IDEA4RC

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

Deliverable 8.1

Rare Cancer Pilots selection

31 May 2023

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, W. 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	M. van Svieten. G. Geleijnse, E. Mezei
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	M. 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	M.L. Clementi, C. Sabelli
25 - UKE	S. Bauer, S. Lang, S. Mattheis, N. Midtank

Revision History

Revision	Date of Issue	Author(s)	Brief Description of Change
0	April 2, 2023	A. Trama (INT)	ToC
1	April 29, 2023	A. Trama	Revised ToC
2	May 1, 2023	A. Trama	First draft
3	May 9, 2023	E. Gaeta (UPM) E. Martinelli, A. Trama (INT) F. Mercalli (MME)	A description of the use cases was added together with the analytics required for major area. Added Abstract.
4	May 16, 2023	E. Martinelli (INT) F. Mercalli (MME)	Added contributions and revisions.
5	May 23, 2023	E. Martinelli (INT) A. Bilbao (UDEU)	Added contribution from UDEU
6	May 30, 2023	E. Gaeta, L. Lopez, I. Alonso (UPM) G. Geleijnse (IKNL) A. Trama, E. Martinelli (INT)	Added contributions
7	June 02, 2023	Coordinator	Final version

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 at the European Commission and appointed experts.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

Abbreviations and definitions

Abbreviation	Definition
DIGICORE	DIGital Institute for Cancer Outcomes REsearch
EBV	Epstein Barr virus
EMRs	Electronic medical records
ERN	European Reference Network
EURACAN	European reference network on rare adult solid cancers
FAIR	Findable Accessible Interoperable Reusable
H&N	Head and neck
HPV	Human Papilloma virus
INT	Fondazione IRCCS Istituto Nazionale dei Tumori, Milan
NLP	Natural Language Processing
OMOP	Observational Medical Outcomes Partnership is a public-private collaboration, chaired by the FDA, which designed the OMOP Common Data Model (CDM), a standardized model for storing Real World Data (RWD), developed to facilitate the generation of scientific evidence through large-scale observational studies.
OSIRIS	Interoperability and data sharing of clinical and biological data in oncology initiative
STS	Soft tissue sarcomas

EXECUTIVE SUMMARY

The main goal of IDEA4RC is to establish a data sharing ecosystem for the EURACAN ERN, which might become a reference for other research networks in healthcare, in particular for rare diseases as well as for cancers. This ambition needs to be sustained by and assessed through measures of quality, usefulness, usability and efficiency, collected in real use scenarios. This deliverable provides a description of the use cases selected by the EURACAN partners in the IDEA4RC consortium and proposes the criteria and performance indicators that will be used to assess the ecosystem.

The IDEA4RC Data Ecosystem will be experienced by the EURACAN centres participating in the project on at least 4 paradigmatic use cases addressing most relevant research questions and unmet needs of researchers and oncologists.

The identification of these candidate research questions is a result of joint works conducted in Task T8.1 and Task T2.1 with the contribution of all actors involved in research and healthcare of the rare cancers addressed by IDEA4RC, soft tissue sarcomas and head and neck cancers.

The four areas of research proposed are:

1. investigation of the natural history of the disease in particular for most challenging cases
2. research and validation of treatment outcome prediction and prognostic factors
3. investigation of diagnostic and/or treatment procedures effectiveness
4. assessment of quality of care.

For each of these areas some exemplary research questions have been proposed and described in chapter 2.3.

The execution of the pilot use cases will depend on data availability and quality across all the eleven participating EURACAN data providers. In this context within Task T8.1 we have identified the “core” datasets that will allow answering at least the most relevant research questions proposed by the pilot use cases. The datasets structure has been derived from ongoing works from other projects and initiatives - including but not limited to the EURACAN clinical registry. In line with the GDPR principles for “data minimization” the “core” datasets will include the most relevant variables. Their preliminary list is presented in Annexes 2 and 3 to this document.

As at present the amount and completeness of data that will be available within the IDEA4RC federated data ecosystem is not fully defined, the selection of the pilot use cases that will experienced in WP8 (Task T8.3) will be consolidated by task T9.2, when all the implementation environments will be defined for each pilot site and the data availability will be assessed.

The criteria and measurements used for the assessment and evaluation of the IDEA4RC ecosystem - in addition to data quality scores that are defined in task T2.4 as part of the metadata associated with each data source and data variable - are presented in chapter 4.

1 ABOUT THIS DOCUMENT

IDEA4RC aims to develop a federated ecosystem for rare adult solid cancers starting from 2 groups of these cancers: soft tissue sarcomas and head and neck cancers.

IDEA4RC has several target groups who have an unmet need regarding insights from real world data, including health professionals, health authorities, researchers, clinicians, patients, citizens. This document aims to describe the pilot data (re)use cases proposed by medical, radiation, surgical oncologists, and researchers to test the ability of the ecosystem to meet their needs. The needs of the other target groups of IDEA4RC are addressed by Task 2.1 and will be reported in Deliverable 2.1 (Data Ecosystem baseline value positions: value analysis and scenarios to guide following work). In addition to listing the data (re) use cases, this document describes the process used to identify them and the information needed to test them.

This document will be relevant for several project activities as follows (see Figure 1):

- the information needed will be used by Task 2.4 to develop the metadata taxonomy and by Task 3.1, 5.1 and 5.3 to map relevant structured and unstructured data;
- the data (re) use cases will contribute to define the federated algorithms needed to ensure data analysis in the distributed data sets (Task 4.3) as well as to align the multimodal interaction framework with the needs of the users of the system (Task 6.1, 6.4).

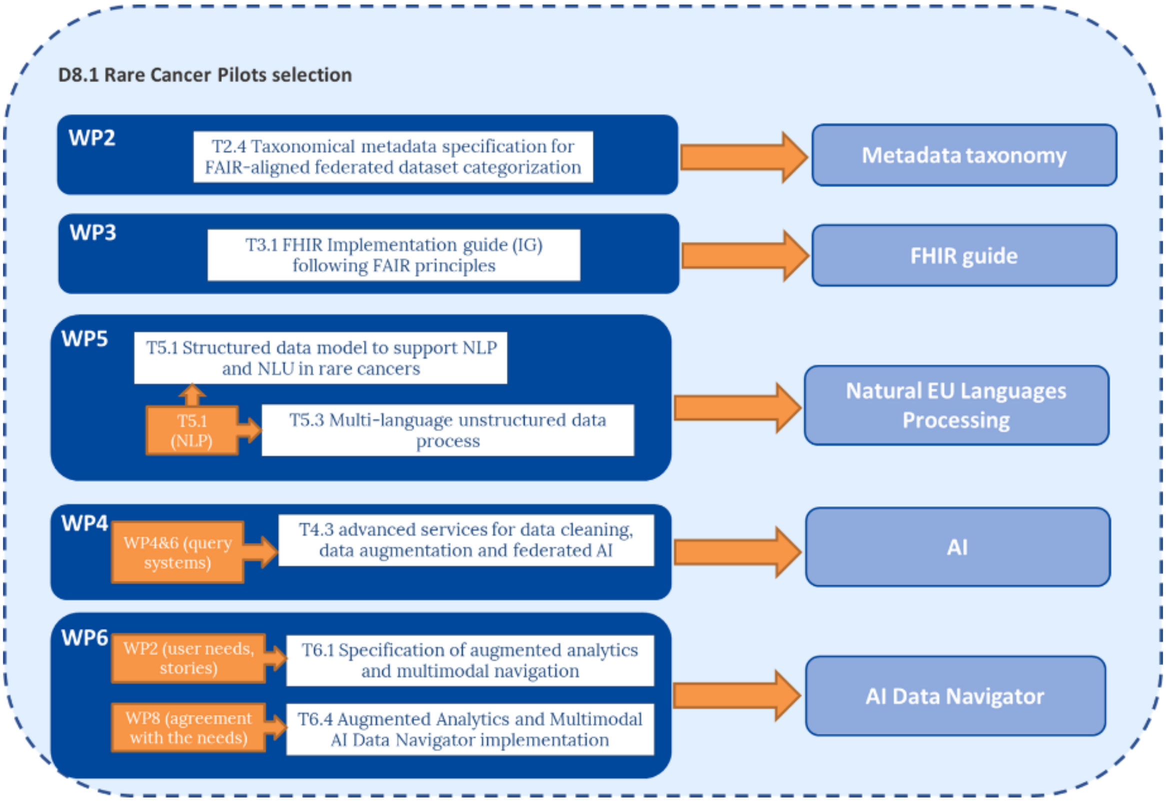
D8.1 Rare Cancer Pilots selection
WP2
T2.4 Taxonomical metadata specification for FAIR-aligned federated dataset categorization
Metadata taxonomy
WP3
T3.1 FHIR Implementation guide (IG) following FAIR principles
FHIR guide
WP5
T5.1 Structured data model to support NLP and NLU in rare cancers
T5.1 (NLP)
T5.3 Multi-language unstructured data process
Natural EU Languages Processing
WP4
WP4&6 (query systems)
T4.3 advanced services for data cleaning, data augmentation and federated AI
AI
WP6
WP2 (user needs, stories)
T6.1 Specification of augmented analytics and multimodal navigation
WP8 (agreement with the needs)
T6.4 Augmented Analytics and Multimodal AI Data Navigator implementation
AI Data Navigator

Figure 1 – Link among D8.1 and associated tasks

2 IDEA4RC DATA (RE) USE CASES

2.1 Introduction to rare cancers

Rare cancers are rare occurrences of a common disease affecting less than 6 per 100,000 individuals a year, approximately one in five new patients diagnosed with cancer. Rare adult cancers comprise a large number of different tumour types[1]. These hundreds of different rare cancer types may affect any of the body's organs, with varying clinical presentations.

Although there are different groups of cancers affecting different organs, all rare cancers share similar challenges in their management which includes:

a) diagnosis and clinical decision-making, due to a lack of available medical expertise and high-quality evidence from clinical research;
b) health care organization, due to difficulties in serving a territory with specialized facilities;
c) clinical research, due to the low number of patients and thus the difficulty to generate high-quality evidence from well powered clinical studies.

IDEA4RC focuses on 2 groups of rare cancers: soft tissue sarcomas (STS) and head and neck (H&N) cancers.

2.1.1 Soft tissue sarcomas

STS is a malignant neoplasm arising from mesenchymal cells. It can be split up into dozens of histological categories, and it can occur in virtually any anatomic site. This gives rise to a huge number of possible combinations of histology (cell type) and primary site which are of clinical importance. The anatomic site influences the therapeutic choice, in particular making surgery more or less viable or even impossible, but histology also influences prognosis and responsiveness to chemotherapy. During the past decade, we have seen the pendulum swinging from a one-size-fits-all treatment paradigm to a more histology-specific treatment recommendation, one that attempts to tailor not only the type and extent of oncologic resection to be performed but also the use and indication of multimodality therapy. This complex management paradigm, combined with the rarity and heterogeneity of the disease, highlights the importance of a multidisciplinary approach.

STS accounts for only 1% of all adult malignancies. As such, generating high-quality evidence for the management of STS is challenging. Despite progress in personalized treatments, the heterogeneity of these tumours has hindered the development of robust, evidence-based treatment strategies. Continued collaborative efforts will allow studies to be both sufficiently large and sufficiently focused to generate evidence that is clinically meaningful in specific STS patient populations.

[1] (https://www.rarecancerseurope.org/what-are-rare-cancers

2.1.2 Head and neck cancers

H&N cancers include cancers originating from the oral cavity, nasal cavity and sinuses, nasopharynx, salivary glands, pharynx, and larynx. Incidence shows large variations across Europe and between sexes. These differences reflect differences in the diffusion of the main risk factors: smoking, alcohol, viruses (HPV, EBV) and occupational exposures. Smoking and alcohol consumption are strong risk factors for larynx and oro-hypopharynx cancers, intestinal-type carcinomas of the nasal cavity and ethmoid cancers have a high attributable fraction due to occupational exposure to wood, leather, dusts, and formaldehyde. Nasopharynx carcinomas are related to EBV infection, while oropharynx carcinomas are related to HPV type 16 infection. Prognosis is very different depending on disease site, and in some cases aetiology (HPV-related cancers have better prognosis if appropriately treated).

Primary treatment varies with the anatomic site and stage of disease. For most early cancers, surgical resection is the cornerstone of treatment. However, for certain anatomic sites such as tonsils, base of tongue and floor of the mouth, as well as for all locally advanced cancers, radiotherapy is used, either alone or combined with surgery. Chemotherapy may be used in addition to radiotherapy. Nasopharynx carcinoma is sensitive to both radiation therapy and chemotherapy. The responsiveness of nasopharyngeal carcinoma to both radiotherapy and chemotherapy distinguishes it from other H&N cancers, which are typically insensitive to chemotherapy.

In brief, also H&N cancers gather a variety of very different malignant diseases, with distinct aetiologies and natural history, requiring expert diagnosis, expert treatments, and prospective collection of clinical data in order to better standardize the treatments.

2.2 Methodology

The IDEA4RC use cases were defined following different steps:

- Interview with STS and H&N cancers expert oncologists and researchers
- Survey to IDEA4RC clinical partners
- Discussion meetings

Before the survey, we had interviews with clinical researchers, oncologists and epidemiologists from the INT to understand: 1) what kind of objectives/studies they need real world data for and 2) how they currently access these data. On the basis of the feedbacks received, in collaboration with Task 2.1, we developed a questionnaire to collect information on the EMRs data available at hospitals and on the expectations of researchers, oncologists and data managers/data scientists regarding the IDEA4RC ecosystem (included in deliverable D2.1). Additionally, the questionnaire asked oncologists and researchers to list at least 2 open research questions for STS and H&N cancers.

The questionnaire was first tested and completed in INT and then distributed to all the other 10 clinical partners of IDEA4RC: Centre de lutte contre le cancer Léon Berard; Assistance publique – Hôpitaux de Paris; Fundación Jimenez Diaz, University Hospital; Sahlgrenska University Hospital, Gothenburg; Maria Sklodowska-Curie National Institute and Oncology Centre; Motol University Hospital; Oslo University Hospital; Masaryk Memorial Cancer Institute; Fundación Profesor Novoa Santos; Universitätsklinik Essen. The survey responses have been included in the IDEA4RC google drive to facilitate the exchange of questions and answers.

Finally, the survey responses were discussed at the second IDEA4RC plenary meeting held in Venice on April 20th 2023.

2.3 Use cases

The main objective of researchers and oncologists was to improve the ability to diagnose and treat all aspects of STS and H&N cancers, with the ultimate goal of improving survival and quality of life for patients with these two rare cancers.

In detail, the research objectives were related to 4 areas:

1. Description of the natural history of STS and H&N cancers (how the rare cancer develops, progress, possible association with other diseases, etc.);
2. Evaluation of factors that influence prognosis (e.g. mortality, survival, progression-free survival) and treatment response;
3. Assessment of the treatment effectiveness (systemic, radiotherapy, surgery, target therapy, immunotherapy and possible combinations);
4. Quality of care (diagnostic and staging procedures, treatment strategies, follow-up etc.).

The research questions address all the different stage of the disease from diagnosis to death/cure of the patient.

Table 1 reports the specific research questions (for STS and H&N cancers) grouped by major objectives.

Description of the natural history of disease
Incidence of skeletal metastases (after diagnosis) in patients with solitary fibrous tumours (SFT) in general and by site of primary SFT (e.g., meningeal versus extra meningeal etc.)
Solitary fibrous tumours (SFTs) are rare soft tissue tumours that can sometimes metastasize or spread to other parts of the body. Skeletal metastases refer to the spread of cancer to bones. The incidence of skeletal metastases in patients with SFTs varies depending on the location of the primary tumour. For example, patients with meningeal SFTs (tumours that originate in the lining of the brain and spinal cord) are more likely to develop skeletal

metastases than those with extra-meningeal SFTs (tumours that originate outside of the brain and spinal cord).

The bones most commonly affected by skeletal metastases in patients with SFTs are the spine, pelvis, and ribs. Other bones, such as the long bones of the arms and legs, can also be affected.

A better understanding of where the metastases will occur and which are the factors associated to the bone metastases, will allow to personalised the follow-up of patients with SFTs.

Relevant variables:

- Patient demographic characteristics (age, gender, etc.)
- Performance status of the patients at diagnosis
- Data of diagnosis of primary cancer
- Histopathological characteristics of the primary tumour and metastases
- Type and location of the primary tumour (meningeal versus extra-meningeal SFTs)
- Presence or absence of skeletal metastases, date of metastases diagnoses
- Sites of skeletal metastases (e.g., spine, pelvis, ribs, long bones)
- Presence or absence of metastases in other organs (e.g., liver, lung)
- Treatment modalities used for primary tumour and metastases
- Patient follow-up and vital status with dates

Incidence of radiation-induced secondary tumours in STS treated with radiotherapy.

Radiation therapy is a common treatment option for sarcomas, which are cancers that develop in the connective tissues of the body, such as bones, cartilage, and muscles. However, one potential complication of radiation therapy is the development of radiation-induced tumours, which are new tumours that can arise within the area that was treated with radiation.

The incidence of radiation-induced tumours in sarcomas treated with radiotherapy is generally low, ranging from 1% to 10%, depending on the dose and duration of radiation therapy. The interval between radiotherapy and the appearance of radiation-induced tumours can vary widely, ranging from a few months to several decades after treatment.

The site of onset of radiation-induced tumours relative to the radiotherapy field can also vary. The histology, of the radiation-induced tumour can also vary including osteosarcoma and malignant fibrous histiocytoma.

Overall, the development of radiation-induced tumours in sarcoma patients is a rare but potentially serious complication of radiation therapy.

A better understanding of the different factors contributing to the development of second primary cancer, could contribute to define a personalised follow-up of patients treated with radiotherapy to ensure early diagnosis of secondary primary cancers.

Relevant variables:

- Patient's age, performance status at diagnosis
- Primary cancer size, side, grading, histology
- Second primary cancer
- Site and date of diagnosis of second primary cancer
- Treatment of primary cancers including site of radiotherapy.
- Follow-up date
- Vital status date
- p53 status and other markers

Survival and incidence of distant metastases in angiosarcomas divided by 1) cutaneous (radio induced versus non) and 2) visceral.

Angiosarcoma is a rare type of cancer that develops in the lining of blood vessels or lymphatic vessels. It can occur in various parts of the body, including the skin and internal organs.

The survival and incidence of distant metastases in angiosarcomas can vary depending on several factors, including the location of the primary tumour and whether it is associated with prior radiation therapy.

Visceral angiosarcomas, which occur in internal organs such as the liver, spleen, and heart, are generally associated with a poor prognosis. The five-year survival rate for patients with visceral angiosarcomas is typically less than 10%, and distant metastases are common.

The incidence of distant metastases in angiosarcomas can also vary depending on the location of the tumour. For example, angiosarcomas of the skin are more likely to metastasize to regional lymph nodes and the lungs, while visceral angiosarcomas are more likely to metastasize to the liver and lungs.

Addressing this question, physicians will be able to tailor pts follow-up time and strategies.

Relevant variables:

- Patient age, performance status, comorbidity, gender
- Primary tumour grade, size, deepness site (cutaneous or visceral + details of the visceral organ)
- Prior radiation therapy
- Primary tumour treatment (surgery, radiotherapy, chemotherapy)
- Distant metastases (yes/no), site, date of diagnosis
- Patients' follow-up and vital status, dates

Identification/validation of prognostic and predictive factors

Validation of the prognostic significance of neutrophils/lymphocytes ratio (NLR) and prognostic index combining serological and inflammatory factors (PISIF) in primary retroperitoneal sarcomas (Voss RK, 2022) (Fiore M, 2023).

Primary retroperitoneal sarcomas develop in the retroperitoneal space, which is located behind the abdominal cavity. They can be difficult to treat and have a poor prognosis, with a five-year survival rate ranging from 15% to 50%.

The NLR is a ratio of the number of neutrophils to the number of lymphocytes in the blood. Studies have found that high NLR is associated with worse overall survival and higher rates of recurrence in patients with primary retroperitoneal sarcomas.

The PISIF is a prognostic index that combines several serological and inflammatory factors, including albumin, C-reactive protein, tumour size, and tumour grade. Studies have found that PISIF can predict the risk of recurrence and overall survival in patients with primary retroperitoneal sarcomas.

Validation of these indexes could help clinicians predict the risk of disease progression and plan appropriate treatment strategies.

Relevant variables:

- Patient's age, performance status at diagnosis, comorbidity, gender
- Primary tumour grade, size, deepness site
- Treatment modality: surgery, chemotherapy, and/or radiation therapy
- Number of neutrophil and lymphocyte in the blood.
- Albumin, C-reactive protein, + other markers of inflammation.

- Recurrence, date
- Follow-up and vital status of the pts with dates

Association of cellularity and myxoid liposarcomas prognostic.

Myxoid liposarcoma is a type of soft tissue sarcoma that is characterized by a mixture of lipoblasts (immature fat cells) and small round cells. The term "cellularity" in this context refers to the density of cells within the tumour.

The correlation of cellularity with myxoid liposarcomas prognosis (e.g., overall survival, progression-free survival) is important because it can provide clinicians with additional information to help guide treatment decisions and predict the likely outcomes for these patients. **For example, if higher cellularity is found to be associated with a worse prognosis, then patients with high-cellularity tumours may be considered for more aggressive treatment or closer follow-up. On the other hand, patients with low-cellularity tumours may be candidates for less intensive treatments.**

<u>Relevant variables:</u>

- Patient's age, performance status, comorbidity, gender
- Tumour size, site, grade, mitotic index, cellularity (i.e. density of cells in the tumour tissue)
- Treatment modality: surgery, chemotherapy, or radiation therapy
- Molecular markers: Certain molecular markers may be associated with myxoid liposarcomas and can potentially impact the prognosis.
- Recurrence, progression, date
- Follow-up and vital status of the patient with date

Identification of risk factors for metastatic potential of chordoma based on genetic alterations.

Chordoma is a rare type of cancer that develops from the notochord, a structure that forms in early embryonic development and plays a key role in the development of the spine. While chordoma is a relatively slow-growing cancer, it is also highly invasive and has a high propensity for metastasis, which can significantly reduce a patient's survival rate.

The identification of risk factors for the metastatic potential of chordoma is an active area of research, with genetic alterations being one of the key factors under investigation. Recent studies have identified several genetic alterations that may be associated with an increased risk of chordoma metastasis. These include TP53, T gene copy number alterations, PI3K/AKT/mTOR pathway alterations.

The feasibility of this use case will depend on the extent to which genetic analyses are current practice for these patient's management.

Definition of different prognostic groups based on location of extra skeletal Ewing sarcoma.

Ewing sarcoma usually affects bones, but it can also occur in soft tissues outside of the bones. When Ewing sarcoma arises in soft tissues outside of the bones, it is called extra-skeletal Ewing sarcoma.

The location of extra-skeletal Ewing sarcoma can vary, and the prognosis for patients with this type of cancer may differ based on the location of the tumour. For example, extra-skeletal Ewing sarcoma that arises in the chest wall or the pelvis may have a worse prognosis compared to tumours that arise in the extremities.

The identification of different prognostic groups of patients with extra-skeletal Ewing sarcoma based on the location of the tumour, could help to predict the prognosis and guide treatment decisions for patients with this rare and aggressive cancer.

Relevant variables:

- Age at diagnosis, Gender, comorbidity, performance status
- Tumour size, site, grading, deepness
- Location of extra-skeletal Ewing sarcoma (e.g. chest wall, retroperitoneum, head and neck)
- Treatment received (e.g. chemotherapy, surgery, radiation therapy)
- Response to treatment
- Recurrence or progression dates

Identification of predictors of outcome after surgical treatment (with respect to both short term morbidity, survival, recurrences and quality of life) in H&N cancers.

Physicians are interested in identifying factors that may predict the outcomes (short-term morbidity, survival, recurrences, and quality of life) of patients who undergo surgical treatment.

Short-term morbidity refers to any negative health effects that occur within the first few weeks or months after surgery. This could include things like pain, infection, or other complications related to the surgery itself. Finally, quality of life refers to how patients feel and function after surgery, including factors such as pain, fatigue, mobility, and emotional well-being.

Possible predictive factors include age, sex, as well as clinical factors such as the site and stage of the disease. Other factors that may be studied include patients' pre-existing medical conditions and lifestyle factors such as smoking or alcohol consumption.

This question is relevant to tailor post-surgery follow-up as well as to define whether adjuvant treatment (chemotherapy or radiotherapy after surgery) will be needed.

Relevant variables:

- Demographic variables such as age, sex, race/ethnicity, socioeconomic status, comorbidity, life style (e.g., smoking, alcohol)
- Tumour site, stage, histological subtype
- Surgical information (e.g., type of surgery performed, margin after surgery, node dissection)
- Other treatment different from surgery (radiotherapy, chemotherapy)
- Surgery complications
- Recurrence, progression (local, distant)
- Follow-up and vital status with dates

Assessment of the association between the mitotic index and the prognosis of solitary fibrous tumour.

The mitotic index (i.e., the ratio between the number of cells in a population undergoing mitosis to the total number of cells in a population). Solitary fibrous tumours are rare neoplasms that arise from mesenchymal cells and can occur in various parts of the body. The number of mitotic figures in a tissue sample can be used as a measure of tumour growth and aggressiveness.

Understanding whether and how the mitotic index is associated to solitary fibrous tumours prognosis could contribute to the definition of primary cancer treatment strategy.

Relevant variables:

- Demographic variables such as age, sex, comorbidity
- Tumour site, grading, size, deepness, mitotic index

- Treatment: surgery, radiation therapy, or chemotherapy
- Recurrence, progression (local, distant)
- Follow-up and life status with dates.

Volume-outcome relationship in retroperitoneal sarcomas treated with curative intent.

Retroperitoneal sarcomas are a rare type of cancer that develops in the soft tissues of the retroperitoneum, which is the space behind the abdominal cavity.

The relationship between the volume of cases treated by a healthcare provider and the outcomes (i.e. overall survival, progression-free survival) of patients with primary retroperitoneal sarcomas who underwent curative-intent surgery has not been proved yet. Other important endpoints that physicians would like to have more information on include progression-free survival, postoperative morbidity, local relapse and distant metastasis if available.

This question is important for ensuring appropriate patients referral and quality of care for these patients.

Relevant variables:

- Demographic variables such as age, sex, comorbidity
- Tumour size, tumour grade, histology, multifocality
- Surgery, completeness of resection (whether the surgical resection was complete or partial), surgery complications.
- Other treatment in addition to surgery (e.g., chemotherapy and radiotherapy)
- The number of cases of retroperitoneal sarcoma that a hospital or surgeon manages per year.
- Volume by surgeon: the number of surgeries performed by the surgeon for retroperitoneal sarcoma, multidisciplinary team discussion
- Local or distance relapse, progression, distant metastasis
- Life status and follow-up with dates

Evaluation of the prognostic significance of different sites of distant metastases in solitary fibrous tumours (i.e., prognosis of patients with metastases to the skeleton is worse than those who have metastases in other sites, e.g., liver or lung).

The prognostic significance of different sites of distant first relapse in patients with SFTs is still unclear. However, studies suggest that patients with first metastatic relapse to the skeleton may have a worse prognosis than those who relapse first in other sites, such as the liver or lung. This may be because skeletal metastases can cause pain, fractures, and other complications that can impact a patient's quality of life and overall survival.

Overall, the incidence of skeletal metastases in patients with SFTs is relatively low, but the prognosis for those who do develop skeletal metastases is generally poor. Further studies are needed to better understand the prognostic significance of different sites of distant relapse in these patients.

Addressing this question, physicians will be able to better define the treatment strategy for patients with these tumours.

Relevant variables:

- Patient demographic characteristics (age, gender, etc.)
- Performance status of the patients at diagnosis
- Data of diagnosis of primary cancer
- Histopathological characteristics of the primary tumour and metastases
- Type and location of the primary tumour (meningeal versus extra-meningeal SFTs)

- Presence or absence of skeletal metastases, date of metastases diagnoses
- Sites of skeletal metastases (e.g., spine, pelvis, ribs, long bones)
- Presence or absence of metastases in other organs (e.g., liver, lung)
- Treatment modalities used for primary tumour and metastases
- Patient follow-up and vital status with dates

Prognosis of STS patients with radiation-induced second primary cancers.

The prognosis for patients with radiation-induced tumours can be poor, with some studies reporting a five-year survival rate of less than 50%. However, the prognosis can vary depending on several factors, including the size and location of the tumour, the patient's age and overall health, and the extent of metastasis.

Studies have also investigated the role of p53 status and other markers in the development of radiation-induced tumours. p53 is a tumour suppressor gene that plays a critical role in preventing the development of cancer. Mutations or abnormalities in the p53 gene can increase the risk of developing cancer, including radiation-induced tumours.

Identify factors that impact on the prognosis of STS patients with radio-induced second primary, could contribute to stratify pts at higher risk of second primary cancer to ensure early diagnosis of secondary primary cancers.

Relevant variables:

- Patient's age, performance status at diagnosis
- Primary cancer size, side, grading, histology
- Second primary cancer
- Site and date of diagnosis of second primary cancer
- Treatment of primary cancers including site of radiotherapy.
- Follow-up date
- Vital status datep53 status and other markers

Feasibility study for radiomics, specifically focusing on if images could be used to predict grade in sarcomas.

Researchers are interested in understanding whether radiomic features extracted from medical images could be used to predict the grade of sarcomas. Radiomic features are mathematical representations of the texture, shape, and intensity of a region of interest in a medical image. They can be extracted using various image processing techniques and can provide quantitative information about the tumour characteristics.

Grade is a known prognostic and predictive factor used, together with other information, to make decision about the treatment.

Relevant variables

The relevant variables for this study are likely related to radiomics features extracted from medical images, such as CT or MRI scans, of sarcomas. These radiomics features could include texture, shape, and intensity measurements, among others. The study may also include clinical variables, such as patient age, gender, and tumour location. The primary outcome variable would likely be the accuracy of the radiomics model in predicting the tumour grade of the sarcomas, which could be compared to the gold standard of histopathology.

Images are outside the scope of IDEA4RC, but we have kept the use case as an example of possible future expansion of IDEA4RC.

Assess treatment and or diagnostic procedures effectiveness

Comparison of fine needle aspiration vs. core biopsy with respect to the pre-surgical diagnosis of salivary gland tumours.

This study aims to compare the diagnostic accuracy of two common biopsy techniques used in the pre-surgical diagnosis of salivary gland tumours: fine needle aspiration (FNA) and core biopsy.

Fine needle aspiration involves using a thin needle to extract a small sample of cells from the tumour for examination under a microscope. Core biopsy, on the other hand, uses a larger needle to remove a small core of tissue from the tumour for examination.

The study will assess the diagnostic accuracy of these two techniques with respect to their ability to identify the type and grade of salivary gland tumours before surgery suggesting the best technique to use in the clinical setting.

Relevant variables

The relevant variables for this study may include

- the biopsy technique used (FNA vs. core biopsy)
- the type and grade of the salivary gland tumour
- the same information coming from the surgical specimen (if available).

Assessment of the outcomes (overall survival, disease free survival) of sino-nasal cancer patients treated with induction chemotherapy.

Evidence on the outcomes of sino-nasal cancer treated with induction chemotherapy is scarce. Induction chemotherapy is a treatment approach where chemotherapy is administered before the main treatment, such as surgery or radiation therapy.

Understanding the impact of induction therapy is important because it could be used to shrink the tumour and make it easier to remove. This could ultimately improve the effectiveness of subsequent treatments.

Relevant variables

- Patient characteristics: age, sex, performance status at diagnosis, comorbidity at diagnosis
- Tumour characteristics: stage at diagnosis and histological sub-type
- Main treatment: surgery, chemotherapy +/- radiotherapy
- Life status and follow-up, dates

Assessment of the role of photon and proton-based radiotherapy on the outcomes (overall survival, disease free survival) of low and intermediate grade mucoepidermoid cancers of salivary gland.

Evidence of the effectiveness of photon and proton-based radiotherapy in patients with low and intermediate grade mucoepidermoid cancers of the salivary gland is scarce and contradictory. Thus, **physicians are interested in having additional evidence on their impact in terms of overall survival and progression free survival to properly define the treatment strategy for patients with these very rare cancers**.

Relevant variables:

- Patient characteristics: age, sex, performance status, comorbidity
- Tumour characteristics: stage at diagnosis and grade
- Main treatment: surgery, chemotherapy +/- radiotherapy including type of radiotherapy, dose and side effects.
- Vital status and follow-up, dates

Assessment of the outcomes (overall survival, disease free survival) of salivary gland cancers treated with surgery + radiotherapy +/- chemotherapy.

Salivary gland cancers are typically treated with a surgical resection of the tumour followed by radiotherapy and/or chemotherapy. The impact on the outcome (survival) of the addition of chemotherapy is still under debate and unclear.

Answering these research questions will contribute to ameliorate treatment and therefore increase survival for patients with these tumours.

Relevant variables:

- Patient characteristics: age, sex, performance status, comorbidity
- Tumour characteristics: stage, histological subtype, site of the tumour, grade
- Treatment: surgery, radiotherapy, and chemotherapy. This variable can be further divided into subcategories such as the type of surgery (e.g., partial or total gland removal) and the specific chemotherapy regimen used.
- Progression, recurrence, dates
- Adverse events: The occurrence of adverse events during treatment, such as radiation-induced side effects or chemotherapy-related toxicity, may also be recorded to assess the safety and tolerability of the different treatment approaches
- Follow-up, vital status, dates

Monitoring quality of care

Adherence to the relevant national and international guidelines for diagnostics and treatment (for both STS and H&N cancers).

By evaluating healthcare provider adherence to national and international guidelines, we can provide insights into whether guidelines are being followed and whether they are effective in improving patient outcomes. It will also help to identify areas where improvements can be made to ensure that patients receive the best possible care and to standardised treatment across centres/countries. Indicators should be developed for critical steps of the patients' management (diagnosis, staging, treatment received accordingly to stage (Trama A, 2019 Aug 28).

Relevant variables:

- Patient characteristics: age, sex, performance status, comorbidity
- Tumour characteristics: stage, histological subtype, site of the tumour, grade
- Treatment: surgery, radiotherapy, and chemotherapy. This variable can be further divided into subcategories such as the type of surgery (e.g., partial or total gland removal) and the specific chemotherapy regimen used
- Progression, recurrence, dates

Describe differences (in term of clinical management, survival, distribution of histotypes, site, age etc.) across countries (for both STS and H&N cancers).

Comparison of a single hospital clinical management against a benchmark (treatment outcomes, adverse events and complications, etc.) can help to identify are for improvement, can increase treatment standardisation reducing inequality in health care and ultimately improving survival of a higher number of patients.

Relevant variables:

- Country of origin
- Patient demographics (age, sex, ethnicity)
- Cancer site, histology, stage

• Treatment: surgery, radiation therapy, chemotherapy, immunotherapy, etc. • Treatment-related complications • Socioeconomic factors (income, education, employment status)

Table 1. Research questions for head and neck cancers and soft tissue sarcomas by major objectives

Some oncologists were interested in possible **methodological innovation** based on data-driven approach applied to the eco-system (i.e., would it be possible to identify reliable proxy of quality of life, safety and outcomes by exploiting the ecosystem? can we improve the predictive and prognostic performance with respect to standard factors by leveraging the ecosystem? In the context of multimodal treatments, database exploration with respect to the single therapeutic approach (e.g., re-interventions, radiotherapy stops, chemotherapy delays, infections).

In addition to the research questions, the oncologists expressed a specific interest in the ecosystem as a tool for **supporting clinical decision making**. In detail, they perceived it as an important tool to retrieve information on patient specific baseline risk and treatment effectiveness, to enable personalized clinical decision-making using knowledge coming from big databases of centres of expertise.

Others envisioned using the ecosystem to address questions raised at the multidisciplinary team meetings on very rare or complex cases (use IDEA4RC as a viewer of the patient's tumour history to provide a comparison with similar cases).

The researchers were also interested in performing simple queries such as those relating to the identification of a specific patient cohort. Last but not least, they were very interested in whether the ecosystem could be used to **automate data collection for the hospital-based cancer registry**.

Most importantly, many of the respondents were interested in understanding the **quality of the ecosystem data** and to what extent the quality and results compare to those of the data manager's hand-filled registries.

The final use cases will be selected among the research questions listed in Table 1 above, considering the data and information that will be made available by the hospitals contributing to the IDEA4RC ecosystem. However, use cases will include some if not all of the following:

- queries for selecting cohorts of STS and H&N cancers of interest and explore data availability for specific research questions (e.g., number of salivary gland cancer patients (any morphology) treated with surgery + radiotherapy +/- chemotherapy with information on stage, sex, age, comorbidities, number of recurrences, life status, late effects)
- queries exploring quality of data for specific research questions (e.g., distribution of salivary gland cancers morphologies within an hospital and across the hospitals contributing to the ecosystem vs expert expectation)
- feasibility of extracting variables to automatize the population of the EURACAN registry

- descriptive analyses reporting on quality of care within and across the hospitals (comparison across the hospitals contributing to the IDEA4RC ecosystem and vs clinical practice guidelines for STS of limbs and H&N cancers)
- prognostic/predictive modelling (research hypothesis- and/or data-driven)
- evaluation of treatment effectiveness

It is worth recalling that IDEA4RC focuses on structured and unstructured data included in free text. Biological samples/data as well as imaging /data are not the focus of IDEA4RC.

2.4 Analyses by use case

Table 2 reports the type of data analyses necessary to answer the request questions identified. This is the results of discussions including statisticians and data scientists. This is intended to contribute to the definition of the federated algorithms that has to be developed and made available in the ecosystem (Task T8.3, Deliverable D8.4).

Major research objectives	Statistical Analyses	Functionalities needed for setting the analyses parameters
Description of the natural history of disease and Monitoring quality of care	Descriptive analysis: - two by two, 3 way contingency table etc.; total and % by row or column; simple counting of selected variables. - Chi square test (to evaluate how likely it is that any observed difference between the sets rose by chance). - Observed survival and cause specific survival (Kaplan Meier method) - Log Rank test (to evaluate differences in survival curves) - Incidence within the selected cohort of specific conditions/characteristic (e.g., recurrence, progression etc.)	Generate new variables from one single variable (e.g. defining specific group of one continuous variable) or merging existing variables (e.g. define multimodal treatment) Propensity score Possibility to select: end point of the analysis (death, progression etc), periods of analysis (begin and end date), end of follow-up, survival by year from diagnosis (1,2,3,4,5 etc.), median, conditional survival. Possibility to visualise residuals of the models and test assumption. Possibility to perform Landmark analyses
Identification/validation of prognostic and predictive factors	Cox proportional hazard models and regression models	
Assess treatment and or diagnostic	Generalised linear models (Multilevel models)	

Major research objectives	Statistical Analyses	Functionalities needed for setting the analyses parameters
procedures effectiveness		(to avoid immortal time bias). Possibility to visualise indicators of the performance of the models (confusion matrix) Possibility to perform Lasso method (least absolute shrinkage and selection operator) Roc curve Being able to divide the dataset into a validation and training set Possibility to select only one centre.

Table 2. Statistical analyses by major research questions

3 CORE CLINICAL DATASETS FOR SOFT TISSUE SARCOMAS AND HEAD AND NECK CANCERS

Based on the research questions listed in the questionnaire and the expectations and comments of oncologists and researchers, we defined 2 core sets of clinical data: 1 for H&N cancers and 1 for STS. These datasets were defined considering the EURACAN registry dataset (ClinicalTrials.gov Identifier: NCT05483374), the OSIRIS minimal set of data (doi: 10.1200/CCI.20.00094), the one million genomes project dataset. This last one was kindly shared with us by our partner DIGICORE.

The core datasets were shared with clinical partners for feedbacks and were discussed and approved in ad hoc meetings. Two meetings were organised: one for H&N cancer and one for STS as the datasets and experts are different.

The STS dataset includes information reported in Table 3 (the details are included in Annex 1).

PATIENTS INFORMATION
Age at diagnosis
Gender
Comorbidities
Genetic syndrome
Occurrence of other cancers
PRIMARY TUMOR
Biopsy - type of biopsy - unplanned excision
Tumour - site - size - morphology - depth - biopsy mitotic count - grading - molecular profiling - stage
TREATMENT OF PRIMARY TUMOR
Surgery
Medical treatment (e.g. chemotherapy, molecular target therapy, immunotherapy etc.)
Radiotherapy
Reason for end of treatment

Treatment response
STATUS OF PATIENT AT LAST FOLLOW-UP
RECURRENCE/PROGRESSION
Type - local - metastatic
Treatment of recurrence - surgery - medical treatment - radiotherapy - reason for end of treatment - treatment response

Table 3. Core dataset for soft tissue sarcomas, by major stage of disease development and progression

The H&N cancers dataset include information reported in Table 4 (the details are included in Annex 2).

PATIENTS INFORMATION
Age at diagnosis
Gender
Race
Country of residence
Comorbidities
Performance status
Smoking
Alcohol
PRIMARY TUMOR
Biopsy
Tumour - site - morphology - grading - clinical and pathological stage - HPV status - EBV status
TREATMENT OF PRIMARY TUMOR
Surgery
Medical treatment (e.g. chemotherapy, molecular target therapy, immunotherapy etc.)

Radiotherapy
Reason for end of treatment
Treatment response
Adverse events
STATUS OF PATIENT AT LAST FOLLOW-UP
RECURRENCE/PROGRESSION
Type - local - regional - metastatic
Treatment of recurrence - surgery - medical treatment - radiotherapy - reason for end of treatment - treatment response

Table 4. Core dataset for head and neck cancers, by major stage of disease development and progression

Cancer stages (i.e., diagnosis, main treatment, recurrence and/or progression) are the same for STS and H&N cancers because they are the general stages of development and progression of any cancer.

Physicians stressed the importance of collecting information on all the different stages from diagnosis to death or recovery. Information on all stages of the disease is also essential to adequately address the main use cases, which in fact focus on the natural history of the disease and the identification of predictive and prognostic factors.

In addition to the information on all the stages of the disease, physicians stressed the importance of the details to collect (please refer to Annex 1 and 2).

These core datasets are intended as a starting point. The final datasets will depend on:

- the number of structured variables already available from hospitals
- the performance of the NLP algorithms for extracting data from the free text
- the quality of each retrieved variable
- the information details available in the hospital's data sources

4 KEY PERFORMANCE INDICATORS

The implementation of the date (re) use cases will contribute to assess the performance of the IDEA4RC ecosystem in relation to its **usability.**
Data usability is intended as the existence of useful and valuable data sets and analysis capabilities available in accessible and convenient forms.

Data usability depends on multiple dimensions including (Figure 2):

- **Relevance:** data address an information need (i.e. data must be fit for purpose or to the extent to which a dataset presents data elements useful to answer a research question). For example, a researcher studying the effectiveness of a new cancer treatment would need data on the patients' status, the tumour, the new treatment and the outcomes.
- **Quality:** data are of acceptable quality for the intended purpose, poor quality data could have negative impacts on the findings generated from these data. In the example above, the information must be complete, accurate and reliable.
- **Coverage and Granularity:** data have adequate coverage and are structured at the right level of granularity. Following the same example, the data should cover the patient population and time period of interest, and be structured at a level of detail that allows for meaningful analysis (e.g., information on dose and regimens of the new treatments, any problems incurred during the treatment etc.).
- **Accessibility and Documentation**: data must be accessible, with sufficient metadata for potential users to understand their derivation and meaning. In the same example, the researcher must have access to the data needed, with sufficient documentation to understand how they were collected and defined (e.g., radiotherapy induced sarcoma: if the new sarcoma is in the field or marginal, the anatomical area that received high radiation dose for the treatment of a prior cancer).
- **Ease of analysis:** appropriate tools must be available to manipulate the data (e.g., filtering, sorting, and aggregating), viewing the data (e.g., mapping and charting) as well as to perform statistical analysis or predictive modelling. In the example above, the researcher must have access to tools for filtering, sorting, aggregating, mapping, charting, and performing statistical analysis or predictive modelling on the data.

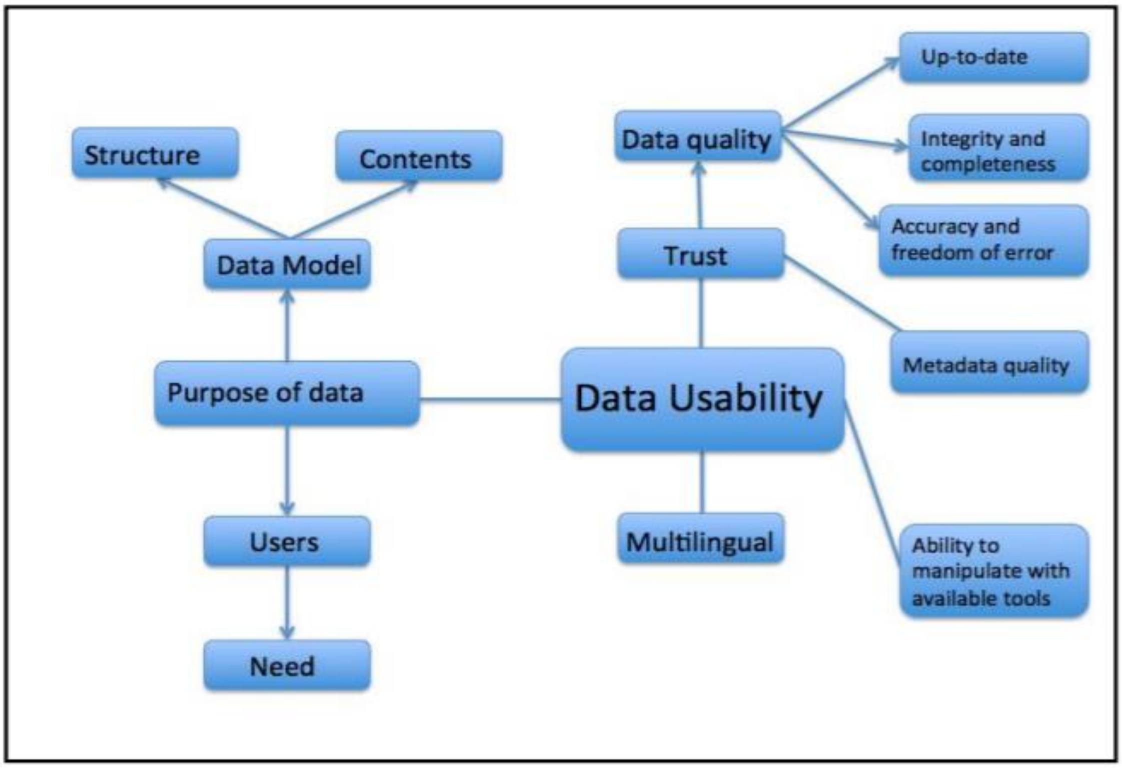

Figure 2. Data usability dimensions

We assess each of the data usability dimensions.

Relevance

Relevance will be assessed by verifying that the variables and number of cases are sufficient and appropriate to answer the main research questions. In detail, the following will be considered:

- Number and type of variables available for each tumour stage, including diagnosis, primary tumour treatment, recurrence, follow-up, and patient's life status.
- Number of cases with STS and H&N cancers with information available for each tumour stage and for all tumour stages until death or cure.
- Number of research questions, among those listed by oncologists and researchers, that can be answered by analysing the data available in the IDEA4RC ecosystem.

Quality

Data quality will assess data conformance, completeness and plausibility in two assessment contexts: verification (not relying on external references) and validation (relying on relevant external benchmarks).

<u>Conformance checks</u> will include: value conformance, relational conformance, and computational conformance. Value conformance verifies whether the values that are present meet syntactic or structural constraints. Relational conformance seeks to determine if the

recorded data elements agree with additional structural constraints imposed by the physical database structures that store data values (i.e. data fields that are allowed to null or must always have a value). Computational conformance determines if computations used to create derived values from existing variables yield the intended results.

For each core set of data (STS and H&N cancers), we will define a "data dictionary," listing the intended format and allowed values for every data element. We will check each variable retrieved from the EMR against the values defined in the data dictionary. In order to proceed to the development of the data dictionary and to the relational conformance assessment we need to first understand how many and which variables out of those included in the core dataset will be in the end retrieved from the EMR of each hospital (Tasks 5.1, 5.2, 3.3 and 9.2). Computational conformance checks will be defined based on the use cases that will be implemented. Most likely computational conformance checks will be performed on derived values such as Charslon comorbidity index, BMI or derived variables such multimodal treatment, TNM stage, grading groups etc.

Completeness checks will assess the absence of data at a single moment over time or when measured at multiple moments over time. Thus, we will assess completeness in all the cancer stages over time and within cohort of STS and H&N cancer patients over time.

Plausibility checks will focus on features that describe the believability or truthfulness of data values. We will assess atemporal plausibility by examining the distribution of values (eg, distribution within an hospital and across the hospitals contributing to the ecosystem of STS morphologies, of H&N cancer site, of STS and H&N cancers stage, STS and H&N cancers treatment of primary cancers, STS and H&N cancers recurrence/progression, STS and H&N cancers overall survival) or by comparing multiple values that have an expected relationship to each other (e.g., distribution within an hospital and across the hospitals contributing to the ecosystem of STS morphologies and treatment, of STS sites and treatment of the primary cancers, of H&N cancers stage and treatment by site and/or by morphologies, H&N cancers and STS treatment by age and sex etc). We will select value distributions based on the use cases. We will assess these distributions over time (temporal plausibility) and we will confront the results of the distribution with domain experts' expectation and with available benchmark as validation tasks (i.e. EURACAN clinical registry of H&N cancers and STS).

Coverage and Granularity

The IDEA4RC ecosystem includes 11 hospitals located in 9 countries. We will estimate the percentage of incident and prevalent cases of STS and H&N cancers managed by each hospital compared with those diagnosed and prevalent annually in the country where the hospital is located.

We will evaluate the selection bias of the IDEA4RC ecosystem by comparing the demographic characteristics and relevant prognostic factors (e.g., age, sex, site, morphologies, stage, treatment) of patients with STS and H&N managed by hospitals contributing to the ecosystem with those of patients with STS and H&N from the same country where the IDEA4RC hospitals are located. We will use literature review and data from population-based cancer registries. (Tasci E, 2022 Jun 12) (Beesley, 2022;) (European Medicines Agency, 2023).

For assessing the risk of bias (ROB), we will consider the use of PROBAST (Prediction model Risk Of Bias ASsessment Tool) (Wolff RF & Group†., 2019). It includes 20 signalling questions across 4 domains: participants, predictors, outcome, and analysis. These signalling questions are designed to highlight potential methodological flaws in the study and help assess the applicability of the model to the intended population and setting. This tool is commonly used to evaluate the ROB and applicability of studies that develop, validate, or update of diagnostic and prognostic prediction model for individualized predictions.

The extent to which the granularity of available information is sufficient will be assessed by considering how many proposed research questions could be answered based on the information included in the ecosystem.

Accessibility and documentation

We will with an iterative approach assess all the FAIR principles:

- (Meta) data are assigned globally unique and persistent identifiers and clearly and explicitly include the identifier of the data they describe.
- (Meta)data are registered or indexed in a searchable resource.
- (Meta)data are retrievable by their identifier using a standardised communication protocol. We will clarify the exact conditions under which the IDEA4RC date will be accessible.
- Metadata should be accessible even when the data is no longer available.
- (Meta)data use a formal, accessible, shared, and broadly applicable language for knowledge representation and meet domain-relevant community standards.
- (Meta)data use vocabularies that follow the FAIR principles.
- (Meta)data include qualified references to other (meta)data.
- (Meta)data are richly described with a plurality of accurate and relevant attributes including provenance (e.g., scope of the data: for what purpose was it generated/collected? particularities or limitations about the data that other users should be aware of, date of generation/collection of the data, who prepared the data, the name and version of the software used, raw or processed data? variable names are explained or self-explanatory (i.e., defined in the research field's controlled vocabulary), workflow that led to your data: Who generated or collected it? How has it been processed? Has it been published before? Does it contain data from someone else that you may have transformed or completed?).
- (Meta)data are released with a clear and accessible data usage licence (we will clarify the conditions under which the data can be used).

The use of FAIR-Aware (https://fairaware.dans.knaw.nl/) or other FAIR assessment tool will be considered.

Ease of Analysis

We will evaluate the satisfaction of oncologists/researchers who have used the ecosystem in terms of ease of use, easy navigation, user-friendliness.

In addition, we will verify:

- The number of research questions analysed for each of the major areas of information needs (i.e. natural history, prediction/prognostication, quality of care).
- The number of researchers and/or oncologists from the contributing hospitals having used the ecosystem at least once before the end of the project.

Finally we will discuss the appropriateness of administering to the users a usability scale https://measuringu.com/sus/.

REFERENCES

A. Althubaiti. (2016, May 4). Information bias in health research: definition, pitfalls, and adjustment methods. J Multidiscip Healthc. *J Multidiscip Healthc.*, pp. 9:211-7. doi: 10.2147/JMDH.S104807. PMID: 27217764; PMCID: PMC4862344.

Beesley, L. M. (2022;). Case studies in bias reduction and inference for electronic health record data with selection bias and phenotype misclassification. . *Statistics in Medicine. doi:10.1002/sim.9579, 41(28):*, 5501– 5516.

European Medicines Agency, .. (2023). *ENCePP Guide on Methodological Standards in Pharmacoepidemiology*. Retrieved from European Network of Centres for Pharmacoepidemiology and Pharmacovigilance: ENCePP Guide on Methodological Standards in Pharmacoepidemiology

Fiore M, L. S. (2023, Feb 1). Preoperative Neutrophil-to-Lymphocyte Ratio and a New Inflammatory Biomarkers Prognostic Index for Primary Retroperitoneal Sarcomas: Retrospective Monocentric Study. *Clin Cancer Res.*, pp. 29(3):614-620. doi: 10.1158/1078-0432.CCR-22-2897. PMID: 36478176.

Guillaudeux, M. R. (2023). Patient-centric synthetic data generation, no reason to risk re-identification in biomedical data analysis. *npj Digit. Med. 6, 37*, https://doi.org/10.1038/s41746-023-00771-5.

JARC. (2019). *RARE CANCER AGENDA 2030 - Ten Recommendations from the EU Joint Action on Rare Cancers.* https://ecpc.org/wp-content/uploads/2020/10/JARC-recommendations_layman-version-final.pdf.

K.J. Jager, C. Z. (2008,). Confounding: What it is and how to deal with it,. *Kidney International*, Volume 73, Issue 3, Pages 256-260, ISSN 0085-2538, https://doi.org/10.1038/sj.ki.5002650.

M.E. Jacob, M. G. (2016). Chapter 1 - Epidemiology for the clinical neurologist,. In F. B. Editor(s): Michael J. Aminoff, *Handbook of Clinical Neurology,* (pp. Volume 138, Pages 3-16, ISSN 0072-9752, ISBN 9780128029732, https://doi.org/10.1016/B978-0-12-802973-2.00001-X.). Elsevier,.

Rare Cancers Europe. (n.d.). *https://www.rarecancerseurope.org/what-are-rare-cancers/definition-of-rare-cancers*.

Sterne JAC, H. M. (Cochrane, February 2022.). Assessing risk of bias in a non-randomized study. In T. J. Higgins JPT, *Cochrane Handbook for Systematic Reviews of Interventions version 6.3* (p. Chapter 25). Cochrane . Retrieved from Cochrane Training.

Tasci E, Z. Y. (2022 Jun 12). Bias and Class Imbalance in Oncologic Data-Towards Inclusive and Transferrable AI in Large Scale Oncology Data Sets. . *Cancers (Basel).* , 14(12):2897. doi: 10.3390/cancers14122897. PMID: 35740563; .

Trama A, B. L.-Ž. (2019 Aug 28). Quality of Care Indicators for Head and Neck Cancers: The Experience of the European Project RARECAREnet. *Front Oncol.* , 9:837.

van der Steen JT, T. R. (2019, Mar. 12). Causes of reporting bias: a theoretical framework. *F1000Res. Mar 12;*, p. 8:280. doi: 10.12688/f1000research.18310.2. PMID: 31497290; PMCID: PMC6713068.

Voss RK, C. D. (2022, Feb 27). Sarculator is a Good Model to Predict Survival in Resected Extremity and Trunk Sarcomas in US Patients. *Ann Surg Oncol.*, pp. doi: 10.1245/s10434-022-11442-2. Epub ahead of print. PMID: 35224688.

Wolff RF, M. K., & Group†., P. (2019, Jan 1). PROBAST: A Tool to Assess the Risk of Bias and Applicability of Prediction Model Studies. *Ann Intern Med.*, pp. 170(1):51-58. doi: 10.7326/M18-1376. PMID: 30596875.

ANNEXES

Annex 1 – Soft Tissue Sarcomas "core" dataset – Variables

Annex 2 – Head and Neck cancers "core" dataset - Variables