR within the IDEA4RC project offers several advantages. OMOP provides a comprehensive data modelling framework for observational healthcare data, enabling the representation and analysis of diverse data elements relevant to rare cancers. FHIR, on the other hand, offers a standardized approach for data interchange and interoperability, facilitating the seamless exchange of healthcare information between different systems and applications. Also seen the ongoing work by the two communities to build ETLs between the two CDMs and align where possible.

By developing a mapping from the IDEA4RC variables to both OMOP and FHIR, the project ensures that the data collected and stored within the ecosystem can be easily transformed and exchanged using either standard. This standard-specific agnostic approach allows each participating centre to choose the most suitable standard based on their existing infrastructure, expertise, and specific use cases.

The utilization of a CDM compliant with both OMOP and FHIR fosters interoperability and data harmonization across centres, facilitating collaborative research, comparative analyses, and the sharing of valuable insights in the field of rare cancers. Moreover, it promotes scalability and

futureproofing, as the CDMs adopted remain adaptable to evolving standards and technological advancements in the healthcare domain. Overall, the compatibility with OMOP and FHIR enhances the usability, versatility, and long-term sustainability of the IDEA4RC data ecosystem, advancing the understanding and management of rare cancers.

5.1 Data model definition methodology

In defining the data model to be adopted within IDEA4RC, a systematic approach was followed, involving the collaboration of clinicians and stakeholders. The initial step involved clinicians agreeing upon the variables and data elements that were deemed essential for capturing and analysing rare cancer data. These variables were then adapted and structured according to the Entity-Relationship Diagram (ERD) framework.

To streamline the data modelling, the variables were further grouped and organized by entities, ensuring a cohesive and logical representation of the data. With the foundational logical representation in place, the next phase focused on the mapping of the variables to two key standards: FHIR (Fast Healthcare Interoperability Resources) and OMOP (Observational Medical Outcomes Partnership). This mapping process enabled the alignment of the CDMs adopted within the IDEA4RC project, ensuring interoperability and compatibility with existing healthcare systems and research initiatives.
In mapping the rare cancers variables to FHIR, the data elements and variables were aligned with the FHIR resources and elements, ensuring consistency and adherence to the FHIR data exchange standard. Similarly, the mapping to the OMOP CDM involved aligning the variables with the corresponding tables and concepts in the OHDSI standard vocabularies, enhancing the standardization and semantic interoperability of the data.

Throughout the development process, regular reviews and feedback sessions were conducted with clinicians. A quick initial review was conducted with a single clinician to ensure the accuracy and relevance of the mappings. Subsequently, a comprehensive review involving all clinicians was carried out, allowing for a thorough evaluation and validation of the adopted CDMs' suitability for capturing and analysing rare cancer data. This iterative process ensured

that the adopted CDMs met the specific requirements and expectations of the clinicians and researchers involved in the IDEA4RC project.
Additionally, it was agreed upon with BLUEBERRY, a specific clinician or clinical team specializing in Sarcoma, to ensure that the adopted CDMs adequately capture the unique aspects and variables relevant to this rare cancer type. This collaboration ensured that the work done within IDEA4RC effectively also addresses the specific needs and considerations of Sarcoma research and care. Note however, that Blueberry handles the inclusion local registries and the ETLs are defined upon them.

By following this systematic approach and engaging clinicians throughout the process, the IDEA4RC project successfully adopted a set of robust and comprehensive CDMs tailored to the unique requirements of rare cancer research. The adoption of data models such as FHIR and OMOP, along with the utilization of specialized vocabularies and collaborative reviews, ensures the interoperability, standardization, and quality of the data, enabling meaningful analysis and research in the field of rare cancers.

5.2 Head & Neck Cancer: Description of the model with tables for the selected core variables, draft of the FHIR implementation guide, entity-relationship diagram.

This section provides a detailed description of the data model specifically adopted for Head & Neck Cancer within the context of the IDEA4RC project. It includes an overview of the selected core variables, accompanied by tables that outline their structure and organization. Also, a draft of the FHIR implementation guide and an entity-relationship diagram are presented to provide a comprehensive understanding of the data model's architecture and implementation approach. This section offers valuable insights into the design and structure of the data model, highlighting its relevance and applicability in the context of Head & Neck Cancer research and analysis.

ERD

The Entity-Relationship Diagram (ERD) presented in this section illustrates the logical structure and relationships between the entities within the Head & Neck Cancer data in the IDEA4RC project. The ERD provides a visual representation of how different entities, such as Patient, CancerEpisode, Treatment (Surgery, Systemic treatment and Radiotherapy), etc. are related to each other through various associations and attributes. This diagram serves as a valuable tool for understanding the interconnections between different data elements. By examining the ERD, researchers and stakeholders can gain insights into the relationships and dependencies, facilitating a comprehensive understanding of the data organization and facilitating effective data analysis and research in the context of Head & Neck Cancer.

This project has received funding from the European Union's Horizon Europe research and innovation programme under grant agreement no. 101057048

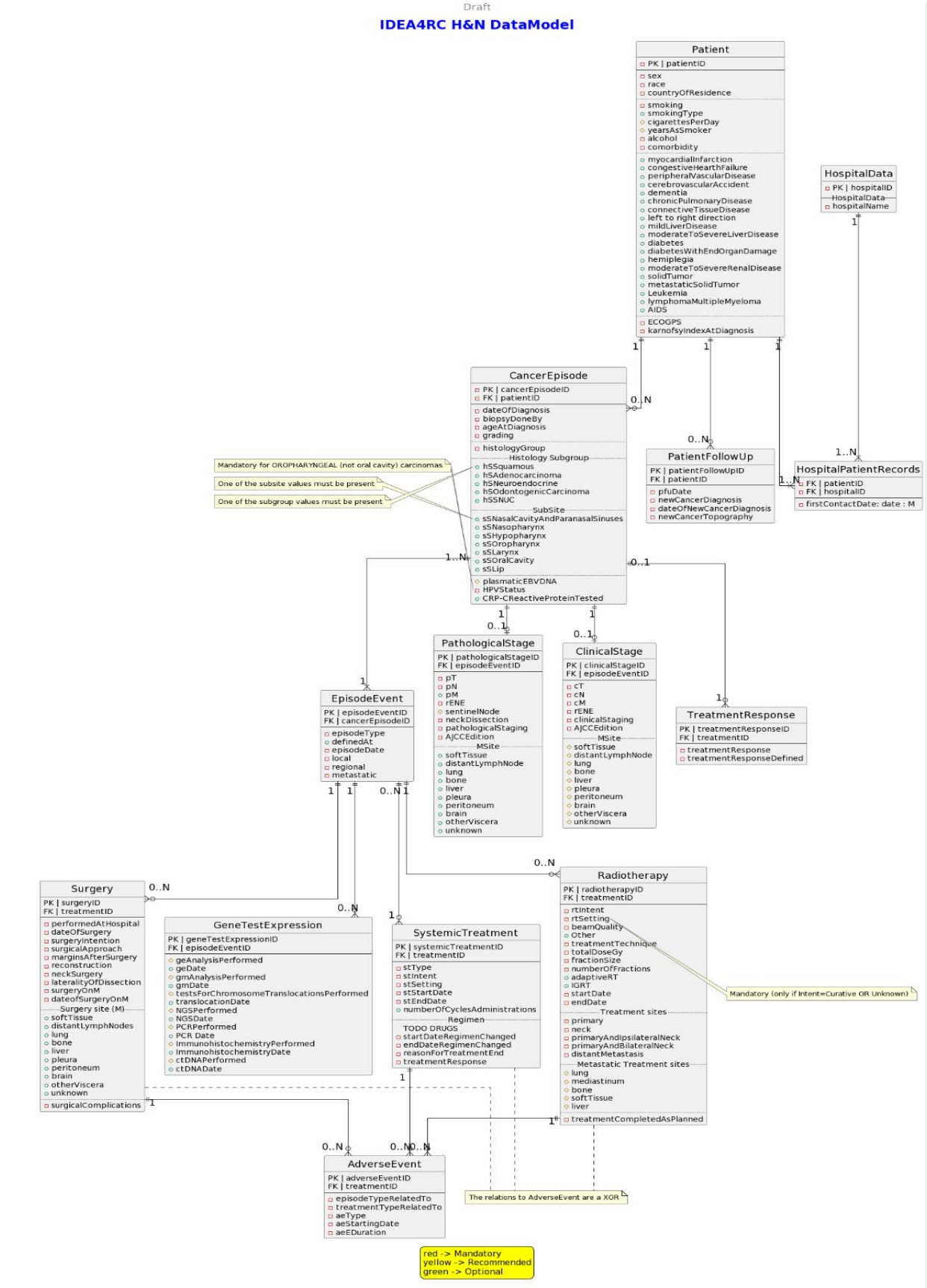


Figure 2 - The ERD.

Entities description:

- Patient: The Patient entity represents an individual who is receiving medical care for cancer. It includes demographic information, such as the age, gender, and relevant clinical characteristics.
- HospitalData: The HospitalData entity encapsulates information related to the healthcare institution where the patient is receiving treatment for cancer. It may include details such as the hospital's name, location, specialized departments, and other relevant administrative information.
- HospitalPatientRecords: The HospitalPatientRecords entity contains comprehensive medical records specific to the patient within the hospital's system.
- PatientFollowUp: The PatientFollowUp entity captures the information related to the ongoing monitoring and follow-up care provided to the patient. It includes details such as the frequency of follow-up visits, examination findings, and any new diagnoses.
- CancerEpisode: The CancerEpisode entity represents a specific episode of the patient's cancer diagnosis and treatment journey. It encompasses a defined period during which the patient undergoes a series of diagnostic procedures, treatment interventions, and monitoring.
- EpisodeEvent: The EpisodeEvent entity captures specific events or milestones that occur within a cancer episode. It includes critical occurrences such as disease progression, treatment response, recurrence, or other significant clinical events that impact the patient's treatment trajectory.
- PathologicalStage: The PathologicalStage entity denotes the stage of cancer determined by examining the tumour tissue and assessing its characteristics under a microscope. It provides information about the extent of the cancer and its progression, assisting in treatment planning and prognostic evaluation.
- ClinicalStage: The ClinicalStage entity represents the stage of cancer determined through clinical assessments, imaging tests, and physical examinations. It provides insights into the size of the tumour, its spread to nearby lymph nodes or distant sites, and other relevant clinical factors.
- Surgery: The Surgery entity captures information related to surgical procedures performed as part of the patient's cancer treatment. It includes details such as the type of surgery,

surgical approach, extent of resection, and any complications or adverse events associated with the procedure.

- Radiotherapy: The Radiotherapy entity represents the administration of radiation therapy as a treatment modality for cancer. It includes details about the radiation dosage, treatment schedule, radiation delivery techniques, and any observed side effects or complications.
- SystemicTreatment: The SystemicTreatment entity encompasses the administration of systemic therapies, such as chemotherapy or targeted therapy, for the treatment of cancer. It includes information about the specific drugs, treatment regimen, dosage, duration, and any associated adverse events.
- AdverseEvent: The AdverseEvent entity captures any unfavourable or unexpected events or reactions experienced by the patient during the course of their cancer treatment. It includes details about the type of adverse event, severity or duration.
- GeneTestExpression: The GeneTestExpression entity represents the results of genetic tests conducted to assess the expression levels or mutations of specific genes associated with cancer. It includes information about the genes tested, the methodology used, and the interpretation of the test results.
- TreatmentResponse: The TreatmentResponse entity captures the patient's response to the overall cancer treatment progression. It includes information about treatment outcomes, such as complete response, partial response, stable disease, or disease progression, providing insights into the effectiveness of the treatment interventions.

DATAMODEL

The specification for the IDEA4RC datamodel is available in Appendix A. We present all the variables and the most remarkable columns related to each entity from the ERD. For ease of reading, we proceed to describe each column:

Table 6 - Columns from the datamodel.

Variable Name (EURACAN file)	DataElementConcept	DataElementConceptDefEN	FormatConceptualDomain	Required	ExpectedValue

- Variable name (EURACAN file): provides the name of the variable as it was given in the original file specified by clinicians for its modelling.

- DataElementConcept: provides the name for the variable without special characters and whitespaces, following the convention {entity}_{variable name}. This helps automatic processing tools to quickly identify the variable and its corresponding entity.
- DataElementConceptDefEN: describes the meaning and application of the variable in English.
- FormatConceptualDomain: specifies the type of variable in computer science terms. Possible values:
    - string
    - integer
    - date
    - Code: a code or set of codes from the OHDSI standard vocabularies (these are integer-like IDs)
    - CustomCode: a code or set of codes from OHDSI standard vocabularies and, or, **new codes** that need to be defined as they were not found in the OHDSI standard vocabularies.
- Required: specifies whether the variable is Mandatory (M), Recommended (R), or Optional (O).
- ExpectedValue: set of possible values identified by clinicians. These needs to be mapped to Codes within the OHDSI standard vocabularies when available.

Following these descriptions as already stated, the IDEA4RC data model is available in Appendix A. Nevertheless, the data model is ongoing work and future updates and changes will be applied.

FHIR IMPLEMENTATION GUIDE

The current iteration of the data model uses a Python script based on the Pandas library to clean and modify the original data model spreadsheet to a FSH[17]-friendly version. In this manner, **T3.1 FHIR Implementation guide (IG) following FAIR principles (Lead: HL7)**, has a fast and automatic workflow for building a draft FHIR Implementation Guide. Afterwards, colleagues from HL7 review and update the final IG.

[17] https://fshschool.org/docs/sushi/

The current version of the IG is accessible[18], and will be updated in the future. Both partners, University of Deusto and HL7 are having ad hoc meetings to improve this workflow and obtain best results.

OMOP MAPPING

The process of mapping the IDEA4RC data to the OMOP CDM involves collaborating with the BLUEBERRY project, which focuses on mapping sarcoma registries to the OMOP CDM. In addition to sarcomas, IDEA4RC aims to integrate the mapping for Head and Neck cancers. The mapping process entails on the one side identifying how each variable in the IDEA4RC project corresponds to specific concepts in the OMOP CDM (semantic mapping). Furthermore, all possible values for a given variable need to be mapped to their respective OHDSI standard vocabularies codes to ensure standardized representation. This mapping process helps ensure that the data from IDEA4RC aligns with the OMOP CDM semantics and facilitates seamless integration and analysis. On the other side, there is the actual ETL which determines where the rare cancers information ends up in the tables of the OMOP CDM (structural mapping). Combining the two and implementing it will ensure that the data already converted in FHIR capsules is also converted into OMOP tables and standard concepts (to the extent possible within the model).

[18] https://build.fhir.org/ig/hl7-eu/idea4rc/index.html

6. IDEA4RC ADOPTED METADATA MODEL V1

Through the initial version of the IDEA4RC Metadata Model, we aim to accomplish several objectives. This metadata model comprises three primary components: quality, findability, and governance metadata. In its development, we have drawn upon the existing works of researchers, as cited in the state-of-the-art review section, to inform our approach.

The primary goal of the metadata model is to establish a structured framework for capturing and managing metadata related to data quality, findability, and governance within the IDEA4RC project. By incorporating these metadata components, we aim to enhance the overall understanding and utilization of shared health data.

It is important to note that the metadata model will undergo further iterations and refinements as the entire data models adoption evolves. As the project progresses and new insights emerge, we anticipate the need for iterative enhancements to the metadata model to align with the evolving requirements and advancements in the field.

Furthermore, it is crucial to emphasize that each data source within the Clinical Centers of Excellence (COEs) may require domain-specific quality indicators. As the metadata model caters to the unique data sources of each COE, we will consider the specific quality dimensions relevant to the respective domains. This tailored approach ensures that the metadata model accurately reflects the data quality considerations and requirements specific to each data source, facilitating robust and contextually relevant analyses.

By adhering to this metadata model, we strive to establish a comprehensive and adaptable framework that enables efficient management, assessment, and utilization of metadata within the IDEA4RC project. The ongoing evolution of the metadata model will ensure its alignment with the evolving project needs, while domain-specific quality indicators will enhance the accuracy and relevance of data analyses within the individual COEs.

6.1 Data quality: Data quality taxonomy at variable, data source, cohort and federated levels.

In the section we are going to present the metadata model for data quality developed within IDEA4RC, our approach is founded upon widely accepted and validated reliability quality dimensions or categories identified by previous research works and the bioinformatics community (most of the analysed research works rely on these primary categories in one way or another). However, considering the federated nature of the project and the requirement of establishing findability mechanisms using cohort characteristics, we have made the decision to expand the existing taxonomies with additional metadata related to data quality. This expansion involves adding new dimensions such as relevance (already proposed by the EMA), different hierarchical levels within which these quality checks will be executed and analyse the feasibility of adding different qualitative categories to rank the quality of the data depending on some predefined criteria.

First, we will define the different hierarchical levels on which the quality checks will be applied. This way, a continuous status of the quality of the different layers of the IDEA4RC ecosystem will be available. Four different hierarchical levels have been defined:

- Variable: at this level, the quality metadata describes the variable's quality through different quality metrics.
- Data source: quality checks performed at the data source level offer valuable insights into the quality of the different data sources that will be utilized to populate the capsules within each of the CoEs. This metadata plays a crucial role in assisting the CoEs in assessing and evaluating the quality of their respective data sources, providing data managers potential areas for improvement.
- Dataset: the quality metadata associated with the dataset level provides a comprehensive description of the quality of the IDEA4RC dataset within each of the capsules present in the CoEs. This dataset is created through an ETL process, utilizing data from the individual data sources of each center.
- Federated: a federated quality metadata component will be established to encompass the collective datasets from each of the centers within the CoEs. This federated quality

metadata will provide a comprehensive assessment of the overall data quality across all datasets, taking into account the combined information from the individual centres. It will allow for an aggregated evaluation of the data quality.

6.1.1 Reliability quality checks

Regarding the reliability metadata that is going to be available in the IDEA4RC ecosystem, as it has been mentioned earlier, the following reliability dimensions will be used.

Completeness. Refers to the extent to which all required and expected data elements or values are present. This quality metric will be available for each of the variables defined in the IDEA4RC CDM. Later, the values will be aggregated to have completeness metrics in the different hierarchical levels.

Table 7 - Description of completeness quality metrics.

Description	Dimension
Site of previous cancer is not specified	Completeness
Year of diagnosis is not specified	Completeness
M site is missing	Completeness
Radiotherapy metastatic site is not specified	Completeness
Age at diagnosis is not specified	Completeness

Conformance. Analyses to the degree to which data values adhere to specified standards, formats, or constraints. It encompasses the evaluation of whether the data values align with the predefined expectations and guidelines for each variable defined within the IDEA4RC CDM (CDM). Furthermore, these conformance metrics will be aggregated and analysed at different hierarchical levels to provide a comprehensive assessment of conformance across the IDEA4RC project. This allows for the evaluation of data conformance not only at the variable level but also at higher levels, such as within data sources, cohorts, or the federated dataset. By aggregating the conformance metrics, we gain insights into the overall adherence of the data

to the specified standards, facilitating a comprehensive understanding of the extent to which the data conforms to the predefined requirements and expectations within each hierarchical level.

Table 8 - Description of conformance quality metrics.

Description	Dimension
Age at diagnosis (automatic) is under 18	Conformance-Value-Verification
Surgery on M performed but clinical or pathological staging is not metastatic	Conformance-Value-Verification
Number of cycles/administrations is greater than 10	Conformance-Value-Verification
Radiotherapy intent is curative but total dose is lower or equal to 40	Conformance-Value-Verification
Treatment site (Distant Metastasis) is not consistent with cM, pM, clinical and pathological staging and site	Conformance-Value-Verification

Temporal plausibility. Refers to the assessment of whether the observed data values and their temporal distributions align with logical expectations and known temporal properties. It involves evaluating if the values of time-varying variables change as expected over time, based on established patterns, clinical knowledge, or external references. It helps identify potential anomalies, inconsistencies, or irregularities in the temporal characteristics of the data, enabling researchers and clinicians to assess the reliability and validity of the temporal information captured within the dataset. To have temporal plausibility values in the different hierarchical levels, aggregated metrics will be used.

Table 9 - Description of possible temporal plausibility quality metrics.

Description	Dimension
Radiotherapy year is prior to previous cancer diagnosis	Temporal plausibility - State transitions.

Description	Dimension
Chemotherapy year is prior to previous cancer diagnosis	Temporal plausibility - State transitions.
Surgery year is prior to previous cancer diagnosis	Temporal plausibility - State transitions.
Radiotherapy intent is Date of diagnosis is prior to year of oncological surgery	Temporal plausibility - State transitions.
Date of diagnosis is prior to year of previous cancer	Temporal plausibility - State transitions.
Date of surgery is not posterior (max 9 months) to date of diagnosis	Temporal plausibility - State transitions.
Date of surgery is not posterior to year of previous surgery	Temporal plausibility - State transitions.
Date of neck surgery is not posterior (max 9 months) to date of diagnosis	Temporal plausibility - State transitions.
Date of neck surgery is not posterior to year of previous surgery	Temporal plausibility - State transitions.
Difference between date of unplanned surgery and date of first surgery is greater than 90 days	Temporal plausibility - Temporal dependencies
Date of unplanned surgery is not posterior to date of first surgery	Temporal plausibility - State transitions.
Start date of systemic treatment is not within 9 months from date of diagnosis	Temporal plausibility - Temporal dependencies
Neo-adjuvant systemic treatment: start date is not prior to surgery	Temporal

Description	Dimension
or radiotherapy	plausibility - State transitions.
Adjuvant systemic treatment: start date is not posterior to surgery or radiotherapy	Temporal plausibility - State transitions
Start date of systemic treatment is not posterior to date of previous systemic treatment	Temporal plausibility - State transitions.
End date of systemic treatment is not posterior to start date of systemic treatment	Temporal plausibility - State transitions.

Atemporal plausibility. Involves assessing whether observed data values and distributions align with local or common knowledge in a non-temporal context. It examines if the data values, regardless of their temporal aspects, agree with logical expectations and known patterns within a specific healthcare context. Through atemporal plausibility checks, the IDEA4RC project ensures that the non-temporal aspects of the health data align with established knowledge and expectations.

Table 10 - Description of possible atemporal plausibility quality metrics.

Description	Dimension
Number of positive lymph nodes (right side) is greater than those remove	Atemporal plausibility - Verification
Number of positive lymph nodes (left side) is greater than those removed	Atemporal plausibility - Verification
pT, pN and pM and pathological staging values are not consistent	Atemporal plausibility - Verification.
Site of surgery on metastasis is not consistent with M site	Atemporal plausibility - Verification

Description	Dimension
Radiotherapy settings are preoperative or postoperative, but surgery is not performed	Temporal plausibility - State transitions.
Radiotherapy settings are preoperative or postoperative concomitant to systemic treatment, but systemic treatment is not performed	Atemporal plausibility - Verification
Radiotherapy intent in progression/recurrence/persistent disease is curative but total dose and/or total high dose is lower or equal to 40	Atemporal plausibility - Verification
Treatment site is Distant Metastasis but patient is not metastatic	Atemporal plausibility - Verification
Adverse event occurrence and related therapy are not consistent with previous data	Atemporal plausibility - Verification

6.1.2 Relevance quality checks

First, it is important to acknowledge that due to the time constraints of the project, it may not be feasible to perform all the relevance quality checks outlined in the EMA's guidelines. However, we have taken a strategic approach by selecting a set of core relevance quality checks that we believe are not only significant from a quality perspective but also play a crucial role in enhancing data findability (see Table 11). These selected checks offer valuable insights into essential characteristics of the dataset, enabling us to assess its relevance effectively.

Table 11 - Initial proposal for relevance quality checks.

Description	Dimension
Data Source	Data source countries
	Data custodian
	Date when the data source was first established
	Data source type
Data elements	Is sociodemographic information available?
	Are lifestyle factors included?
Quantitative descriptors	Population size
	Population size by age
	Median time between first and last available records for unique individuals captured
	Is the data ETL-ed to a CDM?
	Data sourced last refresh

Nevertheless, it is essential to emphasize that the list of relevance quality checks we have compiled is not exhaustive and subject to evolution as the Centers of Excellence (CoEs) further clarify their specific needs and requirements. As the project progresses and we gain a deeper understanding of the data landscape and user needs, additional relevance quality checks may be identified and incorporated into our assessment process.

7. CONCLUSIONS AND FUTURE WORK

In this deliverable, we present the initial release of our comprehensive metadata and data models investigation, in which we recognize their intrinsic relationship and mutual impact on each other. To lay the foundation for our work, we conducted a review of existing models commonly used in the oncological domain, including OMOP, FHIR, DICOM, OpenEHR, SNOMED-CT, and findings from the OSIRIS project, among others. This analysis allowed us to extract valuable insights into the strengths and limitations of each model.

Additionally, we explored metadata taxonomies that could be effectively applied to these data models. We focused on three essential categories of metadata: data quality, data governance, and data findability. By examining the state of the art in these domains, we gained valuable knowledge that has guided our own metadata model design. We also considered the FAIR principles, understanding how their implementation can enhance data and metadata models. With the FAIR principles as a guiding framework, we ensure that our models promote data accessibility, interoperability, and reusability.

Having examined the state of the art, we described the first version of the adopted CDMs for the IDEA4RC project, which specifically targets H&N Cancers. The subsequent version of the model will incorporate the rare cancers data. Our methodology for designing the model has been documented, including the Entity-Relationship Diagram (ERD) design, allowing for a comprehensive understanding of the items relationships within the H&N data. In-depth discussions and specifications of the model, including variable descriptions and requirement levels, have been added in Appendix A.

Simultaneously, we introduced the initial iteration of the metadata model, concentrating on data quality. Recognizing its critical role in the ETL (extract, transform, load) process used to populate the FHIR capsules, we focused on data quality metrics and checks to calculate the metadata. Various context levels will be incorporated into this metadata model, further enriching its capabilities.

Looking forward, upcoming versions of this deliverable will encompass the CDMs adoption for rare cancers. This extension will include the incorporation of relevant variables, making the

model even more comprehensive and applicable. Additionally, we will also include data governance and data findability metadata, completing the trio of essential metadata categories.

With this ongoing work, we aim to establish a robust and adaptable framework that optimizes data management in the health domain, contributing to enhanced research and clinical outcomes in the field of oncology.

This project has received funding from the European Union's Horizon Europe research and innovation programme under grant agreement no. 101057048

ANNEX A

HospitalData:

Variable Name (EURACAN file)	DataElementConcept	DataElementConceptDefEN	FormatConceptualDomain	Required	ExpectedValue
Hospital name	HospitalData_hospitalName	Hospital where the patients is included in the registry	String	M	Text

This project has received funding from the European Union's Horizon Europe research and innovation programme under grant agreement no. 101057048

HospitalPatientRecords:

Variable Name (EURACAN file)	DataElementConcept	DataElementConceptDefEN	FormatConceptual Domain	Required	ExpectedValue
Patient	HospitalPatientRecords_patient	Patient element containing the data regarding the patient followed by the hospital	ElementReference	M	A patient element data
Hospital	HospitalPatientRecords_hospital	Hospital element containing the data regarding the patient followed by the hospital	ElementReference	M	A hospital element data
Date of first contact with the hospital	HospitalPatientRecords_firstContactDate	Date of the first contact of the patient with the hospital registering the data. The hospital will record information on the patient's entire disease trajectory, thus also on procedures and/or treatments performed in another hospital. The "date of first contact" will be crossed with other dates to better understand which parts of the disease path were managed by the hospital that registered the patient.	Date	M	Date

This project has received funding from the European Union's Horizon Europe research and innovation programme under grant agreement no. 101057048

Patient:

Variable Name (EURACAN file)	DataElementConcept	DataElementConceptDefEN	FormatConceptualDomain	Required	ExpectedValue
Sex	Patient_Sex	Describes biological sex as recorded in the patient's identity document or in the hospital record. In the absence of documentation, the one declared by the patient will be recorded	Code	M	Male; Female; Unknown.
Race	Patient_Race	Describes race as recorded in the hospital record, the one declared by the patient,, otherwise, the onerecognized by the observer	Code	M	Unknown; White; Black; Asians/Pacific Islanders; American Indian/Alaska Native
Country of Residence	Patient_CountryOfResidence	Country of residence at the time of diagnosis	Code	M	Value from the code list of countries
Smoking	Patient_Smoking	Describes tobacco smoker habits within the options proposed	CustomCode	M	Current tobacco smoker; Former smoker (at least for 12 months); Never smoker; Unknown
Smoking type	Patient_SmokingType	Describes type of tobacco	Code	O	Cigarettes; Cigar; Unknown
Cigarettes/cigars smoked per day	Patient_CigSmokedPerDay	Number of cigarettes or cigars smoked in one day. Together with the information of number of years as a smoker, these information will allow to automatically calculate the pack year.	Integer	R	numeric
Number of years as a smoker	Patient_YearsAsSmoker	Number of years the person has smoked	Integer	R	numeric

This project has received funding from the European Union's Horizon Europe research and innovation programme under grant agreement no. 101057048

Alcohol	Patient_Alcohol	Describes alcohol habits within the options proposed	CustomCode	M	Current; Former (at least for 12 months); Never; History of alcohol dependence; Unknown
Comorbidity	Patient_Comorbidity	Describes wherther the patient was diagnosed before treatment of at least one of the comorbidities listed next or not	Code	M	Yes; No; Unknown
Myocardial infarction	Patient_MyocardialInfarction	Describes comorbidities reported or assesed before treatment. More than one choice is allowed. Please do not include the current cancer in this calculation, only the previous cancer.	Boolean	O (only if ACE-27 variable is ADDED)	
Congestive heart failure	Patient_CongestiveHeartFailure		Boolean	O (only if ACE-27 variable is ADDED)	
Peripheral vascular disease	Patient_PeripheralVascularDisease		Boolean	O (only if ACE-27 variable is ADDED)	
Cerebrovascular accident (except hemiplegia)	Patient_CerebrovascularAccident(ExceptHemiplegia)		Boolean	O (only if ACE-27 variable is ADDED)	
Dementia	Patient_Dementia		Boolean	O (only if ACE-27 variable is ADDED)	

This project has received funding from the European Union's Horizon Europe research and innovation programme under grant agreement no. 101057048

Chronic pulmonary disease	Patient_ChronicPulmonaryDisease		Boolean	O (only if ACE-27 variable is ADDED)	
Connective tissue disease	Patient_ConnectiveTissueDisease		Boolean	O (only if ACE-27 variable is ADDED)	
Ulcer	Patient_Ulcer		Boolean	O (only if ACE-27 variable is ADDED)	
Mild liver disease	Patient_MildLiverDisease		Boolean	O (only if ACE-27 variable is ADDED)	
Moderate to severe liver disease	Patient_ModerateToSevereLiverDisease		Boolean	O (only if ACE-27 variable is ADDED)	
Diabetes (without complications)	Patient_Diabetes(WithoutComplications)		Boolean	O (only if ACE-27 variable is ADDED)	
Diabetes with end organ damage	Patient_DiabetesWithEndOrganDamage		Boolean	O (only if ACE-27 variable is ADDED)	

Hemiplegia	Patient_Hemiplegia		Boolean	O (only if ACE-27 variable is ADDED)	
Moderate to severe renal disease	Patient_ModerateToSevereRenalDisease		Boolean	O (only if ACE-27 variable is ADDED)	
Solid tumor (non metastatic)	Patient_SolidTumor(NonMetastatic)		Boolean	O (only if ACE-27 variable is ADDED)	
Metastatic solid tumor	Patient_MetastaticSolidTumor		Boolean	O (only if ACE-27 variable is ADDED)	
Leukemia	Patient_Leukemia		Boolean	O (only if ACE-27 variable is ADDED)	
Lymphoma, Multiple myeloma	Patient_Lymphoma,MultipleMyeloma		Boolean	O (only if ACE-27 variable is ADDED)	
AIDS	Patient_Aids		Boolean	O (only if ACE-27 variable is ADDED)	

This project has received funding from the European Union's Horizon Europe research and innovation programme under grant agreement no. 101057048

Eastern Cooperative Oncology Group performance status (ECOG PS) at diagnosis	Patient_EasternCooperativeOncologyGroupPerformanceStatus(EcogPs)AtDiagnosis	Eastern Cooperative Oncology Group performance status (ECOG PS) at diagnosis	Code	M (at least one of the two)	numeric; only if already available at the health care provider level
Karnofsy index at diagnosis	Patient_KarnofsyIndexAtDiagnosis	Karnofsy index at diagnosis	Code	M (at least one of the two)	numeric; only if already available at the health care provider level

This project has received funding from the European Union's Horizon Europe research and innovation programme under grant agreement no. 101057048

PatientFollowUp:

Variable Name (EURACAN file)	DataElementConcept	DataElementConceptDefEN	FormatConceptualDomain	Required	ExpectedValue
Patient	PatientFollowUp_Patient	Patient element	Patient	M	Patient object or id (see Patient sheet)
Status at last follow-up	PatientFollowUp_StatusAtLastFollow-Up	Describes the status at last follow-up	CustomCode	M	Alive, No Evidence of Disease (NED); Dead of Disease (DOD); Dead of Other Cause (DOC); Dead of Unknown Cause (DUC) ; Alive With Disease (AWD)
Patient Follow Up date	PatientFollowUp_PatientFollowUpDate	Date of the clinical follow-up	Date	M	YYYY-MM-DD
New cancer diagnosis	PatientFollowUp_NewCancerDiagnosis	identifies whether the patient has developed a subsequent primary cancer	CustomCode	M	Yes; No; Unknown
Date of new cancer diagnosis	PatientFollowUp_DateOfNewCancerDiagnosis	date of subsequent primary cancer diagnosis	Date	M	YYYY-MM-DD
New cancer topography	PatientFollowUp_NewCancerTopography	clarifies the site of the subsequent primary cancer (from a predefined list of sites)	CustomCode	M	TBD

This project has received funding from the European Union's Horizon Europe research and innovation programme under grant agreement no. 101057048

CancerEpisode:

Variable Name (EURACAN file)	DataElementConcept	DataElementConcept DefEN	FormatConceptualDomain	Required	ExpectedValue
Patient	CancerEpisode_Patient	Patient element containing the data regarding the patient followed by the hospital	ElementReference	M	A patient element data
Date of diagnosis (biopsy or surgical piece)	CancerEpisode_DateOfDiagnosis(BiopsyOrSurgicalPiece)	Date of the procedure from which the specimen was obtained that allowed the histological diagnosis.	Date	M	YYYY-MM-DD
Biopsy done by	CancerEpisode_BiopsyDoneBy	Describes the institution where diagnostic procedure was performed	CustomCode	M	The hospital; A different hospital
Age at diagnosis	CancerEpisode_AgeAtDiagnosis	Age of the patient at the time of the diagnosis.	Integer	M	Whole number greater than 0
Grading	CancerEpisode_Grading	Grading of the cancer	Code	O	Check this

Histology (WHO 2017) group	CancerEpisode_Histology(Who2017)Group	Histology of prymary tumour according to WHO 2017 clasification.	Code	M	WHO 2017 H&N Classification Squamous; Adenocarcinoma; Neuroendocrine;Adenosquamous carcinoma; Teratocarcinosarcoma; NUT carcinoma; HPV-related Multiphenotypic; Olfactory neuroblastoma (esthesioneuroblastoma, esthesioneurocytoma, esthesioneuroepithelioma, Olfactory placode tumor); Odontogenic carcinoma; Sinonasal undifferentiated Carcinoma(SNUC);Carcinoma /Carcinoma undifferentiated
Histology (WHO 2017) subgroup					

Histology subgroup Squamous	CUS_HistologySquamous	Specifies the histological subgroup for squamous cancers	String	M	Keratinizing squamous cell carcinoma, epidermoid carcinoma; Non-keratinizing squamous cell carcinoma; Non-keratinizing squamous cell carcinoma: SMARCB1 (INI-1)-deficient Sinonasal Carcinoma ;Non-keratinizing squamous cell carcinoma: Transitional (cylindrical cell, Schneiderian) carcinoma ;Spindle cell (sarcomatoid) squamous cell carcinoma;Spindle cell (sarcomatoid) squamous cell carcinoma: SMARCB1 (INI-1)-deficient Sinonasal Carcinoma ;Lymphoepithelial carcinoma, lymphoepithelioma like carcinoma;Basaloid squamous cell carcinoma;Squamous cell carcinoma: conventional, NOS, clear cell, microinvasive, adenoid, acantholytic, pseudoglandular, giant

					cell ;Verrucous squamous cell carcinoma: NOS, cuniculatum carcinoma/Ackerman tumor; Papillary squamous cell carcinoma; Squamous cell carcinoma; Squamous cell carcinoma: HPV-positive; Squamous cell carcinoma: HPV-negative

Histology subgroup Adenocarcinoma	CUS_HistologyAdenocarcinoma	Specifies the histological subgroup for adenocarcinomas	String	M	Intestinal-type (sinonasal) adenocarcinoma; NOS, non–intestinal-type (sinonasal), Endolymphatic sac low grade, Intestinal-type (salivary gland), cystoadenocarcinoma, mucinous, Ceruminous (only in ear); Nasopharyngeal papillary adenocarcinoma, thyroid like low grade nasopharingeal papillary adenocarcinoma; Adenoid cystic carcinoma; Adenoid cystic carcinoma: solid type (> 30% solid); Mucoepidermoid carcinoma; Polymorphous, Cribriform of minor salivary glands, Polymorphous (low grade), terminal duct carcinoma, lobular carcinoma; Acinic cell carcinoma;Clear cell carcinoma, hyalinising clear cell carcinoma; Basal cell adenocarcinoma, malignant dermal analog tumor;Salivary duct carcinoma, high grade

This project has received funding from the European Union's Horizon Europe research and innovation programme under grant agreement no. 101057048

					ductal carcinoma;Salivary secretory adenocarcinoma (mammary analog, MASC); Secretory carcinoma; Myoepithelial carcinoma, malignant myoepithelioma; Carcinoma ex pleomorphic adenoma: NOS, Intracapsular, minimally invasive, largely invasive; Sebaceous adenocarcinoma, Sebaceous lymphadenocarcinoma; Carcinosarcoma; Oncocytic carcinoma, oxyphilic carcinoma, oncocytic adenocarcinoma, oncocytic malignant oncocytoma; Salivary gland intraductal carcinoma (cribriform low grade adenocarcinoma)

Histology subgroup Neuroendocrine	CUS_HistologyNeuroendocrine	Specifies the histological subgroup for neuroendocrine cancers	String	M	Small cell neuroendocrine carcinoma (SmCC),Poorly differentiated neuroendocrine carcinoma, small cell (grade 3);Large cell neuroendocrine carcinoma (LCNEC), Poorly differentiated neuroendocrine carcinoma, large cell (grade 3);Well-differentiated neuroendocrine carcinoma, Middle ear carcinoid tumor;Moderately differentiated neuroendocrine carcinoma
Histology subgroup Odontogenic carcinoma	CUS_HistologyOdontogenicCarcinoma	Specifies the histological subgroup for odontogenic carcinomas	String	M	Odontogenic carcinoma, NOS, Ameloblastic carcinoma (primary, secondary intraosseous, secondary peripheral), Primary intraosseous carcinoma, Intraosseous carcinoma developed on odontogenic cyst, sclerosing odontogenic carcinoma; Clear cell odontogenic carcinoma; Gosht cell odontogenic carcinoma

Histology subgroup Sinonasal undifferentiated carcinoma (SNUC)	CUS_HistologySNUC	Specifies the histological subgroup for sinonasl undifferentiated carcinomas	String	M	SMARCB1 (INI-1)-deficient Sinonasal undifferentiated Carcinoma;Sinonasal SMARCA4 deficient carcinoma;IDH2-mutated sinonasal undifferentiated neoplasm
Subsite					
Nasal cavity and paranasal sinuses subsite	CUS_NasalCavityAndParanasalSinusesSubsite	Specifies the subsite for cancers occurred in nasal cavity and paranasal sinusess	String	M	AJCC 8th Edition Cancer Staging Manual Nasal cavity;Maxillary sinus;Ethmoid sinus;Frontal sinus;Sphenoid sinus
Nasopharynx subsite	CUS_NasopharynxSubsite	Specifies the subsite for cancers occurred in nasopharynx	String	M	AJCC 8th Edition Cancer Staging Manual Superior wall of nasopharynx;Posterior wall of nasopharynx;Lateral wall of nasopharynx;Anterior wall of nasopharynx

Hypopharynx subsite	CUS_HypopharynxSubsite	Specifies the subsite for cancers occurred in hypopharynx	String	M	AJCC 8th Edition Cancer Staging Manual Postcricoid region; Hypopharyngeal aspect of aryepiglottic fold; Posterior wall of hypopharynx;Pyriform sinus
Oropharynx subsite	CUS_OropharynxSubsite	Specifies the subsite for cancers occurred in oropharynx	String	M	https://link.springer.com/book/9783319406176 Base of tongue, NOS; Soft palate NOS (excludes Nasopharyngcal surface C11.3); Uvula; Tonsillar fossa; Lingual tonsil; Tonsillar pillar; Vallecula; Anterior surface of epiglottis; Lateral wall oropharynx; Posterior wall oropharynx; Branchial cleft (site of neoplosm);
Larynx subsite	CUS_LarynxSubsite	Specifies the subsite for cancers occurred in larynx	String	M	AJCC 8th Edition Cancer Staging Manual Glottis;Supraglottis;Subglottis;Laryngeal cartilage

Oral cavity subsite	CUS_OralCavitySubsite	Specifies the subsite for cancers occurred in oral cavity	String	M	AJCC 8th Edition Cancer Staging Manual Dorsal surface tongue, NOS; Border of tongue; Ventral surface of tongue NOS; Anterior 2/3 of tongue NOS; Upper gum; Lower gum; Anterior floor of mouth; Lateral floor of mouth; Overlapping lesion of floor of mouth; Hard palate; Cheek mucosa; Vestibule of mouth; Retromolar area; Overlapping lesion of other and unspecified parts of mouth;
Lip subsite	CUS_LipSubsite	Specifies the subsite for cancers occurred in lip	String	M	AJCC 8th Edition Cancer Staging Manual External lower lip;External upper lip;External lip, NOS;Mucosa of upper lip;Mucosa of lower lip;Mucosa of lip,

					NOS;Commissure of lip
Plasmatic EBV DNA at baseline	CUS_PlasmaticEbvDnaAtBaseline	Describes the result of EBV DNA plasma testing before treatment in NPC type II and III (WHO)	String	R	Positive; Negative; not tested;
HPV status	CUS_HpvStatus	Describes the result of HPV tumor testing in oral carcinoma	String	M for OROPHARYNGEAL (not oral cavity) carcinomas	Positive; Negative; Not tested;
CRP – C reactive protein tested	CUS_Crp–CReactiveProteinTested	Describes the result of C reactive protein testing	String	O	Positive; Negative; Not tested;

This project has received funding from the European Union's Horizon Europe research and innovation programme under grant agreement no. 101057048

EpisodeEvent:

Variable Name (EURACAN file)	DataElementConcept	DataElementConcept DefEN	FormatConceptualDomain	Required	ExpectedValue
Cancer Episode Reference	EpisodeEvent_CancerEpisodeReference	CancerEpisode element containing the data regarding the patient's cancer	ElementReference	M	CancerEpisode
Event type	EpisodeEvent_EventType	Type of event registered	CustomCode	M	Baseline; Progression; Recurrence; Persistent disease
Defined At	EpisodeEvent_DefinedAt	Whether or not the progression/recurrnce or persistent disease was performed at the registering hospital or another hospital.	CustomCode	O	the hospital; a different hospital
Date of episode	EpisodeEvent_DateOfEpisode	Start date of progression/recurrence or persistent disease	Date	O (M if not baseline)	YYYY-MM-DD
Is local	EpisodeEvent_IsLocal	Describes if the progression /recurrence or persistent disease is local	Code	O (M if not baseline)	Yes; No; Unknown
Is regional	EpisodeEvent_IsRegional	Describes if the progression /recurrence or persistent disease is regional	Code	O (M if not baseline)	Yes; No; Unknown

Is metastatic	EpisodeEvent_IsMetastatic	Describes if the progression /recurrence or persistent disease is metastatic	Code	O (M if not baseline)	Yes; No; Unknown
Site of metastasis_soft tissue	EpisodeEvent_SiteOfMetastasis_SoftTissue	Describes if site of metastasic disease is soft tissue	Boolean	O (M if not baseline)	True; False
Site of metastasis_distant lymph nodes	EpisodeEvent_SiteOfMetastasis_DistantLymphNodes	Describes if site of metastasic disease is dystan lymph node	Boolean	O (M if not baseline)	True; False
Site of metastasis_lung	EpisodeEvent_SiteOfMetastasis_Lung	Describes if site of metastasic disease is lung	Boolean	O (M if not baseline)	True; False
Site of metastasis_bone	EpisodeEvent_SiteOfMetastasis_Bone	Describes if site of metastasic disease is bone	Boolean	O (M if not baseline)	True; False
Site of metastasis_liver	EpisodeEvent_SiteOfMetastasis_Liver	Describes if site of metastasic disease is liver	Boolean	O (M if not baseline)	True; False
Site of metastasis_pleura	EpisodeEvent_SiteOfMetastasis_Pleura	Describes if site of metastasic disease is pleura	Boolean	O (M if not baseline)	True; False
Site of metastasis_peritoneum	EpisodeEvent_SiteOfMetastasis_Peritoneum	Describes if site of metastasic disease is peritoneum	Boolean	O (M if not baseline)	True; False
Site of metastasis_brain	EpisodeEvent_SiteOfMetastasis_Brain	Describes if site of metastasic disease is brain	Boolean	O (M if not baseline)	True; False
Site of metastasis_other viscera	EpisodeEvent_SiteOfMetastasis_OtherViscera	Describes if site of metastasic disease is other viscera	Boolean	O (M if not baseline)	True; False

Site of metastasis_unknown	EpisodeEvent_SiteOfMetastasis_Unknown	Describes if site of metastasic disease is unknown	Boolean	O (M if not baseline)	True; False

This project has received funding from the European Union's Horizon Europe research and innovation programme under grant agreement no. 101057048

GeneTestExpression:

Variable Name (EURACAN file)	DataElementConcept	DataElementConceptDefEN	FormatConceptualDomain	Required	ExpectedValue
Episode Event Reference	EpisodeEvent_CancerEpisodeReference	EpisodeEvent element containing the data regarding the patient's cancer	ElementReference	M	EpisodeEvent
Gene expression analysis performed	EpisodeEvent_CancerEpisodeReference	Clarifies whether a gene expression analysis is performed	Boolean	R	True; False
Date of Gene expression	EpisodeEvent_CancerEpisodeReference	Date of the gene expression analysis	Date	O	YYYY-MM-DD
Gene mutation analysis performed	EpisodeEvent_CancerEpisodeReference	Clarifies whether a gene mutation analysis is performed	Boolean	R	True; False
Date of Gene mutation	EpisodeEvent_CancerEpisodeReference	Date of the gene mutation analysis	Date	O	YYYY-MM-DD
Tests for chromosome translocations performed	EpisodeEvent_CancerEpisodeReference	Clarifies whether a tests for chromosome translocations is performed	Boolean	R	True; False
Date of traslocation	EpisodeEvent_CancerEpisodeReference	Date of the Chromosome translocation test	Date	O	YYYY-MM-DD
Next generation sequencing (NGS) performed	EpisodeEvent_CancerEpisodeReference	Clarifies whether a NGS analysis is performed	Boolean	R	True; False

This project has received funding from the European Union's Horizon Europe research and innovation programme under grant agreement no. 101057048

Date of NGS	EpisodeEvent_CancerEpisode Reference	Date of the NGS analysis	Date	O	YYYY-MM-DD
Polymerase chain reaction (PCR) test performed	EpisodeEvent_CancerEpisode Reference	Clarifies whether a PCR analysis is performed	Boolean	R	True; False
Date of PCR	EpisodeEvent_CancerEpisode Reference	Date of the PCR analysis	Date	O	YYYY-MM-DD
Immunohistochemistry performed	EpisodeEvent_CancerEpisode Reference	Clarifies whether a immunohiostochemestry analysis is performed	Boolean	R	True; False
Date of immunohistochemistry	EpisodeEvent_CancerEpisode Reference	Date of the immunohiostochemestry analysis	Date	O	YYYY-MM-DD
Circulating Tumour DNA (ctDNA) performed	EpisodeEvent_CancerEpisode Reference	Clarifies whether a ctDNA analysis is performed	Boolean	R	True; False
Date of ctDNA	EpisodeEvent_CancerEpisode Reference	Date of the ctDNA analysis	Date	O	YYYY-MM-DD

This project has received funding from the European Union's Horizon Europe research and innovation programme under grant agreement no. 101057048

ClinicalStage:

Variable Name (EURACAN file)	DataElementConcept	DataElementConceptDefEN	FormatConceptualDomain	Required	ExpectedValue
Episode Event Reference	ClinicalStage_EpisodeEventReference	EpisodeEvent element containing the data regarding the patient's cancer	ElementReference	M	
cT	ClinicalStage_Ct	Specifies the clinical T	Code	M	Tx; Tis; T0;T1;T2; T3; T4; T4a; T4b;
cN	ClinicalStage_Cn	Specifies the clinical N	Code	M	Nx; N0;N1;N2;N2a;N2b;N2c;N3; N3a;N3b;
Radiological Extra-nodal extension (rENE)	ClinicalStage_RadiologicalExtra-NodalExtension(Rene)	Describes the presence or absence of radiological signs of extracapsular extension, as defined in the AJCC 8th Ed	CustomCode	M	ENE-; ENE+;
cM	ClinicalStage_Cm	Specifies the clinical M	Code	M	M0; M1;

This project has received funding from the European Union's Horizon Europe research and innovation programme under grant agreement no. 101057048

Clinical staging	ClinicalStage_ClinicalStaging	Specifies the clinical TNM	Code	M	0;I;II;III;IV;IVA;IVB;IVC;
Ajcc edition	ClinicalStage_AjccEdition	Describe the edition of the AJCC used for staging	Code	M	8th,9th,10th,11th
Soft tissue	ClinicalStage_SoftTissue	Describes if site of metastasic disease is soft tissue	Boolean	R	True; False
distant lymph node	ClinicalStage_DistantLymphNode	Describes if site of metastasic disease is dystan lymph node	Boolean	R	True; False
lung	ClinicalStage_Lung	Describes if site of metastasic disease is lung	Boolean	R	True; False
bone	ClinicalStage_Bone	Describes if site of metastasic disease is bone	Boolean	R	True; False
liver	ClinicalStage_Liver	Describes if site of metastasic disease is liver	Boolean	R	True; False
pleura	ClinicalStage_Pleura	Describes if site of metastasic disease is pleura	Boolean	R	True; False
peritoneum	ClinicalStage_Peritoneum	Describes if site of metastasic disease is peritoneum	Boolean	R	True; False

brain	ClinicalStage_Brain	Describes if site of metastasic disease is brain	Boolean	R	True; False
other viscera	ClinicalStage_OtherViscera	Describes if site of metastasic disease is other viscera	Boolean	R	True; False
unknown	ClinicalStage_Unknown	Describes if site of metastasic disease is unknown	Boolean	R	True; False

This project has received funding from the European Union's Horizon Europe research and innovation programme under grant agreement no. 101057048

PathologicalStage:

Variable Name (EURACAN file)	DataElementConcept	DataElementConceptDEN	FormatConceptualDomain	Required	ExpectedValue
Episode Event Reference	PathologicalStage_EpisodeEventReference	EpisodeEvent element containing the data regarding the patient's cancer	ElementReference	M	
pT	PathologicalStage_Pt	Specifies the pathological T	Code	M (for patients receiving surgery for primary tumor)	Tx; Tis; T0;T1;T2; T3; T4; T4a; T4b; unknown
pN	PathologicalStage_Pn	Specifies the pathological N	Code	M (for patients receiving surgery for regional lymph nodes)	Nx; N0;N1;N2;N2a;N2b;N2c; N3; N3a;N3b; unknown
pM	PathologicalStage_Pm	Describes whether capsular extension is present by histopathologic examination or not	Code	M (for patients receiving surgery for regional lymph nodes)	M0; M1; unknown
Extranodal extension (ENE)	PathologicalStage_ExtranodalExtension(Ene)	Describes extent of extranodal extension. This number must be explicitly referred to in the pathological report, otherwise it is unknown	CustomCode	M (for patients receiving surgery for regional lymph nodes)	ENE-; ENE+; unknown.
Extranodal Extent	PathologicalStage_ExtranodalExtent	Describes whether a sentinel node procedure was performed or not. Sentinel lymph node biopsy is considered a diagnostic procedure, therefore, per se, the neck is not considered to have been	CustomCode	R	< 2mm; >=2mm; unknown

		treated if it does not lead to a neck dissection.			
Sentinel node	PathologicalStage_SentinelNode	Describes whether a protocoled and standardized en block resection of lymphatic tissue is performed or not	CustomCode	M	Yes; No; Unknown.
Neck dissection	PathologicalStage_NeckDissection	Specifies the pathological M	CustomCode	O	Yes; No; Unknown.
Pathological staging	PathologicalStage_PathologicalStaging	Specifies the pathological staging	Code	M	0;I;II;III;IV;IVA;IVB;IVC;Unknown
Ajcc edition	PathologicalStage_AjccEdition	Describe the edition of the AJCC used for staging	Code	M	8th,9th,10th,11th
Soft Tissue	PathologicalStage_SoftTissue	Describes if site of metastasic disease is soft tissue	Boolean	O	True; False
distant lymph node	PathologicalStage_DistantLymphNode	Describes if site of metastasic disease is dystan lymph node	Boolean	O	True; False
lung	PathologicalStage_Lung	Describes if site of metastasic disease is lung	Boolean	O	True; False
bone	PathologicalStage_Bone	Describes if site of metastasic disease is bone	Boolean	O	True; False

liver	PathologicalStage_Liver	Describes if site of metastasic disease is liver	Boolean	O	True; False
pleura	PathologicalStage_Pleura	Describes if site of metastasic disease is pleura	Boolean	O	True; False
peritoneum	PathologicalStage_Peritoneum	Describes if site of metastasic disease is peritoneum	Boolean	O	True; False
brain	PathologicalStage_Brain	Describes if site of metastasic disease is brain	Boolean	O	True; False
other viscera	PathologicalStage_Other Viscera	Describes if site of metastasic disease is other viscera	Boolean	O	True; False
unknown	PathologicalStage_Unknown	Describes if site of metastasic disease is unknown	Boolean	O	True; False

This project has received funding from the European Union's Horizon Europe research and innovation programme under grant agreement no. 101057048

SystemicTreatment:

Variable Name (EURACAN file)	DataElementConcept	DataElementConceptDefEN	FormatConceptualDomain	Required	ExpectedValue
Episode Event reference	SystemicTreatment_EpisodeEventReference	EpisodeEvent element containing the data regarding the patient's cancer	ElementReference	M	
type of systemic treatment	SystemicTreatment_TypeOfSystemicTreatment	Select the type of systemic treatment administered. It is possible to directly select the single treatment as appropriate.	CustomCode	M	Chemotherapy; Immunotherapy; Target therapy; Unknown
Intent	SystemicTreatment_Intent	Clarifies the reasons why systemic therapy is administered • Curative chemotherapy is chemotherapy administered with the goal of achieving a complete remission and preventing the recurrence of cancer. • Palliative chemotherapy refers to any chemotherapy administration that is not curative but administered simply to decrease tumor load and increase life expectancy. It has been defined also as "…treatment in circumstances where the impact of intervention is insufficient to result in major survival advantage, but does affect improvement in terms of tumor-related symptoms…"	Code	M	Palliative; Curative; Unknwon

Setting	SystemicTreatment_Setting	clarifies the context / how the therapy was administered alone or in conjunction with other treatments • Neoadjuvant: treatment given as a first step to shrink a tumor before the main treatment, which is usually surgery, is given. Examples of neoadjuvant therapy include chemotherapy, radiation therapy, and hormone therapy. It is a type of induction therapy. • Adjuvant: additional cancer treatment given after the primary treatment to lower the risk that the cancer will come back. Adjuvant therapy may include chemotherapy, radiation therapy, hormone therapy, targeted therapy, or biological therapy. • Concomitant/concurrent: A treatment that is given at the same time as another (es. Chemotherapy + radiotherapy).	Code	M	Neo-adjuvant; Concomitant; Adjuvant; Systemic tretament alone; Unknown;
Start date systemic treatment	SystemicTreatment_StartDateSystemicTreatment	Specifies when systemic treatment begins	Date	M	dd/mm/yyyy
End date systemic treatment	SystemicTreatment_EndDateSystemicTreatment	Specifies when systemic treatment ends	Date	M	dd/mm/yyyy
Number of cycles/ administrations	SystemicTreatment_NumberOfCycles/Administrations	clarifies how many times the treatment was administered. A cycle of treatment is a period of treatment followed by a period of rest (no treatment). For example, treatment given for one week followed by three weeks of rest is one cycle of treatment. A cycle can be repeated multiple times.	Float	O	numeric
Regimen	SystemicTreatment_				

This project has received funding from the European Union's Horizon Europe research and innovation programme under grant agreement no. 101057048

	Regimen				
Drugs 1	SystemicTreatment_Drugs1		Code	M	ATC list
Drugs 2	SystemicTreatment_Drugs2		Code	M	ATC list
Drugs 3	SystemicTreatment_Drugs3		Code	M	ATC list
Start date regimen changed	SystemicTreatment_StartDateRegimenChanged	specifies when the new systemic treatment begins, if a combination please specify the start of the first drug	Date	M	dd/mm/yyyy;
End date regimen changed	SystemicTreatment_EndDateRegimenChanged	specifies when the new systemic treatment ends,if a combination please specify the end of the last drug	Date	M	dd/mm/yyyy;
Reason for end of treatment	SystemicTreatment_ReasonForEndOfTreatment	Clarifies the reasons why the treatment ended or was interrupted	Code	M	Completion; Toxicity; Comorbidity; Patient intolerance; Patients decision; Death; Unknown.
Treatment response (based on imaging alone; no recist or other criteria)	SystemicTreatment_TreatmentResponse(BasedOnImagingAlone;NoRecistOrOtherCriteria)	Measures how well a cancer patient responds to treatment. RECIST criteria should not be applied. The definition of Complete response; Partial response; Stable disease; Progression, should be based on the clinical judgement based on imaging. Only when setting=neoadiuvant or palliative	Code	M	Complete response; Partial response; Stable disease; Progression; Unknown.

This project has received funding from the European Union's Horizon Europe research and innovation programme under grant agreement no. 101057048

Surgery:

Variable Name (EURACAN file)	DataElementConcept	DataElementConceptDefEN	FormatConceptual Domain	Required	ExpectedValue
Episode Event reference	Surgery_EpisodeEventReference	EpisodeEvent element containing the data regarding the patient's cancer	ElementReference	M	
Surgery	Surgery_Surgery	Whether or not a surgical procedure was performed and whether it was performed at the registering hospital or another hospital. Diagnostic procedures (biopsy) are not included.	CustomCode	M	Yes done at the hospital; Yes done at a different hospital; Not Done; Unknown.
Date of surgery	Surgery_DateOfSurgery	Date of the surgery for primary tumor with or without neck surgery	Date	M	dd/mm/yyyy
Surgery intention	Surgery_SurgeryIntention	Palliative: surgery performed with the intent of improving quality of life or relieving symptoms caused by advanced disease. Curative: surgery performed with the intend of oncologic cure, regardless of its result (R0 / R1/R2)	Code	M	Palliative; Curative; Unknown
Type of surgical approach on Tumour	Surgery_TypeOfSurgicalApproachOnTumour	Describes the approach to tumor resection whether it includes skin incision (extenal or open), or it 's aproached through a natural orifice (Transnasal/transorbital/transoral) or if the approach combines two or more of the previous ones	Code	M	endoscopic….

Margins after surgery	Surgery_MarginsAfterSurgery	The R0 ("no residual rumor") category applies only to cases in which residual tumor cannot be detected by conventional diagnostic methods. A more exact definition would read "no detectable residual tumor."This category corresponds to surgical resection for cure. The R1 category is reserved exclusively for cases in which residual tumor is found by histologic examination. This category may apply to biopsy sampling of the regional tissue at the site of resection or of a distant site at the time of surgery. It also applies to microscopic examination of the resection margins of the surgical resection specimen by the pathologist. R2 applies to cases with macroscopically visible residual tumor that is detected either clinically or pathologically.	Code	M	R0 (microscopic negative); R1 (microscopic positive); R2 (macroscopic positive); Unknown
Reconstruction	Surgery_Reconstruction	Local flap / regional pedicled flap / free flap	Code		Yes, no, unknown
Neck surgery	Surgery_NeckSurgery	Describes whether a surgical procedure to treat and address the neck was performed or not.	Code	M	Yes; No; Unknown.
Date of Neck surgey	Surgery_DateOfNeckSurgey	Date of the surgery on the neck	Date	M	

Laterality of the dissection	Surgery_LateralityOfTheDissection	Describes laterality of the neck surgical procedure: Ipsilateral when only the neck ipsilateral to the tumor has been treated Contralateral when only the neck contralateral to the tumor has been treated Bilateral: when both sides of the neck have been treated	Code	M	Ipsilateral; Bilateral; Controlateral; Unknown.
Surgery on M	Surgery_SurgeryOnM	Describes whether surgery is performed to treat the Metastasis	Code	M	Yes; No; Unknown.
Date of surgery on M	Surgery_DateOfSurgeryOnM	Date of the surgery on the metastasis	Date	M	dd/mm/yyyy
Site of surgery on metastasis_soft tissue	Surgery_SiteOfSurgeryOnMetastasis_SoftTissue	Describes if site of surgery on metastasis is soft tissue	Boolean	O	flag
Site of surgery on metastasis_distant lymph nodes	Surgery_SiteOfSurgeryOnMetastasis_DistantLymphNodes	Describes if site of surgery on metastasis is dystan lymph node	Boolean	O	flag
Site of surgery on metastasis_lung	Surgery_SiteOfSurgeryOnMetastasis_Lung	Describes if site of surgery on metastasis is lung	Boolean	O	flag
Site of surgery on metastasis_bone	Surgery_SiteOfSurgeryOnMetastasis_Bone	Describes if site of surgery on metastasis is bone	Boolean	O	flag
Site of surgery on metastasis_liver	Surgery_SiteOfSurgeryOnMetastasis_Liver	Describes if site of surgery on metastasis is liver	Boolean	O	flag
Site of surgery on metastasis_pleura	Surgery_SiteOfSurgeryOnMetastasis_Pleura	Describes if site of surgery on metastasis is pleura	Boolean	O	flag
Site of surgery on metastasis_peritoneu	Surgery_SiteOfSurgeryOnMetastasis_Perito	Describes if site of surgery on metastasis is peritoneum	Boolean	O	flag

This project has received funding from the European Union's Horizon Europe research and innovation programme under grant agreement no. 101057048

m	neum				
Site of surgery on metastasis_brain	Surgery_SiteOfSurgeryOnMetastasis_Brain	Describes if site of surgery on metastasis is brain	Boolean	O	flag
Site of surgery on metastasis_other viscera	Surgery_SiteOfSurgeryOnMetastasis_OtherViscera	Describes if site of surgery on metastasis is other viscera	Boolean	O	flag
Site of surgery on metastasis_unknown	Surgery_SiteOfSurgeryOnMetastasis_Unknown	Describes if site of surgery on metastasis is unknown	Boolean	O	flag
Surgical complications (Clavien-Dindo Classification)	Surgery_SurgicalComplications(Clavien-DindoClassification)	Describes presence and grade of complications after a surgical procedure,	Code	M	No complication; Grade I-V; unknown.

This project has received funding from the European Union's Horizon Europe research and innovation programme under grant agreement no. 101057048

Radiotherapy:

Variable Name (EURACAN file)	DataElementConcept	DataElementConceptDefEN	FormatConceptual Domain	Required	ExpectedValue
Episode Event reference	Radiotherapy_EpisodeEventReference	EpisodeEvent element containing the data regarding the patient's cancer	ElementReference	M	
Radiotherapy	Radiotherapy_Radiotherapy	Whether radiotherapy was delivered to a patient, either curatively or palliatively and whether it was performed at the registering hospital or another hospital.	CustomCode	M	Yes done at the hospital; Yes done at a different hospital; Not Done; Unknown.
Intent	Radiotherapy_Intent	Radiotherapy intent refers to whether the intention of treatment is to cure the patient or to treat symptoms and palliate	Code	M	Palliative; Curative; Unknwon
Setting	Radiotherapy_Setting	Whether radiotherapy is delivered as the main treatment modality (definitive) or if it is delivered before or after another treatment such as surgery	Code	M (only if "Intent=Curative OR Unknown")	Preoperative; Preoperative concomitant to systemic treatment; Postoperative;Postoperative concomitant to systemic treatment; Definitive; Definitive concomitant to systemic treatment; Unknown

Beam quality	Radiotherapy_BeamQuality	Describes the type of radiation therapy given. If external beam, please specify if delivered with Photons (most common), electrons, carbon, or protons.	Code	M	External beam RT Photons; External beam RT Electrons; External Beam RT Carbons; External Beam RT Protons; Brachytherapy interstitial endocavitary contact; Radionuclide therapy; Boron neutron capture Therapy; other; unknown
Other; specify	Radiotherapy_Other;Specify		String	O	Text
Treatment technique	Radiotherapy_TreatmentTechnique	Refers to the type of radiotherapy treatment delivered	Code	M	2D; 3D; IMRT CONVENTIONAL; VMAT; Tomotherapy; SBRT; FLASH THERAPY; PASSIVE SCATTERING; SINGLE BEAM OPTIMIZATION; IMPT; OTHER; unknown
Total Dose (TD) Gy	Radiotherapy_TotalDose(Td)Gy	Refers to the total dose delivered to the patient in Gy	Float	M	Number
Fraction Size (FS)	Radiotherapy_FractionSize(Fs)	Refers to the Dose per fraction delivered to the patient.	Float	M	Number
Number of fractions	Radiotherapy_NumberOfFractions	Refers to the total number of fractions delivered to the patient	Float	M	Number

Adaptive RT	Radiotherapy_AdaptiveRt	Refers to whether treatment planning was changed or adapted after the initial radiation plan was developed. This could be due to a change in the patient's anatomy or if the tumor changed in size.	Code	O	Yes; No; Unknown.
IGRT (image guide radiotherapy)	Radiotherapy_Igrt(ImageGuideRadiotherapy)	Refers to whether image guided radiotherapy was used for delivery of radiotherapy and to check the patient set up. This includes MV, KV, or Cone Beam CT imaging.	Code	O	Yes; No; Unknown.
Start date	Radiotherapy_StartDate	Date when the first radiation treatment was delivered	Date	M	dd/mm/yyyy
End date	Radiotherapy_EndDate	Date when the last radiation treatment ended	Date	M	dd/mm/yyyy
Treatment Sites:	Radiotherapy_TreatmentSites:	Refers to the areas that the radiation is targeting. This could include the primary tumor , the neck lymph nodes, the ipsilateral neck and the primary, the bilateral neck and the primary, or a distant metastatic lesion			
Primary	Radiotherapy_Primary		Boolean	M (suggest to modify the label into "Primary only")	flag
Neck	Radiotherapy_Neck		Boolean	M (suggest to modify the label into "Neck only")	flag
Primary and Ipsilateral Neck	Radiotherapy_PrimaryAndIpsilateralNeck		Boolean	M	flag

Primary and Bilateral Neck	Radiotherapy_PrimaryAndBilateralNeck		Boolean	M	flag
Distant Metastasis	Radiotherapy_DistantMetastasis		Boolean	M	flag
Metastatic Treatment Sites:	Radiotherapy_MetastaticTreatmentSites:	Designates which treatment sites were irradiated. Lung Vs Mediastinum Vs Bone Vs soft tissue vs liver Vs other.			
Lung	Radiotherapy_Lung		Boolean	R	flag
Mediastinum	Radiotherapy_Mediastinum		Boolean	R	flag
Bone	Radiotherapy_Bone		Boolean	R	flag
Soft Tissue	Radiotherapy_SoftTissue		Boolean	R	flag
Liver	Radiotherapy_Liver		Boolean	R	flag
Treatment Completed as Planned?	Radiotherapy_TreatmentCompletedAsPlanned?	Refers to whether patient completed all treatment as planned or if it had to be interrupted due to several reasons including toxicity, a co-morbidity preventing the delivery of radiation (pulmonary embolism, failure to thrive during RT), death due to progression of the cancer or patient decision	Code	M	Completion; Toxicity; Comorbidity; Patient intolerance; Patient decision; Death; Unknown.

This project has received funding from the European Union's Horizon Europe research and innovation programme under grant agreement no. 101057048

TreatmentResponse:

Variable Name (EURACAN file)	DataElementConcept	DataElementConceptDefEN	FormatConceptualDomain	Required	ExpectedValue
Cancer Episode reference	TreatmentResponse_CancerEpisodeReference	CancerEpisode element containing the data regarding the patient's cancer	CancerEpisode	M	
Treatment response (based on imaging alone; no recist or other criteria)	TreatmentResponse_TreatmentResponse(BasedOnImagingAlone;NoRecistOrOtherCriteria)	It refers to the response to the entire therapy administered to the patient. It measures how well a cancer patient responds to treatment. RECIST criteria should not be applied. The definition of Complete response; Partial response; Stable disease; Progression, should be based on the clinical judgement based on imaging.	Code	M	Complete response; partial response; stable disease; progression; unknown
Treatment response defined/done	TreatmentResponse_TreatmentResponseDefined/Done	refers to whether overall treatment reponse was assessed at the registering hospital or another.	CustomCode	M	At the hospital; At a different hospital

This project has received funding from the European Union's Horizon Europe research and innovation programme under grant agreement no. 101057048

AdverseEvent:

Variable Name (EURACAN file)	DataElementConcept	DataElementConceptDefEN	FormatConceptual Domain	Required	ExpectedValue
Episode Event reference	AdverseEvent_EpisodeEventReference	EpisodeEvent element containing the data regarding the patient's cancer	EpisodeEvent	M	
Adverse event type (CTCAE Term)	AdverseEvent_AdverseEventType(CtcaeTerm)	the Common Terminology Criteria for Adverse Events (CTCAE) is used to identify the adverse events. It includes details of the adverse event type and grade	Code	M	Sons of CTCAE grades
Occurred at	AdverseEvent_OccurredAt	specifies which phase (baseline, progression) of the disease the adverse event is related to		M	baseline; progression/recurrence /persistent disease from i=(1…10)
Adverse event related to	AdverseEvent_AdverseEventRelatedTo	specifies which treatment the adverse event is related to		M	Chemotherapy; Radiotherapy; Immunotherapy; Target therapy; Unknown
Adverse event starting date	AdverseEvent_AdverseEventStartingDate	specifies when adverse events begins	Date	M	dd/mm/yyyy
Adverse event duration	AdverseEvent_AdverseEventDuration	specifies the duration of the adverse event	CustomCode	M	Less than one week; More than one week but less than a month; More than a month but less than 3 months; More than 3 months; Unknown.

Adverse event	1/2/0999			M	