



Project acronym: EOSC4CANCER
Grant Agreement Number: Id: 101058427
Project full title: EOSC4CANCER
Call identifier: HORIZON-INFRA-2021-EOSC-01

D2.1

SOPs for minimal clinical, image and omics data exchange

Version:	2.0
Status:	Version for submission
Dissemination Level:	Public
Deliverable Type:	Report
Due date of deliverable:	31.01.2024
Actual submission date:	26.01.2024
Work Package:	WP2 Distributed harmonisation efforts across domains and jurisdictions to enable reproducible cancer research
Lead partner for this deliverable:	EMBL, Lygature, BSC, Erasmus MC
Partner(s) contributing:	CRG, MU, CNR, NKI, BBMRI-ERIC, VHIO

Main author(s):


Tolganay Kabdullayeva	EMBL
Robin Navest	Lygature
Sergi Aguilo	BSC
Martijn Starmans	Erasmus MC

Other author(s):

Romina Royo	BSC	Beatriz Carvalho	NKI
-------------	-----	------------------	-----

Teresa García Lezana	CRG	Carlo Senore	CPO
Jelle Evers	IKNL	Kurt Majcen	BBMRI-ERIC
Peter Prinsen	IKNL	Remond Fijneman	NKI
Pavel Čupr	MU	Cristina Viaplana	VHIO
Soňa Skřídlová	MU	Raquel Comas	VHIO
Ondřej Mikeš	UZIS	Mireia Sanchis	VHIO
Maria Francesca Vitale	NA3CR	Nadia Saoudi	VHIO
Concetta Ambrosino	CNR	Alexander Ing	EMBL
Laurien Ulfman	UU	Sophie Huisjes-Berends	Lygature
Eva García Álvarez	BBMRI-ERIC		

Revision History

Version	Date	Changes made	Author(s)
0.1	23.10.2023	First draft	Tolganay Kabdullayeva (EMBL)
0.2	17.11.2023	Outline and structure of deliverable	Tolganay Kabdullayeva (EMBL)/Robin Navest (Lygature)/Martijn Starmans (Erasmus MC)/Romina Royo (BSC)
0.3	05.12.2023	Inclusion of first nine SOPs	Tolganay Kabdullayeva (EMBL)/Robin Navest (Lygature)/Martijn Starmans (Erasmus MC)
0.5	12.12.2023	First data type description based on available SOPs	Tolganay Kabdullayeva (EMBL)/Robin Navest (Lygature)/Martijn Starmans (Erasmus MC)/Sergi Aguiló (BSC)
0.8	03.01.2024	Inclusion of four additional SOPs and finalisation of data type descriptions	Tolganay Kabdullayeva (EMBL)/Robin Navest (Lygature)/Sergi Aguiló (BSC)
1.0	10.01.2024	Version submitted for internal review	Robin Navest (Lygature)/Martijn Starmans (Erasmus MC)/Sergi Aguiló (BSC)/Romina Royo (BSC)
1.1	19.01.2024	Processed feedback from internal review	Robin Navest (Lygature)/Martijn Starmans (Erasmus MC)/Sergi Aguiló (BSC)/Romina Royo (BSC)
2.0	26.01.2024	Version for su  Funded by the European Union	Robin Navest (Lygature)/Martijn Starmans (Erasmus MC)/Sergi

Aguilo (BSC)/Romina Royo
(BSC)

Statement of originality

This deliverable contains original unpublished work except where clearly indicated otherwise. Acknowledgement of previously published material and of the work of others has been made through appropriate citation, quotation or both.

TABLE OF CONTENTS

EXECUTIVE SUMMARY	5
Background	6
Methodology	7
Dataset SOP collection	7
Data type overview	7
Standard Operational Procedures	9
1.1 Exposome	9
1.2 Cancer registries	10
1.3 Screening	11
1.4 Clinical	12
1.5 Genomics	14
1.6 Radiology	15
1.7 Pathology	17
Discussion	20
Next steps	20
Conclusion	22
Glossary of international coalitions on data harmonisation and integration	23
AACR	23
AACR GENIE	23
Beacon	23
BigPicture	23
B1MG	23
cBioPortal	24
dbGaP	24
EGA	24
EIRENE-RI	24
ENCR	24
EUCAIM	25
EU Cancer Screening Campaign	25
EXPANSE	25
Health-RI	25
HL7-FHIR	25
IMPACT-Data	26
Molecular Tumour Board Portal (MTBP)	26
OMOP-CDM	26
openEHR	26
RDMkit	26
SNOMED-CT	27
TCIA and TCGA	27
XNAT	27
Annex A - NCR SOP	28
Annex B - CNCR SOP	31
Annex C - NA3CR SOP	33

Annex D - BCR SOP	36
Annex E - EXPANSE Exposome-NL SOP	39
Annex F - EIRENE-CZ SOP	41
Annex G - ARPAC SOP	43
Annex H - ARPAB SOP	46
Annex I - mtFIT SOP	49
Annex J - Piedmont regional screening program SOP	51
Annex K - Catalan screening program SOP	53
Annex L - BBMRI-ERIC CRC-Cohort SOP	55
Annex M - CAIRO5 SOP	59
Annex N - PROVENC3 SOP	65
Annex O - mCRC-VHIO	69

LIST OF TABLES

Table 1. Overview of key EOSC4Cancer datasets for which SOPs were collected	9
---	---

EXECUTIVE SUMMARY

In recent years, partially due to the rise of AI, the trend in cancer research is to combine data from different sources and different modalities. For combining data from different data types within one modality, the main challenges include a lack of standardisation in the data format, data model, metadata model, and data access procedures. Hence, there is a clear need for harmonisation and interoperability. When combining and integrating data from different modalities, these challenges become even more pressing.

In EOSC4Cancer, we focus on all these aspects in the context of the use cases that cover the patient journey from cancer prevention to diagnosis to treatment, laying the foundation of data trajectories and workflows for future cancer mission projects. To address the above challenges, in this deliverable, we formulate standard operating procedures (SOPs) per data type for data access, data models, and data interoperability. To this end, we first defined SOPs for all datasets within the use-cases of EOSC4Cancer. Additionally, we inventorized which standards or SOPs from relevant other international consortia already exist or are being developed. Per data type, we looked at commonalities to formulate data type SOPs and provide an overarching view of the current practices for data exchange.

The results are SOPs for seven data types widely used in cancer research: exposome, cancer registry, screening, clinical, genomic, radiology, and pathology data. Instead of niche SOPs for the EOSC4Cancer datasets, we provide more general considerations and guidelines, so they can be used by the broader community. Lastly, we provide some insights in next steps to improve these SOPs.

Background

The inherent complexity of cancer necessitates the integration of advanced research data across national boundaries to enable progress. The better we organise cancer data across Europe, the better and faster we can bring the fruits of new biological and technical innovations to the benefit of EU citizens/patients. EOSC4Cancer will make cancer genomics, imaging, medical, clinical, environmental and socio-economic data accessible, using and enhancing existing federated and interoperable systems for securely identifying, sharing, processing and reusing FAIR (Findable, Accessible, Interoperable, Reusable) cancer data across borders, and it will offer them via community-driven analysis environments.

To enable progress in cancer research, data, knowledge and digital services - accessible across the European Research Area through federated infrastructures - are key. In EOSC4Cancer, we focus on all these aspects in the context of the EOSC4Cancer use cases (4.1¹, 4.2², 4.3³, 4.4⁴, and 4.5⁵) that cover the patient journey from cancer prevention to diagnosis to treatment, laying the foundation of data trajectories and workflows for future cancer mission projects.

Uniting a consortium of 29 organisations from 13 countries, EOSC4Cancer brings together a diverse array of stakeholders, including cancer research centres, research infrastructures, leading research groups, hospitals, and supercomputing centres. The project's commitment to sustainability is evident in its strategic leveraging of partners' research infrastructures and active engagement with the broader EOSC ecosystem.

In light of our mission to make data accessible, using and enhancing existing federated and interoperable systems for securely identifying, sharing, processing and reusing FAIR cancer data across borders, we provide in this deliverable standard operating procedures (SOPs) for different data types. In these SOPs, we focus on two key aspects. The first is data access, which procedures vary between datasets, which hampers reuse. Second, we address the different data models and interoperability. Once data can be accessed, datasets are often not interoperable and require mapping before they can be used to enrich each other. Based on an overview of the data models, vocabularies/ontologies, and dictionaries/codebooks used by the key datasets in EOSC4Cancer, an interoperability analysis was performed per data type. To ensure a holistic and impactful approach, we took harmonisation efforts by other international coalitions, such as GDI (European Genomic Data Infrastructure), ENCR (European Network of Cancer Registries), EIRENE-RI, EUCAIM (European Federation for CAncer IMages), and BigPicture, into account for our recommendations (See [Glossary](#)).

In this effort, we will create an overview of the SOPs for the key datasets in EOSC4Cancer and identify commonalities per data type to formulate data type SOPs, which can be re-used by others. Additionally, we describe what steps we would like to take based on this report as WP2 in EOSC4Cancer to improve these SOPs. Finally, based on the collected information, we suggest researchers and policymakers consider selecting data structures with appropriate formats, metadata, documentation, and access points according to the data type and FAIR principles.

¹ <https://eosc4cancer.eu/cancer-registries-and-environmental-data/>

² <https://eosc4cancer.eu/screening-programmes/>

³ <https://eosc4cancer.eu/multi-omics/>

⁴ <https://eosc4cancer.eu/circulating-dna/>

⁵ <https://eosc4cancer.eu/clinical-decision-support-systems/>

Methodology

The approach we used to construct this report consisted of two steps. First, we collected SOPs, including a step-by-step data access procedure and information about the data models, vocabularies/ontologies and dictionaries/codebooks used by the key datasets in EOSC4Cancer. Subsequently, we grouped the collected information per data type and investigated their similarities and connected it to current work in our project as well as other initiatives. Based on this analysis, we wrote recommendations per data type.

Dataset SOP collection

We created a dataset SOP template and set up a pilot session with BBMRI-ERIC, where we filled out the template for their colorectal cancer cohort (CRC-Cohort)⁶ together to resolve ambiguities and improve the template. Before sending out the updated template to the dataset representatives we identified, we prefilled the SOPs as much as possible with information previously collected in M4.1 (available upon request) and the EOSC4Cancer catalogue⁷.

Data type overview

After most of the dataset SOPs were collected, the data access and interoperability between the datasets containing the same data type were compared to create an overarching overview per data type. Based on this overview, differences were identified and potential harmonisation could be suggested. The collected dataset SOPs are listed below in Table 1.

Dataset name (acronym)	Data type(s)	SOP
Netherlands Cancer Registry (NCR)	Registry	Annex A
Czech National Cancer Registry (CNCR)	Registry	Annex B
Cancer Registry Naples 3 South (NA3CR)	Registry	Annex C
Basilicata Cancer Registry (BCR)	Registry	Annex D
EXPANSE Exposome-NL	Exposome	Annex E
EIRENE-CZ	Exposome	Annex F
Regional Agency for the Protection of the Environment of Campania (ARPAC)	Exposome	Annex G
Regional Agency for the Protection of the Environment of Basilicata (ARPAB)	Exposome	Annex H
Multitarget FIT study (mtFIT)	Screening	Annex I
Piedmont regional screening program (Prevenzione Serena)	Screening	Annex J

⁶ <https://www.bbmri-eric.eu/scientific-collaboration/colorectal-cancer-cohort/> (latest seen 26 January 2024)

⁷ <https://data-catalogue.molgeniscloud.org/catalogue/ssr-catalogue/EOSC4Cancer>

Catalan screening program	Screening	Annex K
BBMRI-ERIC Colorectal Cancer Cohort (CRC-Cohort)	Genomic, Clinical, and Pathology	Annex L
Treatment strategies in colorectal cancer patients with initially unresectable liver-only metastases: CAIRO5	Genomic, Clinical, and Radiology	Annex M
PROgnostic Value of Early Notification by Ctdna in colon cancer stage 3 (PROVENC3)	Genomic and Clinical	Annex N
mCRC-VHIO	Genomic and Clinical	Annex O

Table 1. Overview of key EOSC4Cancer datasets for which SOPs were collected

Standard Operational Procedures

1.1 Exposome

1.1.1 Description/introduction

The exposome aims to capture the non-genetic influences on health and disease. This concept describes environmental exposures that an individual encounters throughout life and how these exposures impact biology and their health. It encompasses both external and internal factors, including chemical, physical, biological, and social factors. An example of an exposome measure could be air pollution, diet or lifestyle. The EIRENE-RI aims to fill the gap in the European infrastructural landscape and to pioneer the first EU infrastructure on human exposome.

Within EOSC4Cancer, exposome data is used at the first step of the patient journey, i.e. cancer risk identification and prevention (Use Case 4.1). A key challenge in cancer prevention is the identification of risk factors. Next to genetic predisposition, exposure to environmental factors and socio-economic characteristics are key determinants. For this purpose, the exposome data is linked to data provided by the cancer registries. Such an approach helps in defining relationships between cancer and environmental contaminants.

1.1.2 Considerations and recommendations for data access

As the exposome data is not personal, the metadata is publicly available. Furthermore, unless under project specific restrictions, the exposome data itself is typically publicly available as well. Nevertheless, we did observe that the data access procedures differ for each exposome dataset used in EOSC4Cancer. Access to the Czech exposome data is provided by the national EIRENE-CZ node upon request. Although Italy and the Netherlands are partner countries of EIRENE and have a national hub (under development), access for these countries is arranged differently. For the Italian exposome data, open access is organised regionally, whereas, in the Netherlands, access requests are organised on a project basis (EXPANSE). Regardless of these differences, it is foreseen that access to exposome data will be managed in a common way by EIRENE-RI in the future.

1.1.3 Considerations and recommendations for (meta)data models and interoperability

At the moment, each exposome dataset used in EOSC4Cancer uses a custom (meta)data model, which hampers (international) interoperability. We expect that the interoperability will improve through the expansion of EIRENE-RI in combination with expected future harmonisation efforts within and between their national nodes.

With EOSC4Cancer Use Case 4.1 in mind, it would be desirable to harmonise, to a certain degree, between the exposome datasets. As the matching of the geo-referenced health data with geo-referenced environmental assessment will play a fundamental role in determining potential relationship(s) between the environmental factors and cancer, it would be suggested to harmonise the geospatial granularity of the exposome measures being linked to cancer registry data. Currently, across different standards (e.g., EIRENE, Expanse), there is no common standard set of environmental factors. Another consideration would be the comparison of exposome measures between the datasets acquired in different countries. As the use of incomparable measurements would make a general recommendation for linkage between relevant exposome and cancer registry data challenging.

1.2 Cancer registries

1.2.1 Description/introduction

Cancer registries are information systems designed for the collection, storage, and management of data on persons with cancer and play a critical role in cancer research, surveillance, cancer prevention and control interventions. These cancer registries can be organised nationally or regionally.

Within EOSC4Cancer, data collected by cancer registries is used at the first step of the patient journey, i.e. cancer risk identification and prevention (Use Case 4.1). A key challenge in cancer prevention is the identification of risk factors. For this purpose, exposome data corresponding to the same region is linked to data provided by the cancer registries. In EOSC4Cancer, cancer registries of particularly polluted areas that use 'geocoding' collection of data were identified to be used for examining spatial patterns of cancer incidence, as well as stage, survival and mortality, and to derive social status indicators (deprivation factors) and environmental characteristics. Such an approach helps in defining relationships between cancer and environmental contaminants.

We noticed that cancer registries are organised differently per country. The Czech and Dutch cancer registries for example are national registries. In Italy, on the other hand, cancer registries are organised locally. The implications for data access and interoperability will be discussed further below.

1.2.2 Considerations and recommendations for data access

All cancer registries present an overview of the cancer types they collect data on and available variables through publicly available metadata.

In the collected SOPs, it was observed that the data access procedure is different for each cancer registry. Furthermore, even between regional cancer registries within the same country, the data access procedures were different due to local regulatory compliance. Access is governed by national laws, thus complicating harmonising access procedures across countries. Hence it is not expected that data access procedures will be harmonised in the lifetime of EOSC4Cancer, and also falls outside of our scope. An international legal agreement on access procedures surpassing or harmonising national regulations is first required.

1.2.3 Considerations and recommendations for (meta)data models and interoperability

The ENCR promotes collaboration between cancer registries, defines data collection standards, provides training for cancer registry personnel and regularly disseminates information on incidence and mortality from cancer in the European Union and Europe.

The Czech National Cancer Registry (CNCR), Basilicata Cancer Registry (CROB), Cancer Registry Naples 3 South (NA3CR), and Netherlands Cancer Registry (NCR) are all members of ENCR. As a consequence, all cancer registry variables provided to EOSC4Cancer that are part of the minimum dataset defined by ENCR adhere to the ENCR recommendations⁸.

With the EOSC4Cancer Use Case 4.1 in mind, it would be desirable to harmonise to a certain degree between the cancer registry datasets. As the matching of the geo-referenced health data with geo-referenced environmental assessment will play a fundamental role in determining the potential relationship between the environmental factors and cancer, it would be suggested to harmonise the geospatial granularity of the cancer registry data being linked to exposome data. According to the ENCR recommendations, this could range from postal code to municipality. Hence, although there are guidelines from the ENCR, these are not strictly enforced, thus there is a lack of a standardised set of attributes. For a harmonisation effort, the NUTS⁹ (Nomenclature of territorial units for statistics) classification defined by EUROSTAT could be used as a standardised reference coding system for residence areas.

⁸ <https://encr.eu/ENCR-Recommendations> (latest seen 26 January 2024)

⁹ <https://ec.europa.eu/eurostat/web/nuts/overview>

1.3 Screening

1.3.1 Description/introduction

Screening programs are used for the early detection of most major tumour types, including colorectal cancer. Early detection and diagnosis play a pivotal role in improving public health outcomes by identifying cancer at its most treatable stages. This allows for more effective and less invasive interventions, significantly increasing the chances of successful treatment and survival. Because of the specific aim and way of acquiring and storing data, screening data is considered as a separate data type, regardless of the data modalities it includes (e.g., clinical data, radiology).

Screening programs involve systematic and regular examinations, such as mammograms, colonoscopies, and Pap smears, tailored to specific cancer types and populations. By identifying abnormalities or precancerous lesions before symptoms manifest, screenings enable timely intervention, reducing the overall burden of the disease and improving the survival rates. Additionally, early detection often translates into more cost-effective healthcare, as treating cancer in its advanced stages tends to be more resource-intensive. Thus, harmonising available representative screening codebooks from Catalunya (Spain), the Netherlands, and Piedmont (Italy) for Use Case 4.2 will contribute to optimised screening strategies, healthier communities, and prolonged, higher-quality lives.

1.3.2 Considerations and recommendations for data access

At the moment this report was written, only the Dutch Multitarget FIT (mtFIT) study had an established data exchange process. This data is available in the EOSC4Cancer reference instance of cBioPortal. Access is granted after the request approval by the Health-RI service desk¹⁰ and the study PI. Unfortunately, a streamlined data access procedure has not yet been established for the Piedmont regional screening program (Prevenzione Serena) and Catalan screening program (ATOS).

Making use of existing resources for data storage and sharing, such as cBioPortal, and linking national coordination centres for data management are key to preserving the collected data for reusability. As commonly there is not a single national coordinating centre, but this varies per specific screening program conducted, a challenge is to harmonise across different screening trials. We can envision that cancer research can be enhanced by making European screening program data accessible using national nodes compliant with local legislation and linking the national nodes into a European network of screening programs enhanced by the EU Cancer Screening Campaign¹¹.

1.3.3 Considerations and recommendations for (meta)data models and interoperability

SOPs from the Catalan screening program (ATOS), the Dutch mtFIT study and Piedmont regional screening program (Prevenzione Serena) were used in the analysis of data model interoperability. All screening programs use custom dictionaries that have overlapping and unshared variables. The main focus of the ongoing efforts on harmonisation is based on the search for commonalities that might lead to the implementation of an ontology inclusive of as many variables and comparable values as possible.

The currently ongoing harmonisation effort of the colon cancer screening programs started by analysing the commonalities between two different studies: mtFIT (NL; 242 variables) and ATOS (ESP, 416 variables). After a careful comparison, a total of 42 variables were identified as common between both studies. The common content was

¹⁰ <https://health-ri.topdesk.net/> (latest seen 26 January 2024)

¹¹ https://cancer-screening.campaign.europa.eu/index_en (latest seen 26 January 2024)

structured following 4 main conceptual blocks: 1) General information, 2) Stool screening (FIT), 3) Colonoscopy (collects information about the colonoscopy procedure and the findings), and 4) Pathology. The work was refined and complemented, including the Italian colon cancer screening program (48 variables). The initial version (V0) of the harmonisation contains 55 variables and is available upon request.

Once the common variables were identified, the second stage was to harmonise the permissible values defined in the data model for each of the screening programs. This work in progress is available upon request as well. With this approach, we identified the commonalities and differences between studies, and we started to work towards achieving a consensus on the final granularity of the data, trying to maintain as much information as possible.

The next stage will cover the identification of the best-suited standard ontologies, the creation of domain-specific terminologies, if needed, as well as different levels of refinement and testing of the model.

1.4 Clinical

1.4.1 Description/introduction

Clinical data encompasses all the information about the patients and their disease, in this case, colorectal cancer. This includes details about their health history, diagnostic procedures, treatments, and outcomes. The data for cancer research is from secondary use, providing privacy protection and anonymity to the patients that conform to it, as well as the ability to be shared with other researchers to accelerate research progress.

In EOSC4Cancer, the nature of clinical data varies based on its origin. Primarily, it constitutes data intended for pheno-clinical analysis, meticulously recorded throughout the patient's cancer journey. These datasets are samples from Electronic Health Records (EHR) and are formatted using data models tailored to the objectives of each study. Furthermore, certain studies within EOSC4Cancer are accessible for visualisation by clinical professionals and researchers through platforms like cBioPortal.

The project identifies four datasets that include clinical data: BBMRI-ERIC CRC-cohort, CAIRO5, PROVENC3, and mCRC-VHIO. The first three are used in use cases 4.3 and 4.4, involving the integration of multi-omics and imaging data (T4.3), and including ctDNA (T4.4). The final study, mCRC-VHIO, serves the purpose of populating the Molecular Tumour Board Portal (MTBP) and other clinical decision support systems.

1.4.2 Considerations and recommendations for data access

The SOPs underscore the pivotal role of a Data Access Committee (DAC) in obtaining access to clinical data. In the clinical data case, it can be found that the data across all studies is either anonymized or pseudo-anonymized. CAIRO5 and PROVENC3 facilitate data access through the DAC, allowing visualisation, exploration, and in-depth analysis of the displayed data in cBioPortal. Conversely, mCRC-VHIO has direct access to the data through the DAC.

A notable distinction arises within CRC-Cohort from BBMRI, where users initially engage with metadata for discovery purposes. Following authentication, users can then query the anonymized dataset. For conclusive access to the data, authorization is required through the DAC, ensuring a structured and secure process for data retrieval. As a centrally deposited cohort (from different contributing data owners) controlled at BBMRI-ERIC users go through an expedited procedure with timely limited vetoing possibility for original data holders to get access.

1.4.3 Considerations and recommendations for (meta)data models and interoperability

Data models play a fundamental role in facilitating the exchange and interoperability of clinical data. And, their crucial role is augmented by the inclusion of metadata models

that sometimes are combined with the data models. Their significance lies in providing a structured framework that, when coupled with comprehensive documentation, grants researchers a clear understanding of the dataset, allowing them to combine it with other datasets. Moreover, through the incorporation of standard ontologies, vocabularies, and accompanying codebook dictionaries, they can be easily converted to various formats, enabling easy sharing and utilisation for machine-readable purposes. Also, the inclusion of metadata models further refines the contextual information associated with clinical data, facilitating more comprehensive analysis and making it more machine-understandable.

Within the clinical studies found in EOSC4Cancer, there are different clinical data models. For further visualisation and exploration of data, the cBioPortal data model in CAIRO5 and PROVENC3 is found. These studies share the same vocabulary standard for tumour classification within the Netherlands, with other vocabularies/ontologies specific to the type of data they are describing. Also, there is a notable overlap in codebook content between them.

The clinical data model of the CRC-Cohort in BBMRI-ERIC allows conversion to OMOP-CDM, openEHR or HL7-FHIR, increasing the interoperability to other formats. This model shares the SNOMED-CT ontology with the PROVENC3 model, a widely recognised standard for diseases and treatments. In contrast, mCRC-VHIO has a custom data model with a flexible vocabulary which makes it more cumbersome for researchers to use it together with other datasets.

In parallel projects related to EOSC4Cancer, such as AACR GENIE, B1MG and IMPaCT-Data, similar clinical data models are present. All these projects demonstrate high interoperability with either OMOP CDM or cBioPortal, offering the potential for adaptation and harmonisation to enhance datasets within the EOSC4Cancer initiative. Nevertheless, each data model has their own focus. Project GENIE by AACR is creating a pan-cancer solid tumour data model (currently including twelve cancer types). There are four working groups represented in the model, i.e. diagnosis, treatment, response/outcomes and social determinants of health. IMPaCT-Data has developed a data model based on publicly available TCGA data to have a consensus among the partners of their project and interoperability with the Beacon model. Consequently, close collaboration is established with these initiatives, engaging the process of clinical data harmonisation based on their proposals.

Finally, outside the scope of EOSC4Cancer, there are other initiatives trying to fit the cancer data in the clinical domain. As an example, there is the Oncology extension from OMOP-CDM that adds the concept of episodes to the table.

1.5 Genomics

1.5.1 Description/introduction

Genomic data is critical in advancing cancer research and the overall patient experience. Studying an individual's genetic makeup through genomics provides crucial insights into the molecular mechanism of cancer, aiding researchers in understanding the specific genetic alterations driving the disease. This knowledge not only facilitates the development of targeted therapies but also enhances the precision and efficacy of treatment strategies, minimising adverse effects. Furthermore, genomic data contributes to the developing field of personalised medicine, enabling healthcare professionals to tailor treatments based on the unique genetic profile of each patient. This personalised approach enhances treatment outcomes while minimising unnecessary interventions. From a patient perspective, access to genomic information empowers individuals with a deeper understanding of their cancer risk, prognosis, and treatment options. This knowledge allows for informed decision-making, fostering a sense of control and involvement in one's healthcare journey. Overall, integrating genomic data in cancer

research and patient care represents a transformative paradigm, promising more effective treatments and improved outcomes.

Genomic data in the context of cancer research belongs to four different categories based on the degree of computational analysis, in ascending order: raw, processed or normalised, interpreted, and summarised¹². While summarised genomic data can be openly disseminated, the first three categories of genomic data are protected by personal data laws. Therefore, the major archival centres, such as EGA in Europe and dbGaP in the USA, use controlled access for authorised researchers only.

In the context of EOSC4Cancer, use cases 4.3 and 4.4 will deliver standardised templates for the complex and longitudinal data handling in studies investigating localised cancer applicable in the framework of experimental precision oncology projects. We focused on datasets from the BBMRI-ERIC CRC cohort, CAIRO5, and PROVEN3 studies to provide data type-specific analysis of data structure and exchange.

1.5.2 Considerations and recommendations for data access

The access to metadata varies considerably from public (non-authenticated) high-level search and browsing access provided by BBMRI-ERIC for their CRC-Cohort, PROVEN3 and CAIRO5 to authenticated access only for the mCRC-VHIO cohort.

For access to anonymized or pseudonymized data, each of the four datasets requires authentication. Once this access has been granted, we observed two different options. The data can either be shared, as is the case for mCRC-VHIO and the BBMRI-ERIC CRC-Cohort, or the data will be shared through cBioPortal. Within cBioPortal the researcher can query and visualise the data.

Besides the processed data from the CAIRO5 and PROVEN3 datasets being available in the EOSC4Cancer reference instance of cBioPortal, the raw sequencing data for CAIRO5 is accessible in EGA as well. Once the raw sequencing data has been imported to EGA (or another) archival repository, this will allow the researchers to investigate the available data to find undiscovered associations across datasets and lead to more discoveries using the existing datasets accessible on one platform.

1.5.3 Considerations and recommendations for (meta)data models and interoperability

Both CAIRO5 and PROVEN3 datasets follow the cBioPortal data model providing processed data in MAF format, while the mCRC-VHIO and BBMRI-ERIC CRC cohorts are structured according to a custom data model. The BBMRI-ERIC data model can be converted to openEHR format for improved interoperability¹³ and a conversion to the cBioPortal data model will be available in the near future. Besides the processed data formats, the raw data in the CAIRO5 is available in the EGA data format (FASTQ) and post mapping format (BAM).

In our opinion, considering the dawn of hybrid sequencing projects thanks to the advancement of long-read sequencing^{14,15,16} raw data from the published projects might become prominent for hybrid data reprocessing. Therefore, the availability of raw datasets is paramount. However, considerations need to be taken into account regarding the quality of the existing data as well as the granularity of the metadata.

1.6 Radiology

1.6.1 Description/introduction

¹² <https://doi.org/10.1101/gad.2017311>

¹³ <https://doi.org/10.1093/jnci/djh034>

¹⁴ <https://doi.org/10.1080/15476286.2023.2220210>

¹⁵ <https://doi.org/10.1515/mr-2021-0013>

¹⁶ <https://doi.org/10.1038/s10038-019-0658-5>

Radiological imaging data (e.g., CT, MRI, PET, Ultrasound) is an essential step in the diagnostic work-up and follow-up of most cancer types. Radiological scans are relatively non-invasive, cheap, quick, and give an overview of a large field of view, being able to capture tumours and their surroundings. Therefore, radiology is nearly always used in diagnostics and is one of the primary modalities for treatment response monitoring. Due to the wide arsenal of imaging options, various characteristics of the patient can be characterised. Moreover, within AI for Healthcare, radiology is at the front, with the majority of FDA-approved solutions currently in radiology.

Within EOSC4Cancer, radiology is only used within one dataset (CAIRO5) within one use case (T4.3). In this study, CT and MRI data are available for patients with colorectal cancer for both diagnostics and follow-up monitoring. Research includes combining imaging data, ctDNA and genomics, and clinical data to create biomarkers. Hence, the main aim of EOSC4Cancer is to connect radiological data to these other data types. Due to its widespread use in oncology, this linkage can majorly impact research in other diseases as well.

1.6.2 Considerations and recommendations for data access

Due to having a widely adopted file and metadata model standard (DICOM), access procedures are quite standardised. Public datasets also exist, for example The Cancer Imaging Archive (TCIA)¹⁷. Access is generally handled on a project / dataset basis, with the PI or a DAC approving the access, including the researcher having to accept a licence or terms of use. Afterwards, imaging data can simply be accessed and downloaded. Different access levels can however be identified, which we illustrate with above examples from EOSC4Cancer.

A first consideration should be given to the level of access to metadata. The lowest level is to access the image only and no metadata, except metadata essential to interpretation of the image such as pixel spacing. To this end, data is commonly converted to a more convenient format such as NIfTI (Neuroimaging Informatics Technology Initiative). Data formats such as NIfTI allow compression and are easier to interpret due to the limited metadata model. Alternatively, a selection of metadata can be shared either in addition to a NIfTI or more conveniently through the DICOM format. Many DICOM headers are standardised and thus can be safely shared. However, various fields are not standardised and/or allow free text, requiring careful curation to prevent privacy or security issues. Since these are vendor-dependent, harmonisation of these tags falls outside the scope of EOSC4Cancer; EUCAIM is making an effort on curation protocols to standardise extraction of information from these tags without damaging security. See also the anonymization process described in section 1.6.3. A selection of DICOM metadata is typically shared through (public) catalogues to allow findability of the imaging data. Additionally, to discover how much data a dataset contains for a specific study and compile the relevant dataset, data controllers can allow authenticated, anonymized (federated) queries based on additional metadata.

A second consideration should be given to the sequences and derived data that can be accessed. In the XNAT repository used for the radiological imaging storage in the CAIRO5 trial, different levels are defined regarding this aspect. The XNAT data model includes a Project - Subject - Experiment (e.g., a CT scanning Session) - Scan structure, corresponding to different access levels. In this way, individual scans can be shared, or all sequences of all subjects in a project, and so on.

In this way, by having flexible control over the amount of metadata and data that can be shared, sharing can be more easily tuned to different use-cases or studies, and thus fostering the possibilities for sharing.

1.6.3 Considerations and recommendations for (meta)data models and interoperability

¹⁷ https://imaging.cancer.gov/informatics/cancer_imaging_archive.htm (latest seen 26 January 2024)

Regarding data models, due to the wide acceptance of DICOM, adopting the DICOM standard for both the file format as well as the repository metadata model is highly recommended. This facilitates FAIR data, as well as interoperability between datasets and repositories. For example, the XNAT repository used in the CAIRO5 study is heavily DICOM-based, with native support for, e.g., exposing DICOM tags through the REST API. Recommendations are also given from the five projects within EUCAIM¹⁸.

Several challenges and discrepancies, however, exist. First, commonly, data is anonymized or pseudonymized before sharing. There is no legal definition currently for when DICOM data is anonymized, and while there are some accepted standards in the field for how to treat certain aspects, there is variety in this aspect. For example, private tags should be removed as these can, in theory, contain any data and thus also privacy-sensitive data. However, these tags can also contain crucial information on the images, e.g., B-values of DWI MRI are commonly stored in private tags. The way such values are stored depends on the vendor, and various parties have worked on defining validated exceptions for retrieving this data without sacrificing patient privacy. Examples are TCIA, which has extensive experience in storing data and thus have adopted its anonymization protocols as such, and the Horizon EUCAIM project, which is working on harmonising the protocols of five other Horizon Projects (EuCanImage, PRIMAGE, CHAMELEON, ProCancerI, and INCISIVE). Both projects use the RSNA-endorsed Clinical Trial Processor (CTP) software¹⁹, defining specific configurations to catch these exceptions (the TCIA has included CTP in their POSDA solution). Health-RI, the partner in EuCanImage, has also published various CTP configurations from various Dutch clinical studies²⁰. We recommend adopting such standards and expanding upon them if other exceptions are encountered.

Second, DICOM does not include any semantic annotations. Most noticeable is the lack of identifying specific imaging protocols, e.g., T2-weighted MRI, T1-weighted MRI, and DWI MRI. Examples of models that try to caption such concepts are RSNA's RadLex²¹ and DICOM-MIABIS²². See also the work done in EUCAIM to map imaging metadata models²³.

1.7 Pathology

1.7.1 Description/introduction

Pathology data obtained from biopsies or resection often serves as the golden standard for diagnosis and treatment response assessment, and is therefore a crucial step in clinical practice. Commonly, the term pathology is used to refer to fresh frozen tissue samples and/or formalin-fixed paraffin-embedded (FFPE) samples. These samples provide a valuable resource for obtaining additional data, such as genetic testing, or applying different stainings for molecular analysis, e.g., hematoxylin and eosin staining. Digital pathology commonly refers to the digitization of glass whole slide imaging (WSI). In recent years, digital pathology has gained increased attention due to the potential use of AI on the digitised images.

In the context of EOSC4Cancer, we will therefore focus our attention on the digital pathology imaging data, which is used in the BBMRI-ERIC colorectal cancer (CRC) cohort. In this study, data of 10,000+ CRC cases has been collected, mostly as part of

¹⁸ <https://doi.org/10.1186/s41747-023-00336-x>

¹⁹ https://mirwiki.rsna.org/index.php?title=MIRC_CTP (latest seen 26 January 2024)

²⁰ https://github.com/CTMM-TraIT/traic_ctp (latest seen 26 January 2024)

²¹ <https://radlex.org/>

²² <https://doi.org/10.1186%2Fs41747-021-00214-4>

²³ <https://doi.org/10.1186/s41747-023-00336-x>

the Horizon 2020 funded ADOPT BBMRI-ERIC project. The dataset serves as a use case for piloting access to European biobanks, and thus can be shared. Part of the pathology data has been digitised and is stored at BBMRI-ERIC. Besides digital pathology, the dataset also contains data on various (manually scored) pathological markers. The CRC cohort also contains clinical and genomic data, which have been addressed in earlier sections of this deliverable (1.4 and 1.5).

1.7.2 Considerations and recommendations for data access

The BBMRI-ERIC CRC-Cohort has an extensive access and data protection policy²⁴. Access can be requested to specific data types, e.g., the tissue samples or digitised pathology data. For the digital pathology data, access is similar to the radiology data, hence we refer to section 1.6.2. For tissue samples, access requests are generally more complex, as physical data needs to be shared, which falls outside the scope of EOSC4Cancer.

1.7.3 Considerations and recommendations for (meta)data models and interoperability

Currently, there is no single accepted data format, data model, or metadata model for digital pathology data. These can vary from more standard image formats such as JPG or Tiff, to vendor specific formats. This poses a major challenge for harmonisation and integration of digital pathology data. Within EOSC4Cancer, we therefore align with the efforts from the Horizon 2020 BigPicture project²⁵, which will create a central digital pathology repository for AI studies.

BigPicture has adopted the DICOM standard from radiology. Recommendations for a unified open digital slide and annotation format and the underlying investigations that have been performed when designing this format are already publicly available²⁶. This format is based on the submission format schema used by the European Genome-phenome Archive (EGA) that has been expanded to include functionality to support pathology imaging using the DICOM standard. Additionally, a Common Mandatory Metadata Structure (CMMS) comprising all metadata bearing entities was developed by BigPicture. Similar to the image format, the CMMS was based on the EGA Metadata Model²⁷, adapted to fit the needs of a repository storing digitised pathological images. It is designed in a way that allows easy and automatic data extraction from various Laboratory Information Management Systems (LIMSs) and Clinical Information Systems (CISs) while at the same time takes manual data entry as well as the usage of data from a researcher perspective into account.

By adopting a common data and metadata model such as the one from BigPicture, harmonisation and interoperability of digital pathology data is more easily facilitated. Additionally, usage of the DICOM format enables usage of DICOM-oriented solutions from radiology, e.g., repositories, integration with other data types. BigPicture also provides a DICOM converter with support of 11+ image format from different vendors, which includes the popular OpenSlide²⁸ library.

²⁴ <https://zenodo.org/doi/10.5281/zenodo.7513755>

²⁵ <https://bigpicture.eu/>

²⁶ https://bigpicture.eu/sites/default/files/2023-04/945358-BIGPICTURE_D4.03_Report%20on%20unified%20open%20digital%20slide%20and%20annotation%20format%20specification.pdf (latest seen 26 January 2024)

²⁷ <https://ega-archive.org/submission/tools/submitter-portal/> (latest seen 26 January 2024)

²⁸ <https://openslide.org/>

Discussion

We found that the data access procedures and data interoperability (e.g. data models, vocabularies/ontologies and dictionaries/codebooks) reported in the collected SOPs vary considerably between datasets, even within the same data type. For most of these data types, data sources from multiple countries were included to capture as much heterogeneity in data access procedures and interoperability as possible. Nevertheless, it is not possible to describe the full scope of the complex and diverse cancer data landscape encompassed by the patient journey with just these datasets. For additional input, we reached out to other European and non-European projects and research infrastructures focused on specific data types. The information collected from these sources external to EOSC4Cancer complemented the recommendations on the data access and interoperability described in the data type SOPs.

Next steps

For future work, two lines of work have been identified. First, a commitment to integrate the accessibility procedures and information collected about data models, vocabularies/ontologies and dictionaries/codebooks for all SOPs into the EOSC4Cancer catalogue provided by MOLGENIS in WP1. Additionally, we are looking into integrating our SOPs in RDMKit Cancer View, which may require a conversion step. Finally and most important, a general effort is made to improve harmonisation and interoperability procedures for each distinct data type. In the lines below, the specific considerations for the different data types are explained.

There is currently no standard access procedure or general interoperability for exposome data. As EIRENE-RI is the research infrastructure dedicated to exposome data, it would be beneficial to align any harmonisation efforts between our respective consortia to avoid duplicate work.

The access procedures for each of the cancer registries differ, and this would be difficult to change due to national and/or regional regulatory compliance. There is an interoperability effort ongoing through the ENCR, which harmonises the basic variables shared by the cancer registries that are part of the ENCR. Harmonisation of additional variables could be useful but should be aligned with ENCR to prevent potential duplicate work.

While the access procedures for the screening data are not established for all available data sets, we would suggest adopting the example of the Dutch mFIT study and their national RI-cBioPortal use. This could further enhance the ongoing harmonisation efforts that will next cover the identification of the best-suited standard ontologies and model testing.

For clinical data, a valuable improvement would be to optimise the interoperability and visualisation in EOSC4Cancer through identified clinical data models. Simultaneously, adaptation of clinical data models from AACR GENIE, BIMG, and IMPaCT-Data will be explored for harmonisation with cBioPortal, each with unique focuses.

The shifting focus of cancer genomic research from processed datasets to raw data might highlight the necessity of raw cancer data accessibility and (meta)data visualisation. In this regard, further efforts should be exerted on the visualisation of data on the cBioPortal platform and controlled access via EGA.

International consensus standards should be developed on the anonymization of DICOM tags and on semantic annotations, thus requiring an extension of the metadata model for radiological data. Addressing these challenges will facilitate harmonisation of radiological data and interoperability with other data types.

Standardisation of the data format and data model remains the main challenge for harmonisation of pathology data. Currently, the BigPicture standard, using DICOM as data model expanded with both a metadata model for DICOM and for the hierarchy of pathology (derived) data seems the most promising solution. Large-scale validation is required to test this standard and potentially expand it to address the current state of diversity in pathology data. Collaboration with vendors is key for this effort to facilitate standardisation at the source.

Conclusion

In conclusion, the outlined strategies for exposome data, cancer registries, screening codebooks, clinical data, genomic data, radiology, and pathology provide a roadmap for enhanced collaboration within and beyond EOSC4Cancer. As we look ahead, the focus on incorporating SOPs into the centralised catalogue and fine-tuning data harmonisation for specific types across institutions in different countries reflects our commitment to advancing interoperability in cancer research. We will continue to align our harmonisation efforts with the projects and initiatives of the relevant communities, such as the ones included in this deliverable. This approach ensures that EOSC4Cancer continues to evolve as a dynamic and collaborative project, fostering innovation and progress in cancer data for all steps of the cancer journey.

Glossary of international coalitions on data harmonisation and integration

This glossary provides an overview of international coalitions on data harmonisation and integration which were investigated in defining the SOPs in this deliverable. These include (clinical) organisations, research infrastructures, (software) communities, consortia, and projects (e.g., EU-funded).

AACR

American Association for Cancer Research (AACR²⁹) is a cancer research organisation that focuses its programs and services on fostering cancer research in cancer and related biomedical science. AACR disseminates new research findings and promotes science education and training. The organisation now has more than 52,000 members in 130 countries and territories.

AACR GENIE

Project GENIE^{30,31} (Genomics Evidence Neoplasia Information Exchange) inside AACR is a publicly accessible cancer registry designed to advance precision medicine in oncology. With data shared among 19 leading international cancer centres, it compiles real-world clinico-genomic information from diverse patients. GENIE prioritises openness, aiming to expedite drug discovery, refine clinical trials, and benefit cancer patients globally by providing a shared resource for the global cancer research community

Beacon

Beacon is an GA4GH standard protocol for implementing data discovery services focused on biomedical genomics, enhancing the privacy of the patient and sharing of data. This allows researchers to check the accessibility of specific genomic variants or clinical data across diverse databases without revealing any individual data.

BigPicture

BigPicture³² is a Horizon Europe funded project working on a central repository of digital pathology slides to boost the development of artificial intelligence. As part of this, BigPicture is working on whole slide imaging (WSI) standards for the image format, data repository, and metadata model. Besides using these within the central repository of the project, their aim is to facilitate harmonisation and integration of other data resources of other researchers.

B1MG

B1MG³³ is a 1+ Million Genomes Initiative (1+MG) initiative project that helps to create a network of genetic and clinical data across Europe. Collaborating with regional, national, and European stakeholders, the initiative seeks to outline the prerequisites for seamless cross-border access to genomics and personalised medicine. Additionally, B1MG aims to add technical specifications and implementation guidelines for the relevant data types involved in the initiative.

²⁹ <https://www.aacr.org/>

³⁰ <https://www.aacr.org/professionals/research/aacr-project-genie/>

³¹ <https://doi.org/10.1158/2159-8290.cd-21-1547>

³² <https://bigpicture.eu/>

³³ <https://b1mg-project.eu/>

cBioPortal

cBioPortal³⁴ is an open-access, open-source resource for interactive exploration of multidimensional cancer genomics data sets. The portal supports and stores data sets of information including non-synonymous mutations, DNA copy-number data, mRNA and microRNA expression data, protein-level and phosphoprotein level data (RPPA or mass spectrometry-based), DNA methylation data, and de-identified clinical data. Within EOSC4Cancer, the Health-RI cBioPortal³⁵ is used.

dbGaP

The database of Genotypes and Phenotypes (dbGaP³⁶) is an NIH-maintained database of datasets that was developed to archive and distribute the results of studies that have investigated the interaction of genotype and phenotype. dbGaP provides two types of data - open access and controlled access.

EGA

The European Genome-phenome Archive (EGA³⁷) is a global network for the permanent archiving and sharing of personally identifiable genetic, phenotypic, and clinical data generated for the purposes of biomedical research projects or in the context of research-focused healthcare systems. EGA is managed by the European Bioinformatics Institute (EMBL-EBI) in Cambridge (UK) and the Centre for Genomic Regulation (CRG) in Barcelona. They aim to advance biomedical research and promote personalised medicine worldwide by enabling discovery of and access to human genomic and health research data.

EIRENE-RI

Environmental Exposure Assessment Research Infrastructure (EIRENE³⁸) is a pioneering EU infrastructure on human exposome that aims to fill the gap in the European infrastructural landscape. EIRENE RI was designed as a geographically balanced network of distributed research infrastructures consisting of 17 National Nodes representing ca. 50 institutions with complementary expertise.

ENCR

The European Network of Cancer Registries (ENCR³⁹) is a framework established in 1990 within the Europe Against Cancer Programme of the European Commission. The ENCR promotes collaboration between cancer registries, defines data collection standards, provides training for cancer registry personnel and regularly disseminates information on incidence and mortality from cancer in the European Union and Europe.

EUCAIM

EUCAIM⁴⁰ is a Horizon funded project focused on establishing a research infrastructure for cancer imaging. It consists of over 70 partners, primarily from a combination of five preceding EU projects on similar projects organised in AI for Health Imaging (AI4HI): EuCanImage, PRIMAGE, CHAIMELEON, Incisive, and ProCancer-I. As part of establishing a research infrastructure, EUCAIM

³⁴ <https://www.cbioportal.org/>

³⁵ <https://www.health-ri.nl/services/cbioportal>

³⁶ <https://www.ncbi.nlm.nih.gov/gap/>

³⁷ <https://ega-archive.org/>

³⁸ <https://www.eirene-ri.eu/>

³⁹ <https://www.encr.eu/>

⁴⁰ <https://cancerimage.eu/>

is working on defining a hyper-ontology and metadata models for harmonising cancer imaging data, primarily radiology.

EU Cancer Screening Campaign

An overarching European initiative with the goal of achieving a 90% screening rate for eligible individuals in breast, cervical, and colorectal cancer by 2025. This campaign is geared towards combating cancer through a multifaceted approach, encompassing prevention, early detection, treatment, and comprehensive care, with a specific emphasis on improving the quality of life post-diagnosis. Additionally, the collaboration with EOSC4Cancer and the established network in this project can enhance the utilisation of common data models and data sharing practices.

EXPANSE

EXPANSE⁴¹ is a five-year European research project that focuses on the urban exposome and involves 20 academic and non-academic partners located in 14 European countries and the USA. The project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 874627 and is coordinated by Utrecht University.

Health-RI

Health-Research Infrastructure (Health-RI⁴²) is a Dutch national coordination centre for agreements on the reuse of health data, promoting collaboration among all stakeholders, and supporting researchers. The cBioPortal instance⁴³ hosted by Health-RI serves as the EOSC4Cancer reference instance and hosts the data made available through use case 4.4. The integration work is currently ongoing in Work Packages 2 and 3.

HL7-FHIR

FHIR⁴⁴ (Fast Healthcare Interoperability Resource) is an interoperability standard made by HL7 (Health Level 7 suite of standards). This standard facilitates basic electronic exchange of healthcare data across diverse systems within the healthcare industry. Notably, FHIR prioritises simplicity in implementation while ensuring the integrity of exchanged data between healthcare applications.

IMPACT-Data

IMPACT-Data⁴⁵ is a Spanish personalised medicine project with the objective of fostering a unified, interoperable, and integrated system for gathering and analysing clinical and molecular data. It leverages the knowledge and resources within the Spanish Science and Technology System to facilitate research inquiries using various clinical and molecular information systems. Moreover, it integrates tools like cBioPortal and Beacon, widely utilised in EOSC4Cancer, to allowing the accessibility and collaboration in cancer research.

⁴¹ <https://expansoproject.eu/>

⁴² <https://www.health-ri.nl/en>

⁴³ <https://cbioportal.health-ri.nl/login.jsp>

⁴⁴ <https://www.hl7.org/fhir/>

⁴⁵ <https://impact-data.bsc.es/>

Molecular Tumour Board Portal (MTBP)

Integrated platform^{46,47} maintained by the Karolinska Institutet, serving as a singular entrance for the interpretation of gene variants' functional and predictive relevance through interactive reports. This portal employs distinct supporting evidence levels, annotating variants using comprehensive knowledge bases, and ensuring alignment with the consensus of clinical experts for molecular tumour analysis.

OMOP-CDM

OMOP CDM⁴⁸ (Observational Medical Outcomes Partnership Common Data Model) is a standardised data model that provides a common framework for organising and representing healthcare data from diverse sources, facilitating interoperability and enabling large-scale observational research. The model ensures consistency in data structure, allowing researchers to analyse healthcare data comprehensively for various studies and analyses.

openEHR

openEHR⁴⁹ is an open standard for electronic health records (EHR) that focuses on creating flexible and interoperable frameworks. It provides a standardised approach to structure, store, and exchange health data, allowing for seamless integration across different healthcare systems.

RDMkit

The ELIXIR Research Data Management Kit (RDMkit⁵⁰) is a collection of guidelines for Data Stewards, Policy makers, Researchers, Research Software Engineers, and Trainers in life sciences in their efforts to manage research data following the FAIR Principles better. RDMkit is based on the various steps of the data lifecycle. The contents are generated and maintained by the ELIXIR community and RDMkit Alliance⁵¹. RDMkit is recommended in the Horizon Europe Program Guide⁵² as the "resource for Data Management guidelines and good practices for the Life Sciences."

SNOMED-CT

SNOMED-CT⁵³ is a comprehensive and standardised clinical terminology used in healthcare. It provides a structured and universal language for capturing, sharing, and exchanging health information across different clinical settings and systems. SNOMED-CT plays a crucial role in enhancing interoperability and communication in the healthcare domain by offering a common vocabulary for describing clinical concepts and relationships.

TCIA and TCGA

The Cancer Imaging Archive (TCIA⁵⁴) and The Cancer Genomics Atlas (TCGA⁵⁵) are two primarily NIH funded American organisations facilitating large scale storage of imaging (TCIA) and genomics

⁴⁶ <https://www.mtbp.org/about.php>

⁴⁷ <https://doi.org/10.1038/s43018-022-00332-x>

⁴⁸ <https://www.ohdsi.org/data-standardization/>

⁴⁹ <https://openehr.org/>

⁵⁰ <https://rdmkit.elixir-europe.org/>

⁵¹ https://rdmkit.elixir-europe.org/rdmkit_alliance

⁵² https://ec.europa.eu/info/funding-tenders/opportunities/docs/2021-2027/horizon/guidance/programme-guide_horizon_en.pdf

⁵³ <https://www.snomed.org/>

⁵⁴ https://imaging.cancer.gov/informatics/cancer_imaging_archive.htm

⁵⁵ <https://portal.gdc.cancer.gov/>

(TCGA) data, the latter including digital pathology. As part of this, both programmes are developing standards for data anonymization, curations, and metadata models. Both also support a catalogue to make data findable.

XNAT

XNAT⁵⁶, the extensible neuroimaging archive toolkit, is an open-source solution designed for sharing and storing medical imaging data, especially radiology data. The XNAT metadata model is heavily DICOM based, thus providing support and functionality for working with DICOM. XNAT was acquired by the company Flywheel in 2022 and has a highly active community, including regular workshops and development of additional plugins by various groups. Within EOSC4Cancer, the Euro-Biomedicine / BMIA XNAT⁵⁷ hosted by Health-RI will be used in the CAIRO5 study.

⁵⁶ <https://doi.org/10.1385/NL:5:1:11>

⁵⁷ <https://xnat.bmia.nl/>

Annex A - NCR SOP

Standard operating procedure for minimal data exchange for the Netherlands Cancer Registry.

Dataset Name (acronym):	Netherlands Cancer Registry (NCR)
Description of the Dataset:	The NCR compiles clinical data of all individuals newly diagnosed with cancer in the Netherlands. Data in the NCR are registered from initial diagnosis up to and including the completion of first-line treatment.
Data Type(s):	Cancer registry
EOSC4Cancer Use Case:	T4.1
Date:	21-11-2023

A. Purpose

This SOP describes the steps required for cancer registry data from the NCR to be accessed and used by researchers in the context of the EOSC4Cancer project and beyond. Based on this information, we will define and suggest broader recommendations for data exchange within and beyond the EOSC4Cancer project. Below you will find an overview of the data models, ontology, vocabulary and dictionaries used by the dataset. Furthermore, data access procedures are described step-by-step.

B. List of definitions and abbreviations

Definitions	
CRC	Colorectal cancer
IKNL	Netherlands Comprehensive Cancer Organisation

C. Data Models and ontology, vocabulary, dictionaries

Registry	Name	Description	References	Is this a standard?	Available conversions for harmonisation
Data Models	OMOP CDM	A part of the NCR is available in OMOP: - ICD-O-3 diagnosis - TNM classification - Location of metastases - First-line treatment - Death	https://as-nkr-catalogus-2023-prod.azurewebsites.net/omop.html	Yes	NA
Vocabulary/ Ontology	OMOP CDM vocabularies	Vocabularies used in the OMOP CDM (for variables available	https://github.com/OHDSI/Vocabulary-v5.0/wiki	Yes	NA

		in OMOP CDM)			
Dictionary/C odebooks	Data dictionary CRC	Data dictionary used for CRC (in Dutch)	https://as-nkr-catalogus-2023-prod.azurewebsites.net/CRC.html	Yes (in the Netherlands)	SNOMED-CT (in English)
Formats	csv	Contains parameters that were requested			

D. Description of the data exchange process

Step	Action	Responsible (Name, email)	Link	Estimated time
1	Fill in the application form and send it to IKNL	Researcher	https://gegevensaanvrage.iknl.nl/gegevensaanvraag	Between 2 weeks and 6 months
2	Go over the details of the research question, methodology and the selection of data	Researcher and IKNL		
3	Submit the data request for approval by the privacy board (CvT) and a scientific committee	IKNL		
4	If necessary, guide the linking process with third parties such as PALGA, DHD or cohorts	IKNL		
5	Make a quotation if necessary	IKNL		
6	Make sure the researcher can formally agree to the terms and conditions	IKNL		
7	Retrieve, transform and format the data and set up a data dictionary (after the researcher agrees to the quotation)	IKNL		
8	Check the data	IKNL		
9	Deliver the data to the researcher through a highly secure environment (with the sole purpose of safe data transfer, i.e. it is not possible to analyse data in this environment)	IKNL		

E. Related documents and links

- Latest version of the data request guidelines: <https://iknl.nl/en/ncr/apply-for-data>

Annex B - CNCR SOP

Standard operating procedure for minimal data exchange for the Czech National Cancer registry.

Dataset Name (acronym):	Czech National Cancer Registry (CNCR)
Description of the Dataset:	Czech National Cancer Registry (National Oncology Registry - NOR) is registry of oncological diseases, which periodically monitor them and their development in time. UZIS is in charge of data collection, verification, storage, protection and processing. The CNCR provides aggregated data for statistical surveys at both national and international levels, as well as for epidemiological studies and medical research. CNCR is a population-wide registry that follows the monitoring of neoplasms in the population of the Czech Republic. It was introduced in the 1950s and has been operated as a population registry of records of individual neoplasms by the UZIS since 1976.
Data Type(s):	Cancer Registry
EOSC4Cancer Use Case:	T4.1
Date:	04-12-2023

A. Purpose

This SOP describes the steps required for cancer registry data from NOR to be accessed and used by researchers in the context of the EOSC4Cancer project and beyond. Based on this information, we will define and suggest broader recommendations for data exchange within and beyond the EOSC4Cancer project. Below you will find an overview of the data models, ontology, vocabulary and dictionaries used by the dataset. Furthermore, data access procedures are described step-by-step.

B. List of definitions and abbreviations

Definitions	
UZIS	Institute of Health Information and Statistics of the Czech Republic
CNCR	Czech National Cancer Registry

C. Data Models and ontology, vocabulary, dictionaries

Registry	Name	Description	References	Is this a standard?	Available conversions for harmonisation
Data Models	NA				
Vocabulary/ Ontology	NA				
Dictionary/Codebooks	Dataset distribution for lung cancer	Custom for standardised incidence of lung cancer	https://data-catalogue.molgeniscloud.org/CNCR/tables/#/DataSources	Yes (in Czech republic)	No
Formats	csv				

D. Description of the data exchange process

Step	Action	Responsible (Name, email)	Link	Estimated time
1	Create a data export request	Researcher	https://www.uzis.cz/index.php?pg=registry-sber-dat--narodni-zdravotni-registry--narodni-onkologicky-registr	
2	Review the request	UZIS		
3	Make a selection from the CNCR according to the requested selection criteria (if the request was approved)	UZIS		
4	Send the exported data securely and encrypted	UZIS		

Annex C - NA3CR SOP

Standard operating procedure for minimal data exchange for the Cancer Registry Naples 3 South.

Dataset Name (acronym):	Cancer Registry Naples 3 South (NA3CR)
Description of the Dataset:	The NA3CR collects clinical data from all individuals newly diagnosed with cancer in their own geographical area of reference. The data in the NA3CR is recorded from the initial diagnosis until healing or death of the patient with biennial follow-ups; (see ENCR recommendation https://encr.eu/sites/default/files/Recommendations/ENCR-Recommendation-standard-dataset_Mar2023.pdf)
Data Type(s):	Registry
EOSC4Cancer Use Case:	T4.1
Date:	18-12-2023

A. Purpose

This SOP describes the steps required for cancer registry data from NA3CR to be accessed and used by researchers in the context of the EOSC4Cancer project and beyond. Based on this information, we will define and suggest broader recommendations for data exchange within and beyond the EOSC4Cancer project. Below you will find an overview of the data models, ontology, vocabulary and dictionaries used by the dataset. Furthermore, data access procedures are described step-by-step.

B. List of definitions and abbreviations

Definitions	
CRC	Colorectal cancer
LC	Lung cancer

C. Data Models and ontology, vocabulary, dictionaries

Cancer Registry	Name	Description	References	Is this a standard?	Available conversions for harmonisation
Data Models	International Classification of Diseases for Oncology (ICD-O-3), third edition	<p>DATA SET NA3CR:</p> <ul style="list-style-type: none"> · Gender · Date of birth · Date of death · Postal code · Patient data · Age at diagnosis · Gender · Vital status <p>TUMOUR:</p> <ul style="list-style-type: none"> · data incidence: year and month · Basis for diagnosis · Topography · Morphology · Behaviour · Stage · TNM - Localisation including sub localisation metastases <p>Addition in lung tumours*</p> <ul style="list-style-type: none"> · Tumour size · Multifocality · Cytogenetic or molecular abnormalities in the tumour · PDL-1 result <p>Addition in colorectal tumours*</p> <ul style="list-style-type: none"> · Multifocal tumour - Molecular diagnostics (BRAF, RAS, MSI) 	<p>it is possible to consult aggregate data from the cancer register by consulting the site:</p> <p>www.registrotumorinapoli3sud.it</p>	Yes	NA
Vocabularies/ Ontology	ICD-O-3			Yes	NA
Dictionaries/Codebooks	ICD-O-3			Yes	NA
Formats	csv	Contains parameters that were requested			

D. Description of the data exchange process

Step	Action	Responsible	Link	Estimated time
The data exchange process is guided in Italy by the GDPR - EU- 2016/679 and by the Ministry of Health Decree 1st August 2023 "National Cancer Registry". The entire and complex procedure will make it very difficult to initiate an exchange of non-aggregated data				

making it impossible to analyse the data outside the Cancer Registry. At the moment the process of transmitting non-aggregated data from the register, but from individual records, is structured in different phases:				
1	Authorization requested from the General Director of the NA3CR for the transmission of data	Researcher		it is not possible to define the times of the entire authorization and data transmission process.
2	Preliminary assessment and authorization by the Data Protection Officer (DPO)	DPO		
3	Subsequent authorization by the General Director.	General Director		
4	Stipulation of an ad hoc operational protocol between the General Director of the ASL and the person in charge of the study project in which one is participating	General Director		

E. Related documents and links

NA

Annex D - BCR SOP

Standard operating procedure for minimal data exchange for Basilicata Cancer Registry.

Dataset Name (acronym):	Basilicata Cancer Registry (BCR)
Description of the Dataset:	<p>The BCR compiles standard dataset (see ENCR recommendation https://encr.eu/sites/default/files/Recommendations/ENCR-Recommendation-standard-dataset_Mar2023.pdf and Italian National Cancer Registry rules https://www.regioni.it/news/2023/09/01/registro-nazionale-tumori-decreto-1-8-2023-gazzetta-ufficiale-n-203-del-31-08-2023-656001/) of all individuals newly diagnosed with cancer in the Basilicata Region. Clinical data are not recorded except for biological factors (when available) and a simplified system of variables linked to the treatment as indicated below (Treatment variables refer to the curative first course of anticancer therapy after diagnosis).</p> <ul style="list-style-type: none"> - Surgery (Surgery): No; Yes, without specification; Yes, local surgery only; Yes, 'operative' surgery; Missing/Unknown - Radiotherapy (Rt): No; Yes, without other specification; Yes, neoadjuvant (pre-operative) radiotherapy; Yes, adjuvant (post-operative) radiotherapy; Unknown/missing - Chemotherapy (Cht): No; Yes, without other specifications; Yes, neoadjuvant (pre-operative) chemotherapy; Yes, adjuvant (post-operative) chemotherapy; Yes, both neoadjuvant and adjuvant chemotherapy; Unknown/missing.
Data Type(s):	Cancer registry
EOSC4Cancer Use Case:	T4.1
Date:	18/12/2023

A. Purpose

This SOP describes the steps required for data from CROB to be accessed and used by researchers in the context of the EOSC4Cancer project and beyond. Based on this information, we will define and suggest broader recommendations for data exchange within and beyond the EOSC4Cancer project. Below, you will find an overview of the data models, ontology, vocabulary and dictionaries used by the dataset. Furthermore, data access procedures are described step-by-step.

B. Data Models and ontology, vocabulary, dictionaries

Registry	Name	Description	References	Is this a standard?	Available conversions for harmonisation
Data Models	OMOP CDM	A part of the BCR is available in OMOP:		Yes	NA

		- ICD-O-3 diagnosis - TNM classification - Death			
Vocabulary/ Ontology	OMOP CDM vocabularies	Vocabularies used in the OMOP CDM (for variables available in OMOP CDM)	https://github.com/OHDSI/Vocabulary-v5.0/wiki	Yes	NA
Dictionary/ Codebooks	International Classification of Diseases for Oncology (ICD-O-3), third edition	Complete collection of ICD-O-3, ICD-O-3.1, ICD-O-3.2 classifications With integrations for Italian cancer registries	https://doi.org/10.19191/2022.064	Yes	NA
Formats	csv	Contains parameters that were requested			

C. Description of the data exchange process

Step	Action	Responsible (Name, email)	Links	Estimated time
1	The request must be addressed to the person in charge of the cancer registry, who will respond according to what is established in the register regulation (Art. 7 – Data communication) and the related laws on the matter.	General Director	https://www.crob.it/wp-content/uploads/2023/09/202300162-1.pdf https://www.crob.it/wp-content/uploads/2023/09/regol-3.pdf https://www.crob.it/wp-content/uploads/2023/09/relazione-alllegate-2.pdf	Between 2 weeks and 6 months

Annex E - EXPANSE Exposome-NL SOP

Standard operating procedure for minimal data exchange for the EXPANSE Exposome-NL.

Dataset Name (acronym):	EXPANSE Exposome-NL
Description of the Dataset:	The External Exposome data inventoried by the Exposome-NL and EXPANSE projects is categorised into built, food, physicochemical, and social environments
Data Type(s):	Exposome
EOSC4Cancer Use Case:	T4.1
Date:	23-11-2023

A. Purpose

This SOP describes the steps required for Exposome data from EXPANSE Exposome-NL to be accessed and used by researchers in the context of the EOSC4Cancer project and beyond. Based on this information, we will define and suggest broader recommendations for data exchange within and beyond the EOSC4Cancer project. Below you will find an overview of the data models, ontology, vocabulary and dictionaries used by the dataset. Furthermore, data access procedures are described step-by-step.

B. Data Models and ontology, vocabulary, dictionaries

Exposome	Name	Description	References	Is this a standard?	Available conversions for harmonisation
Data Models	NA				
Vocabulary/ Ontology	NA				
Dictionary/ Codebooks	Exposome Surfaces dictionary	Description of available variables in the Exposome Surfaces	https://surfdrive.surf.nl/files/index.php/s/yaAe9JCqqDMQa2U	No	No
Formats	csv				

C. Description of the data exchange process

Step	Action	Responsible (Name, email)	Link	Estimated time
1	Browse the available exposures on the “Data” tab of the platform	Researcher	https://exposome.dataplatform.nl/#/data	

2	Create an account on the exposome platform	Researcher		
3	Translate address information (of your cohort data you wish to link to the exposome data) into coordinates in the correct projection	Researcher		
4	Fill out and submit the data request form	Researcher	https://surfdrive.surf.nl/files/index.php/s/PhbCHIuqvjASVFe	Maximum 3 weeks (The proposal is considered approved if no objection was raised within this time)
5	Secure upload of your generated cohort coordinates	Researcher	https://surfdrive.surf.nl/files/index.php/s/b33zjSBPmrdRnPp	
5.1	Generate your public key using Kleopatra	Researcher	https://surfdrive.surf.nl/files/index.php/s/b33zjSBPmrdRnPp	
5.2	Upload cohort file and public key	Researcher	https://surfdrive.surf.nl/files/index.php/s/yaAe9JCqqDMQa2U	
5.3	Decrypt result file using Kleopatra	Researcher	(section 6.2)	

D. Related documents and links

- Instruction Manual Exposome Maps platform
<https://surfdrive.surf.nl/files/index.php/s/b33zjSBPmrdRnPp>
- Table with information about available Exposome Surfaces
<https://surfdrive.surf.nl/files/index.php/s/yaAe9JCqqDMQa2U>
- Description of Environmental variables available through Exposome Maps
<https://surfdrive.surf.nl/files/index.php/s/uqUORDrd428H2F9>

Annex F - EIRENE-CZ SOP

Standard operating procedure for minimal data exchange for the Czech Exposome data EIRENE-CZ.

Dataset Name (acronym):	EIRENE-CZ
Description of the Dataset:	Exposome variables consist of various themes. This dataset focuses on selected air pollutants and other related exposomic factors, which will be available for appropriate health outcomes timeframe as well as the estimated socio-economic factor to be used in the analyses. This research infrastructure benefits from the collaboration of several key institutions: Research Centre for Toxic Compounds in the Environment at Masaryk University (RECETOX-MU), the Czech Hydrometeorological Institute (CHMI), the National Institute of Public Health (SZU), Czech Statistical Office (CSO) and the EXPANSE project.
Data Type(s):	Exposome data
EOSC4Cancer Use Case:	T4.1
Date:	6/12/2023

A. Purpose

This SOP describes the steps required for Exposome data from EIRENE-CZ to be accessed and used by researchers in the context of the EOSC4Cancer project and beyond. Based on this information, we will define and suggest broader recommendations for data exchange within and beyond the EOSC4Cancer project. Below you will find an overview of the data models, ontology, vocabulary and dictionaries used by the dataset. Furthermore, data access procedures are described step-by-step.

B. List of definitions and abbreviations

Definitions	
EIRENE-CZ	Environmental Exposure Assessment Research Infrastructure

C. Data Models and ontology, vocabulary, dictionaries

Exposome	Name	Description	References	Is this a standard?	Available conversions for harmonisation
Data Models	NA				
Vocabulary/ Ontology	NA				
Dictionary/Co debook	Exposome data variables	Variables provided by EIRENE-CZ for linkage with cancer registry data	https://www.eiren-e-ri.eu/	Yes (for EIRENE-RI)	NA
Formats	csv				

Step	Action	Responsible (Name, email)	Link	Estimated time
1	Identify the available exposure data	Researcher		
2	Create and submit the data request by email to eirene@recetox.muni.cz (ensure to include "RECETOX - EIRENE RI" in the email subject).	Researcher	https://www.eirene-ri.eu/contact-us	
3	Review the data request	RECETOX, MU	https://www.recetox.muni.cz/hear/contacts	Maximum 3 weeks (The proposal is considered approved if no objection was raised within this time)
4	Share the data upon approval	RECETOX, MU		

D. Description of the data exchange process

E. Related links

NA

Annex G - ARPAC SOP

Standard operating procedure for minimal data exchange for the ARPAC.

Dataset Name (acronym):	Regional Agency for the Protection of the Environment of Campania (ARPAC)
Description of the Dataset:	The environmental dataset is critical for the assessment of potential exposure risk to support epidemiological assessments. The dataset has been designed to have the widest adherence on a European scale. It is built up of registering variables determined in potentially polluted sites (e.g. Contaminated and brownfield sites, Industrial activities) and analytical variables (e.g. air and water quality).
Data Type(s):	Exposome
EOSC4Cancer Use Case:	T4.1
Date:	12/07/2023

A. Purpose

This SOP describes the steps required for exposome data from ARPAC ISZM to be accessed and used by researchers in the context of the EOSC4Cancer project and beyond. Based on this information, we will define and suggest broader recommendations for data exchange within and beyond the EOSC4Cancer project. Below you will find an overview of the data models, ontology, vocabulary and dictionaries used by the dataset. Furthermore, data access procedures are described step-by-step.

B. List of definitions and abbreviations

Definitions	
ARIR	Companies at Risk of Major Accidents

C. Data Models and ontology, vocabulary, dictionaries

Exposome	Name	Description	References	Is this a standard?	Available conversions for harmonisation
Data Models	Air quality	Regional data is available in multiple forms, e.g. raw hourly updated or validated daily updated	https://dati.arpacampania.it/dataset/?groups=rqa-qualita-aria	yes	Yes, compatible with EIRENE-RI
	Groundwater quality	Includes analytical outcomes of chemical monitoring and other quantitative measures	https://dati.arpacampania.it/organizzazione/arpac?group=s=acque-sotteranee	yes	Yes, compatible with EIRENE-RI
	Contaminated and brownfield	Contaminated sites are those areas in which, due to past or ongoing	https://www.arpacampania.it/web/guest/siti-	no	Yes, compatible with EIRENE-RI

	sites	human activities, pollution of environmental matrices has been determined	contaminati		
	ARIR	Monitors major accidents involving dangerous substances	https://dati.arpacampania.it/organizzazione/arpac?organizzazione=arpac&groups=aia	yes	Yes, compatible with EIRENE-RI
Vocabulary/ Ontology	NA				
Dictionary/ Codebook	Exposome Maps inventory	Description of available variables in the Exposome Surfaces	Available upon request through the EOSC4Cancer shared drive	No	No
Formats	csv	Air quality			
	csv	Groundwater quality			
	pdf	Contaminated and brownfield sites			
	csv	ARIR			

D. Description of the data exchange process

Step	Action	Responsible (Name, email)	Link	Estimated time
1	The air quality, groundwater quality and ARIR data are in csv format and are downloaded from the Agency's open data by accessing the link provided	NA	https://dati.arpacampania.it/organizzazione/arpac?organizzazione=arpac	1 day
2	The pdf files of the contaminated sites and brownfield sites are downloaded at the link provided. Once downloaded, they are converted to csv format	NA	https://www.arpacampania.it/web/guest/siti-contaminati	1 day

E. Related documents and links

1. [Directive 2008/50/EC]

2. [Directive 2000/60/EC]
3. [Directive 2012/18/EU]

Annex H - ARPAB SOP

Standard operating procedure for minimal exposome data exchange for the ARPAB.

Dataset Name (acronym):	Regional Agency for the Protection of the Environment of Basilicata (ARPAB)
Description of the Dataset:	<p>The ARPAB carries out the technical-scientific activities connected to the exercise of public functions for the protection of the environment referred to in article 1 of the legislative decree of 4 December 1993, n. 496, converted, with amendments, by law 21 January 1994, n. 61 (Urgent provisions on the reorganisation of environmental controls and establishment of the national ARPAB for the protection of the environment) and as indicated in L. R. n.1 of 20 January 2020 ss.mm.ii: Reorganisation of the regulations of the Regional Protection Agency of the Environment of Basilicata (A.R.P.A.B.)</p> <p>https://www.arpab.it/arpab/wp-content/uploads/2023/10/1476.pdf</p> <p>The mandatory institutional activities are those defined in the LEPTA (Essential Levels of Environmental Technical Performance) and present in the national Catalog of Services and Performances of the SNPA (national system for environmental prevention).</p> <p>https://www.snambiente.it/chi-siamo/atti-fondamentali/catalogo-nazionale-dei-servizi-e-prestazioni-snpa/</p>
Data Type(s):	Exposome
EOSC4Cancer Use Case:	T4.1
Date:	18/12/2023

A. Purpose

This SOP describes the steps required for exposome data from ARPAB to be accessed and used by researchers in the context of the EOSC4Cancer project and beyond. Based on this information, we will define and suggest broader recommendations for data exchange within and beyond the EOSC4Cancer project. Below you will find an overview of the data models, ontology, vocabulary and dictionaries used by the dataset. Furthermore, data access procedures are described step-by-step.

B. Data Models and ontology, vocabulary, dictionaries

Exposome	Name	Description	References	Is this a standard?	Available conversions for harmonisation
Data Models	Air	Information about air quality and emissions is available. Additionally, an air quality index (IQA), which describes the state of the atmospheric environment, correlating air	https://www.arpab.it/temi-ambientali/aria/	yes	NA

		quality to human health risk levels, is calculated.			
	water	Surface water (intended for production of drinking water), groundwater, marine/coastal bathing areas, water (e.g. waterways, lakes and ponds) suitable for fish life and wastewater are monitored	https://www.arpa.b.it/temi-ambientali/acqua/	yes	NA
	soil and waste	Monitors production and management of urban and special waste	https://www.arpa.b.it/temi-ambientali/suolo-e-rifiuti/	yes	NA
	Industrial risks	Investigation of projects of industries at risk of major accidents	https://www.arpa.b.it/temi-ambientali/rischi-industriali/	No	no
Vocabulary/ Ontology	NA				
Dictionary/ Codebook	Exposome Maps inventory	Description of available variables in the Exposome Surfaces	Available upon request through the EOSC4Cancer shared drive	No	No
Formats	csv				
	pdf				

C. Description of the data exchange process

Step	Action	Responsible (Name, email)	Link	Estimated time
1	Open (non-personal) data can be downloaded from the links provided in the table above.	Researcher	See table above	

Annex I - mtFIT SOP

Standard operating procedure for minimal data exchange for the mtFIT study

Dataset Name (acronym):	Multitarget FIT study (mtFIT)
Description of the Dataset:	The mtFIT study is a cross-sectional intervention study where subjects for the Dutch FIT-based national CRC screening program are invited to perform mtFIT in addition to FIT. If either one or both tests are positive, subjects are referred for colonoscopy.
Data Type(s):	Screening
EOSC4Cancer Use Case:	T4.2
Date:	5-12-2023

A. Purpose

This SOP describes the steps required for screening data from the mtFIT cohort to be accessed and used by researchers in the context of the EOSC4Cancer project and beyond. Based on this information, we will define and suggest broader recommendations for data exchange within and beyond the EOSC4Cancer project. Below you will find an overview of the dataset's data models, ontology, vocabulary and dictionaries. Furthermore, data access procedures are described step-by-step.

B. List of definitions and abbreviations

Definitions	
FIT	faecal immunochemical test
CRC	Colorectal cancer

C. Data Models and ontology, vocabulary, dictionaries

Screening	Name	Description	References	Is this standard?	Available conversions for harmonisation
Data Models	NA				
Vocabulary/ Ontology	NA				
Dictionary/ Codebook	mtFIT Dictionary	Custom dictionary used for mtFIT study	Available upon request through the EOSC4Cancer shared drive	no	no
Formats	NA				

D. Description of the data exchange process

Step	Action	Responsible (Name, email)	Link	Estimated time
1	Email the Health-RI servicedesk and request a Health-RI account (necessary to submit the cBioPortal request access to existing study form)	Researcher	servicedesk@health-ri.nl	<1 hour
2	Health-RI sends Terms of Use to requesting researcher	Health-RI first line operator		1 day
3	Requesting researcher agrees to Terms of Use	Researcher		<2 weeks
4	Login to the Self Service Portal and fill in the form for requesting access to an existing study in cBioPortal For the mtFIT study, select “mtFIT_LancetOncol2023” from the dropdown menu	Researcher	Self Service Portal Requesting access to an existing study in cBioPortal	<1h
5	The PI will evaluate the request and inform Health-RI first-line operator on the decision access request	PI (data owner)		1 week
6	In case access may be granted, Health-RI first-line operator will contact requesting researcher and set up Health-RI cBioPortal account	Health-RI first line operator		2 days
7	Requesting researcher supplies requested information to set up account (google-linked authentication)	Researcher		1 day
8	Health-RI first-line operator grants access to requested study in the Health-RI cBioPortal. The researcher will see the study in cBioPortal		https://cbioportal.health-ri.nl/login.jsp	<1 day

Annex J - Piedmont regional screening program SOP

Standard operating procedure for minimal data exchange for the Piedmont regional screening program.

Dataset Name (acronym):	Piedmont regional screening program (Prevenzione Serena)
Description of the Dataset:	Regional population base CRC screening program. Eligible residents aged 58 to 69 are invited every two years to perform FIT (single sample; 20 µg Hb/gr faeces). Subjects with a positive test result are invited to perform a colonoscopy
Data Type(s):	Screening
EOSC4Cancer Use Case:	T4.2
Date:	12-12-2023

A. Purpose

This SOP describes the steps required for screening data from the Piedmont regional screening program to be accessed and used by researchers in the context of the EOSC4Cancer project and beyond. Based on this information, we will define and suggest broader recommendations for data exchange within and beyond the EOSC4Cancer project. Below is an overview of the dataset's data models, ontology, vocabulary and dictionaries.

B. List of definitions and abbreviations

Definitions	
FIT	faecal immunochemical test
CRC	Colorectal cancer

C. Data Models and ontology, vocabulary, dictionaries

Screening	Name	Description	References	Is this standard?	Available conversions for harmonisation
Data Models	NA				
Vocabulary/ Ontology	NA				
Dictionary/ Codebook	Prevenzione Serena Dictionary	Custom dictionary used for the Prevenzione Serena datawarehouse	Available upon request through the EOSC4Cancer shared drive	no	no
Formats	NA				

D. Description of the data exchange process

Currently, it is not possible to share the data from the Piedmont regional screening program.

Annex K - Catalan screening program SOP

Standard operating procedure for minimal data exchange for the Catalan screening program.

Dataset Name (acronym):	Catalan screening program (ATOS)
Description of the Dataset:	Cancer screening dataset form the Hospital Clínic in Barcelona
Data Type(s):	Screening
EOSC4Cancer Use Case:	T4.2
Date:	09-01-2024

A. Purpose

This SOP describes the steps required for screening data from the Catalan screening program to be accessed and used by researchers in the context of the EOSC4Cancer project and beyond. Based on this information, we will define and suggest broader guidelines for data exchange within and beyond the EOSC4Cancer project. Below you will find an overview of the data models, ontology, vocabulary and dictionaries used by the dataset. Furthermore, data access procedures are described step-by-step.

B. List of definitions and abbreviations

Definitions	
None	None

C. Data Models and ontology, vocabulary, dictionaries

Screening	Name	Description	References	Is this a standard?	Available conversions for harmonisation
Data Models	ATOS	Custom-made for hospital	Available upon request through the EOSC4Cancer shared drive	No	No
Vocabulary/ Ontology	ATOS vocabulary	Custom-made for hospital		No	No
Dictionaries /Codebooks	ATOS tables	Custom-made for hospital		No	No
Formats	csv				

D. Description of the data exchange process

Step	Action	Responsible (Name, email)	Link	Estimated time
------	--------	---------------------------	------	----------------

1	Not available for the moment			
---	------------------------------	--	--	--

E. Related documents and links

NA

-

Annex L - BBMRI-ERIC CRC-Cohort SOP

Standard operating procedure for minimal data exchange for the BBMRI-ERIC Colorectal Cancer Cohort.

Dataset Name (acronym):	BBMRI-ERIC colorectal cancer cohort (CRC-Cohort)
Description of the Dataset:	The CRC-Cohort collection is a joint long-term European endeavour, with contributions from existing, well-established biobanks. Currently, the CRC-Cohort numbers over 10,000 datasets from across Europe
Data Type(s):	Genomic, Clinical, and Pathology
EOSC4Cancer Use Case:	T4.3
Date:	5-12-2023

A. Purpose

This SOP describes the steps required for Genomic, Clinical, and Pathology data from CRC-Cohort to be accessed and used by researchers in the context of the EOSC4Cancer project and beyond (e.g. raw data access and download and/or via analysis and visualisation platforms like cBioPortal). This SOP is created for data scientists who want to access the datasets shared within the EOSC4Cancer project. Based on this information, we will define and suggest broader recommendations for data exchange within and beyond the EOSC4Cancer project. Below, you will find an overview of the data models, ontology, vocabulary and dictionaries used by the dataset. Furthermore, data access procedures are described step-by-step.

B. List of definitions and abbreviations

Definitions	
HNPCC	Hereditary nonpolyposis colorectal cancer (also known as Lynch syndrome)
MSI	Microsatellite instability

C. Data Models and ontology, vocabulary, dictionaries

Genomic	Name	Description	References	Is this standard?	Available conversions for harmonisation
Data Models	Data model for CRC-Cohort	The data model of the CRC-Cohort includes Molecular markers, like <ul style="list-style-type: none"> • Microsatellite Instability • Mismatch Repair Gene Expression–IHCarrayfordifferent genes • Risk Situation • RAS mutation status • BRAF, PIC3CA.HER2 mutation status 	data model https://doi.org/10.5281/zenodo.7930536 See “Data Model for CRC-Cohort” (appendix B / section Molecular markers) in BBMRI-ERIC Colorectal Cancer Cohort (CRC-Cohort): Data Protection Policy	no, custom	Converted to format openEHR Conversion to cBioPortal data model

Vocabulary/ Ontology	Revised Bethesda Guidelines for HNC and Microsatellite Instability	Bethesda Guidelines for identifying individuals at risk for HNPCC; and recommend criteria for MSI testing.	https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2933058	no	No
Dictionary/ Codebook	NA				
Formats	csv				

Clinical	Name	Description	References	Is this standard?	Available conversions for harmonisation
Data Models	Data model for CRC-Cohort	<p>The data model of the CRC-Cohort includes</p> <ul style="list-style-type: none"> • Sex • Participation In Clinical Study • Age At Primary Diagnosis • Known Risk Factors for CRC • Time Of Recurrence (metastasis) and availability of biological material from recurrence • Vital Status And Survival Information • Surgery • Pharmacotherapy Targeted therapy • Radiation therapy • Response to therapy • Diagnostic Exam 	<p>https://doi.org/10.5281/zenodo.7930536</p> <p>See “Data Model for CRC-Cohort” (appendix B) in BBMRI-ERIC Colorectal Cancer Cohort (CRC-Cohort): Data Protection Policy</p>	no, custom	Converted to formats openEHR / OMOP / HL7-FHIR
Vocabularies / Ontologies	SNOMED CT	SNOMED Clinical Terms	http://purl.bioontology.org/ontology/SNOMEDCT/423493009	yes	NA
	Phenotype And Trait Ontology	An ontology of phenotypic qualities (properties, attributes or characteristics)	http://purl.obolibrary.org/obo/PATO_0020000	no	No

Dictionary/ Codebook	NA				
Formats	csv				

Pathology	Name	Description	References	Is this standard?	Available conversions for harmonisation
Data Models	Data model for CRC-Cohort	The data model of the CRC-Cohort includes the “Histopathology part”.	https://doi.org/10.5281/zenodo.7930536 See “Data Model for CRC-Cohort” (appendix B / section Histopathology part) in BBMRI-ERIC Colorectal Cancer Cohort (CRC-Cohort): Data Protection Policy	No	No
Vocabulary/ Ontologies	NA				
Dictionary/ Codebook	ICD-10	The ICD-10 is the 10th version of the International Statistical Classification of Diseases and Related Health Problems (ICD), a medical classification list of the World Health Organisation (WHO)	https://icd.who.int/browse10/2019/en	yes	NA
Formats	dicom	Image format		Work in progress by BigPicture	Compatible with XOpac
	tiff	Image format			

D. Description of the data exchange process

Step	Action	Responsible (Name, email)	Link	Estimated time
0	Public (non-authenticated) search and browsing access to meta-data describing the CRC-Cohort for discovery and select CRC-Cohort in BBMRI-ERIC Directory	Researcher	https://directory.bbmri-eric.eu	

1-5	<p>See BBMRI-ERIC Policy for general process on Access to and Sharing of Biological Samples and Data (appendix H of BBMRI-ERIC Colorectal Cancer Cohort (CRC-Cohort): Data Protection Policy) for a step-by-step description</p> <p>The CRC-Cohort has an expedited procedure directly between BBMRI-ERIC as a controller and the requester. That replaces Step-3 of the general procedure at BBMRI-ERIC and the details (with a 2 week veto possibility for contributing biobanks) can be found in Annex I.</p>	<p>Researcher, Access Committee (at BBMRI-ERIC) and biobanks contributing to CRC-Cohort</p>	<p>https://doi.org/10.5281/zenodo.7930536</p>	<p>approx. 1 month</p>
-----	--	---	--	------------------------

E. Related documents and links

- BBMRI-ERIC Colorectal Cancer Cohort (CRC-Cohort): Data Protection Policy (V1.5: <https://doi.org/10.5281/zenodo.7930536>)

Annex M - CAIRO5 SOP

Standard operating procedure for minimal data exchange for the CAIRO5 cohort.

Dataset Name (acronym):	Treatment strategies in colorectal cancer patients with initially unresectable liver-only metastases: CAIRO5
Description of the Dataset:	Various data types including genomic data, plasma ctDNA (also includes phenotype and longitudinal data, plus digital pathology slides and radiology images)
Data Type(s):	Genomic, Clinical, and Radiology
EOSC4Cancer Use Case:	T4.4 and T4.3
Date:	5-12-2023

A. Purpose

This SOP describes the steps required for Genomic, Clinical and Radiology data from CAIRO5 to be accessed and used by researchers in the context of the EOSC4Cancer project and beyond (e.g. raw data access and download and/or via analysis and visualisation platforms like cBioPortal). Based on this information, we will define and suggest broader recommendations for data exchange within and beyond the EOSC4Cancer project. Below you will find an overview of the data models, ontology, vocabulary and dictionaries used by the dataset. Furthermore, data access procedures are described step by step.

B. List of definitions and abbreviations

Definitions	
CRC	Colorectal cancer
PLCRC	Prospective Dutch CRC cohort
ACT	Adjuvant chemotherapy
DDD	Defined daily dose (for drugs)
IKNL	Netherlands Comprehensive Cancer Organisation
NCR	Netherlands Cancer Registry (provided by IKNL)
DAC	Data access committee

C. Data Models and ontology, vocabulary, dictionaries

Genomic	Name	Description	References	Is this a standard?	Available conversions for harmonisation
Data Models	cBioPortal data model	different file formats for clinical, sample and (gen)omics data described in cBioPortal's file format page	https://docs.cbioportal.org/file-formats/	yes (for cBioPortal)	No
	File-formats (gen)omics data, raw data in European Genome-phenome Archive	different files described for uploading raw (gen)omics data and accompanying metadata	https://ega-archive.org/submitting/metadata/ega-schema/	Yes (for EGA)	No
Vocabulary/ Ontology	NA				
Dictionary/ Codebook	NA				
Formats	MAF	cBioPortal file format for (processed) mutation data	https://docs.cbioportal.org/file-formats/#mutation-data	yes (for cBioPortal)	
	FASTQ	EGA file format for (raw sequencing) mutation data (raw data, available in European Genome-phenome Archive,)	https://ega-archive.org/studies/EGAS00001006695	yes (for EGA)	

Clinical	Name	Description	References	Is this a standard?	Available conversions for harmonisation
Data Models	cBioPortal data model	different file formats for clinical, sample and (gen)omics data described in cBioPortal's file format page	https://docs.cbioportal.org/file-formats/	yes (for cBioPortal)	
	File-formats (gen)omics data, raw data in European Genome-phenome Archive...	different files described for uploading of raw (gen)omics data and accompanying metadata	https://ega-archive.org/submitting/metadata/ega-schema/		

Vocabulary/ Ontology	IKNL tumorindeling	Tumour classification provided by the Dutch cancer registry	https://iknl.nl/getmedia/2a1cf6d0-9285-40dd-bce3-209a31541ec9/Nederlandsse-Kankerregistratie-indeling-tumorsoorten_IK_NL.pdf%20target%20=1	yes (within the Netherlands)	OMOP?
Dictionary/ Codebook	CAIRO5 dictionary	Custom dictionary used for CAIRO5	Available upon request through the EOSC4Cancer shared drive	no	Codebook overlaps/is compatible in part with the PLCRC codebook of the PROVENC3 study
Formats	tsv	cBioPortal file format for (processed) clinical/sample data	https://docs.cbioportal.org/file-formats/#clinical-data		
	tsv/csv	EGA file format for subject/sample metadata	https://ega-archive.org/submission/metadata/ega-schema/		

Radiology	Name	Description	References	Is this standard?	Available conversions for harmonisation
Data Models	DICOM standard	Internationally recognized standard	https://www.dicomstandard.org/	Yes	Conversion to other file formats possible
	XNAT data model	Standard XNAT model. Structure projects - subjects - experiments (i.e., scan sessions) - scans. Furthermore, it is heavily DICOM based.	https://wiki.xnat.org/documentation/understanding-the-xnat-data-model https://xnat.bmja.nl/ https://www.health-nl/en/services/xnat	Yes	The standard model cannot be changed, but is quite flexible in the sense that it can be expanded and other file formats added. Using the EMCs Structure file for a FAIR data point in XNAT, we can define any structure while making sure data is FAIR.
	EuCanImage data model	For non-raw DICOM files, e.g., segmentations, converted images	https://eucanimag.e.eu/	No	Yes, based on flexible XNAT data model
Vocabulary/ Ontology	NA				
Dictionary/ Codebook	NA				
Formats	DICOM for images	See above	See above	See above	See above

	Others, e.g. segmentations	Depending on needs from CAIRO5. In principle, XNAT can facilitate anything.	NA	NA	NA
--	----------------------------	---	----	----	----

D. Description of the data exchange process

Clinical and genomic data made available through the Health-RI cBioPortal				
Step	Action	Responsible (Name, email)	Link	Estimated time
1	Email the Health-RI servicedesk and request a Health-RI account (necessary to submit the cBioPortal request access to existing study form)	Researcher	servicedesk@health-ri.nl	<1 hour
2	Health-RI sends Terms of Use to requesting researcher	Health-RI first line operator		1 day
3	Requesting researcher agrees to Terms of Use	Researcher		<2 weeks
4	Login to the Self Service Portal and fill in the form for requesting access to an existing study in cBioPortal For the CAIRO5 study, select “CAIRO5_ClinCancerRes2023” from the dropdown menu	Researcher	Self Service Portal Requesting access to an existing study in cBioPortal	<1h
5	The PI will evaluate the request and inform Health-RI first-line operator on the decision access request	PI (data owner)		1 week
6	In case access may be granted, Health-RI first-line operator will contact requesting researcher and set up Health-RI cBioPortal account	Health-RI first line operator		2 days
7	Requesting researcher supplies requested information to set up account (google-linked authentication)	Researcher		1 day
8	Health-RI first-line operator grants access to requested study in the Health-RI cBioPortal. The researcher will see the study in cBioPortal		https://cbioportal.health-ri.nl/login.jsp	<1 day

Raw sequencing data, available in the European Genome-phenome Archive (EGA)

Step	Action	Responsible (Name, email)	Link	Estimated time
1	Register yourself as an EGA user	Researcher	https://ega-archive.org/register/	< 10 min
2	Validate your account	Researcher		< 10 min
3	Log into the EGA page	Researcher		< 5 min
4	Go to the dataset	Researcher	https://ega-archive.org/studies/EGAS00001006695	< 1 min
5	Click on “Request access”	Researcher		< 5 min
6	Add a comment requesting the desired ‘EGA datasets’; EGAS00001006695	Researcher		<1 hour
7	Submit the EGA access request to NKI’s internal registration and evaluation system. Provide the information requested in: ega.nki.nl (supply the EGAS study number and the EGA Dataset numbers in your request as well)	Researcher	https://ega.nki.nl/	1 hour
8	The Data Access Committee (DAC) and the NKI-AVL Internal Review Board (IRB) will review the request and determine whether the planned study and its goals are competing with other study interests. The DAC and IRB can approve or decline your request or request a modification.	DAC/IRB		1-2 months
9	Upon approval by the IRB, the researcher will receive a Data Transfer Agreement (DTA) that needs to be filled out and signed by the legal representative of your institute/company. The Knowledge Transfer & Contracting Office of NKI will finalise the DTA and send a double-signed copy to the researcher requesting it.	Knowledge Transfer & Contracting Office (KT&C)		2-4 months
10	A study DAC member grants access to the requesting researcher through the EGA DAC portal	DAC member	DAC portal	(new procedure, untested yet)
11	Download the data through the pyEGA3 download client (once access is granted)	Researcher	https://ega-archive.org/access/download/files/pyega3/	

Radiology				
Step	Action	Responsible (Name, email)	Link	Estimated time
1	Contact a CAIRO5 PI / DAC member with a Health-RI XNAT account to request an XNAT account for you. Note that you have to provide an institutional email address. The requester includes access to the CAIRO5 XNAT project (name to be determined) in the request.	Researcher / DAC Member	servicedesk@health-ri.nl Request new XNAT user as existing user	<1 hour
2	Health-RI sends Terms of Use to requesting researcher	Health-RI first line operator		1 day
3	Requesting researcher agrees to Terms of Use	Researcher		<2 weeks
4	Login on the Health-RI BMIA XNAT and download the data from the CAIRO5 project.	Researcher	https://xnat.bmia.nl/	

E. Related documents and links

1. The latest version of the EGA data access request procedure: <https://ega-archive.org/access/request-data/how-to-request-data/>
2. For NKI's internal review policies related to EGA requests, a version of the EGA data access request for NKI can be found here: <https://ega.nki.nl/>

Annex N - PROVENC3 SOP

Standard operating procedure for minimal data exchange for the PROVENC3 cohort.

Dataset Name (acronym):	PROgnostic Value of Early Notification by Ctdna in colon cancer stage 3 (PROVENC3)
Description of the Dataset:	Nationwide cohort study in the Netherlands with inclusion of patients with non-metastatic CRC treated with ACT
Data Type(s):	Genomic and Clinical
EOSC4Cancer Use Case:	4.4
Date:	5-12-2023

A. Purpose

This SOP describes the steps required for Genomic and Clinical data from the PROVENC3 cohort to be accessed and used by researchers in the context of the EOSC4Cancer project and beyond (e.g. data access, download and/or via analysis and visualisation platforms like cBioPortal). Based on this information, we will define and suggest broader recommendations for data exchange within and beyond the EOSC4Cancer project. Below you will find an overview of the data models, ontology, vocabulary and dictionaries used by the dataset. Furthermore, data access procedures are described step by step.

B. List of definitions and abbreviations

Definitions	
CRC	Colorectal cancer
PLCRC	Prospective Dutch CRC cohort
ACT	Adjuvant chemotherapy
DDD	Defined daily dose (for drugs)
IKNL	Netherlands Comprehensive Cancer Organisation
NCR	Netherlands Cancer Registry (provided by IKNL)

C. Data Models and ontology, vocabulary, dictionaries

Genomic	Name	Description	References	Is this standard?	Available conversions for harmonisation
Data Models	cBioPortal data model	different file formats for clinical, sample and (gen)omics data described in	https://docs.cbioportal.org/file-formats/	yes (for cBioPortal)	

		cBioPortal's file format page			
Vocabulary/ Ontology	NA				
Dictionary/ Codebook	NA				
Formats	MAF	cBioPortal file format for (processed) mutation data	https://docs.cbioportal.org/file-formats/#mutation-data	yes (for cBioPortal)	

Clinical	Name	Description	References	Is this standard?	Available conversions for harmonisation
Data Models	cBioPortal data model	different file formats for clinical, sample and (gen)omics data described in cBioPortal's file format page	https://docs.cbioportal.org/file-formats/	yes (for cBioPortal)	No
Vocabulary/ Ontology	IKNL tumorindeling	Tumour classification provided by the Dutch cancer registry	https://iknl.nl/getmedia/2a1cf6d0-9285-40dd-bce3-209a31541ec9/Nederlandsse-Kankerregistratie_indeling-tumorsoorten_IK_NL.pdf%20target%20NL.pdf	yes (within the Netherlands)	No
	WHO ATC/DDD	- The ATC classification system classifies active substances in drugs - The DDD is a unit of measurement assigned per ATC	https://www.who.int/publications/m/item/atc-ddd-in-dex/	No	No
	ICD-0-3	Used for - Topography - Morphology - TNM classification	https://seer.cancer.gov/icd-o-3/	yes	NA
	SNOMED-CT	Used for the pathology data from the national PALGA archive		yes	NA
Dictionary/ Codebook	PLCRC dictionary	Custom dictionary used for PLCRC studies	Available upon request through the EOSC4Cancer shared drive	no	Codebook overlaps in part with the CAIRO5 study codebook
Formats	tsv	cBioPortal file format for (processed)	https://docs.cbioportal.org/file-formats/#clinical-		

		clinical/sample data	data		
--	--	----------------------	----------------------	--	--

D. Description of the data exchange process

Clinical and genomic data, processed data made available through the Health-RI cBioPortalGenomic				
Step	Action	Responsible (Name, email)	Link	Estimated time
1	Email the Health-RI servicedesk and request a Health-RI account (necessary to submit the cBioPortal request access to existing study form)	Researcher	servicedesk@health-ri.nl https://www.health-ri.nl/en/health-ri-service-desk	<1 hour
2	Health-RI first line operator sends Terms of Use to requesting researcher	Health-RI first line operator		1 day
3	Requesting researcher agrees to Terms of Use	Researcher		<2 weeks
4	Login to the Self Service Portal and fill in the form for requesting access to an existing study in cBioPortal For the PROVEN3 study, select “PROVEN3” from the dropdown menu	Researcher	Self Service Portal Requesting access to an existing study in cBioPortal	<1h
5	The PI will evaluate the request and inform the Health-RI first line operator on the decision for the access request	PI (data owner)		1 week
6	In case that access may be granted, Health-RI first line operator will contact requesting researcher and set up a Health-RI cBioPortal account	Health-RI first line operator		2 days
7	Requesting researcher supplies requested information necessary to set up account (google-linked authentication)	Researcher		1 day
8	Health-RI first line operator grants access to requested study in the Health-RI cBioPortal. The researcher will see the study in cBioPortal.		https://cbioportal.health-ri.nl/login.jsp	<1 day

Annex O - mCRC-VHIO

Standard operating procedure for minimal clinical data exchange for the mCRC-VHIO.

Dataset Name (acronym):	Metastatic Colorectal Cancer cohort VHIO (mCRC-VHIO)
Description of the Dataset:	Cohort containing patients with metastatic colorectal cancer
Data Type(s):	Genomic and Clinical
EOSC4Cancer Use Case:	4.5
Date:	28/11/2023

A. Purpose

This SOP describes the steps required for genomic and clinical data from mCRC-VHIO to be accessed and used by researchers in the context of the EOSC4Cancer project and beyond (e.g. raw data access and download and/or via analysis and visualisation platforms like cBioPortal). Based on this information, we will define and suggest broader guidelines for data exchange within and beyond the EOSC4Cancer project. Below you will find an overview of the data models, ontology, vocabulary and dictionaries used by the dataset. Furthermore, data access procedures are described in a step-by-step way.

B. List of definitions and abbreviations

Definitions	
MSI	Microsatellite instability
PD	Progression Disease
pT, pN, pM	Pathological T (Tumour) N (Node) M (Metastasis)
G1, G2, G3	Grade of tumour
Mut	Mutation
WT	Wild Type

C. Data Models and ontology, vocabulary, dictionaries

Genomic	Name	Description	References	Is this standard?	Available conversions for harmonisation
Data Models	Data model for CRC-Cohort from VHIO	The data model of the CRC-Cohort includes Clinical information and Molecular markers, like <ul style="list-style-type: none"> • Mismatch Repair Gene Expression–IHCarrayfordifferent genes • RAS and BRAF mutation status 	NA	no, custom	No
Vocabularies/ Ontology	NA				
Dictionaries /Codebooks	NA				
Formats	CSV				

Clinical	Name	Description	References	Is this standard?	Available conversions for harmonisation
Data Models	Data model for CRC-Cohort	The data model of the CRC-Cohort includes Molecular markers, like <ul style="list-style-type: none"> • Gender • Birthdate • Age At Primary Diagnosis • Date Diagnosis • Tumour location • Adjuvance • Neoadjuvance • Relapse date and location • Vital Status And Survival Information • Surgery date • Treatment dates and regimen 	NA	no, custom	No
Vocabularies/ Ontologies	NA				
Dictionaries/ Codebooks	NA				
Formats	CSV				

D. Description of the data exchange process

Step	Action	Responsible (Name, email)	Link	Estimated time
1	Send a data access request to the principal investigator in colorectal cancer of the VHIO	Researcher		
2	The principal investigator will review the request	Principal investigator		15 days
3	Data Transfer Agreement Signature	Legal Department from VHIO		1-2 months
4	Anonymization process	Research support technician		21 days

E. Related documents and links

NA