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

Funded by the European Union

eosc | cancer

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

Draft Version 18 November 2024

Main Authors:
Alfonso Valencia (Barcelona Supercomputing Center)
Salvador Capella-Gutierrez (Barcelona Supercomputing Center)
Daniel Barrowdale (ELIXIR Hub)
Carola Schulz (empirica Technology Research)
Jointly with the EOSC4Cancer consortium and stakeholder forum

Document Authors

Main Authors

Salvador Capella-Gutierrez	Barcelona Supercomputing Center
Alfonso Valencia	*Barcelona Supercomputing Center*
Daniel Barrowdale	*ELIXIR Hub*
Carola Schulz	*empirica Technology Research*

Additional Authors

Jan-Willem Boiten	*Lygature*
Robin Navest	*Lygature*
Eivind Hovig	*University of Oslo*
Romina Royo	*Barcelona Supercomputing Center*
Lifang Liu	*Health-RI*
Josephine Mosset	*Cancer Patients Europe*
Sergi Aguiló-Castillo	*Barcelona Supercomputing Center*
David Marshall	*Instruct*
Gerrit Meijer	*NKI*
Munazah Andrabi	*University of Manchester*
Sophie Huiskes-Berends	*Lygature*
Fotis Psomopoulos	*CERTH*
Sarah Morgan	*EATRIS*
Maria Alexandra Rujano	*ECRIN*

Funded by the European Union

Macha Nikolski	*CNRS*
Eva Garcia Alvarez	*BBMRI-ERIC*
Griselda Marku	*empirica Technology Research*
Rabea Richter	*empirica Technology Research*
Veli Stroetmann	*empirica Technology Research*

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

Funded by the European Union

Table of Contents

Executive summary

EOSC4Cancer is a European funded project designed to make diverse types of cancer data accessible: genomics, imaging, medical, clinical, environmental and socio-economic. 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 community-driven analysis environments.

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, and 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 dataspace 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 a future Cancer Digital 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 Cancer Digital 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 Cancer Digital 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, building upon the work carried out by EOSC4Cancer.

Funded by the European Union

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 a 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 a Cancer Digital Platform 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.

Chapter 7 features additional considerations for the Cancer Digital Platform, including the importance of patient engagement in projects and the role of patient organisations, and the effect of sex and gender bias in cancer research.

Ending with a comprehensive timeline in Chapter 8, this section looks forward over the next six years for the key milestones in European cancer initiatives, legal acts and infrastructures.

Funded by the European Union

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 Digital Platform (CDP) 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 the CDP will therefore be relevant to the different stages of the cancer patient journey: prevention, screening, diagnosis, treatment, recovery, 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, particular 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 Centers and National Cancer Data Nodes.

To make the design robust and sustainable the CDP'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; 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 CDP 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 CDP 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.

Funded by the European Union

Acronyms and Abbreviations

Acronym	Meaning
API	Application Programming Interface
CDP	Cancer Digital Platform, a future platform intended to facilitate data reuse and lead to new innovations in cancer research
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
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

Funded by the European Union

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 Cancer[1] and Europe's Beating Cancer Plan[2] for cancer data, as well as the European Health Data Space[3] (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 8 (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 Registry[5] (2022), the agreement by the European Parliament and Council on the Regulation for a European Health Data Space (2024)[6] and 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.

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 the use and reuse of cancer-related data (from both healthcare and research). This 'Cancer Digital Platform' (CDP) 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] ttps://commission.europa.eu/strategy-and-policy/priorities-2019-2024/promoting-our-european-way-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].

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 primary purposes: 1) primary prevention; 2) secondary prevention; 3) diagnostics; 4) treatment. We are extending to a fifth use case 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.[9]

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-roadmap-2024-02-26_en

[9] 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

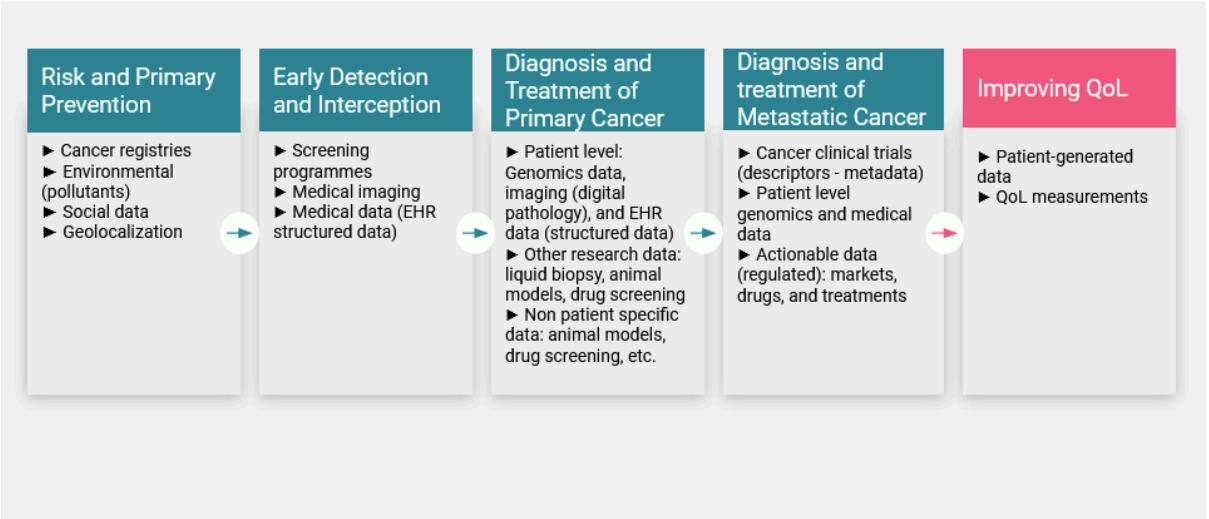

Figure 1. Adopted from the EOSC4Cancer project. Updated graphical depiction of the five main stages of the Cancer Patient Journey.

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:

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 the EIRENE exposome network[10], EXPANSE project[11]) to investigate the relation between cancer incidence and environmental factors. The result will be a report outlining guidelines, methodologies, challenges, and recommendations on how best to perform such data integration. It also highlights the national differences in how such data integration efforts are conducted.

[10] https://eirene.eu/
[11] https://expanseproject.eu/

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 data availability for (next generation) polygenic risk scores would need to be developed.

2. **Data driven optimization 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 is currently being performed for the screening data of the Czech Republic.

   EOSC4Cancer works now in the harmonisation of codebooks and converted to OMOP, allowing broad use in the future and enabling meaningful comparisons of the available data (e.g. relevant also for performing HTA analyses).

   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 detection[12].

3. **Data-driven treatment selection for localised tumours using multiple patient-derived data types**

   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 will facilitate secure and seamless access, analysis, and sharing of clinical, genomic, and other patient-specific information, fostering better collaboration and insights.

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

[12] (see ESCALATION project: https://www.nki.nl/research/research-groups/gerrit-meijer/escalation-study/)

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 4 and 5: 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/greatest 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. This use case, 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**

This use case focuses on the necessary data infrastructures and format specifications for analysing tumour data from patients through a 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.

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 GENIE[13] 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.eu platform. Some specific ideas on how to move forward with each of these use cases have already been added 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 a future CDP 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. It 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.
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 long and health 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)

[13] 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 Intended future UNCAN.eu 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 a sixth 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.

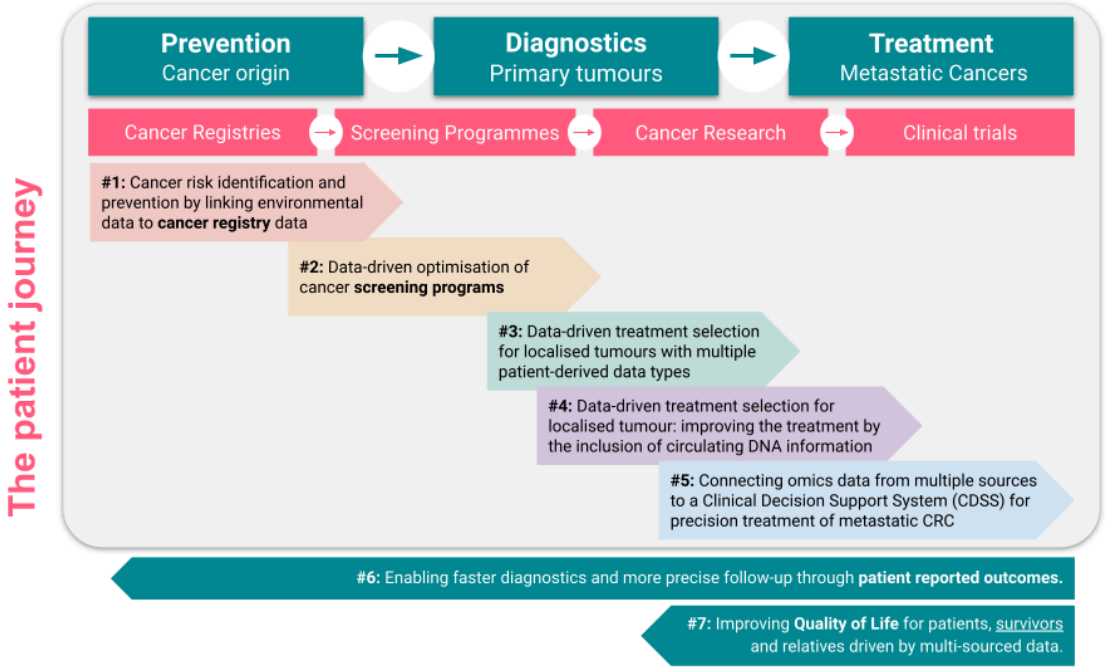

Figure 3. **Extended EOSC4Cancer use cases:** Addition of at least two use cases to represent the inclusion of patient-reported outcomes and how data can be potentially used to increase the quality of life of patients, survivors and relatives.

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.

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 CDP.

b) **Access**:
Secure and distributed 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 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 the 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 at least 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 acting as coordinators of the local cancer community, connecting 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 CDP will depend on its ability to build relationships and encourage engagement with its future users and stakeholders. The CDP 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 CDP 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 and cancer survivors**
This covers a broad range of users, such as current cancer patients, cancer survivors, 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 CDP. This 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 CDP, 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 CDP. 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 RDMkit[14].

The RDMkit Cancer Data Management page 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 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 RDMkit Cancer Data Management page is in progress and we expect the initial version to be completed by the end of the EOSC4Cancer project, at which point the wider community will be able to continue to improve this resource.

4 European perspective on cancer research

Three high level political initiatives guide the European perspective on cancer research as far as 2030: the EU Mission on Cancer[15], the Europe Beating Cancer Plan[16] and the European Health Data Space[17]. All three initiatives provide rich implementation pathways towards facilitating cancer research and ultimately reduce the burden of cancer.

[14] https://rdmkit.elixir-europe.org

[15] 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

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

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

This chapter outlines the “playing field” shaped by these initiatives. We put special emphasis on the European Health Data Space and its links to cancer research, as this is the least defined element as of October 2024.

EU Mission on Cancer

The EU Mission on Cancer[18] supports 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
- Engaging with European citizens, including patients
- 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:

- The UNCAN.eu data exchange platform will provide a better understanding of the development and progression of cancer.
- The European Cancer Patient Digital Centre (ECPDC) is a data platform targeted to helping cancer patients, survivors, and their families navigate the cancer care pathway and, when possible, contribute to research

EOSC4Cancer prepares the technical infrastructure for the EU Mission on Cancer, by organising cancer data more efficiently, enhancing access to innovations for EU patients.

Europe’s Beating Cancer Plan

Europe's Beating Cancer[19] Plan 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’s 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 includes efforts to ensure equal access to cancer diagnosis and treatment among member states, promoting training programmes for healthcare professionals and improving cancer medicine as a whole. Thanks to improvement in early diagnosis rates, effective specialised therapies, and appropriate care, there are an estimated 12 million cancer survivors around Europe. They require access to follow-up care and

[18] 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

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

support services, from healthcare to financial resources, which increase their quality of life. Europe's Beating Cancer Plan is working towards making these services accessible for all survivors and their families.

European Health Data Space

Clarifying the link between infrastructures created by EOSC4Cancer and the EHDS is a common request by different stakeholders[20]. The European Health Data Space Regulation was approved in April 2024 and is expected to be published in the EU Official Journal in early 2025. . This EU Regulation 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[21].

Use of cancer data for research is considered secondary use of health data according to the European Health Data Space Regulation, and any virtual platform for cancer research will most likely be a data holder. Article 33 mentions data categories especially crucial for cancer research, that should be made available for secondary use: genomic, multi-omic and biobank data. Thus, **the federated digital platform to advance cancer research** should provide access to this data for secondary use. 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.[22]
Cancer Data initiatives feed into the EHDS by developing infrastructures for specific types of health data - specifically for genomic data (e.g. via GDI[23], B1MG[24]) and Cancer Imaging (e.g. via EUCAIM[25]).

The main national institution for secondary use of health data in the EHDS are Health Data Access Bodies. As outlined in Article 35-37 of the EHDS Regulation, these bodies are responsible for managing and deciding on requests for health data access for secondary use, process health data for that purpose, ensure confidentiality and IP rights and manage a national dataset catalogue.
The figure below depicts the interplay between the federated digital platform to advance cancer research and the Health Data Access Bodies.

[20] As of August 2024.
[21] https://health.ec.europa.eu/ehealth-digital-health-and-care/european-health-data-space_en
[22] https://health.ec.europa.eu/document/download/6e26bad9-5722-4c95-8bc5-4c21d8e370dd_en?filename=policy_com-2024-206_fr.pdf
[23] https://gdi.onemilliongenomes.eu/
[24] https://b1mg-project.eu/
[25] https://digital-strategy.ec.europa.eu/en/policies/cancer-imaging

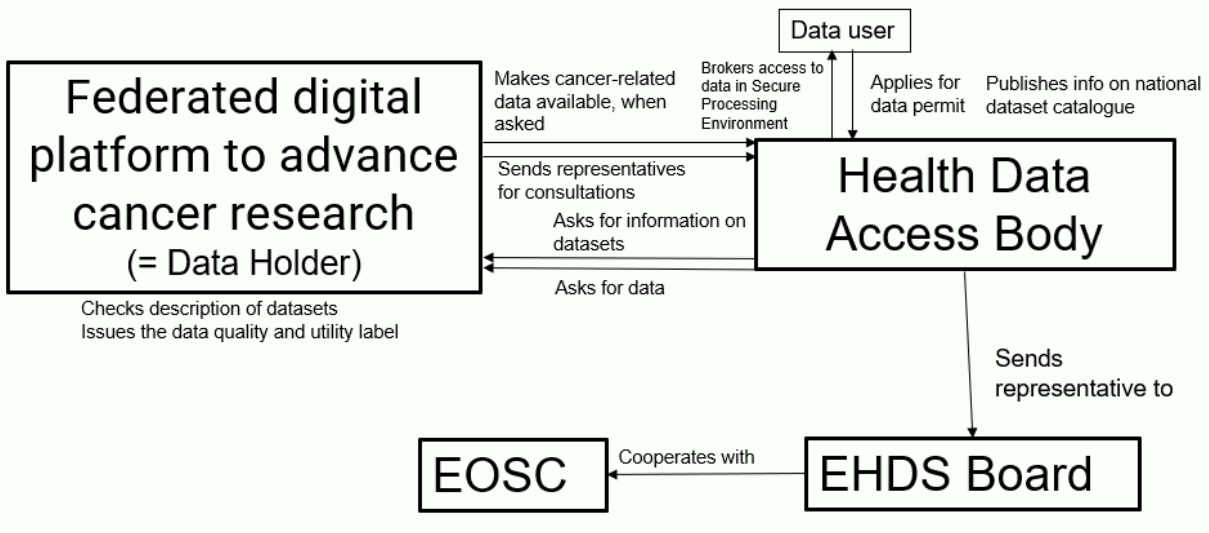

Figure 4. Tentative depiction of interaction with Health Data Access Body in EHDS: Own elaboration based on EHDS Regulation proposal.

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

- 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

HealthData@EU and TEHDAS2

In this context, it is important to consider the outcomes of TEHDAS2[26] and HealthData@EU Pilot[27]. 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.
Also the outcomes of HealthyCould should be considered in this regard. Its Draft Strategic Agenda[28] recommends both a HealthData@EU community interface service and an EOSC sensitive data users service. The HealthData@EU pilot also has a use case on cancer genomics, which could take up EOSC4Cancer's results.
The UNCAN.eu platform will federate a network of cancer data nodes on national level, suggesting also ways to connect this structure to the EHDS.[29]

[26] https://tehdas.eu/
[27] https://ehds2pilot.eu/
[28] https://zenodo.org/records/7331832#.Y9KUYnbMKUI
[29] https://ec.europa.eu/info/funding-tenders/opportunities/portal/screen/opportunities/calls-for-proposals?callIdentifier=HORIZON-MISS-2024-CANCER-01

A more granular and timed link between the federated digital platform to advance cancer research and the major EHDS milestones is outlined in the chapter on the timeline.

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 contact with project EOSC projects with a health focus – such as Sci-Lake[30], RAISE[31] and BY-COVID[32]. Insights include:

- Differently from other EOSC data use cases, health data is sensitive and thus requires stricter compliance on data protection.
- An important institution in the upcoming roadmap is the EOSC Health Data Taskforce[33], which focuses on interoperability and sharing of health data between the European Health Data Space and EOSC.
- Upcoming EOSC governance structure is still being defined. An efficient infrastructure from a cancer data point of view would allow the EU level to link to the UNCAN.eu platform and the national level to link to cancer data nodes and to cancer mission hubs

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.eu - and the services that the EOSC EU Node is creating.[34]

5 National vs European levels for implementation

Via the upcoming UNCAN.eu research data platform, the European Commission aims to gain a better understanding of cancer by fostering collaboration between Member States and unlocking the vast amount of preexisting data.

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 the data subjects. These aims align with the upcoming European Health Data Space legislation.

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

[30] https://scilake.eu/
[31] https://raise-science.eu/
[32] https://by-covid.org/
[33] https://eosc.eu/wp-content/uploads/2024/07/Health-Data-TF-ToR.pdf
[34] https://open-science-cloud.ec.europa.eu/news/european-commission-announces-eosc-eu-nodes-transition-full-production

of National Cancer Data Nodes (NCDNs) linked to UNCAN.eu, with each country or region having its own node. These nodes will function as central hubs in their own jurisdiction, with responsibility for standardising and harmonising data procedures, whilst also providing the digital tools needed for the entire cancer research community to collaborate on joint research projects.

These NCDNs need not start from scratch, many projects and initiatives are already working on federated data infrastructures. The Genomic Data Infrastructure 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 efforts. This potential could spread to other areas, such as harmonised patient consent forms to facilitate cross border legal agreements.

However, the implementation of a Cancer Digital Platform does not only need to aim for interoperability across Member States, but also between the Member State and EU level. It is important for efforts to be made in parallel to avoid incompatible solutions being implemented, or confusion around overarching governance. With EHDS still in the making, Health Data Access Bodies should ideally connect on a national level to the national cancer data nodes, with EHDS leading the overall governance. Consideration should also be given to the AI Act, as well as the Data Governance Act.

Forming a National Cancer Data Node

While all NCDNs share the goal to improve interoperability in multiple dimension of the (re)use of cancer data by standardisation at the organisational and operational levels as much as possible, both nationally and transnationally, ultimate implementations of NCDN's may vary across member states, e.g. as a consequence of already existing organisation of the cancer data field at regional or national levels. As a consequence, in terms of their operations, their setup may vary to best suit the situation in each Member State. Examples below provide some illustrations of this.

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 regional/local 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.

A successful and efficient implementation of the vanguard nodes will require support and guidance from European-level funded projects setting out plans, for example maturation models to guide their set up in manageable stages, and centrally implemented tools and services that will be required for the interoperability between the nodes. These projects that involve and align with the European Research Infrastructures will ensure their expertise in data management, FAIRification and cross border research is passed onto the fledgling nodes. Once the network has been implemented the reliance on central funding should fade away, with the expectation that Member States fund their own node also in the connection to the services implemented for the EHDS. This will lead to a decentralisation of the infrastructure, implying that national roadmaps along the lines of any produced maturation model will become more important.

NCDNs should evidently align with other Member State Level Structures - namely Cancer Mission Hubs, Health Data Access Bodies, ECPDC nodes, Comprehensive Cancer infrastructures and Centers.

6 Data spaces and actionable research software

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

- Experience with data flows structured according to cancer patient journey (beyond genomics)
- Learnings on what is transferable

- Data discovery mechanisms adequate for new data types, with Beacons
- Learnings on how to adapt RDMKit, RSQ Kit, IDTK to cancer
- Visualisation in cBioPortal
- Demonstrators on secure access

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 datasets available to cancer researchers.

Importance of standardisation

Thus, 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 UNCAN.eu 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, primary cancer and metastatic cancer), we aimed to identify and address potential data interoperability gaps through standardisation and harmonisation protocols. EOSC4Cancer delivered Standard Operating Procedures for the data types included in the existing use cases, explained in the following sub-chapters.

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 of data linked to cancer registry information.

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.

Cancer registries setup varies by country, for example, in EOSC4Cancer the Czech and Dutch registries are national, while Italy has local registries. Data access procedures differ by registry and are influenced by national and regional regulations, making harmonisation of access procedures challenging. An international legal agreement would be required to standardise access across countries.

The European Network of Cancer Registries (ENCR)[35] 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's suggested to harmonise geospatial granularity of registry data linked to exposome data, using NUTS classification[36] as a standard reference.

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

[35] https://encr.eu/

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

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. Additionally, the harmonised codebook was converted to OMOP CDM to enhance reuse of this data.

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 . Harmonisation of data models is ongoing, with a focus on identifying common variables and agreeing on data granularity to retain as much information as possible. The next step involves refining these data models and integrating appropriate ontologies to enhance interoperability.

Clinical

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 by 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. In EOSC4Cancer, this reuse happened primarily through visualisation and further exploration through cBioPortal instances.

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 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)[37], they will be available for reuse by other researchers. 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.

[37] https://ega-archive.org/

Efforts in standardising Genomics data should ensure alignment with the ongoing work in the Genomic Data Infrastructure (GDI)[38] 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 monitoring.

These images are typically stored in a repository, such as XNAT[39]. 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 EUCAIM[40] 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 pathology imaging. Access to this data is governed by data protection policies, requiring specific requests for different data types.

The lack of a universally accepted data or metadata format for digital pathology presents challenges for harmonisation and integration. To address this, EOSC4Cancer adopts the DICOM standard used with radiology data, improving interoperability and the utility of digital pathology in research. It will also consider the developments of the BigPicture[41] project.

Synthetic Data

Synthetic Data can be a valuable option to have access to realistic data to test technical infrastructures, with largely reduced privacy issues. Thus, EOSC4Cancer proposes the generation and use of synthetic data.

EOSC4Cancer has developed strategies for the generation of synthetic data that can be used as demonstrators of the technical developments within and beyond the project. The project has generated initial synthetic data from colorectal cancer patients, as well as synthetic genomes, trying to mimic real cancer data.

Until spring 2025, cancer-specific synthetic cohorts including different modalities (i.e., genomic data together with clinical data based on OMOP) will be created, delivered, and stored in publicly available repositories (e.g., EGA).

[38] https://gdi.onemilliongenomes.eu/
[39] https://xnat.org/
[40] https://cancerimage.eu/
[41] https://bigpicture.eu/

Actionable Research Software

EOSC4Cancer has made significant strides in advancing data integration and accessibility for cancer research. It has extended the external resource linkage functionality in cBioPortal with a pathology (XOpat) and radiology (XNAT) viewer as examples.

Furthermore, EOSC4Cancer has paved the way for integrating cBioPortal with environments like Galaxy, enabling researchers with seamless link to the analysis, thus facilitating a more holistic view of patient data across genomics, imaging, and clinical attributes. Such integration is crucial for advancing federated data access and supporting the broader cancer research community. This integration is further being developed towards full interconnectivity between cBioPortal and Galaxy through API connections to trigger workflows in Galaxy with results provided back to cBioPortal. In this way it provides access to novel data types within cBioPortal, as well as improvements to mitigate existing data upload complexities and to enhance the analysis capabilities. As the relevant genomic data are considered sensitive, thereby requiring restricted access, through the use of Trusted Research Environments (TREs), software portability and standalone installation and function may be required. In this context, the software cBioPortal/Galaxy combination will be available using containerisation and enabled for simple configuration and deployment setup, ensuring full support for the diverse data and analytical needs of the research community.

Another major outcome is the enhancement of interoperability between clinical trial databases and Clinical Decision Support Systems. The Molecular Tumour Board Portal (MTBP) and TrialMatchAI[42] were central developments, automating the integration and interpretation of diverse molecular and clinical data to support precision cancer medicine across European cancer centres. The MTBP, developed at Karolinska Institutet, streamlines the capture and analysis of next-generation sequencing data, linking functional genomic alterations with clinical outcomes and trial opportunities. Complementarily, TrialMatchAI leverages state-of-the-art artificial intelligence, particularly large language models, to automate the matching of patient profiles with clinical trials, focusing on genomic biomarkers and other relevant patient data to generate tailored clinical trial recommendations. APIs ensure seamless data flow and integration with Clinical Decision Support Systems, aiming to impact the landscape of cancer treatment and research by providing a framework for data-driven orientation of patients to clinical trials.

In many genomic analyses with cancer data, where analytical power comes from the pooling of larger cohorts, federated analysis may be a requirement. Different parts of the data reside at different locations, and often different national legal systems. Interoperable federated analysis and learning at scale will be undertaken. A phase of gathering technical requirements for distributed/federated analysis, including infrastructure, software, and mechanisms for accessing and orchestrating federated data resources, will be followed by a deployment phase, integrating researcher identity and authorization schemas, and incorporating selected reference workflows and software.

A major effort will be on AI tools to enhance all aspects of data analysis, with support for the full cancer data lifecycle, including tools for data harmonisation. Advanced workflow systems

[42] https://github.com/cbib/TrialMatchAI

will be put to work: Integration of standardised data management and analysis workflows systems like Galaxy, Nextflow, and Snakemake, ensuring that supported workflows are registered as FAIR resources (e.g. WorkflowHub, DockStore) with consideration to the GA4GH[43] standards Tools Registry Service (TRS), Workflow Execution Service (WES) and Task Execution Service (TES). Standardised APIs will abstract the technical details of executing workflows from the researcher, while allowing maximal compatibility across workflow systems, workflow specification languages and the deployed compute infrastructures. Evidence of secure executions will be captured as Research Object Crates (RO-Crate). Developed software solutions will be evaluated for functionality, relevance, integrative ability (APIs), and performance (e.g., through OpenEBench) to ensure maximum usability.

Regarding data integration, EOSC4Cancer also provides first test results from the use of cBioPortal, identifying bottlenecks in data upload.

7 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.

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 cancer digital 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.

[43] https://www.ga4gh.org/

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.

Sex/Gender Bias

A sustainable cancer research ecosystem should also address current sex and gender biases. One of the EU Mission on Cancer's guiding principles is equity and access to knowledge research and care, linked to the understanding why some people develop certain cancers compared to others. Also Europe's Beating Cancer Plan mentions sex/gender differences and importance to understand these in cancer research as a priority. AIt also explains that differences in survivorship and access to care are often directly related to gender differences.

Sex/Gender bias manifests itself in several areas - with one very relevant example being the sex/gender data gap. Not only in cancer, but in medicine in general, significant data gaps in biomedical research can be seen through and have been caused by the male-centric bias in

the past. Traditionally, research has been predominantly conducted on male subjects, including male cells, animals, and male study participants which excluded female representation in research.[44]

- Sex disaggregated data are missing on a large scale. If this lack is continued in the major upcoming EU Dataset Catalogue (foreseen in the EHDS), this problem will persist.
- Specific cancers occur in different ways in females and males - both in incidence rates and characteristics. E.g. EOSC4Cancer's use case of colorectal cancer, females are more likely to develop it on the right side, which has different molecular characteristics compared to left-sided colon cancer. Diagnostic methods of bloody stool, as well as colonoscopies do not work the same way with women. Women, on average, have a longer and straighter colon.[45]

In the recent EU project landscape, a focus on *females and women* in cancer was mainly related to projects focusing on breast cancer or cervical cancer (in relation to HPV) (e.g. GlycoMap, MammoScreen,). Some projects without this focus at least mention sex/ gender dimension (e.g. UNCAN.eu blueprint, PROTECT-EUROPE, EU-CanIneq, and Endeavor).

8 Timeline for implementation

The final utility of EOSC4Cancer's work is 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.
This timeline reaches until 2030 - a year that foresees the full implementation of many cancer flagship initiatives and the EHDS.

2025

Fully operational Network of Comprehensive Cancer Centres, including dedicated platform[46] (Europe's Beating Cancer Plan)

Dialogue with UNCAN.eu on lessons learnt for platform implementation, establishment of entities on Member State level and bridge between care and research.

[44] 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

[45] Caroline Criado Perez: Invisible Women. 2020.

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

Start of UNCAN.eu platform implementation (Europe's Beating Cancer Plan)

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

1) Linking use cases to technical infrastructures

- Experience in breaking silos
- Awareness raising on involved people on both sides
- Well curated data sets relevant to cancer research
- Experiences related to four EOSC4Cancer use cases and related data handling
- Ideas for further use case building

2) Proof of concept for analysis and management of heterogeneous data, especially regarding:

- Work from use case on screening programme
- Different types of data: exposome, imaging, genomic etc.
- EOSC4Cancer provides the foundational work, showing what is possible and where problems arise
- Work on converting codebooks to OMOP, data merging
- Preparatory work, conceptual work, hands-on work with data

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

Close dialogue with responsible actors for EHDS2 infrastructure to make sure that cancer research initiatives like UNCAN use as much infrastructure synergies as possible.

Launch of a Comprehensive Cancer Infrastructure

Constant dialogue 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 available

Make sure these results use EOSC4Cancer's experience in the Task Force's main working areas, such as:

- Making 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: 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

Funded by the European Union

2026

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

Consideration of EOSC4Cancer input on imaging data. Consideration of main implementing acts of EHDS - e.g. regarding pseudonymisation. UNCAN.eu to connect with national cancer image repositories and support AI-driven cancer diagnosis and treatment solutions.

European Cancer Patient Digital Centre launched, rollout of federated network of national infrastructures (e.g. “national nodes”)

Ensure ECPDC is known and trusted by patients across Member States, receiving input from National Cancer Data Nodes and translating research from UNCAN.eu into accessible information for patients.

Deadline main implementing acts European Health Data Space

Consider the final definition of key concepts 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. 6), Identification Mechanisms (Art. 9), MyHealth@EU (Art.12), Specifications for Conformity of EHR systems (Art. 23), Enforcement by Health Data Access Bodies (Art. 43), Templates for Data Access Application (Art. 47), Secure Processing Environments (Art. 50), IT Architecture of HealthData@EU (Art. 52), Dataset Description and Catalogue (Art. 55), Data Quality and Utility Label (Art. 56), Minimum Dataset Specifications for Secondary Use (Art. 58), EHDS Board (Art. 64).

Knowledge Center on Cancer: launch a collaborative network linking Member States

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

Comprehensive Cancer Infrastructure: full integration of CCCs and ERNs into a single, harmonised cancer care and research ecosystem

Constant dialogue to incorporate learnings from ERNs in platforms like UNCAN.eu

2027

Most Health Data Access Bodies Operational

If applicable, integration or at least collaboration with other structures with national node level - e.g. Cancer Data Nodes, ECPDC nodes, Cancer Mission Hubs, Comprehensive Cancer Centers. Collaboration with EOSC Nodes.

Funded by the European Union

European Cancer Imaging Initiative: Scaling up the initiative to cover a broader range of cancer types

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

2028

UNCAN.eu: expansion of platform's data collection to include real-time patient data and broader genomic datasets

Consider achievements by GDI and B1MG. Definition of what is achievable regarding real-time patient data.

Knowledge Center on Cancer: achieve full interoperability of cancer data across all EU countries

Consider the achievements of previous cancer projects and in the EHDS2 infrastructure.

2030

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

Full consideration of assets and work developed in GDI and related use cases in EOSC4Cancer and UNCAN.eu. Establishment of reasonable link to primary use of health data and EHR data.

European Cancer Imaging Initiative: full integration of AI-powered imaging tools across Europe

UNCAN.eu: full operationalisation of the UNCAN.eu platform as a central hub for cancer research in Europe

UNCAN.eu is fully integrated with the Cancer Digital Platform, to provide 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 the EU Mission on Cancer.
UNCAN.eu cancer data nodes on Member State level are mostly integrated with Cancer Mission Hubs, Comprehensive Cancer Centres, Comprehensive Cancer Infrastructures and Health Data Access Bodies in most cases.

Complete EU-wide access to HealthData@EU infrastructure

Access to cancer-related data via the frontend of the Cancer Digital Platform is integrated with HealthData@EU on the backend.

Comprehensive Cancer Infrastructures: a fully operational CCI across the EU

Capacity building by comprehensive cancer infrastructures has reduced the inequalities in cancer care and research across Member States. Work results feed into the UNCAN.eu platform.

European Cancer Patient Digital Centre: fully operational

The European Cancer Patient Digital Centre provides a wide range of information materials to patients, combined with a user-friendly AI bot. ECPDC nodes are 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.eu and the CPD, ensures that those research results are translated into information for patients - and that patient needs are considered in research.