



Project acronym: EOSC4CANCER
Grant Agreement Number: 101058427
Project full title: A European-wide foundation to
accelerate Data-driven Cancer Research
Call identifier: HORIZON-INFRA-2021-EOSC-01

D6.3 Roadmap towards a federated digital platform for advancing cancer research: Leveraging current efforts and projects for a sustainable ecosystem

Version: 3.0
Status: Final
Dissemination Level: Public
Due date of deliverable: 31.05.2025
Actual submission date: 30.05.2025
Work Package: WP6 Dialogue with patients, researchers, and EU
initiatives to shape project execution, outcomes and
long-term alignment
Lead partner for this deliverable: ELIXIR/EBI/EMBL
Partner(s) contributing: BSC, EMPIRICA (for others see below)



Document Authors

Main Authors

Salvador Capella-Gutierrez	Barcelona Supercomputing Center
Alfonso Valencia	<i>Barcelona Supercomputing Center</i>
Daniel Barrowdale	<i>ELIXIR Hub</i>
Carola Schulz	<i>empirica Technology Research</i>

Additional Authors

Jan-Willem Boiten	<i>Lygature</i>
Robin Navest	<i>Lygature</i>
Eivind Hovig	<i>University of Oslo</i>
Romina Royo	<i>Barcelona Supercomputing Center</i>
Lifang Liu	<i>Health-RI</i>
Josephine Mosset	<i>Cancer Patients Europe</i>
Sergi Aguiló-Castillo	<i>Barcelona Supercomputing Center</i>
David Marshall	<i>Instruct</i>
Gerrit Meijer	<i>NKI</i>
Munazah Andrabi	<i>University of Manchester</i>
Sophie Huiskes-Berends	<i>Lygature</i>
Fotis Psomopoulos	<i>CERTH</i>
Sarah Morgan	<i>EATRIS</i>
Maria Alexandra Rujano	<i>ECRIN</i>
Macha Nikolski	<i>CNRS</i>
Eva Garcia Alvarez	<i>BBMRI-ERIC</i>
Griselda Marku	<i>empirica Technology Research</i>
Rabea Richter	<i>empirica Technology Research</i>
Veli Stroetmann	<i>empirica Technology Research</i>

We deeply thank the members of the EOSC4Cancer Stakeholder Forum for their input

Table of Contents

Document Authors	2
Table of Contents	3
Acronyms and Abbreviations	4
Executive summary	5
1 Vision/Ambition	7
Current situation and environment	8
EOSC4Cancer's long term goal	8
2 Cancer Patient journey as the driver	9
Existing use-cases	10
Extending the existing use cases	13
Researcher journey	14
3 User perspective	17
Platform user profiles	17
Capacity building	18
Establishing an RDM knowledge base	18
4 European perspective on cancer research	19
EU Mission on Cancer	19
Europe's Beating Cancer Plan	20
European Health Data Space	20
TEHDAS2 and HealthData@EU	22
Link to EOSC	22
The eCancer thematic group	23
The EU AI Act	23
5 National vs European level for implementation	24
UNCAN-CONNECT	25
National Cancer Data Nodes	25
EU Network of National Comprehensive Cancer Centres	27
European Life Science Research Infrastructures	27
6 Resources ready for the UNCAN platform	29
Data Sources	29
Importance of standardisation	30
Actionable Research Software	34
7 Data biases	36
Sex/Gender Bias	37
Tackling the sex and gender biases	38
8 Assembling a sustainable ecosystem	39
Patient engagement	39
9 Timeline for implementation	40
2025	41
2026	42
2027	43
2028	43
2029	44
2030	44

Acronyms and Abbreviations

Acronym	Meaning
API	Application Programming Interface
DAC	Data Access Committee
ECPDC	European Cancer Patient Digital Center
EGA	European Genome-phenome Archive
EHDS	European Health Data Space
EHR	Electronic Health Records
EOSC	European Open Science Cloud
EOSC4Cancer	A European-wide project to accelerate data-driven cancer research, and author of this roadmap.
FAIR	Findable, Accessible, Interoperable and Reusable. By making data outputs FAIR researchers can enhance their utility by making it easier for other researchers to build on their work.
GDI	Genomic Data Infrastructure, a project aiming to build a federated platform for accessing genomic data across Europe.
GDPR	General Data Protection Regulation
HTA	Health Technology Assessment
LS RI	Life Science Research Infrastructure
NCDN	National Cancer Data Nodes
OMOP CDM	Observational Medical Outcomes Partnership ((Common Data Model)
QoL	Quality of Life
RDM	Research Data Management
RI	Research Infrastructure
TRE	Trusted Research Environments
UNCAN	UNderstanding CANcer, a future cancer research digital platform

Executive summary

EOSC4Cancer is a European funded project aiming to facilitate future cancer research. It does this by using and enhancing federated and interoperable systems for securely identifying, sharing, processing and reusing FAIR data across borders and offering them via popular community-driven analysis environments. By integrating digital tools, data analytics, and advanced Artificial Intelligence and Machine Learning capabilities, the project supports more efficient and insightful analysis of cancer data. In addition, EOSC4Cancer makes diverse types of cancer data more accessible, such as genomics, imaging, medical, clinical, environmental and socio-economic data.

Cancer's complex nature requires integration of advanced research data across national boundaries to enable progress. The Horizon Europe mission board for cancer has identified access to data, knowledge, and digital services – accessible across the European Research Area through federated infrastructures – as a key enabling condition for success. The better organised cancer data is across Europe, the better and faster it can bring the fruits of new biological and technical innovations to the benefit of EU citizens and patients.

EOSC4Cancer's five selected use cases cover the patient's trajectory from cancer prevention, diagnosis, treatment, to medical management. With the use cases following the patient journey through cancer care they interact with different data types and sources, which become relevant at different stages of the journey. These data are systematically organised and made relevant for use in translational research, medical practice, and health outcomes. Colorectal cancer was chosen as a working case for representing a tumour type with abundant data and ample cross border collaborations.

With high-quality cancer data across Europe, more efficient biological and technical innovations will reach the citizens of the EU. In this context, we produced a roadmap, looking at the future of the European cancer data space beyond the timeframe of the project. It is now possible to move from developing research instruments to implementing systems for using and reusing cancer-related data (from both healthcare and research) based on existing robust technical developments. This concerted effort should constitute the foundations of the digital framework for the future UNCAN Platform. This transnational infrastructure will be the pillar to enable access to distributed and heterogeneous cancer-related data, as well as to deploy the software needed for the federated analysis required for cancer research. To guarantee sustainability and adoption, these efforts should be made in collaboration with major stakeholders, including research centres, hospitals and national cancer authorities.

To work effectively as a federated structure, the UNCAN Platform will require National Cancer Data Nodes to be set up in each Member State. These Nodes would need to act as coordinators of the local cancer community, connecting university hospitals, national registries, funders and governmental departments to develop their own national health data infrastructure. By design, the UNCAN Platform and the Nodes will comply with all existing regulations, above all aligning with the European Health Data Space (EHDS).

In this document, Chapter 1 sets out the core vision and ambition for advancing cancer research via a large-scale federated Cancer Digital Platform, a role likely to be filled by the Cancer Mission's UNCAN platform, building upon the work carried out by EOSC4Cancer.

Chapter 2 provides an overview of the cancer patient journey, with the corresponding use cases that were selected by EOSC4Cancer as demonstrators for each stage. Suggested future extensions to these use cases and new use cases to consider in future work follow this. The chapter ends by considering the steps in the researcher journey, and how these would need to be considered in the future platform.

Based on the above, Chapter 3 details the perspective of various user types foreseen to have an interest in the platform, how they can benefit from it and what the platform needs to take into consideration in its design to meet these aims. The chapter ends with an overview of the work done in EOSC4Cancer on training users and designing a RDMKit page dedicated to cancer data management.

Chapter 4 looks at three major political initiatives that guide the European perspective on cancer research, and drive much of the work in this space: the EU Mission on Cancer, the Europe Beating Cancer Plan and the European Health Data Space.

Following on from this, Chapter 5 considers how to implement UNCAN at a European but also National level. For such a federated system to thrive it is broadly recognised that a network of National Cancer Data Nodes will need to be set up, following a framework of recommended and manageable stages that are flexible enough to suit the needs of the different Member States.

Cancer data types, sources and improvements needed in their interoperability are covered in Chapter 6. Several data types are used in cancer research, covering molecules in the body, test results from screening programmes, as well as clinical, imaging and treatment data. The chapter finishes with advances made in integrating software within the EOSC4Cancer project. The resources generated by EOSC4Cancer will be ready for implementation by UNCAN from day one.

With Chapter 7 we consider the issue of biases that can affect cancer research and the results generated. In EOSC4Cancer we focussed on sex and gender biases, and report our findings here along with recommendations to mitigate these in future.

Chapter 8 features additional considerations for the future UNCAN Platform, including the importance of patient engagement in projects and the role of patient organisations, and the sustainability considerations for the outputs coming from EOSC4Cancer, so that they can continue to aid cancer research in the long term.

Ending with a comprehensive timeline in Chapter 9, this section looks forward over the next six years for the key milestones in European cancer initiatives, legal acts and infrastructures.

1 Vision/Ambition

We envision a federated digital platform for advancing European cancer research, leveraging EOSC4Cancer's outcomes and linking to work of current and upcoming synergy initiatives.

This Cancer Research Digital Platform (likely to be fulfilled by the Cancer Mission's future UNCAN platform) will be a computational environment containing the necessary software to process and analyse data in a federated fashion. Scientists and clinical researchers will be able to access high quality cancer data to drive research and innovation. As a user-friendly one-stop-shop, it will lower the barriers to work with heterogeneous data sources required for gaining a better understanding of cancer.

The design of this platform will therefore be relevant to the different stages of the cancer patient journey: prevention, screening, diagnosis, treatment, survivorship, as well as the triad of diagnostics, treatment and outcome. Thus, it links to the four pillars of the EU Mission on Cancer, leading to improved cancer care outcomes across all Member States.

The platform will be flexible enough to incorporate new developments in the cancer and data management fields as cancer research continuously produce new methods and instruments, particularly those to be produced by initiatives implemented in the EU Mission on Cancer and Europe's Beating Cancer Plan - from EC-funded projects such as the Genomic Data Infrastructure (GDI), to Member State efforts to set up the Network of Comprehensive Cancer Centres and National Cancer Data Nodes. The changing landscape with the European Open Science CCloud (EOSC) will also be important to incorporate and work with.

To make the design robust and sustainable the platform's development will need to involve multiple stakeholders to adapt and develop the platform to their needs and requirements, especially: a) researchers, to ensure that the platform suits their needs; b) clinicians, to find the right data for professional research; c) policy makers, to ensure alignment with current and future regulations and support HTA decision making; d) technology-oriented companies and professionals, to favour interoperability across heterogeneous systems; e) patients, to illustrate how data is being used to accelerate the understanding of cancer.

To work effectively as a federated structure, the platform will require National Cancer Data Nodes (NCDNs) to be set up in each Member State. These Nodes would need to act as coordinators of the local cancer community, connecting university hospitals, national registries, funders and governmental departments to develop their own national health data infrastructure. By design, the platform and the Nodes will comply with all existing regulations, above all aligning with the European Health Data Space (EHDS). The platform is important beyond cancer research in this context, by acting as an exemplar for the advancement of federated systems for research based on health data.

Current situation and environment

Cancer treatment and an associated increase in life expectancy in cancer patients has improved tremendously over recent years, largely thanks to the progress in cancer research. There has been a fast development of new technologies during this period, such as scanning/imaging, digital pathology, faster drug screening, new model organoids, animal models etc. This very fast development would not have been as effective without the very large community of computational biologists that have built software and developed research based on this new data.

To realise the full potential of these technologies and applications, the European Commission has launched a series of major initiatives, aiming to facilitate the reuse of health data for research: EU Mission on Cancer¹ and Europe's Beating Cancer Plan² for cancer data, as well as the European Health Data Space³ (EHDS) for health data in general. Each of these initiatives consists of numerous flagship initiatives and projects. Among the three actions, the years 2021-2027 cover major milestones likely to benefit cancer data use. A full list of these milestones is included in Chapter 9 (timeline). Since 2021, a basis for these major initiatives has been set with the launch of the Knowledge Centre on Cancer (2021)⁴ and the European Cancer Inequalities Registry⁵ (2022), the agreement by the European Parliament and Council on the Regulation for a European Health Data Space (2024)⁶ and the initiation of projects towards a fully populated Cancer Image Europe Platform (2024)⁷.

The EOSC4Cancer project fits among these milestones, by providing the basis of the technical infrastructure for the EU Mission on cancer, making diverse types of cancer data accessible and laying the foundation for data trajectories for future EU Cancer Mission projects.

EOSC4Cancer's long term goal

We aim for EOSC4Cancer's outcomes to feed into a future federated digital platform for cancer, which would connect recent technical developments in data handling and analysis with the needs of data and systems for cancer research. To make this possible, we need solid and sustainable implementations for secondary use of cancer-related data from any source (e.g. healthcare or research). This platform should be able to evolve to support future developments - adapting to new discoveries, data, and technological innovations.

This platform would offer access to high quality cancer data to scientists and clinical

¹

https://research-and-innovation.ec.europa.eu/funding/funding-opportunities/funding-programmes-and-open-calls/horizon-europe/eu-missions-horizon-europe/eu-mission-cancer_en

²

https://commission.europa.eu/strategy-and-policy/priorities-2019-2024/promoting-our-european-way-of-life/european-health-union/cancer-plan-europe_en

³ https://health.ec.europa.eu/ehealth-digital-health-and-care/european-health-data-space_en

⁴ https://knowledge4policy.ec.europa.eu/cancer_en

⁵ https://knowledge4policy.ec.europa.eu/cancer_en

⁶ https://health.ec.europa.eu/ehealth-digital-health-and-care/european-health-data-space_en

⁷ <https://digital-strategy.ec.europa.eu/en/policies/cancer-imaging>

researchers, in a user-friendly manner.

The platform will incorporate data sources relevant to different stages of cancer patient journey: prevention, diagnostics, treatment and survivorship, thus linking to the four pillars of the EU Mission on Cancer.

Its development should consider the work of initiatives that implement the EU Mission on Cancer and Europe's Beating Cancer Plan - on EU and Member State level. This relates above all to the European Initiative to Understand Cancer (UNCAN.eu), and European Cancer Patient Digital Center (ECPDC) platforms, as well as the platform implemented for the European Cancer Imaging Initiative^{8,9}.

An enabling factor of this goal have been the advancements in research data and software - namely a) adoption of FAIR principles; b) privacy-preserving technologies for remote data access; c) advanced machine learning models and AI techniques; d) High-Performance Computing-based solutions for the massive processing of cancer-related data.

Ultimately, the platform would aim to enable basic and clinical cancer researchers to discover, access, and integrate data from different domains, analyse it, interpret the results, and publish their findings.

2 Cancer Patient journey as the driver

Cancer-related data are quite diverse, often noisy with varying quality and come from different sources and domains. It is heterogeneous, distributed, complex (interlinked), semantically rich and very specialised in interpretation, i.e. requires subject-specific experts.

EOSC4Cancer uses one guiding theme to classify these data for its project work: the Cancer Patient Journey. Thus, the data follow the patients as they navigate their care along four stages: 1) primary prevention; 2) secondary prevention; diagnostics and treatment of 3) primary and 4) metastatic cancers. We are extending to a fifth stage of survivorship. Thus, it aligns perfectly with the four pillars of the EU Mission on Cancer, namely: 1) Understanding of cancer, 2) Prevention and early detection, 3) Diagnosis and treatment, 4) Quality of life (QoL) for patients and their families.¹⁰

The figure below, adapted from the EOSC4Cancer project, illustrates the various stages along the Cancer Patient Journey and offers a condensed view of the data sources used.

⁸

https://health.ec.europa.eu/latest-updates/updated-europes-beating-cancer-plan-implementation-road-map-2024-02-26_en

⁹ <https://dashboard.eucaim.cancerimage.eu/>

¹⁰

https://research-and-innovation.ec.europa.eu/funding/funding-opportunities/funding-programmes-and-open-calls/horizon-europe/eu-missions-horizon-europe/eu-mission-cancer_en

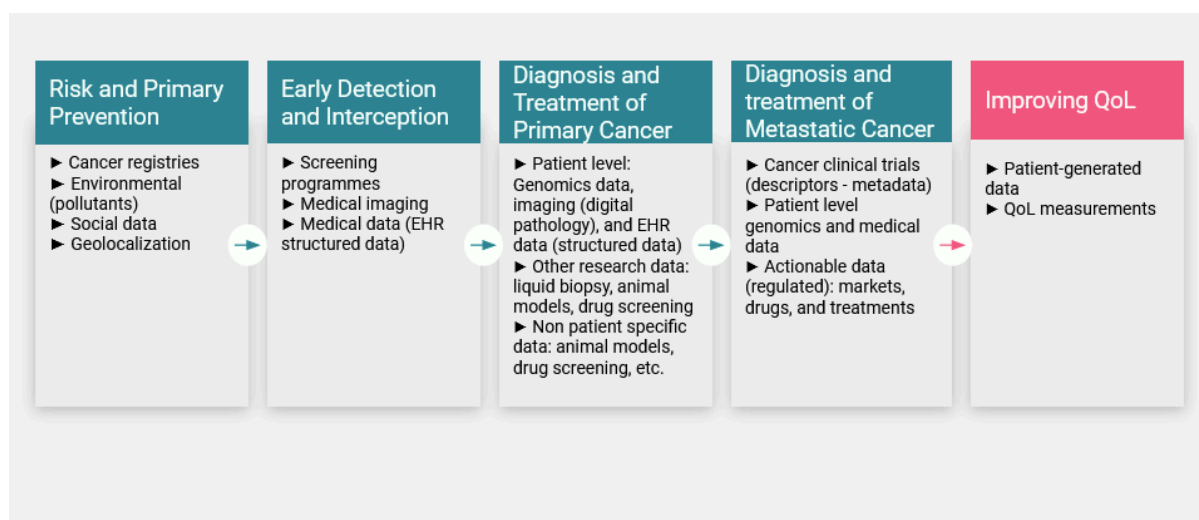


Figure 1. Adopted from the EOSC4Cancer project. Updated graphical depiction of the five main stages of the Cancer Patient Journey.

Answering specific questions along these stages implies, in many cases, the need of integrating data from different sources with different levels of granularity.

Existing use-cases

The existing EOSC4Cancer use cases were selected as prototypic demonstrators from each of the four main stages of the Cancer patient journey to illustrate the challenges and opportunities that arise when bringing together different datasets from across Europe. These use cases were also used to help identify the necessary software components e.g. virtual research environments, platforms, and clinical support systems, and to detect potential gaps in terms of data interoperability. More information on these aspects can be found in Chapter 6.

While this patient journey applies to virtually all tumour types, the EOSC4Cancer project initially focussed on colorectal cancer as a prototypic use case split into five use cases along the patient journey. Here we describe the project use cases, and how they could be extended in future:

1. Cancer risk identification and prevention by linking environmental data to cancer registry data (“primary prevention”)

Cancer registries are 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. In this use case, data from three different national cancer registries (Italian, Czech, and Dutch) were integrated with exposome data (collected from e.g. the EIRENE exposome network¹¹, EXPANSE project¹²) to investigate the relation between cancer incidence and environmental factors. The result is a report outlining guidelines, methodologies, challenges, and

¹¹ <https://eirene.eu/>

¹² <https://expansoproject.eu/>

recommendations on how best to perform such data integration. It also highlights the national differences in how such data integration efforts are conducted.

Extending this use case: The use cases could extend to more tumour types and higher volume, granularity and data quality. Extending this effort to additional data sources and countries would require substantial effort in professionalising cancer registries across Europe. For environmental risk factors, the link with EIRENE and national nodes thereof could be further developed. For genetic risk factors, the use and availability of data for (next generation) polygenic risk scores would need to be developed.

2. Data driven optimisation of cancer screening programs (“secondary prevention”)

Early detection is critically important to improve survival rates in most major tumour types which has prompted nation-wide screening programmes in many European countries. These programmes produce highly relevant data sets for further (data-driven) research on early cancer diagnostics, yet analysis of already existing data is hampered by the heterogeneity with which each country and study reports their results. This use case has focused on harmonising the codebook variables for transnational use to facilitate future studies at this stage of the patient journey. For Catalunya (Spain), the Netherlands and Italy this standardisation has been achieved, and conversion and remapping of their original codebook towards the standardised model was performed for the screening data of the Czech Republic as validation. A publication summarising these standardisation efforts is under preparation.

Extending this use case: For the future, various steps could be taken in this use case. First of all the data should be made accessible across EU cancer screening programmes. Data collection should also be tuned towards early detection beyond the common cancer currently screened for. In addition, we should prepare for the challenge of multi-cancer early detection¹³. Another step would be to convert the harmonised codebook into a widely used clinical data model, such as FHIR or OMOP, to allow for wider interoperability, standardized analytics and federated analysis.

3. Data-driven treatment selection for localised tumours using multiple patient-derived data types (“diagnostics” and “treatment”)

This use case, focused on data-driven treatment selection for localised tumours, addresses the treatment decision-making stage of the patient journey. The use case involves integrating multiple patient-derived data types which are organised through standardised and generalisable templates for managing complex, longitudinal data in studies investigating localised cancers. The use case aims to prepare data for a molecular tumour board, enabling more precise diagnostics, improved decision-making, personalised treatment options, and enhanced care for oncology patients. This will be achieved by consolidating all relevant data in a single, integrated platform, using the broadly used open-source cBioPortal tool standardising on a solution also used for use case 4. The platform facilitates secure and seamless access, analysis, and sharing of clinical, genomic, and other patient-specific information, fostering better collaboration and insights.

¹³ (see ESCALATION project:
<https://www.nki.nl/research/research-groups/gerrit-meijer/escalation-study/>)

4. Data-driven treatment selection for localised tumour: improving the treatment of colorectal cancer by the inclusion of circulating tumour DNA information (“treatment”)

This use case has integrated the clinical, biosample and omics data for localised tumours in the EOSC4Cancer reference implementation of cBioPortal installed at the Dutch national Health-RI instance. Data sources included national registries such as the Netherlands Cancer Registry and the Dutch pathology registry PALGA. The longitudinal ctDNA data results were modelled in a standardised manner, facilitating data capturing as well as data-integration and visualisation in cBioPortal. Through its intuitive and user-friendly interface, non-bioinformaticians will be capable of viewing, querying and analysing the data in cBioPortal.

In order to facilitate easy data dissemination allowing the collected data to be reused by others, an interface from a cBioPortal instance to the European Genome-phenome Archive (EGA), is being created. This allows for published results to be made available to improve the scientific route from bench to bedside: peer review of research results, quicker translation time of meaningful results to implementation in the clinics etc.

Extending use cases 3 and 4: The multimodality of treatment (surgery, radiotherapy, systemic, IO, ATMPs, etc) needs to be addressed further. Significant extension could be achieved through systematic use of health data (in particular through EHDS mechanisms) which will require “cohort level” quality. We should also consider “trials within cohorts” as a default option for real world investigations. For the molecular data, the EOSC4Cancer proposal is to make raw panel sequencing data across (inter)national labs available for periodically repeating the latest data analysis pipeline for variant (or signature / other signal) calling and classification, facilitating the update and evaluation of the diagnosis-treatment-outcome results with the latest scientific data. Next Generation Sequencing panels are generated in many different molecular pathology labs. These use cases, as well as the whole EOSC4Cancer process, lends itself for a federated approach; leave the raw data on premise, send the updated algorithm to these sites and then return the calls and classification to a central repository, e.g. cBioPortal.

5. Connecting omics data from multiple sources to a Clinical Decision Support System for precision treatment of metastatic Colorectal Cancer (“treatment”)

This use case focuses on the necessary data infrastructures and format specifications for analysing tumour data from patients through Clinical Decision Support Systems. Molecular Tumour Board Portals designed to guide biomarker-driven precision medicine interventions, will require direct access to up-to-date information about the functional significance of genetic alterations in a given patient's tumour and to be able to pinpoint standard-of-care drug biomarkers as well as investigational treatment options. In this context, EOSC4Cancer developed recommendations for this process, utilising a cohort of metastatic Colorectal cancer patients from three centres (Vall D’Hebron Institute of Oncology, Karolinska Institutet, and Netherlands Cancer Institute) as representative examples of real-world data. Also the usage of Large Language Models to systematically extract the required clinical annotation from Electronic Health Record (EHR) systems was investigated.

Extending the use case: For future developments, a number of critical issues should be addressed to improve the utility and interoperability of the molecular tumour board installed in Europe. The primary focus should be on providing access to enhanced information, including original clinical trial data. Additionally, there should be an emphasis on creating frameworks that facilitate the development of Tumour

Board Portals with compatible technologies and functionalities, creating environments of tools and data that will enhance monitoring in real-world settings to determine whether new drugs deliver the outcomes indicated in trials. Establishing a pan-European repository for all panel sequencing data, modelled after the successful large-scale AACR GENIE¹⁴ project in the U.S., will be an additional key step. Finally, maintaining this platform through real-time updates of variant calling and classification across this comprehensive EU resource will be crucial for its ongoing effectiveness.

Extending the existing use cases

The EOSC4Cancer use cases were carefully selected along the cancer patient journey ensuring representativeness for a broad range of tumour types. They will need to be preserved and maintained, preferably in the context of the upcoming UNCAN platform. Some specific ideas on how to move forward with each of these use cases have already been suggested above in each of the use case descriptions.

Apart from the use case extension specific to the use cases mentioned above, the future scale-up of these use cases in the future UNCAN platform should take place along multiple axes:

1. Generalise to other tumour types than colorectal cancer, including addressing the cross-tumor type data challenge, recognising the large differences in knowledge and treatment between cancer types, in particular cancer types with poor diagnosis and paediatric cancer. Patients often suffer from different tumour types, while clinical and research communities tend to be organised around specific tumour types.
2. Extend adoption of the prototypic use cases developed within EOSC4Cancer (more users) and populate these use cases with more granular data and substantially higher volumes. The latter is a critical requirement to support AI model development and validation, and the development of specific AI foundation models for oncology.
3. As the technology evolves and patients become more engaged in research activities, it is necessary to extend existing prototypical use cases to incorporate new data sources, e.g. the ones reported directly by patients through the use of different personal devices. Adding new (alternative) data sources also offers opportunities to truly engage patients as active participants in the research process rather than only as “data subjects”. This would require dedicated user interfaces for patients to manage their own health data, as well as suitable legal conditions and data quality control.
4. Data collection and usage across the patient journey should cover three different levels:
 - a. Societal, covering topics like health economics / HTA, Health policy making, and public health
 - b. Individual/organism, focusing on topics such as long and healthy living, reduce the loss of health (including support in coping with health loss), and the longitudinal collection of personal health data (to be organised in the context of EHDS)

¹⁴ <https://www.aacr.org/professionals/research/aacr-project-genie/>

- c. Cellular, which is very focused on understanding cancer in basic cancer research using techniques such as multi-omics, single cell spatial profiling, unravelling the dark genome, etc.
- 5. New use cases will arise from projects funded to further mature the future UNCAN platform.

The recovery of cancer patients and the quality of life after recovery is an important additional stage in the patient journey that should be added as an additional use case moving forward. Longitudinal data collections add vast value in cancer, following up survivors for several years to collect any relapse data and also their quality of life with various treatment options. Additional related data would include social data (diet, lifestyle factors, sleep, mental health, etc.) and also looking at prehab data (preparing in advance of cancer treatment) alongside rehab. Obtaining digital biomarkers from wearables such as smart watches provides another source of longitudinal data.

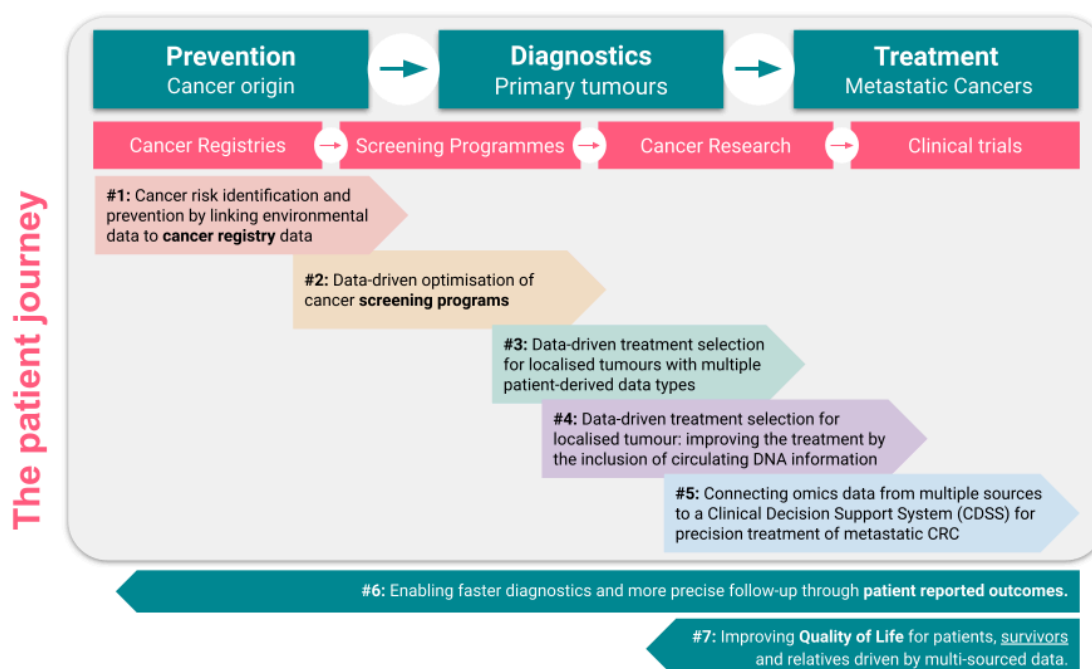


Figure 2. Extended EOSC4Cancer use cases: Addition of at least two use cases to represent the inclusion of patient-reported outcomes and how data can be potentially used to increase the quality of life of patients, survivors and relatives.

Researcher journey

Orthogonal to the patient journey, there is also a researcher journey through each of the use cases, which also needs to be supported at each of its stages. Most oncology studies, whether basic research, translational research or clinical studies, are passing through very similar stages: conception, grant application, data workflow planning, ethical approval, study preparation, study execution and analysis, dissemination of results, and finally the archival of data for future reuse. If use cases operate at later stages of the clinical research pipeline then additional steps come into play such as HTA analysis and regulatory pathways. The

EOSC4Cancer integral view should cover solutions for all, or at least most, of the stages of this researcher journey.

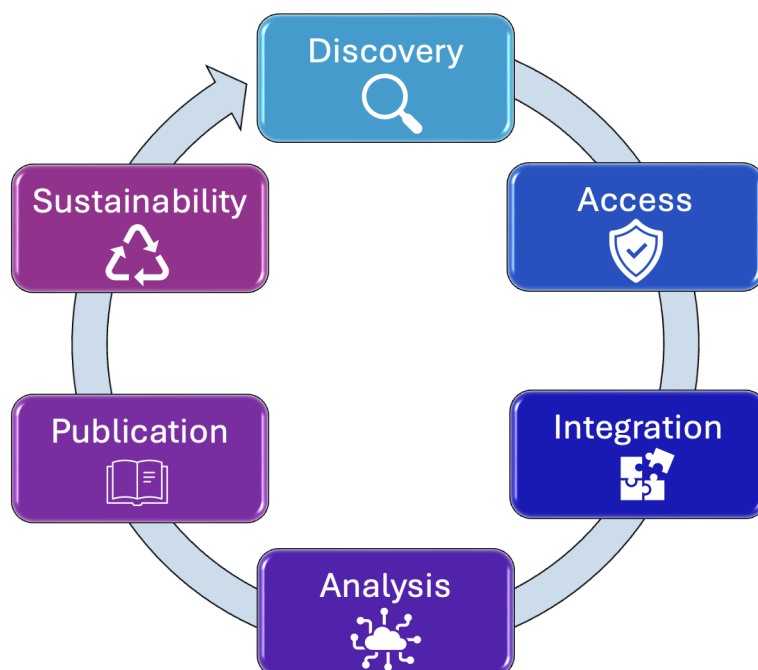


Figure 3. Researcher journey: The stages of the researcher journey when considering the reuse of data and working across national boundaries.

Taking into account the different steps that researchers often do as part of their scientific activities, we can dive into the challenges and opportunities associated with each of these stages:

a) **Discovery:**

Cancer-related data need to be discoverable across various systems, making sure that previously generated data becomes available for further reuse. This includes making datasets coming from healthcare persistent once they are extracted for secondary use. Such data should be at least adequately pseudonymized or anonymized to protect the patients' identity and sensitive personal data. Researchers bringing new datasets into the ecosystem need clear guidelines on how to do that systematically. As an example, applicable guidance is posed by the FAIR principles, and should also be applied within the UNCAN platform.

b) **Access:**

Secure and distributed data access using standard access protocols, including the automatic evaluation of the access rights and permissions, as well as the approval by the corresponding ethical committees. Establishment of ethical and legal frameworks to govern data access and ensuring compliance with relevant regulations are needed to safeguard data privacy, ideally following the framework that will be set out by the EHDS. Further, robust mechanisms such as sensitive data

controlling, authentication, and encryption to prevent unauthorised access need to be implemented.

c) **Integration:**

Requires well-labelled data/metadata that will be syntactically and semantically interoperable, including data from very different experimental sources and analytical procedures. Harmonisation of data formats, standards and interoperability protocols support better integration across different data settings. This aligns with the Beating Cancer Plan and the Cancer Mission, aiming at leveraging insights from multidimensional datasets, including clinical, genomic, epidemiology, and patient-reported data, to increase personalisation of cancer care. Data integration is a key enabler in cancer research, diagnosis, and treatment from diverse information sources.

d) **Analysis:**

Trusted Research Environments (TREs) or Secure Processing Environments (SPEs) in EHDS terms will be vital here, providing the users with a single location to not only access datasets but also with analytical tools they require for their analysis when working with sensitive data and with access to sufficient computational and storage capacity. The software and algorithmic needs of researchers within these TREs/SPEs will vary wildly depending on the research stage and question. EOSC4Cancer addresses this diversity by containerising relevant software to provide portability of software solutions to TREs. More detail of the work of EOSC4Cancer in this area can be found in Chapter 6.

e) **Publication & reuse:**

After completion of the analysis the researcher will publish the results. Traditionally, this is done in scientific journals where data is only published as supplementary data (if at all). The EOSC4Cancer approach promotes making data systematically available for reuse, which is supported with various tools and pipelines. This requires that data should be made discoverable in persistent repositories (e.g. European Genome-phenome Archive). The underlying patient-level or raw research data should be archived in a reusable manner for a minimum of 5-10 years, preferably in easily accessible solutions such as cBioPortal.

f) **Sustainability:**

The framework has to guarantee the continuity of the implementations and the persistence of the data, developed in a sustainable environment associated with entities that will guarantee open access to the resources. This is in line with the perspective of Europe's Beating Cancer Plan and the EU Mission on Cancer, which focus on long-term, sustainable frameworks and systems to promote cancer research, prevention and care, in a holistic and equitable approach. A key component of sustainability will be the National Cancer Data Nodes, which will form a network of nodes across the Member States and associated countries acting as coordinators of the local cancer research community, connecting all relevant stakeholders (e.g. university hospitals, national registries, funders and governmental departments) to

develop their own national health data infrastructure. These are further described in Chapter 5.

3 User perspective

The success of the future UNCAN platform will depend on its ability to build relationships and encourage engagement with its future users and stakeholders. The platform must be able to encourage and facilitate data deposition, and enable data reuse, doing so via a user-friendly interface with clear guidance and training available to future users.

EOSC4Cancer's work on cancer data user profiles and their training needs provides us with important input on how to achieve this user-friendly design.

Platform user profiles

The platform we envision should serve different categories of users:

1. Policy makers:

These users are responsible for setting policy and regulation on a national / Member State level. Policy makers will want to be able to extract information from research to aid in building policies or strategic frameworks. For example a governing board of a public health directorate may wish to access data to help implement a national cancer task force.

2. Researchers:

Arguably the main focus of the platform, these users will want to obtain, process, create, store and share research data. Cancer researchers should be able to browse datasets to be able to find those suitable for carrying out their own study, which may focus on any element of the patient journey. They may need tools and suitable processing environments to process the data, and finally be well signposted to the most suitable location to store their results for others to reuse in future.

3. Clinicians:

Whilst some clinicians will be 'clinician researchers', which fit into the category above, others will be looking for more nuanced data when dealing with rare cancers for example. Others may be involved in making anonymised health data accessible via a national plan for secondary reuse of health data, and will benefit from guidance of recommended ontologies.

4. Patients, cancer survivors, caregivers and patient organisations:

This covers a broad range of users, such as current cancer patients, cancer survivors, caregivers, citizens with an interest in cancer and its associated risk factors, and respective organisations representing these groups. Such users may have different interests in a cancer digital platform, for example to find the latest information on a

particular cancer type, to understand how to take part in research themselves or to see how many datasets (perhaps of a specific cancer type) are being made available through Member State efforts.

5. Technology professionals:

The platform needs to provide more than just data, it also needs to provide well tested and interoperable tools for processing data. IT experts designing, implementing, and maintaining software will be able to make their tools available for use by researchers. The platform should also provide a rich source of real world data required for training AI tools effectively.

Capacity building

All users will benefit from a well designed portal, ensuring that the information required to use and understand the platform is prominent, and that it is built with all types of users in mind.

Users will need to know how to interact with the platform. This training could be done centrally, but this aspect could also leverage national expertise via the National Cancer Data Nodes (see chapter 5). Efforts will be required to ensure that the skills acquired via a curated training portfolio are transferable within the platform, across countries for example.

One way to do this would be to have a standard definition of bespoke Learning Paths, consistently mapping the skills and expected roles/users for the platform. This approach should assist in maintaining the various learning pathways, as well as help identify any potential gaps in the training offered.

The content of the training resources should include both documentation on the resources available, but also training on how to use them effectively to get the most from these resources.

Establishing an RDM knowledge base

The research data management (RDM) Knowledge base will be a gateway for the broader cancer scientific community to come together and create RDM best practices for cancer data according to FAIR principles and open science standards. This will meet the community's evolving needs and serve as a distinctive reference point, helping to prevent duplication of information and effort across the community. The data management guidelines will span across patient journeys and the multiple data types produced during cancer diagnostics and treatment. These best practices and guidelines will be showcased in the RDMkit¹⁵.

The RDMkit Cancer Data Management page¹⁶ begun by EOSC4Cancer will be a community-driven, coordinated effort, encouraging project members to share their expertise in building a shared knowledge base. It will act as a collaborative platform where stakeholders can actively contribute to the creation and refinement of RDM guidelines and standards. Through an organised

¹⁵ <https://rdmkit.elixir-europe.org>

¹⁶ https://rdmkit.elixir-europe.org/cancer_data

contribution process, members can offer best practices within their specific areas of expertise, enriching the platform with insights grounded in practical experience. Contributor roles will be recognised, and each contribution will be appropriately acknowledged. The work for the initial RDMkit Cancer Data Management page was completed at the end of the EOSC4Cancer project, after that point the wider community is able to continue to improve this resource on an ongoing basis, via annual contentathons for example.

4 European perspective on cancer research

Three high level political initiatives guide the European perspective on cancer research until 2030 and beyond: the EU Mission on Cancer¹⁷, the Europe Beating Cancer Plan¹⁸ and the European Health Data Space¹⁹. This chapter outlines the implementation pathways shaped by these initiatives.

EU Mission on Cancer

The EU Mission on Cancer²⁰ works in synergy with the implementation of Europe's Beating Cancer Plan and links Research & Innovation policies by:

- Generating knowledge and further evidence in understanding of cancer, prevention, diagnosis, treatment, and quality of life of cancer patients
- Engaging with European citizens, including patients and cancer survivors
- Establishing national cancer hubs in Member States and Associated Countries
- Delivering a sound basis and scientific evidence for the overall implementation of Europe's Beating Cancer Plan.

There are several initiatives under the European Commission working towards accomplishing the Cancer Mission's ambitions - including UNCAN.eu platform and the European Cancer Patient Digital Centre (ECPDC), also highlighted in chapter 5. EOSC4Cancer prepares the technical infrastructure for the EU Mission on Cancer, by organising cancer data more efficiently, enhancing access to innovations for EU patients.

¹⁷

https://research-and-innovation.ec.europa.eu/funding/funding-opportunities/funding-programmes-and-open-calls/horizon-europe/eu-missions-horizon-europe/eu-mission-cancer_en

¹⁸

https://commission.europa.eu/strategy-and-policy/priorities-2019-2024/promoting-our-european-way-of-life/european-health-union/cancer-plan-europe_en

¹⁹ https://health.ec.europa.eu/ehealth-digital-health-and-care/european-health-data-space_en

²⁰

https://research-and-innovation.ec.europa.eu/funding/funding-opportunities/funding-programmes-and-open-calls/horizon-europe/eu-missions-horizon-europe/eu-mission-cancer_en

Europe's Beating Cancer Plan

Europe's Beating Cancer Plan²¹ aims to improve the lives of more than 3 million people by 2030 by enhancing prevention, early detection, diagnostics, therapeutics, and quality of life. It is currently estimated that around 40% of cancer cases are preventable with the implementation of adequate cancer prevention strategies. One hallmark of Europe's Beating Cancer Plan is improving early diagnosis of cancer through screenings, increasing the chance for recovery and rehabilitation.

Europe's Beating Cancer Plan also includes efforts to ensure equal access to cancer diagnosis and treatment among member states, as well as better services for cancer survivors (see for example the Flagship 6: The new 'Cancer Diagnostic and Treatment for All' Europe's Beating Cancer Plan). These efforts are supported by scientific evidence and tools generated and gathered by the EU Mission on Cancer.

Additionally, several specialised working groups operate under the umbrella of Europe's Beating Cancer Plan - e.g. the Subgroup on Cancer under the Expert Group on Public Health.

European Health Data Space

The European Health Data Space (EHDS) Regulation entered into force in February 2025²². It aims to empower individuals through better access to their health data, supporting free movement of health data with people and outlining rules for use of health data for research, innovation and policy making. Many stakeholders have urged us to clarify the link between the EHDS and cancer data infrastructures - created by EOSC4Cancer and beyond.

Use of cancer data for research is considered secondary use of health data in the EHDS Regulation. Any virtual platform for cancer research will most likely be a data holder. Article 51 mentions priority data categories for secondary use - including some especially crucial for cancer research: genomic, multi-omic and biobank data. Thus, the federated digital platform to advance cancer research should provide access to this data. This will also contribute to goals communicated in the EC Communication on the European Health Union of May 2024: leveraging both the EHDS and specialised infrastructures to enable early detection, prevention and treatment, as well as having a critical mass of genomic data to enable secure access without transferring highly sensitive data.²³

²¹

https://commission.europa.eu/strategy-and-policy/priorities-2019-2024/promoting-our-european-way-of-life/european-health-union/cancer-plan-europe_en

²² https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=OJ:L_202500327

²³

https://health.ec.europa.eu/document/download/6e26bad9-5722-4c95-8bc5-4c21d8e370dd_en?filename=policy_com-2024-206_fr.pdf

Cancer Data initiatives, by developing infrastructures for specific types of health data - specifically for genomic data (e.g. via GDI²⁴, B1MG²⁵) and Cancer Imaging (e.g. via EUCAIM²⁶) need to align with EHDS.

Health Data Access Bodies are the main national institution for secondary use, as outlined in Article 55-59 of the EHDS Regulation. These bodies are responsible for managing and deciding on requests for health data access, process health data, ensure confidentiality and IP rights and manage a national dataset catalogue.

The figure below depicts the interaction in the cross-border secondary use infrastructure.

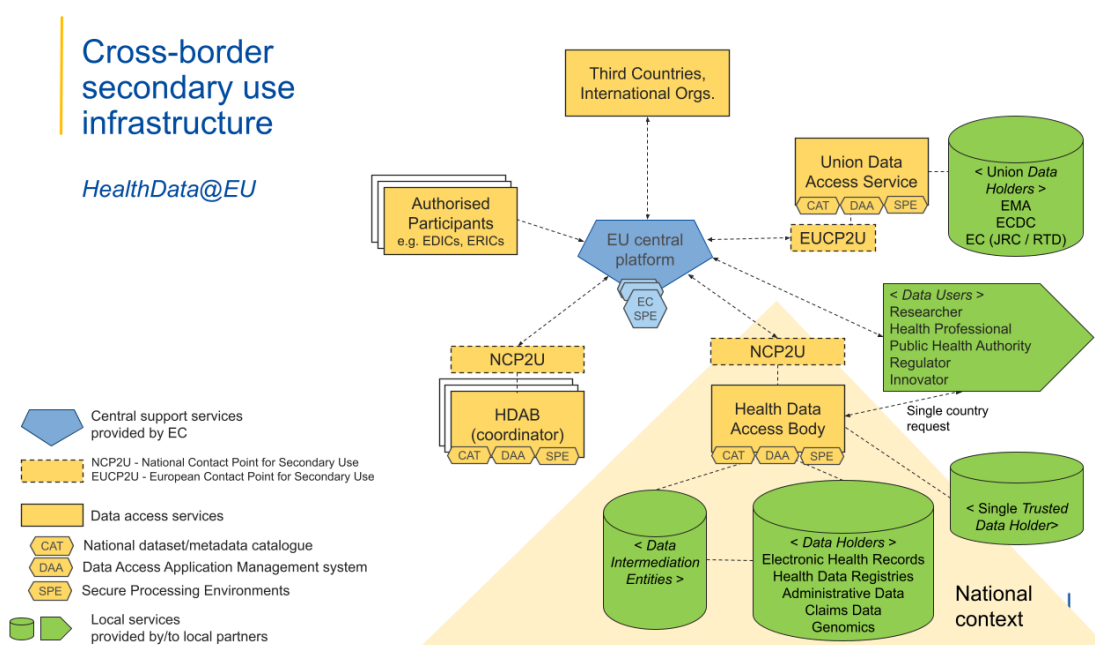


Figure 4. Depiction of cross-border secondary use infrastructure (Source: European Commission, presentation at Roadmap Consultation Workshop 2 December 2024).

The federated digital platform to advance cancer research could also contribute best practices on the ongoing capacity building in HDABs on the following topics²⁷:

- Expertise in data harmonisation
 - Metadata standards to be used in national data set catalogues
- Developed artefacts from cancer projects
 - Technical capacity to analyse multi-omic data
- Sharing of specific types of data
 - Large data sets
 - Sensitive data
 - Imaging data

²⁴ <https://gdi.onemilliongenomes.eu/>

²⁵ <https://b1mg-project.eu/>

²⁶ <https://digital-strategy.ec.europa.eu/en/policies/cancer-imaging>

²⁷ Part of these tasks could be performed by the National Cancer Data Nodes - see chapter 5.

EOSC4Cancer outcomes can inform the upcoming specification laws the European Commission is preparing on some of these topics.

TEHDAS2 and HealthData@EU

In this context, it is important to consider the outcomes of TEHDAS2²⁸ and HealthData@EU Pilot²⁹, that are building the EHDS for secondary use. Respectively, these projects will specify the upcoming infrastructure for secondary use of health data and provide first results on proof of concept of implementing it.

An EOSC4Cancer and TEHDAS2 workshop “European Health Data Space Meets Cancer Data” 30 January 2025 in Warsaw explored the relationship between EHDS secondary and cancer data in depth. The cancer research community was highlighted as a leading example for EHDS implementation: with a lot of high-quality data, structured data collections and prior experiences with data sharing initiatives. TEHDAS2 plays a central role in bridging the gap between these cancer research practices and the EHDS. Its guidelines help researchers find suitable datasets. Its metadata catalogue, stakeholder collaboration models and interpretation of EHDS legal requirements also guide researchers on the conditions under which they can use it. Yet, existing TEHDAS2 guidelines need to be complemented by Standard Operating Procedures (SOPs).³⁰

Also the outcomes of HealthyCloud should be considered in this regard. Its Draft Strategic Agenda³¹ recommends both a HealthData@EU community interface service and an EOSC sensitive data users service.

Link to EOSC

EOSC4Cancer’s legacy is also well positioned to help establish the link between EHDS and the EOSC infrastructure. Throughout the project, we fostered the exchange on health data and the EU Mission on Cancer within the EOSC ecosystem. This happened especially in specific exchange with EOSC projects with a health focus – such as Sci-Lake³², RAISE³³ and BY-COVID³⁴. Insights include:

- Different from other EOSC data use cases, health data is sensitive and thus requires stricter compliance on data protection.
- The EOSC Health Data Taskforce³⁵ provides an important link on interoperability and sharing of health data between the EHDS and EOSC.
- The upcoming EOSC governance structure is still being defined. From the perspective of cancer data, an effective infrastructure would enable connections at multiple levels: at the EU level, linking EOSC to the UNCAN.eu platform; and the national level, connecting to individual cancer data nodes and cancer mission hubs.

²⁸ <https://tehdas.eu/>

²⁹ <https://ehds2pilot.eu/>

³⁰ Source: Workshop European Health Data Space Meets Cancer Data 30 January 2025

³¹ <https://zenodo.org/records/7331832#.Y9KUYnbMKUI>

³² <https://scilake.eu/>

³³ <https://raise-science.eu/>

³⁴ <https://by-covid.org/>

³⁵ <https://eosc.eu/wp-content/uploads/2024/07/Health-Data-TF-ToR.pdf>

Another point is the alignment of EOSC Nodes with cancer research infrastructures - both on EU and on national level. This means avoiding redundancies between cancer research flagship initiatives - such as UNCAN-CONNECT (see Chapter 5) - and the services that the EOSC EU Node is creating.³⁶

The eCancer thematic group

The eCancer Expert Group is a thematic group formed by the European Commission. The secretariat of this group is led jointly by DG RTD (Cancer Mission) and DG SANTE. The eCancer group provides advice on the implementation and sustainability of UNCAN and ECPDC, and their integration under the European Health Data Space (EHDS). The experts provide insights into national views on digital platforms, national data functions while also assessing potentially arising barriers in establishing national cancer data nodes. Further, they provide advice on governance and financing of digital infrastructure.³⁷

The EU AI Act

The European Union Artificial Intelligence Act (EU AI Act, (Regulation (EU) 2024/1689)³⁸) is the world's first comprehensive legal framework designed to regulate artificial intelligence. Adopted in June 2024, it aims to ensure the safe, transparent, and ethical development and use of AI technologies across Europe. By employing a risk-based approach, the Act categorizes AI systems into four levels: unacceptable, high, limited, and minimal risk, each subject to tailored regulatory measures. The Act prohibits harmful applications like social scoring while imposing strict requirements on high-risk systems, such as those used in hiring or biometric identification. The legislation also promotes human oversight, environmental sustainability, and data privacy, positioning Europe as a leader in trustworthy AI innovation. The AI Act entered into force on 1 August 2024, and will be fully applicable on 2 August 2026, with some exceptions. The Act is expected to have significant impacts and benefits for European cancer research, particularly in advancing precision medicine and improving diagnostic and treatment outcomes.

Potential impacts include:

- The Act provides a clear regulatory framework that helps researchers balance innovation with legal compliance in cancer research. Since many AI systems used in this field, such as imaging tools and predictive models, are classified as “high-risk”, they must meet stringent requirements, including robust data governance, risk mitigation systems, transparency measures, and human oversight. This reduces the likelihood of non-compliance and associated penalties, creating a more secure environment for AI development³⁹. The Act also emphasizes ethical AI use and strict adherence to privacy standards, minimizing liabilities related to data misuse or breaches when handling sensitive health information.

³⁶

<https://open-science-cloud.ec.europa.eu/news/european-commission-announces-eosc-eu-nodes-transition-full-production>

³⁷ See Sub-group on Cancer (E03884/1) in the Register of Commission Expert Groups and Other Similar Entities (<https://ec.europa.eu/transparency/expert-groups-register/screen/home?lang=en>).

³⁸ <https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A32024R1689>

³⁹ https://health.ec.europa.eu/ehealth-digital-health-and-care/artificial-intelligence-healthcare_en

Importantly, it includes exemptions for AI systems developed exclusively for scientific research, offering researchers flexibility while still ensuring ethical and legal accountability⁴⁰.

- The Act supports initiatives like EUCAIM, which will enable the development of AI tools to improve cancer diagnosis, treatment, and predictive medicine, benefiting patients across Europe⁴¹. Beyond imaging, initiatives integrating diverse datasets, such as molecular, genomic, clinical, and radiomic data, will foster the creation of multimodal AI-powered clinical prediction models that enhance physicians' decision-making and empower patients in shared treatment decisions. By addressing challenges like data fragmentation and standardisation, these initiatives unlock the full potential of AI algorithms for advancing precision oncology⁴².

In terms of benefits,

- The Act encourages investment in trustworthy AI tools while minimizing administrative burdens for researchers and small enterprises, promoting innovation in cancer diagnostics and treatment technologies.
- By aligning regulatory measures with the needs of researchers and healthcare systems, the AI Act supports transformative cancer care across Europe. It harmonizes AI regulations, fosters cross-border collaboration, and ensures equitable access to AI-powered healthcare solutions. In doing so, the Act positions Europe as a leader in ethically driven, innovative cancer research and care⁴³.
- AI applications such as medical image analysis, generation of data for synthetic control arms in clinical trials, and patient identification systems for clinical trials, hold transformative potential for improving diagnostic accuracy, optimising treatments and streamlining clinical research in the field of cancer. These tools can detect subtle disease markers, simulate control groups using real-world data, and identify eligible patients with unprecedented precision. Given their significant impact on patient outcomes, regulatory decisions, and data privacy, such systems are classified as “high-risk” under the EU AI Act. This classification ensures compliance with stringent requirements while enabling the responsible use of cutting-edge technologies that can revolutionise cancer care across Europe.

5 National vs European level for implementation

Focussing on interoperability across the Member States and EU level is crucial to avoid incompatible solutions or confusion around overarching governance. For this to succeed it is vital to create the capacity to share interoperable data from different sources securely. Cancer related data are often personal and sensitive data, relating to health characteristics of individuals, and must therefore follow strict rules and use, security and sharing to protect

⁴⁰ <https://doi.org/10.1016/j.healthpol.2024.105152>,
<https://eur-lex.europa.eu/legal-content/EN/ALL/?uri=CELEX:32024R1689>

⁴¹ <https://cancerimage.eu/>

⁴²

https://health.ec.europa.eu/document/download/6e4f5ecb-ea9c-445a-a661-7147809aa255_en?file_name=policy_20241126_js02_en.pdf

⁴³ https://health.ec.europa.eu/ehealth-digital-health-and-care/artificial-intelligence-healthcare_en

the data subjects. These aims align with the recently approved European Health Data Space legislation. Important players here will be the Health Data Access Bodies, the future UNCAN platform and the National Cancer Data Nodes.

UNCAN-CONNECT

Building on the previous 4.UNCAN.eu CSA blueprint, the upcoming UNCAN-CONNECT project (2025-2030)⁴⁴ will build a large-scale federated cancer data network. This infrastructure for federated query and analysis will allow cancer researchers and clinicians to identify new biomarkers, stratify patient groups, and design better-targeted therapies by analyzing data patterns on a large scale. The platform will aim to employ common data models, interoperable formats and AI-driven tools. The project will have a number of use cases to cover a range of cancer types: a) pediatric tumors, b) lymphoid malignancies, c) pancreatic cancer, d) ovarian cancer, e) lung cancer, and f) prostate cancer. This will cover a broad range of attributes such as rarity, gender specific disease, younger age groups & hard to treat cancers.

The UNCAN-CONNECT platform will need to integrate genomic, clinical, and molecular data for these cancer types, to demonstrate the system's broad applicability. The platform can build on EOSC4Cancer's technical infrastructures, use case experiences and stakeholder consultation outcomes.

National Cancer Data Nodes

For this European level plan to work, the complexities and differences between the Member States must be understood and accounted for. This can best be done by building a network of National Cancer Data Nodes (NCDNs) that will link to UNCAN-CONNECT. These nodes will be federated hubs, responsible for standardising and harmonising data procedures within their country/region, and providing the digital tools for research collaboration - all aligned with national EHDS roles. The CANDLE project will help guide the formation of these NCDNs and support effective implementation mechanisms.

These NCDNs need not start from scratch, many projects and initiatives are already working on federated data infrastructures. The GDI project aims to enable access to genomic, and related phenotypic and clinical data, held in databases across Europe by establishing a secure federated infrastructure to access such data. The project already incorporates a use case for cancer related research to facilitate the future needs of these researchers.

Additionally, pre-existing networks working across member states can greatly benefit these nodes, for example the ENCR (European Network of Cancer Registries) - see more in chapter 6. With one national organisation/node coordinating activities within their jurisdiction the potential for large scale collaborative projects across member states improves drastically. Common data standards can be more easily adopted across valuable data sources, such as national healthcare systems, cancer registries and national research

⁴⁴ See e.g. <https://www.muni.cz/vyzkum/projekty/74328>
<https://www.presseportal.de/pm/154764/5944789>

efforts. This potential could spread to other areas, such as harmonised patient consent forms to facilitate cross border legal agreements.

While all NCDNs share the goal to improve interoperability in multiple dimensions of the (re)use of cancer data, ultimate implementations of NCDNs may vary across member states e.g. as a consequence of the already existing organisation of the cancer data field at regional or national levels. Examples below provide some illustrations of how existing resources could be leveraged:

1. NCDNs that oversee national research on behalf of their government in cancer prevention, diagnostics, treatment, and rehabilitation, whilst also representing regional healthcare services and medical universities, and establishing a National Cancer Strategy.
2. NCDNs that are built on a preexisting national node of a European Research Infrastructures (e.g. ESFRIs) or national health data initiatives. This model benefits from best practices set up by ESFRIs, following the same hub-node structure as European infrastructures. The NCDN can serve as national coordinator of the cancer community, connecting (university) hospitals, national registries, funders, and the government to develop a national health data infrastructure
3. NCDNs stemming from a leading research institute/university with a long-standing track record in facilitating multidisciplinary collaborations between AI experts, Digital Health, Data Science with patients, health care professionals and researchers, ultimately gaining governmental support to become the nominated organisation for NCDN.
4. Regional nodes depending on the specific organisation of the research and health system of the Member State, i.e. the NCDNs may operate at regional level in federally organised countries.

The resulting network of national nodes can only thrive when working closely together, by sharing best practices in data management and recommended IT infrastructure. This would allow later nodes to seamlessly join this network of early 'vanguard' nodes benefitting from the lessons learned in these nodes avoiding the issues they already encountered. Moreover, multinational oncology research projects will substantially benefit from the joint support of this node network to overcoming transnational data sharing issues at the technical and governance level.

Again, efficient engagement with existing infrastructures is crucial here. These include: eCancer thematic group, Health Data Access Body (HDAB), National Cancer Mission Hub, National Cancer Registries, Comprehensive Cancer Center, etc.

EOSC4Cancer's and TEHDAS2's joint workshop 'European Health Data Space Meets Cancer Data', on the 30th January 2025 in Warsaw, explored the emerging concept of NCDNs and their potential role in shaping a pan-European cancer data infrastructure. They highlighted NCDNs' role as connective tissue between national data landscapes and the broader EHDS vision. Main discussion points are summarised in the bullets below:

- NCDNs should act as connective infrastructures, not monolithic entities.

- There is a need to define better the NCDNs: a) shared understanding of their role; b) clarification whether they can be data holders; c) embedment of existing infrastructures (e.g. EUCAIM, GDI); d) possible merger with national cancer registries.
- Focusing on cancer encourages collaboration among domain experts and supports integration of specialized tools and datasets like those from EOSC4Cancer.
- NCDNs could act as intermediaries between HDABs and researchers, offering e.g. support on data linkage, regulatory compliance, data catalogue curation.
- A critical function for NCDNs could be to manage multimodal data (genomics, clinical, imaging, other omics) and ensuring interoperability.
- It is likely that NCDNs and national cancer registries will coexist, with NCDNs serving as service providers to the research community, supporting data access and discoverability.
- Lessons can be drawn from other EU networking initiatives organized as distributed local nodes, which share aggregated results (not raw data) for meta-analyses, offering useful insights for structuring NCDNs.

EU Network of National Comprehensive Cancer Centres

As part of the Europe Beating Cancer Plan, the EU Network of National Comprehensive Cancer Centres initiative aims to connect and enhance quality and sustainability of comprehensive cancer centres across Europe. The concluded Joint Action CraNE⁴⁵ laid the groundwork for this infrastructure, while the JA EUnetCCC will advance the implementation of the network⁴⁶. Within the EHDS structure, the Centres are most likely data holders.

European Life Science Research Infrastructures

In 2020, BBMRI, EATRIS, ECRIN, and ELIXIR jointly affirmed their unique capacity to contribute to the success of the Horizon Europe Mission on Cancer⁴⁷. They highlighted their role in providing access to cutting-edge facilities, technologies, expertise, data, biological samples and analytical tools, while fostering trusted environments for the secure and cost-effective sharing of data and samples. However, these four RIs are only part of a much broader ecosystem.

Multiple European Life Science Research Infrastructures (LS-RIs) provide critical services relevant to cancer research, including biobanking, imaging, data management, and clinical research support. The canSERV initiative exemplifies how these infrastructures can collaborate effectively, bringing together world-class European RIs such as BBMRI, EATRIS, ECRIN, ELIXIR, EMBRC, EU-IBISBA, EU-OPENSOURCE, EURO-BIOIMAGING, EuroPDX, INFRAFRONTIER, INSTRUCT and MIRRI. Collectively, they span the entire oncology development pipeline and can interconnect technologies to guide users through the entire translational value chain.

⁴⁵

https://health.ec.europa.eu/document/download/d0fac20a-70d8-44dc-bf84-ee09e9bcaf18_en?filename=fs_cancer_crane.pdf

⁴⁶ <https://ecc-cert.org/health-service-research/eunetccc/>

⁴⁷

https://ecrin.org/sites/default/files/ckeditor-inlines/Cancer-Mission-Statement-BBMRI-EATRIS-ECRIN-ELIXIR-Sept-2020_0.pdf

Key contributions of LS-RIs to the federated platform include:

- LS-RIs provide researchers with expertise, state-of-the-art facilities, technologies, biological and chemical samples, animal models, data analysis tools and support.
- LS-RIs can support sample and data sharing in a trustworthy research environment where patient data and samples can be shared safely and cost-effectively.
- By integrating datasets, tools, and expertise into the federated platform, RIs can support data exchange and collaboration.
- LS-RIs can collaborate in the development and deployment of new technologies tailored to the needs of cancer researchers.
- By offering training in data management, interoperability, and advanced research methods, RIs help equip stakeholders, especially in underrepresented regions, with the skills to effectively use the platform.
- With a long-standing commitment to open access, LS-RIs encourage the adoption of best practices for transparency and collaboration across the research community.
- LS-RIs also promote and support international interoperability standards by working alongside global initiatives such as Global Alliance for Genomics and Health (GA4GH)

Integrating LS-RIs into the federated digital platform represents a critical step in creating a comprehensive, collaborative, and sustainable ecosystem for cancer research. To achieve this, robust technical and governance frameworks must be developed and implemented.

Technical requirements:

- Implement a common authentication and authorization framework to enable access across infrastructures via a single log in.
- Develop standardised access procedures across RIs to manage access, track usage, and ensure policy compliance.
- Define shared data exchange standards and implement standardised APIs to enable interoperability between the platform and individual RIs.
- Enable privacy-preserving federated query mechanisms that allow researchers to discover and analyse data across multiple repositories without moving the data.

Governance recommendations:

- Establish joint and coordinated access committees to harmonise access policies and procedures, and ensure consistent evaluation of access requests.
- Form specialised working groups to address integration of diverse data types and tackle technical and operational challenges.
- Maintain a shared, comprehensive service catalogue, building on models like the one developed in canSERV, that aggregates and categorises services from all participating RIs.

By effectively integrating LS-RIs, the federated platform can provide access to specialised technologies and expertise, support end-to-end research workflows, and foster interdisciplinary collaboration. Leveraging the strengths of Europe's RIs offers a solid foundation for driving innovation, bridging regional disparities, and accelerating the translation of research into clinical impact.

6 Resources ready for the UNCAN platform

This chapter explores how EOSC4Cancer's technological insights can be used as a stepping stone for the future cancer data infrastructure, namely the UNCAN platform, from day one. The technical infrastructures developed by EOSC4Cancer will help data trajectories for the future Cancer Mission implementation by providing:

- Experience with data flows combining multiple data types structured according to cancer patient journey
- Learnings on what is transferable
- Data discovery mechanisms adequate for some new data types, with Beacons
- Learnings on how to adapt RDMKit, RSQ Kit, IDTK to cancer
- Visualisation in cBioPortal
- Demonstrators on secure access

Specific outputs from EOSC4Cancer are referenced throughout this chapter, however all the public deliverables from the project can be found in the EOSC4Cancer Zenodo community⁴⁸.

Data Sources

There are many data sources (potentially) available for advancing the understanding of Cancer. However, there is still a long journey to the systematic and integrated use and reuse of such data across Europe. It is possible to have a greater impact on prospectively generated data as it can be made FAIR at source from the very beginning. Thus, efforts for improving existing datasets should be only conducted for those cases with recognized high value for research activities.

- ▶ Cancer registries
- ▶ Environment (pollutants)
- ▶ Social data
- ▶ Geolocalisation
- ▶ Results from Screening programmes
 - ▶ Medical imaging
 - ▶ Medical data (EHR structured data)
- ▶ Genomics and other omics data
- ▶ Imaging (digital pathology)
- ▶ Real World Data (EHR non-structured data)
- ▶ Other research data: liquid biopsy, drug screening, Wearables, IoT devices.
- ▶ Pre-clinical data: in-vitro and in-vivo data
- ▶ Cancer clinical trials (descriptors - metadata)
- ▶ Relevant databases and knowledge bases (approved drugs, indications, etc)
- ▶ Patients reported outcomes and data collected by patient associations.
- ▶ Real World Data (social media and other sources)

Table 1. Adopted from the EOSC4Cancer project. Summary of the existing data types available to cancer researchers.

⁴⁸ <https://zenodo.org/communities/eosc4cancer>

During the project, EOSC4Cancer created a living map of cancer data resources⁴⁹, including high level descriptors, access procedures and interfaces, and data use requirements. This metadata catalogue was developed using the MOLGENIS platform and enables discovery, documentation, and harmonisation of cancer-related datasets across Europe. It supports the FAIR data principles (Findable, Accessible, Interoperable, Reusable) by allowing researchers and clinicians to identify relevant cancer data resources without accessing sensitive data directly. It enables variable-level interoperability mapping, supports cohort documentation, and offers pathways for federated analysis and harmonisation workflows.

Importance of standardisation

Harmonisation and interoperability are crucial to enable integration of cancer-related data from diverse sources - especially considering upcoming developments, like the UNCAN platform. Recommendations emphasise the use of common data models, the adoption of widely-extended controlled vocabularies and ontologies, and the importance of modelling data and metadata before capturing it. These challenges are even more critical when attempting to combine and integrate data across different modalities, underscoring the critical need for standardised practices.

Specifically, EOSC4Cancer contributes its implementation of FAIR data at the source for cancer data to the future UNCAN platform, contributing to an API federation without data duplication.

Driven by use cases derived from all stages of the cancer patient journey (i.e. primary prevention, secondary prevention, diagnostics, treatment and survivorship), we aimed to identify and address potential data interoperability gaps through standardisation and harmonisation protocols. EOSC4Cancer explored the standardisation needs within the community⁵⁰ and delivered Standard Operating Procedures for the data types based on the existing use cases, explained in the following sub-chapters. More details can be found in EOSC4Cancer Deliverable 2.1⁵¹.

EOSC4Cancer's standardisation and harmonisation efforts align closely with broader European initiatives focused on enabling cross-border, interoperable cancer research. Notably, synergies exist with the 1+ Million Genomes (1+MG) initiative, particularly its cancer dataset harmonisation activities⁵² supported through projects like B1MG, which aim to standardise genomic and clinical data models for pan-European use. Similarly, the IHI Data

⁴⁹ <https://data-catalogue.molgeniscloud.org/catalogue/ssr-catalogue/EOSC4Cancer>

⁵⁰ <https://zenodo.org/records/10067296>

⁵¹ <https://zenodo.org/records/10829319>

⁵² <https://www.nature.com/articles/s41588-024-01721-x>

Sharing Playbook⁵³ offers valuable guidance on data governance and sharing practices that are consistent with EOSC4Cancer's FAIR data principles. Furthermore, connections with initiatives such as EUCAIM (for radiology imaging), BigPicture (for digital pathology), IMPaCT-Data (for clinical-genomic integration in Spain), and the AACR GENIE project provide a framework for harmonising SOPs, ontologies, and metadata models across modalities. By aligning with these efforts, EOSC4Cancer contributes to a broader ecosystem of interoperable research infrastructures and ensures its outputs are reusable and compatible with future Cancer Mission projects.

A key consideration for the long-term sustainability and impact of EOSC4Cancer is the ability to efficiently integrate data from healthcare systems into research environments. This involves bridging operational clinical standards like HL7 FHIR, widely used for exchanging data in healthcare settings, with research-oriented models such as OMOP CDM. Mapping data flows from FHIR to OMOP enables the secondary use of clinical data for research purposes, while preserving semantic consistency and interoperability. Several EOSC4Cancer datasets already demonstrate compatibility with both formats, and this dual-standard approach is increasingly adopted by related initiatives such as IMPaCT-Data and B1MG. Strengthening these FHIR-to-OMOP transformation pipelines, ideally through reusable, open-source tools, would not only support EOSC4Cancer use cases but also contribute to broader ambitions of the European Health Data Space and cross-project data harmonisation.

Exposome

Exposome data is used at the primary prevention stage of the patient journey. In the EOSC4Cancer use case on cancer risk identification and prevention, this type of data is linked with cancer registry data to investigate relationships between environmental factors and cancer.

Since exposome data is non-personal, both data and metadata are generally accessible, under specific conditions. Access procedures vary across countries like Italy, the Netherlands, and the Czech Republic, but EIRENE-RI aims to standardise how the data will be managed in the future.

In EOSC4Cancer, each exposome dataset utilises a custom (meta)data model. To achieve interoperability between countries, harmonisation is needed, particularly for the geospatial granularity.

Cancer registries

Cancer registries are information systems designed for collection, storage and management of data on persons with cancer. They are mandatory in all EU Member States.

In EOSC4Cancer we experienced that cancer registries are organised in various ways, e.g. the Czech and Dutch registries are national, while Italy has local registries. As a result, data

⁵³

https://www.ihl.europa.eu/sites/default/files/uploads/Documents/ProjectResources/IMI_IHI_DataSharingPlayBook_2024.pdf

access procedures differ by registry and are influenced by national and regional regulations, making harmonisation of access procedures challenging.

The European Network of Cancer Registries (ENCR)⁵⁴ promotes collaboration, sets data collection standards, and supports cancer registries across Europe. While many cancer registries adhere to ENCR's minimum dataset recommendations, full harmonisation is lacking. For interoperability, it is suggested to harmonise geospatial granularity of registry data linked to exposome data, using NUTS classification⁵⁵ as a standard reference.

A current priority in this area is to align cancer registry data with the standards of the European Cancer Information System (ECIS). The upcoming Joint Action CancerWatchJA will focus on improving quality and timeliness in this area, jointly with direct grants to Member States.⁵⁶

Screening

Screening programs are considered secondary prevention and consist of regular, systematic examinations like mammograms or colonoscopies to detect cancer at its earliest stages. In EOSC4Cancer, a harmonised codebook for colorectal cancer screening was developed based on screening studies in Catalunya (Spain), Piedmont (Italy) and the Netherlands. This codebook was validated against the Czech screening codebook.

During the project, only the Dutch Multitarget FIT (mtFIT) study had an established data access process, but future access could be changed through platforms like cBioPortal due to its simplicity on data storage and sharing .

Clinical data

Clinical data is all patient information related to their disease, including health history, diagnostics, treatments, and outcomes, primarily derived from Electronic Health Records (EHR) for secondary use.

Access to this clinical data is managed by Data Access Committees (DACs), with procedures varying per dataset.

EOSC4Cancer promotes the use of OMOP CDM. During the project conversions from various data models, some of which use standardised vocabularies for tumour classification, to OMOP were created. From this common ground, clinical data could be more easily reused.

Genomics

Genomics data are used to understand the genetic alterations of the patients, to understand the underlying mechanisms of a particular cancer enabling clinicians to provide the correct

⁵⁴ <https://encr.eu/>

⁵⁵

<https://www.europarl.europa.eu/factsheets/en/sheet/99/common-classification-of-territorial-units-for-statistics-nuts->

⁵⁶ European Cancer Information System (ECIS)

treatments and improve the patient outcomes. In EOSC4Cancer, these data are classified into four types: raw, processed, interpreted, and summarised.

While summarised data can be openly shared, the other types are protected by personal data laws and require controlled access (approval must be given by the relevant data access committee before sharing can occur, with relevant legal agreements in place). Once raw data are deposited in an archive, such as the European Genome-Phenome Archive (EGA)⁵⁷, they will be available for reuse by other researchers through a standardised access procedure. Processed and interpreted data can be made available through cBioPortal using their access procedure.

The raw data are available in the standard FASTQ and BAM formats, while processed and interpreted data follows the cBioPortal data model in MAF format or custom models, with plans to align custom models with cBioPortal in the future. Efforts in standardising Genomics data should ensure alignment with the ongoing work in the Genomic Data Infrastructure (GDI)⁵⁸ project, which is creating and deploying the technical capacity for accessing genomic data across the EU.

Radiology

Radiological imaging data, such as CT, MRI, PET, and Ultrasound, is crucial for cancer diagnosis and (response) monitoring.

These images are typically stored in a repository, such as XNAT⁵⁹. Access to these images is managed on a project or dataset basis through specific Data Access Committees.

The majority of the radiological data follows the standardised DICOM format. The ongoing EUCAIM⁶⁰ project, the cornerstone of the European cancer imaging initiative, is working on protocols to standardise uncurated fields within the DICOM model and improve the semantic annotation of imaging protocols.

Pathology

Pathology data, derived from biopsies or resections, is essential for cancer diagnosis and assessing treatment responses. In EOSC4Cancer, the focus is on digital whole slide pathology imaging.

Access to the digital pathology data is managed on a project or dataset basis through specific Data Access Committees as this is personal data. For tissue samples, access requests are generally more complex, as physical data needs to be shared, which falls outside the scope of EOSC4Cancer.

⁵⁷ <https://ega-archive.org/>

⁵⁸ <https://gdi.onemilliongenomes.eu/>

⁵⁹ <https://xnat.org/>

⁶⁰ <https://cancerimage.eu/>

In EOSC4Cancer, in alignment with the BigPicture⁶¹ project, the DICOM standard in combination with the unified open digital slide and annotation format specification⁶² were adopted for whole slide imaging data.

Synthetic Data

Synthetic Data can be a valuable option to have access to realistic data to test technical infrastructures, with largely reduced privacy issues.

Multimodal synthetic cohorts were generated in EOSC4Cancer. This includes clinical data based on colorectal cancer patients, which follows the OMOP CDM. Additionally, genetic genomes were generated to mimic real cancer data, with details available in a project deliverable report⁶³.

The generated synthetic data was stored in publicly available repositories (e.g., EGA), so the standardised access procedure of the repository could be used.

Actionable Research Software

Vision and Strategic Importance

Actionable research software should be a cornerstone of the future UNCAN Platform. It enables the FAIR (Findable, Accessible, Interoperable, and Reusable) use of heterogeneous cancer data across institutional and national boundaries while upholding privacy, regulatory compliance, and analytical reproducibility. Software solutions that support federated, privacy-preserving workflows are essential to transforming multi-modal data—such as clinical, genomic, imaging, and pathology records—into actionable insights. Through the EOSC4Cancer project, substantial progress has been made in designing and implementing modular, interoperable tools that can operate securely within national infrastructures and scale to meet the ambitions of the European Health Data Space (EHDS) and UNCAN-CONNECT.

1. Integrated Analysis Environments

EOSC4Cancer focused on improving the connectivity and usability of analytical environments to streamline cancer data analysis. A key achievement was the integration between cBioPortal—a widely adopted platform for exploring clinical and genomic cancer data—and Galaxy, an open-source environment for bioinformatics workflows. Additionally, EOSC4Cancer has extended the external resource linkage functionality in cBioPortal with a pathology (XOpat) and radiology (XNAT) viewer as examples. To enable reuse of this development an image integration manual⁶⁴ was created.

To enable the integration between cBioPortal and Galaxy, an intermediary server was developed to manage API connections. This allows users to trigger analysis workflows

⁶¹ <https://bigpicture.eu/>

⁶²

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

⁶³ [10.5281/zenodo.10847696](https://zenodo.org/record/10847696)

⁶⁴ <https://doi.org/10.5281/zenodo.14900294>

directly from cBioPortal and retrieve results back into the same interface, providing a seamless analysis loop across genomics, imaging, and clinical data domains.

EOSC4Cancer developed custom Galaxy tools for facilitating the transfer of data from Galaxy to cBioPortal and for transferring mutation data into Galaxy. The project also involved wrapping existing bioinformatics tools for execution within the Galaxy environment. Corresponding enhancements to cBioPortal's frontend support direct interaction with Galaxy-based analyses. This pipeline improves access to novel data types within cBioPortal, including processed omics and digital pathology data, while mitigating complexities in data upload and reuse. Notably, EOSC4Cancer also provides first test results from the use of cBioPortal, identifying bottlenecks in data upload.

To ensure portability and reproducibility, the cBioPortal-Galaxy system has been containerised into several modular components: cBioPortal, the intermediary server, and Galaxy with the integrated tools. This enables straightforward configuration and deployment in local infrastructures or within secure environments (see section 3). A complete deployment solution is already available, and a more complete overview can be found on the GitHub page⁶⁵.

2. Decision Support and Trial Matching Systems

EOSC4Cancer also developed software tools that directly support clinical decision-making and precision oncology workflows, focusing on both tumour board facilitation and clinical trial matching.

The Molecular Tumour Board Portal (MTBP), developed at Karolinska Institutet, demonstrates a streamlining of the capture and analysis of next-generation sequencing (NGS) data. It links functional genomic alterations with clinical outcomes and available therapeutic or trial options, thus supporting clinicians in evidence-based treatment planning. MTBP exemplifies how harmonised, structured molecular data can be applied in real-world clinical settings.

Complementing this, the project delivered TrialMatchAI⁶⁶, developed at CNRS, an AI-powered recommendation system designed to automate patient-to-trial matching. TrialMatchAI processes heterogeneous clinical data—including structured fields and unstructured physician notes—using fine-tuned open-source large language models (LLMs) in a retrieval-augmented generation (RAG) framework. The TrialMatchAI system performs end-to-end normalization of critical biomedical entities, employs a hybrid retrieval strategy combining both semantic and lexical approaches to identify relevant clinical trials, and conducts detailed eligibility assessments through medical Chain-of-Thought reasoning.

TrialMatchAI supports the GA4GH Phenopackets standard for patient data, allowing integration with other systems through standard APIs. It is designed for modularity, data privacy, and lightweight local deployment, making it suitable for integration into hospital or national infrastructures. Importantly, the system is modular, allowing for replacement of LLM components as more powerful or specialised models become available.

⁶⁵ <https://github.com/elixir-oslo/cbioportal-docker-compose>

⁶⁶ <https://github.com/cbib/TrialMatchAI>

As an important outcome, TrialMatchAI should be maintained and scaled post-project, particularly within the UNCAN platform and future EHDS-compliant infrastructures. It represents a reproducible, ethical, and high-impact application of generative AI in precision oncology.

3. Federated and Secure Deployment Mechanisms

To meet Europe's regulatory and ethical standards for working with sensitive health data, EOSC4Cancer placed a strong emphasis on federated analysis, modular software deployment, and secure execution environments.

Many cancer data types—especially genomic, clinical, and imaging data—are subject to legal restrictions and cannot be moved across jurisdictions. Therefore, interoperable federated analysis was treated as a core capability. EOSC4Cancer began by gathering technical requirements for such systems, including infrastructure, software, and mechanisms for accessing and orchestrating federated data resources. The upcoming deployment phase will integrate researcher identity and authorization schemas with selected reference workflows and software libraries.

To support privacy-preserving workflows, all components—including cBioPortal, Galaxy, intermediary servers, and TrialMatchAI—have been containerised for flexible deployment in Trusted Research Environments (TREs) or Secure Processing Environments (SPEs) as defined by EHDS. These containers allow researchers to analyse sensitive data in-place, without compromising security, compliance, or interoperability.

EOSC4Cancer also advanced the development of FAIR-compliant workflow systems by integrating platforms such as Galaxy, Nextflow, and Snakemake. These workflows are registered on platforms like WorkflowHub and Dockstore, following GA4GH standards for workflow execution and tool description:

- TRS (Tools Registry Service)
- WES (Workflow Execution Service)
- TES (Task Execution Service)

Standardised APIs abstract away infrastructure-specific details, enabling maximum compatibility and portability. To ensure transparency and reproducibility, workflow executions are tracked using Research Object Crates (RO-Crate), a metadata packaging format that supports provenance capture.

All developed software components will be subject to performance, relevance, integrative ability (APIs), and performance testing (e.g., through OpenEBench), including evaluation across usability, reproducibility, and community standards to ensure maximum usability.

Outlook

The tools developed by EOSC4Cancer—ranging from federated infrastructure components to AI-based decision support systems—form a robust and extensible software stack to power the UNCAN Platform. They provide the building blocks for federated, secure, and scalable

research infrastructures capable of linking cancer patients, researchers, and clinicians across Europe.

Future success depends on sustaining and expanding these tools within platforms like UNCAN, aligning with EHDS regulations, and supporting emerging AI governance under the EU AI Act. With continued investment and adoption, actionable research software will become a critical enabler of equitable, data-driven cancer research and care across the continent.

7 Data biases

Data biases can significantly flaw cancer research results. One of the EU Mission on Cancer's guiding principles is equity and access to knowledge research and care - e.g. by understanding why some people develop certain cancers compared to others. The EHDS regulation does not address data biases or groups more likely to be discriminated against.⁶⁷

Recent research is dedicated to different types of bias in the data (historical, representation, measurement bias), model (aggregation, evaluation) and deployment.⁶⁸ This chapter will focus on sex/gender biases as an example, since it was the most studied kind of bias in EOSC4Cancer's work.⁶⁹

Sex/Gender Bias

Sex and gender bias still influences cancer research, especially regarding availability of data. Europe's Beating Cancer Plan mentions sex/gender differences in cancer research as a priority - especially regarding differences in survivorship and access to care.

The following aspects are important to consider in this regard:

- Significant data gaps in biomedical research exist, caused by past male-centric bias. Research has been predominantly conducted on male, white, subjects, including male cells, animals, and male study participants.⁷⁰
- Sex disaggregated data are missing on a large scale.

⁶⁷ Article 54 does mention discrimination in the context of prohibited use.

⁶⁸ See e.g. Suresh, Harini and Gutttag, John. 2021. "A Framework for Understanding Sources of Harm throughout the Machine Learning Life Cycle." Equity and Access in Algorithms, Mechanisms, and Optimization. Available at: <https://dspace.mit.edu/bitstream/handle/1721.1/143588/3465416.3483305.pdf?sequence=2&isAllowed=y>

⁶⁹ We acknowledge the existence of other biases, such as socioeconomic status, race/ethnicity, and neighborhood characteristic.

⁷⁰ Catuara-Salarz, Cirillo, Guney (2022): Introduction: The relevance of sex and gender in precision medicine and the role of technologies and artificial intelligence. In: Sex and Gender Bias in Technology and Artificial Intelligence. Available at: <https://www.sciencedirect.com/science/article/pii/B9780128213926000030#s0010>

- Specific cancers occur in different ways in females and males - including incidence, mortality, treatment efficacy and toxicity, survivorship.⁷¹
- For colorectal cancer (EOSC4Cancer's use case), females are more likely to develop it on the right side, which has different molecular characteristics. Diagnostic methods of bloody stool, as well as colonoscopies do not work the same way with women. Women have a longer and straighter colon.⁷² 5-year survival rates are lower in females.⁷³
- In the recent EU project landscape, a focus on women in cancer was mainly related to projects focusing on breast cancer or cervical cancer.
- Gender biases can also be created by the choice of computational methods and ML.
- For EOSC4Cancer:
 - Cancer Patient Journey might differ in men and women due to lifestyle factors, attitudes towards screening, psychosocial factors etc.
 - Project use cases might have different effects on men and women: different cancer risks through occupational exposure, multi-omics analysis might uncover gender-specific molecular signatures in colorectal cancer, etc.
 - Cancer Research infrastructures are likely rather developed by men, which may lead e.g. to underrepresentation of women's health concerns in algorithms.

EOSC4Cancer and BioInfo4Women held a workshop on gender biases (December 2024) in close collaboration with BioInfo4Women. It explored gender gaps and biases in cancer research, considerations for study design, AI-biases, microbiome research and more. Participants wished for standards to follow to include sex/gender in cancer research and more leadership and investment in that area. For EU-funded research, participants recommended diverse stakeholder consultations and more follow-up on the Horizon Europe gender/sex requirements: these are often in place, but their implementation is not verified.⁷⁴

Determining the degree to which the data sources have the specific sex and gender metadata correctly annotated helps make gender data gaps visible. This is part of a responsible machine learning pipeline, in the phase of data acquisition and exploration. It provides the basis for a risk management towards the possible impact of biases.

Good practice examples from cancer research include the Hipàtia Community⁷⁵ research toolkit. It guides researchers on sex or gender in basic science, clinical, health systems and

⁷¹ EOSC4Cancer Sex and Gender Bias Workshop 12 December 2024.
<https://zenodo.org/records/15380682>

⁷² Caroline Criado Perez: Invisible Women. 2020.

⁷³ EOSC4Cancer Sex and Gender Bias Workshop 12 December 2024.
<https://zenodo.org/records/15380682>

⁷⁴ EOSC4Cancer Sex and Gender Bias Workshop 12 December 2024.
<https://zenodo.org/records/15380682>

⁷⁵ The Hipàtia Community of Practice was created after the first Gender Equality and Women's Leadership in Biomedical and Health Sciences Meeting, celebrated in Girona on October 22, 2019, with the participation of Catalan biomedical and health sciences research centres.

population health studies with a list of questions analysing the importance and representation of the sex- and gender-differences in research topics.⁷⁶

Tackling the sex and gender biases

Awareness is the first step towards less biased data. We see the European Cancer Information System (ECIS) in a major role here, since it provides the latest information on indicators that quantify cancer burden, and is explicitly intended for policy making. ECIS already provide sex-aggregated data and analysis in some instances - notably in its cancer fact sheets⁷⁷ and data explorer⁷⁸. While these are valuable sources, we recommend having more standard visualisations of sex/gender disaggregated data covering the related indicators like, incidence, survival, and outcomes of treatment.

For Horizon Europe projects, we recommend concrete handbooks on how to incorporate sex/gender dimension in cancer research - similar to the one by EOSC-Life regarding life sciences.⁷⁹

For the upcoming EU Dataset Catalogue in the EHDS (Art.79, 96), we recommend a metadata structure that allows sex-disaggregated data and informs whether sex-disaggregated data is available. Given the presumed size and importance of the EU Dataset Catalogue, missing at least this basic functionality would contribute significantly to widening the gender data gap.

For Horizon Europe projects, we recommend concrete handbooks on how to incorporate sex/gender dimension in cancer research - similar to the one by EOSC-Life regarding life sciences.⁸⁰

With regards to the EU cancer initiatives - especially the upcoming UNCAN platform we recommend to integrate gender-specific data collection and analysis in all use cases and implement gender-sensitive approaches in patient engagement and communication strategies.

8 Assembling a sustainable ecosystem

We want to ensure that EOSC4Cancer's outcomes can enrich the ecosystem in a way that data and use cases are transversal to different cancer types. Sustainability has been considered from the proposal writing stage of EOSC4Cancer, where it was decided to focus efforts on improving and aligning existing tools and resources rather than 'reinvent the wheel'. This has led to a sustainability model that can largely rely on pre-existing sustainable products. The consortium also includes a number of Life Science Research Infrastructures,

⁷⁶

https://aquas.gencat.cat/web/.content/minisite/aquas/publicacions/2022/toolkit_perspective_sex_gender_research_aquas2022.pdf

⁷⁷ <https://ecis.jrc.ec.europa.eu/factsheets>

⁷⁸ <https://ecis.jrc.ec.europa.eu/data-explorer/#/>

⁷⁹ <https://zenodo.org/records/13767883>

⁸⁰ <https://zenodo.org/records/13767883>

which are specialised in maintaining and exploiting resources across Europe.

The next stage of the Cancer Mission's UNCAN plan would benefit from using a similar model when possible, building on existing work for European scientific data sharing infrastructure. Two key examples related to cancer would be GDI (Genomic Data Infrastructure), with their work on sharing genetic data including cancer use cases, and EUCAIM that focuses on cancer image sharing. These pan European projects are tackling the technical and legal barriers for secondary use of data for research across national boundaries, the resulting solutions will be of great use to cancer research that often relies on combining several data types as well as cohorts originating in numerous countries.

A central part of the sustainable ecosystem is to effectively leverage the unique selling points of EHDS, EOSC, Europe's Beating Cancer Plan and EU Mission on Cancer flagship initiatives. This entails, above all, avoiding redundancies among the respective structures:

- Gain a clear understanding of what the EHDS defines and provides - e.g. which data is within and out of its scope
- Structures and guidelines that will be set out by EHDS and implementing acts should not be duplicated in other structures or projects
- Complementary use of existing funding structures

Patient engagement

From a patient's perspective, a sustainable cancer data ecosystem is one they can trust, actively contribute to, benefit from, and understand. For EOSC4Cancer and a future UNCAN platform to thrive and provide benefits back to the patient, it should ensure their involvement at every stage—from design and implementation to long-term sustainability—while ensuring that their data are used for the purposes they shared it, such as advancing research. While patients should not be expected to become IT or data experts, their insights are essential in shaping the future direction and ensuring that their needs remain central throughout the process and beyond. A key aspect of this sustainability is scaling up the EOSC4Cancer's five use cases, to capture a wide range of tumour types and cross-tumour conditions. The research purpose should be explained to cancer patients in a clear and accessible language.

In this context, patients take on a dual role as both contributors and consumers of data. The EOSC4Cancer data ecosystem should integrate various sources, including structural biology data, demographic insights, and crucially, patient-generated data like lifestyle and behavioural information. By involving patients from different backgrounds and geographical areas, cancer research infrastructures can ensure that data collection reflects real-world outcomes and addresses quality of life (QoL) considerations, building a system patients can trust.

Patient engagement in this ecosystem must go beyond simple data contribution—it must be holistic, incorporating the social and emotional dimensions of health, especially for survivors. This includes not just clinical data, but post-treatment rehabilitation, mental and physical health, nutritional status, and socioeconomic factors. Patients must feel empowered to manage their health data, facilitated by wearable devices and apps that provide valuable information integrated with traditional datasets like electronic health records and biobanks.

However, ensuring data quality through accreditation processes for these technologies is crucial to maintain patient trust.

From a patient's perspective, a promising additional use case would include physiological data from wearables and QoL measurements, as described in chapter 2. This expansion enhances the relevance of clinical trials, particularly for rare conditions, while enabling long-term tracking of cancer patients' outcomes. Longitudinal data collection will provide critical insights into relapse patterns and post-treatment quality of life, helping guide policymaking and resource allocation for survivorship care.

Patient organisations serve as the primary leaders in fostering engagement, providing communication pathways and platforms for direct patient involvement. These organisations can facilitate focus groups and other initiatives where patients can share their experiences and help shape future data collection and research strategies. The European Cancer Patient Digital Centre (ECPDC) also plays a key role by offering a trusted, central platform that bridges the gap between patients and data-driven research. It provides a secure space where patients can contribute their data while retaining control over its use. By fostering trust and transparency, both patient organisations and the ECPDC will help advance a sustainable, patient-centred cancer data ecosystem.

9 Timeline for implementation

EOSC4Cancer aims to provide a technical infrastructure for the EU Mission on Cancer and Europe's Beating Cancer Plan, as well as showing ways forward to integrate with the EHDS. Thus, milestones in these three initiatives are the milestones that mark our cancer data space roadmap.

It aims to show in which moments and ways EOSC4Cancer's outputs can be used by others, demonstrating a thorough exploitation of EOS4Cancer's outcomes. It also suggests ways to create synergies across other initiatives, not directly involving EOSC4Cancer's outputs. This timeline reaches until 2030 - a year that foresees the full implementation of many cancer flagship initiatives and the EHDS.

This implementation timeline considers our best state of knowledge as of May 2025 - based on both the quoted sources and bilateral exchange with the implementers of the respective initiatives. Thus, the timelines may differ from the published sources.

2025

Fully operational Network of Comprehensive Cancer Centres, including dedicated platform⁸¹

We suggest for the Comprehensive Cancer Centres to maintain a close dialogue with UNCAN-CONNECT on lessons learnt for platform implementation. Exchange with CANDLE on establishment of entities on Member State level and bridge between care and research.

⁸¹

https://health.ec.europa.eu/non-communicable-diseases/cancer/europes-beating-cancer-plan-eu4health-financed-projects/projects/crane_en

Start of UNCAN-CONNECT platform implementation

EOSC4Cancer's experience should be considered for the platform implementation in various aspects.

- 1) Linking use cases to experiences in technical infrastructures
 - Breaking silos, applying common data models, interoperable formats and groundwork for use of AI tools
 - Lessons learnt from applying infrastructure to four EOSC4Cancer use cases and related data handling
- 2) Proof of concept for analysis and management of heterogeneous data, especially regarding:
 - Work experience on different types of data: exposome, imaging, genomic etc.
 - Possibilities and challenges on building an infrastructure around different use cases
 - Work on converting codebooks to OMOP, data merging

Initial rollout of the HealthData@EU infrastructure for secure cross-border data sharing, interoperability, patient-centric approach, data for research and policy

We recommend a close dialogue with responsible actors for EHDS2 infrastructure to make sure that cancer research initiatives like UNCAN use as much infrastructure synergies as possible. The EHDS Regulation enters into force. This starts off the transition period.

Launch of a Comprehensive Cancer Infrastructure

We recommend a constant dialogue for UNCAN-CONNECT, CANDLE and EU-CIP to gain insights from their work on cancer education and care and to support their work on cancer research.

First Results of EOSC Health Data Task Force available

EOSC4Cancer has contributed to the Task Force results via members involved in both initiatives. Thus, there is continuity to make sure these results take the project's experience into account, in area such as:

- Making health and cancer data FAIR
- Special (legal and ethical) requirements: e.g. EOSC4Cancer's experience with synthetic data
- Interoperability with upcoming EHDS services for secondary use of health data, with regards to EOSC nodes: e.g. avoidance of redundancies with EHDS services such as metadata catalogues
- Network between EHDS and EOSC communities: EHDS and EOSC user journeys and outcome of stakeholder synergy outreach, to EHDS2 projects (TEHDAS2 etc.) and other EOSC projects with health use cases
- Input on how EOSC can support EHDS adoption

2026

Creation of federated, interoperable platform across MS (European Cancer Imaging Initiative)

Via a memorandum of understanding between EOSC4Cancer and EUCAIM, we have been ensuring ongoing consideration of EOSC4Cancer input on imaging data. We also recommend considering main implementing acts of EHDS - e.g. regarding pseudonymisation. It would further be beneficial for UNCAN-CONNECT to connect with national cancer image repositories and support AI-driven cancer diagnosis and treatment solutions.

Launch a collaborative network linking Member States (Knowledge Center on Cancer)

Establish a link between the Knowledge Center on Cancer to ECPDC (to get a realistic patient perspective) and research initiatives like UNCAN-CONNECT (to have an information flow from research to the Knowledge Center).

Full integration of Comprehensive Cancer Centres and ERNs into a single, harmonised cancer care and research ecosystem (Comprehensive Cancer Infrastructure)

We recommend a constant dialogue to incorporate learnings from Comprehensive Cancer Centres and ERNs into platforms like UNCAN-CONNECT and ECPDC. These links to patient care and to public-facing information strengthen the holistic picture of the cancer ecosystem.

2027

Scale up of the European Cancer Imaging Initiative to cover a broader range of cancer types

We recommend making responsible initiative leaders aware of related work in EOSC4Cancer, to contribute to the scale-up: experiences from use cases and with imaging data.

2028

Deadline main implementing acts European Health Data Space

We recommend for the upcoming projects to consider the final definition in the Implementing Acts and their impact on the Cancer Data Space - such as availability of real world data for cancer research, mechanisms and infrastructures that could be used. Relevant implementing acts include: the European Electronic Health Record Exchange Format (Art. 15), Identification Mechanisms (Art. 16), MyHealth@EU (Art.23), Specifications for Conformity of EHR systems (Art. 36), Enforcement by Health Data Access Bodies (Art. 63), Templates for Data Access Application (Art. 67), Secure Processing Environments (Art. 73), IT Architecture of HealthData@EU (Art. 75), Dataset Description and Catalogue (Art. 77),

Data Quality and Utility Label (Art. 78), Minimum Dataset Specifications for Secondary Use (Art. 77), EHDS Board (Art. 92).

Most Health Data Access Bodies operational

Based on our collaboration with TEHDAS2, we recommend close collaboration between HDABs and other structures with national node level - e.g. National, Cancer Data Nodes, ECPDC nodes, Cancer Mission Hubs, EOSC Nodes.

Expansion of UNCAN-CONNECT's data collection to include real-time patient data and broader genomic datasets

We recommend to consider achievements by GDI and B1MG and define what is achievable regarding real-time patient data. EOSC4Cancer provided input to the cancer use cases in GDI. Based on the experience from the cancer use cases, GDI can support the collection of genomic datasets for UNCAN-CONNECT

Full interoperability of cancer data across all EU countries (Knowledge Center on Cancer)

EOSC4Cancer's recommendations regarding different cancer data types, as outlined in chapter 6, could be very relevant here. We also recommend considering the achievements of previous cancer projects and in the EHDS2 infrastructure.

2029

EHDS Regulation fully applicable in all Member States. Fully operational
HealthData@EU Infrastructure

EOSC4Cancer's work and lessons learned can help guide all cancer data actors to comply with the applicable rules. Access to cancer-related data via the frontend of UNCAN-CONNECT is integrated with HealthData@EU on the backend.

Launch of the Info Portal for the European Cancer Patient Digital Centre and rollout of federated network of national infrastructures

It is crucial for EU-CIP to ensure that ECPDC is known and trusted by patients across Member States, receiving input from National Cancer Data Nodes and translating research from UNCAN-CONNECT into accessible information for patients. It provides a user-friendly AI bot. Ideally, ECPDC nodes should be in close contact with cancer data nodes, Digital Health Agencies, HDABs, Comprehensive Cancer Centres and Infrastructures and EU Cancer Mission Hubs. A constant dialogue with these entities, as well as with UNCAN-CONNECT, ensures that those research results are translated into information for patients - and that patient needs are considered in research.

EOSC4Cancer results on patient engagement, as well as on bias in cancer data could be useful here.

2030

Implementation of selected categories of secondary use of health data: genomics, omics, wellness applications

We recommend for the EHDS to fully consider assets and work developed in GDI and related use cases in EOSC4Cancer and UNCAN-CONNECT. We also recommend for the secondary use infrastructure to reasonably link to primary use of health data and EHR data.

Full operationalisation of the UNCAN-CONNECT platform as a central hub for cancer research in Europe

UNCAN-CONNECT provides a one stop shop to access cancer data from heterogeneous sources. It allows for searching for cancer data across the cancer patient journey, thus uniting relevant results from EOSC4Cancer and other EU Mission on Cancer projects.

National cancer data nodes are mostly integrated with Cancer Mission Hubs, Comprehensive Cancer Centres, Comprehensive Cancer Infrastructures and Health Data Access Bodies in most cases.

Fully operational Comprehensive Cancer Infrastructures across the EU

Capacity building by comprehensive cancer infrastructures is operational, contributing to reducing inequalities in cancer care and research across Member States. EOSC4Cancer's WP5 results on the Cancer View of the RDMKit can be useful here.