ONCOSCREEN

D1.3 DATA MANAGEMENT PLAN (FIRST VERSION)

June 30, 2023

Funded by the European Union

Submission date: M6

Due date: M6

DOCUMENT SUMMARY INFORMATION

Grant Agreement No	101097036	**Acronym**	ONCOSCREEN
Full Title	A European "shield" against colorectal cancer based on a novel, more precise and affordable risk-based screening methods and viable policy pathways		
Start Date	1 January 2023	**Duration**	48 months
Project URL	https://OncoScreen.health/		
Deliverable	D1.3 Data Management Plan (First version)		
Work Package	WP1 Project and Technical Management		
Type	Report (RE)	**Dissemination Level**	Public
Lead Beneficiary	TIMELEX		
Contributions	All partners and tasks		
Authors	Magdalena Gad-Nowak (TIMELEX)		
Co-authors	N/A		
Reviewers	Paul Torke (MUG), Michael Papazoglou (SERVTECH)		

DISCLAIMER

COPYRIGHT MESSAGE

DOCUMENT HISTORY

Version	Date	Changes	Contributor(s)	Comments
V0.1	10.05.2023	ToC prepared	TIMELEX	Initial ToC
V0.2	23.05.2023	First draft	TIMELEX with the contribution from all partners	Contribution to individual section
V0.3	13.06.2023	Updated Second Draft	EXUS	Prefinal version ready for internal review
V0.4	19.06.2023	Internal Peer Review	SERVTECH, MUG	Internal review comments
V0.5	22.06.2023	Updated version with addressed comments	TIMELEX	Final Version ready for submission
V1.0	30.06.2023	Final version	EXUS	Quality assurance and submission to the EU portal

PROJECT PARTNERS

Partner	Acronym
EXUS SOFTWARE MONOPROSOPI ETAIRIA PERIORISMENIS EVTHINIS	EXUS
UNIVERSITAETSMEDIZIN DER JOHANNES GUTENBERG-UNIVERSITAET MAINZ	UMC-Mainz
INSTITUTE OF COMMUNICATION AND COMPUTER SYSTEMS	ICCS
FIRALIS	Firalis
UNIVERSITATSKLINIKUM SCHLESWIG-HOLSTEIN	UKSH
UNIVERSITAET zu LUEBECK	UzL
LIETUVOS SVEIKATOS MOKSLU UNIVERSITETAS	LSMU
MEDIZINISCHE UNIVERSITAT GRAZ	MUG
INSTITUTO PORTUGUES DE ONCOLOGIA DO PORTO FRANCISCO GENTIL, EPE	IPO
INSTITUTUL ONCOLOGIC PROF. DR. ALEXANDRU TRESTIOREANU BUCURESTI	IOB
TECHNION - ISRAEL INSTITUTE OF TECHNOLOGY	TECHNION
UNIVERSIDADE DO MINHO	UMINHO
UNIVERSITEIT VAN TILBURG	TLBG
VLAAMSE INSTELLING VOOR TECHNOLOGISCH ONDERZOEK N.V	VITO
ETHNIKO KENTRO EREVNAS KAI TECHNOLOGIKIS ANAPTYXIS	CERTH
INNOVATION SPRINT	iSPRINT
SCIENTIFIC ACADEMY FOR SERVICE TECHNOLOGY EV	SERVTECH
AINIGMA TECHNOLOGIES	AINIGMA
CATALINK LIMITED	CATALINK
KONNEKT ABLE TECHNOLOGIES LIMITED	KT
BEIA CONSULT INTERNATIONAL SRL	BEIA
UNIVERSIDAD DE LA RIOJA	URIOJA
TIME.LEX	time.lex
CARR COMMUNICATIONS LIMITED	CARR
MINISTRY OF HEALTH	MoHGR
PAGALBOS ONKOLOGINIAMS LIGONIAMS ASOCIACIJA	POLA LT
EUROPACOLON PORTUGAL- ASSOCIACAO DE LUTA CONTRA O CANCRO DO INTESTINO	ECPT
ELLINIKI ETAIREIA OGKOLOGIAS PEPTIKOU	HSGO
EUROPEAN SOCIETY OF DIGESTIVE ONCOLOGY	ESDO
FUNDATIA YOUTH CANCER EUROPE	YCE
MEDIZINISCHE UNIVERSITAT INNSBRUCK	MUI
LIETUVOS RESPUBLIKOS SVEIKATOS APSAUGOS MINISTERIJA	MoH-LT
EY ADVISORY SPA	EY
AGENCIA ESTATAL CONSEJO SUPERIOR DE INVESTIGACIONES CIENTIFICAS	CSIC
UNIVERSITE DE FRANCHE-COMTE	UFC
ROZENBAUM KONSULTING	ROSENBAUM
GIE AXA	GIE AXA
ASSOCIATION GERCOR	GERCOR
LOUWEN ROGIER	CC RL

SANNE VOOGD - CCassured	CC SV

LIST OF ABBREVIATIONS

Abbreviation	Description
CA	Consortium Agreement
D	Deliverable
GA	Grant Agreement
IPR	Intellectual Property Rights
WP	Work Package
CRC	Colorectal Cancer

Executive Summary

The ONCOSCREEN project is funded under the European Union's Horizon Europe research and innovation programme under **grant agreement no. 101097036**. The ONCOSCREEN Consortium consists of 39 partner organisations representing 15 countries across Europe, **EXUS AI Labs** being the coordinator of the project.

ONCOSCREEN seeks to develop new methods and technologies for cancer screening and early detection. It aims to develop a risk-based, population-level stratification methodology for CRC, to account for genetic prevalence, socio-economic status, and other factors. It complements this methodology by a) developing a set of novel, practical, and low-cost screening technologies with high sensitivity and specificity, b) leveraging AI to improve existing methodologies for CRC screening, allowing for the early detection of polyps and provision of a personalized risk status stratification, and c) providing a mobile app for self-monitoring and CRC awareness raising. Furthermore, ONCOSCREEN develops an Intelligent Analytics dashboard for policy makers facilitating effective policy making at regional and national levels.

The core idea is to orchestrate large amounts of heterogenous data and, with the help of novel as well as AI-assisted screening methods, design risk-based CRC screening approaches providing specific evidence-based recommendations end-user target groups (i.e. High Risk Citizens/Patients, Clinicians and Policy Makers).

This deliverable provides **an initial Data Management Plan** (DMP) for ONCOSCREEN, taking into account the data to be processed within the project. Since ONCOSCREEN is in its early stages, most of the data described in this document are subject to change as they are still under development. This version of the DMP outlines the current understating about the handling of the research data collected or generated in ONCOSCREEN. The DMP will also describe the standards and methodologies that will be followed for data collection, storage and sharing. This version of the document should be regarded as a living document, that will require regular updates throughout the project's duration.

TABLE OF CONTENTS

Executive Summary 7
1 Introduction 10
1.1 The need for a Data Management Plan in ONCOSCREEN 10
1.2 Deliverable objectives 10
1.3 Relationship with other deliverables and tasks 10
1.4 Deliverable structure 11
2 Data Summary 12
2.1 Research data 12
2.2 Origin of data and re-use of existing data 12
2.3 Data providers 14
2.4 Data recipients 14
2.5 Types and formats of data collected/generated 15
2.6 Data collection/generation purposes in relation to the project's objectives 55
2.7 Expected size of the data 56
2.8 Data utility 56
3 FAIR Data 58
3.1 Making data findable, including provisions for metadata 58
3.2 Making data openly accessible 59
3.2.1 Repository 59
3.2.2 Data 60
3.2.3 Metadata 61
3.3 Making data interoperable 62
3.4 Increased data re-use 64
4 Management of other research outputs 66
5 Allocation of resources 67
6 Data Security 68
7 Legal and ethical aspects 71
7.1 General 71
7.2 Ethical and GDPR compliance 71

7.3 DPIA 72
7.4 Use of AI 72
7.5 Transfers outside the EEA 73
7.6 Data sharing agreement 73
7.7 Collection of human tissue and cells 73
7.8 Intellectual property 74
8 Other Issues 75
9 Joint Data Management Strategies of Cluster Projects 76
9.1 “Prevention and early detection” cluster 76
9.2 FAIR data management 77
10 Next Steps 80
Conclusion 81

1 Introduction

1.1 The need for a Data Management Plan in ONCOSCREEN

Data Management Plans (DMPs) are considered to be a key element to sound data management. In Horizon Europe, a great importance is attached to research data management and data management plans are mandatory in all projects.

1.2 Deliverable objectives

A DMP defines a trustworthy and ethical data management policy of the project, and describes the data management life cycle for the data to be collected, processed and/or generated by a project. As part of making research data findable, accessible, interoperable and re-usable (FAIR), a DMP should include, in particular, information on:

- The handling of research data during and after the end of the project
- What categories of research data will be collected, processed and/or generated during the project
- Which methodology and standards will be applied in respect of the FAIR data principles for handling research data during and after the project
- Whether research data will be shared/made publicly accessible (and if so – following which open access model) and
- How data will be curated and preserved during the project and after it ends
- Any additional safeguards that will be implemented to ensure respect of the FAIR data principles in terms of allocation of resources, data security, research ethics and intellectual property rights.

The goal of this document is to set the initial DMP for the ONCOSCREEN Project. It contains guidelines that will be used by the ONCOSCREEN Consortium Partners with regards to all the data that will be generated by the Project.

1.3 Relationship with other deliverables and tasks

This deliverable receives input from all project partners. The present document (D1.3) constitutes the first version of the DMP and will be submitted in month six (M6) of the Project. It will not be a fixed document, but it will evolve and will gain more precision and substance during the Project's implementation. Information will be made available on a finer level of granularity through updates of the DMP as the implementation of the project progresses and when significant changes occur, such as (but not limited to) the inclusion of new data, new possibilities for aggregation and anonymisation, changes in consortium policies (e.g. new innovation potential), changes in consortium position and external factors (e.g. new consortium members joining). As such, more detailed versions hereof will follow at later stages of the Project (i.e. within

the review reports of M18, M36 and M48) and will be concluded with the final Deliverable D1.4 - Data Management Plan (final version). Furthermore D1.3 will feed D2.1 ONCOSCREEN clinical knowledge base (First version), D2.2 ONCOSCREEN clinical knowledge base (Final version) and D3.3 ONCOSCREEN bio bank that will take into considerations the provisions for data governance described in this deliverable. Finally, the D4.1 ONCOSCREEN Co-Designed System Architecture (First Version) and D4.2 ONCOSCREEN Co-Designed System Architecture (Final Version) will accommodate special provisions in the architecture ensuring the security by design aspects mentioned in this deliverable. Finally, a common approach in regards to data management with the Mission Cancer cluster "Prevention, including screening" of projects will be conducted on the basis of this deliverables and more specifically chapter 9 "Joint Data Management Strategies of Cluster Projects".

1.4 Deliverable structure

This DMP is structured following the precise template, provided by the European Commission for Horizon Europe projects, which contains a set of detailed questions to be answered by the beneficiaries with an appropriate level of detail (where no information is available, the acronym N/A as in "non-applicable" will be used; where no information is available at this stage of the project, a statement that "The information will be provided at a later time" will be included).

In addition, this Data Management Plan will include a common chapter on the "Prevention, including Screening" cluster. The aim of said chapter is to address commonalities in data standards, data validation, data protection and to foster data exchange between different cluster projects.

2 Data Summary

This section describes what data will be collected and generated within ONCOSCREEN.

2.1 Research data

From a general perspective, research data can be described as the evidence used to inform or support research conclusions (University of Sheffield n.d.). The tangible forms this 'material' may take are e.g. "facts, observations, interviews, recordings, measurements, experiments, simulations, and software; numerical, descriptive and visual; raw, cleaned up and processed" (Van Berchum & Grootveld, 2017).

According to the European Commission's "Guidelines to the Rules on Open Access to Scientific Publications and Open Access to Research Data in Horizon 2020" (Version 3.2, 21 March 2017), the notion of "*research data*" refers to "*information, in particular facts or numbers, collected to examined and considered as a basis for reasoning, discussion or calculation*".

Examples of research data include, amongst others, statistics, results of administrative procedures, measurements relating to exchanges, database contents, data captured in transaction logs – possibly curated to remove or reduce personal data – survey results, contents of applications, interview recordings and images. Trade secrets, commercially sensitive information and confidential information are however not considered to be research data. The implications of this scoping will be outlined in the following sections.

2.2 Origin of data and re-use of existing data

The ONCOSCREEN project (also referred to as "**the Project**") both leverages existing ("**retrospective**") data and generates new ("**prospective**") data, as explained below in Table 1. During the Project's course, various research data and results will be collected and generated.

Table 1 ONCOSCREEN data

Existing (retrospective) data collected during ONCOSCREEN	New (prospective) data generated during ONCOSCREEN
• Data stemming from previous and ongoing studies on CRC screening and early detection in National /EU Healthcare • General medical and clinical knowledge of healthcare providers • Existing open data repositories/bases • Existing scientific literature and publications	• Each patient's personalized background data (data received from the mobile app, such as profile information, behavioural data collected from 3rd party activity wearables; data collected through PROMS/PREMs questionnaires) • Data linked to research activities (collected during the clinical trials) including in particular formation from diagnostic solutions (such as liquid biopsy, breath

• Existing clinical partners' data (previous/ongoing studies implemented by medical partners)[1] • Socio-economic status data (socio-economic background, educational background, geographic areas, obesity/BMI index, income, history of smoking, alcohol consumption, hyperlipidaemia) • Environmental/exposomics data (environmental stressors/risks: air/water pollution, unhealthy life-style) • Medical history status (Electronic Health Records; entries of pre- and post- cancer periods for understanding the medical history (including heredity) • A database of retrospective colonoscopy sessions • Financial data needed to perform (a) the overall assessment of the clinical trials (health technology assessment, economics, lessons learnt, medical findings, feedback from actors/end-users, etc.) and (b) to perform the evaluation/benchmarking of ONCOSCREEN solutions against standard of care (health economics, affordability, cost benefit and cost effectiveness)	biopsy, stool tests, colonoscopy images/videos, histopathological images) • Other data: e.g. eye tracking data of the pathologists, when they observe a Whole Slide Image during the diagnosis process; CRC risk level calculation results obtained through the risk stratification engine

[1] For example, the following internal datasets of MUG and MUI are going to be used:

Colorectal Cancer (CRC) Whole Slide Images – Clinical Annotation Cohort: Patients: 1000; Samples:1800

Colorectal Cancer (CRC) Whole Slide Images – Survival Cohort: Patients: 6400; Samples:180.000 WSIs

OncoTrack Cohort: Patients: 106; Samples: 100

Furthermore, MUG will utilize its experience in AI algorithms in histopathological slides from its established collaboration with Google where ~500.000 slides (1.5 PB) were analysed.

2.3 Data providers

All clinical (medical) partners will play a major role in the Project as providers of retrospective data, as well as data from clinical trials (hosted in their respective sites) and from the use of the diagnostic WP3 tools.

Table 2. Data providers

Clinical partners	
1.	UMC-Mainz
2.	UzL
3.	LSMNU
4.	MUI
5.	UFC
6.	GERCOR
7.	IPO
8.	IOB
9.	HSGO
10.	ESDO
11.	ROSENBAUM

2.4 Data recipients

Technical and supporting partners will, amongst others:

- seek data from existing open data repositories (ECIS, GCO, European Cancer Inequalities Registry etc.), cancer registries and publications both at an EU and world-wide level.
- collect and process data on behalf of the clinical partners (e.g. behavioural data such as answers to questionnaires, activity measurements etc. collected via the mobile app) to later on provide it to other ONCOSCREEN components (i.e. the data fusion engine)
- receive patient data (breath samples, blood samples, urinal samples) from the medical sites, for the purpose of running the diagnostics using the WP3 diagnostic tools
- integrate and fuse heterogenous data from multiple sources (including structured, semi-structured, and unstructured data) to generate comprehensive and meaningful insights, to be stored in a repository that can be used for further analysis and decision-making.
- Create a risk-based calculation tool considering all factors, stressors, the results from WP3 Diagnostic Tools, faecal tests, colonoscopy adenoma/polyp classification, tissue biopsy adenoma/polyp the background of individuals and the pool of expert rules to

automatically identify dependencies and reveal correlations among a variety of features concerning the clustering of citizens/patients into their respective risk-level classification.

Table 2. Data recipients

Technical partners		
No.	**Partner's name**	**OncoScreen tools/solutions[2]**
1.	EXUS	For ONCO-RISTE
2.	ICCS	For ONCO-AICO
3.	FIRALIS	For the biomarker part
4.	UzL	For the ONCO-NMR
5.	MUG	For ONCO-BIOBA, ONCO-AITI
6.	TECHNION	For the ONCO-VOC
7.	CCASSURED	For the ONCO-CRISPR
8.	UMINHO	For the ONCO-CTC
9.	TLBG	For the privacy preservation
10.	VITO	For the ONCO-EVIDA
11.	CERTH	For the Data Fusion
12.	ISPRINT	For the ONCO CAWA
13.	SERVTECH	For ONCO-CLIDE
14.	AINIGMA	For ONCO-CLIDE
15.	CTL	Data Fusion, ONCO-EVIDA
16.	KONN	For ONCO-CLIDE, For ONCO-EVIDA
17.	BEIA	Supporting ONCO-RISTE dealing with wearables.
18	MUI	Supporting MUG in biobanking
19.	AXA	For ONCO-EVIDA
20.	MoHGR	For ONCO-EVIDA
21.	MoHLT	For ONCO-EVIDA

Please note that the activities of the below partners do not fall within the scope of this deliverable. These partners neither produce data nor have access to research data or research data processing activities and have a supportive role in regards to the project activities.

- Timelex (a legal and ethics support partner)
- Carr Communications (communication and dissemination partner).

2.5 Types and formats of data collected/generated

Tables below present details on the types of data, their origin/provenance, the partner who is providing or processing this data in the context of the ONCOSCREEN as well as, in particular, the estimated size and formats of the data. Where appropriate, the table indicates the potential interdependencies or correlations/links with other tasks and tools.

[2] All of the ONCOSCREEN solutions are descibed in tables below and in Section 3.3. hereof.

Partner	**CCAssured**
Partner's role	CCAssured – leading partner
Solution/Tool/Task	**T3.2 - ONCO-CRISP** screening tool ONCO-CRISPR can interact with ONCO-AITI, ONCO-CTC, ONCO-BIOBA, ONCO-RISTE and ONCO-EVIDA.
Data collected/produced/processed	1) Experimental methods and validation data; 2) Gel electrophoresis data containing RPA and (RT-)PCR data, sequencing data, positive and negative results on test strips or by colour change (stored as pictures), test results (positive/negative) of control samples and patient samples (excel), true and false positives and negatives, sensitivity and specificity data (excel). In-silico data obtained from the GEOdatabase (NCBI) and literature (pubmed.gov) to identify potential biomarkers that can be used to validate our CRISPR-Cas prototypes and synthetic generated DNA or RNA of biomarker regions suited for the CRISPR-Cas POCT technology in CRC
Purpose of the collection/generation/processing	Developing and validating a CRISPR-Cas based POCT prototype, which will include determining the reliability of the new investigated biomarkers and developed PCR and Point-of-care Test protocols, compared to existing biomarkers and test protocols.
Expected Size	The exact size will depend on the sample numbers that are going to be tested, as well as on the quality of the tested biomarkers (black/white versus significant difference). It is expected that the minimal used data size/volume will oscillate somewhere between 1-5 GBs
Format	At least the following formats will be used: Images (digital) (BMP/PNG), word documents, text, seq (Wordpad), applied biosystems sequence trace (Snapgene), GRAPHpad prism files, PDFs and excel files

	(digital) with data (pos/neg, % specificity and % sensitivity)
Metadata	CCassured will not provide any metadata
Will the data be made available for re-use	Yes. Except for the patent related data and information. After patent filing the data can be made available in a publication and presentation, so that other scientist can validate the findings. The embargo period for patent related data will be at least 1,5 years.
How will the security of the data be ensured?	Data will be stored secure and will be regularly being stored on a backup system that is only connected to the internet when a backup is being made.

Partner	**MUG, MUI, CERTH, ICCS**
Partners' role	MUI, MUG – leading partners
Solution/Tool/Task	**T3.3 ONCO-AITI (AI-Assisted Tissue Image Analysis)** Interdependable with ONCO-BIOBA, ONCO-AICO.
Data collected/produced/processed	Whole Slide Images (WSI), obtained mainly from MUG and MUI, which will be used to train an AI algorithm and to observe pathologists during the diagnosis process. Additionally, ONCO-AITI will have access to the metadata, which describes the information of the WSIs. The produced data will consist of annotations for the WSIs, which are performed by pathologists or trained user, which can define organs or tissue type in a WSI. The following internal datasets of MUG and MUI will be used: **Colorectal Cancer (CRC) Whole Slide Images – Clinical Annotation Cohort:** • Patients: 1000 • Samples:1800

	Colorectal Cancer (CRC) Whole Slide Images – Survival Cohort: • Patients: 6400 • Samples:180.000 WSIs **OncoTrack Cohort:** • Patients: 106 • Samples: 100 The produced data will consist of annotations for the WSIs, which are performed by pathologists or trained user, which can define organs or tissue type in a WSI. The tool will also collect Eye-tracking data of the pathologists, when they observe a WSI.
Purpose of the collection/generation/processing	The aim of AITI is to develop a software including AI algorithms to use them for the training of junior pathologists, to improve their training efforts and to support their decision process in diagnostics for histopathological slides. In order to understand the training process of a pathologist, we will observe junior and expert pathologists during the diagnosis process, so as to be able to recognize possible differences between juniors and experts. For that purpose, the eye tracking data of the pathologist (when they observe the WSI) will be used as additional information for the AI algorithm to improve the software, which aims to train junior pathologists.
Expected Size	Since the complexity of the tool cannot be estimated at this point in time, it is difficult to make a statement. However, the storage space of the developed tool will be in the gigabyte range.
Format	The data of the annotations will be stored in different formats, such as qpdata (QuPathdata), png, jpg, mrxs and ndpi (slides). The Eye Tracking data and clinical information will be stored as csv.
Metadata	The AI algorithm will not produce meta data in the proper sense. However, since WSIs will be used for training, at least this data will provide the metadata ((sex, age, T, N,

	M stage and subtype) in a machine-readable and interpretable format (csv, plain text, word or pdf documents).
Will the data be made available for re-use	Yes. Stored annotations and Eye tracking data can be re-used with the corresponding WSI. However, the data itself (WSI) will not be made openly available nor for re-use.
How will the data security be ensured?	MUG: Data security will be ensured through the procedure in MUG IT like firewalls, VPN, access protocols, only access for participants named in the study protocol. More in general, the access to patient data is restricted only to authorized personnel who need the information to perform a defined task. This will be accomplished through secure user authentication for example. MUI: The data will be stored locally in a directory at the MUI where daily backups take place on the one hand. On the other hand, the slides might be made available using a SFTP server of the MUI. Only authorized users (employee and cooperation partners) could have access to download the slides. Person identifying information on the label (Case IDs, names) will be removed, furthermore the label could be blinded. The created metadata of the slides itself will be checked for sensitive information and the sensitive information will be removed.

Partner	**CERTH, ICCS**
Partners' role	CERTH, ICCS - leading Partners, AINIGMA, KT - supporting partners
Solution/Tool/Task	**T3.3 ONCO-AICO (Real-time AI-Assisted Colonoscopy platform)**

	Not dependant on any other technology or data from other technologies, but there will be collaboration with ONCO-AITI.
Data collected/produced/processed	CERTH will gather annotations made by users on retrospective colonoscopy videos, including the location and classification of areas of interest. The data will be collected through the platform's user interface.
Purpose of the collection/generation/processing	The purpose of collecting the data is to track user performance in identifying areas of interest in colonoscopy videos, and to calculate scores based on the accuracy of the annotations made by the AI algorithms. This will help trainee colonoscopists to improve their skills and increase their accuracy.
Expected Size	The size of the dataset will depend on the number of users and the number of annotations made on the videos. We expect the dataset to grow over time as more users engage with the platform.
Format	The data will be collected through the platform's user interface and stored in a secure database. The data will be stored in a structured format that can be easily queried and analysed.
Metadata	Metadata will be provided to ensure that the data can be easily queried and analysed. The metadata will include information on the video, selected dataset, selected algorithm as well as on the annotations made by user
Will the data be made available for re-use	Yes.
How will the security of data be ensured?	We will ensure that appropriate data protection policies are in place to protect the confidentiality and integrity of the data. This will include appropriate authentication and authorization mechanisms. We will also ensure that any third-party software or services used to process or store the data comply with relevant security and data protection standards.

Partner	**MUG**
Partner's role	MUG - Leading partner; all technical and medical partners – supporting partners
Solution/Tool/Task	**T3.4 ONCO-BIOBA** (ONCOSCREEN's bio data bank to be used during the project and its clinical trials) Interdependencies with: T4.2 Privacy-preserving multi-sourced data fusion & system integration T4.1 Functional, non-functional requirements and complaints to standards & secure architectures T1.6 Data management T2.4 Retrospective CRC screening, Data collection and analysis
Data collected/produced/processed	Already existing data (e.g. data cohorts) will be summarized in a **data catalogue**. ONCO-BIOBA will not be a DATA WAREHOUSE but rather a digitalized catalogue containing descriptions of all ONCOSCREEN data sets on an aggregated level and linking to those data sets. In case new data is collected retrospectively, histopathological sections will be scanned with high-resolution scanners and digitized as Whole Slide Images (WSI). Additionally, corresponding metadata will be provided to the data cohorts, which may consist of different formats. This metadata can be used to describe the cohort, which is listed in the data catalogue.
Purpose of the collection/generation/processing	To provide a data catalogue that can offer the user an overview of existing data sets. The creation of a data catalogue is also related to task 3.3 which aims to develop new AI algorithms to use them for the training of junior pathologists, to improve their training efforts and to support their decision process in diagnostics. The data catalogue will provide data to the planned AI algorithm.

Expected Size	The amount of data collected can exceed the range of gigabytes and depends on the amount of registered data bio banks and cohorts in the catalogue.
Format	The collected metadata will be available in commonly used data formats. Such as csv, word or pdf documents. The sample information should be harmonized according to the MIABIS standard, where MIABIS (Minimum Information About Biobank data Sharing) aims to standardize data elements, which are used to describe biobanks.
Metadata	The data catalogue will provide metadata, which may be stored in different formats, depending on the corresponding data cohort. The catalogue will provide the metadata in a machine-readable and interpretable format. This rich metadata can be harvested by different search engines.
Will the data be made available for re-use	Yes, since the data catalogue will follow the FAIR principles. In general, the catalogue will provide data as open as possible and as closed as necessary. At this point in time, we cannot estimate exactly to what extent the data can be made publicly available.
How will the data security be ensured?	Data security will be ensured through the procedure in MUG IT like firewalls, VPN, access protocols, only access for participants named in the study protocol. Access to patient data is restricted only to authorized personnel who need the information to perform a defined task. This will be accomplished through secure user authentication for example. Patient data is stored in secure facilities that are designed to prevent unauthorized access, like secured servers.

Partner	**CERTH**
Partner's role	Leading partner
Solution/Tool/Task	**T4.2 DATA FUSION ENGINE**
Data collected/produced/processed	The data fusion engine will fuse heterogeneous information from 3rd party sources, especially environmental, socio-economic and behavioural data with patient data from hospitals, open databases in EU and world-wide that will be stored in the ONCOSCREEN data lake. The output of the multi-source data fusion tool will be a comprehensive repository of various types of data, including structured, semi-structured, and unstructured data, relevant to the ONCOSCREEN project. The repository will contain processed and analysed data generated by the tool, which will be used to generate insights and identify patterns related to cancer risk and diagnosis.
Purpose of the collection/generation/processing	The objective is to utilize the combination of different sources of data; aby analysing this information, new correlations and knowledge can be discovered for the early detection of CRC, as well as to assess risk and categorize patients based on this information. • To gain a deeper understanding of cancer risk and diagnosis. • To generate insights and identify patterns related to cancer risk and diagnosis; • to discover new knowledge and correlations for CRC early detection, risk analysis and stratification of patients based on the multi-modal fusion of heterogeneous sources of data.
Expected Size	Open datasets from European CRC Registries, clinical evaluation protocols, and interventions' outcomes, lifestyle and environmental data: ~500 GBs Data & information from hospitals: ~500 MBs

	Electronic Health Records (EHR) entries for pro- and post-cancer periods, omics, environmental data of specific regions and related health issues: ~10 TBs
Format	The input will be collected mainly in electronic format. The output will likely be presented in various formats, depending on the needs of the intended users. For example, clinicians may need a user-friendly interface that summarizes patient data and provides risk-based stratification results, while researchers may require access to the raw data for further analysis.
Metadata	Metadata will be provided to enhance usability and discoverability of the data, promote collaboration and interoperability, and ensure that the data is used appropriately and in compliance with data usage policies.
Will the data be deposited in a trusted repository?	Data will be stored in the ONCOSCREEN Data Lake repository developed by CERTH.
Will the data be made available for re-use	The data will be published under the CC-BY Creative Commons License that allows users to reuse the data with proper attribution to the project ONCOSCREEN and the European Commission. CERTH will publish a subset of the data within the Zenodo repository service, providing easy access to research results via an innovative viewing option, integration with existing online services, using persistent identifiers, Digital Object Identifiers (DOIs).
How will the data security be ensured?	ONCOSCREEN's fused data-lake developed by CERTH will offer receiving and storing options, metadata querying and visualization options, while by design catering to scaling, tuning, recovering and security functionalities. A key point of this task will be the privacy-preserving technology capable of incrementally allowing the data fusion to take place (a) incrementally, taking steps to ensure the most useful data points are integrated first as well as (b) iteratively, making sure that the usefulness of the resulting fused (big) dataset is preserved while its anonymization is guaranteed and finally (c) in a time-

	efficient manner, namely, delivering usable, even partial-results, as soon as possible and continuously. Security by design provisions will be considered following the System architecture (D4.1, D4.2) and supported by the privacy preservation tool KGEN tool of TLBG for ensuring the anonymity of patients prospective data.

Partner	**iSPRINT**
Partner's role	Leading partner
Solution/Tool/Task	**T4.4 ONCO-CAWA (personalized mobile app)** Linked with ONCO-CLIDE, Data Fusion, ONCO-EVIDA, ONCO-RISTE.
Data collected/produced/processed	All behavioural data of the patients, i.e., the data collected (measured via 3rd party activity COTS wearables or self-reported via questionnaires (PROMs/PREMs) at home. E.g. physical activity, sleep, heart and other vitals measurements, quality of life self-assessments and the like. The data includes all aspects of the everyday life of the patients that the ONCOSCREEN clinicians deem necessary, as long as such data can be self-measured or self-reported by the patients.
Purpose of the collection/generation/processing	The collected behavioural data is to complete the clinical data, towards understanding of the patient, especially in their everyday setting, while away from the healthcare institutions.
Expected Size	The volume depends on (and will vary depending on) the exact attributes to be collected, the frequency of the collection, the number of patients enrolled and the duration of the studies.
Format	ONCO-CAWA provides the data via endpoints that will be exposed, allowing ONCOSCREEN platform to ingest it.
Metadata	The data attributes will be thoroughly documented.

Will the data be made available for re-use	This will be decided by the clinicians. However, since this data is of highly sensitive nature and pertains to the patients' health – it will not be made available for re-use outside of the project, other than in an anonymised form. Fully anonymized versions of the ONCO-CAWA data can be made available outside ONCOSCREEN, e.g., for research purposes, always after the explicit permission of the study controllers.
How will the data security be ensured?	Healthentia, the system empowering ONCO-CAWA is already a class I medical device under the Medical Device Directive (MDD), soon to be a class IIa medical device under the Medical Device Regulation (MDR). All strict security guidelines of MDR are already in place and are successfully audited.

Partner	**Konnektable (KT)**
Partner's role	Leading partner
Solution/Tool/Task	**T4.5 AI-based Clinical Decision Support System (cDSS) ONCO-CLIDE** on precise CRC detection for clinicians Interacts with ONCO-RISTE
Data collected/produced/processed	cDSS will use the aggregated patient data from the diagnostic tools i.e. breath biopsy (ONCO-VOC), liquid biopsy (ONCO-CRISP, ONCO-NMR, ONCO-CTC), AI-assisted colonoscopies (ONCO-AICO), AI-assisted histopathological images (ONCO-AITI) along with faecal tests, and background information from ONCO-CAWA mobile app. These datasets will be acquired from the ONCOSCREEN fused data lake. The data produced by the cDSS through the use of machine learning models will contain suggestions for clinical experts as well as the estimated adenoma-carcinoma level.
Purpose of the collection/generation/processing	To train the models of the cDSS that will assist clinical experts.

Expected Size	The expected size of the dataset cannot be calculated at this stage. However, since the cDSS will use all collected data from the diagnostic tools it is expected to be very large.
Format	cDSS will provide its outcomes in JSON format, and it will store them in the Data Lake component or a dedicated database. This will be defined in the following months, driven by the project's architecture.
Metadata	N/A
Will the data be made available for re-use	The data will be re-trained by the cDSS, in order to provide better accuracy and also train any new incoming data. Moreover, the cDSS outcomes will be able to be used by other technical providers as their input.
How will the security of the data be ensured?	Through encryption techniques

Partner	**BEIA**
Partner's role	**Supporting partner**
Solution/Tool/Task	WP4 and WP5
Data collected/produced/processed	BEIA contributes mainly in WP4 and WP5. More specifically it is responsible analyse personalised behaviours based on commercial of the self-wearables (COTS), selectively given to enrolled population of ONCOSCREEN clinical trial PHASE A, within task T4.3 in order to provide additional intelligence to the risk stratification engine. It will be also involved in T4.4 together with iSPRINT, CERTH and EXUS on the development activities of the ONCOSCREEN mobile app. Also, it contributes heavily on the training activities of WP5 to the end-users leading T5.2, especially in the Romanian clinical trial cluster and facilitates the clinical trials in the Romanian clinical cites (T5.4) proving support.

Purpose of the collection/generation/processing	To provide decision support models for selecting a course of treatment as well as predicting models for the presence of CRC.
Expected Size	No more than 1GB
Format	The information will be stored in a local repository and subsequently subjected to processing, involving the collection of particular indicators and measures. Following this, the processed data will be disseminated to the other project partners. The data in question is stored in a comma-separated values (CSV) file with a retro style.
Metadata	Yes, if it is determined that it is significant.
Will the data be made available for re-use	Certainly, the models that are generated from the data will be accessible for future utilization.
How will the security of the data be ensured?	By applying the rules established internally within BEIA and the set of rules that will be agreed upon with the consortium. Moreover, the data to be processed and utilized in the project has already undergone anonymization.

Partner	**UNI MAINZ**
Partner's role	Leading partner
Solution/Tool/Task	**T2.2 - Study for Target Population groups identification for CRC screening accessibility**
Data collected/produced/processed	Retrospective medical data about CRC prevalence factors (e.g. age, gender, socio-economic status, environmental stressors, behavioural factors) from different hospitals (for example UMC)
Purpose of the collection/generation/processing	The purpose of the collection of data is to define the CRC high-risk target population groups.

Expected Size	8 GB
Format	Data will be only collected in digital format and stored in excel, csv, and sav files.
Metadata	The retrospective medical data will be transferred in metadata with regard to the CRC Prevalence factors. If possible, categories are formed within the different CRC prevalence factors.
Will the data be made available for re-use	Yes.
How will the security of the data be ensured?	Only anonymized data will be collected from the hospitals, so that no traceability is possible. Data will be deposited in a trusted repository and are only accessible to the ONCOSCREEN project team.

Partner	**UNI MAINZ**
Partner's role	Leading partner
Solution/Tool/Task	**T5.4 Clinical Trials Implementation and validation**
Data collected/produced/processed	UNI MAINZ will mainly gather socio-demographic information as well as data in the form of biomarkers related to CRC by means of the chosen technologies (non-invasive test) for recruitment purposes: either via ONCO-VOC, ONCO-NMR, ONCO_RISTE, FIT, ONCO-AITI. The collection of the data will be based on the screening SOPs, containing risk factors, medication data on physical evaluation (e.g., vital signs) and psychometrics. New data will be obtained by measuring psycho and socio-metric variables and biomarkers. Samples will be forwarded to a further laboratory for analysis.
Purpose of the collection/generation/processing	The purpose of the collection of data is as follows:

	• To evaluate and validate all system's components in a comprehensive set of lab tests and clinical trials in hospital settings • To perform the overall assessment of the clinical trials (health technology assessment, economics, lessons learnt, medical findings, feedback from actors/end-users, etc.) • To train/upskill end users and all involved actors during the clinical trials.
Expected Size	Approx. 16 GB (including analyses for validation purposes)
Format	The data will be collected in non-digital (e.g., questionnaires from the patients/participants) digital format (biomarker from the used technologies) and will be stored in the institute's file system as well as in word and excel-files.
Metadata	The metadata will be openly available and licensed under a public domain dedication CC0, as per the Grant Agreement. The guarantee of availability will be discussed with the research partners.
Will the data be made available for re-use	Yes. The data will be made available among the ONCOSCREEN project team (consortium) and re-used according to WP7 and open access statement as per grant agreement. For accuracy purposes, the availability of the data can be stated after project termination.
How will the security of the data be ensured?	All collected data will be assigned a code number and thus pseudonymized for storage. A conclusion on study participants is only possible with a password-protected key, which is only available to the head of the study, Prof. Katja Petrowski and Prof. Markus Möhler. Pseudonymized means that personal data (e.g., name, date of birth and address) are replaced by a value-neutral code (e.g., XYZ01). Using this code, study data can be assigned without personal data being publicly available. Access to the coding list and thus to potential reverse coding lies solely with the study director involved with the study. For data analysis, the data obtained (e.g., biometric data) are

	subsequently anonymized and do not contain any personal information. In this step, it is no longer possible to draw conclusions about study participants, which also means that the study data can no longer be assigned to the generated pseudonym. Deletion of a specific data set after completion of the above-mentioned procedure can no longer be guaranteed. The pseudonymization procedure used in the enclosed study describes a list procedure. In this process, data records are assigned to randomly selected pseudonyms on the basis of a table. These also contain a number. The list is numbered; however, this follows neither a temporal order nor an alphabetical order. The pseudonymized list is stored separately from the anonymized data record.

Partner	**UNI MAINZ**
Partner's role	Leading partner
Solution/Tool/Task	**T1.3 – Clinical and medical steering**
Data collected/produced/processed	In these tasks the collected information refers to internal communication of the ONCOSCREEN research partners for coordination purposes.
Purpose of the collection/generation/processing	Coordination of the medical and clinical aspects of the project (Clinical and medical steering)
Expected Size	Approx. 6 GB
Format	The data is collected via e-mail and Microsoft applications.
Metadata	N/A
Will the data be made available for re-use	No. N/A. The data regarding coordination and implementation purposes will only be available among the ONCOSCREEN team and partners.
How will the security of the data be ensured?	The data is confidential, the access to PCs at the workplace are secured by a personal code and personal password.

Partner	**SERVTECH**
Partner's role	Leading partner
Solution/Tool/Task	**T2.4 Retrospective Data Collection & Analysis**
Data collected/Produced	SERVTECH will not generate any data but will be responsible for helping medical partners to shape the CRC data models and the data repositories that will be used during development, testing and clinical validation in T2.4. All medical partners will contribute by providing (or making available) the necessary retrospective meta-data, database schemas & descriptions. Currently **a federated data base model is being considered.**
Purpose of the collection/generation	Use of an advanced knowledge-based model to homogenise and describe retrospective data by means of using advanced meta-data techniques and standardised data descriptions to achieve data homogenisation on the basis of upcoming standards such as the Fast Healthcare Interoperability Resource (FHIR) a data standard developed and nurtured by HL7 International.
Expected Size	N/A
Format	SERVTECH will assume availability of structured data and schemas stored in relational formats and databases. To support interoperability, it will map these structured data to expressive meta-data and knowledge-based descriptions that can be processed by AI tools. It will store only metadata and knowledge-based descriptions for the data sets to support interoperability. It will use a federated database approach to support data exchange between diverse CRC data sources while respecting privacy and security concerns.
Metadata	T2.4 will work with meta-data that will be openly available and made discoverable, stored and re-used. Metadata will be protected and will be accessed based on user and tool roles and access privileges.

Will the data be made available for re-use	Yes.
How will the security of the data be ensured?	Digital signature and cryptographic hashing functionality could be provided for data and meta data integrity.

Partner	**Technion, ICCS**
Partner's role	Leading partner
Solution/Tool/Task	**T3.1 ONCO-VOC** (breath biopsy non-invasive screening based on exhaled VOCs biomarkers) Linked to ONCO-RISTE, ONCO-CAWA, ONCO-CLIDE
Data collected/produced/processed	In this task, breath samples from known patients and healthy volunteers will be collected through the Gas Chromatography-Mass Spectrometry (GC-MS) instrument, to establish a VOC signal database (a list of VOCs that are significant in the breath samples of CRC-diagnosed patients.). Then using AI-modules for VOC's biomarkers a pattern recognition method for CRC screening and early detection will be established.
Purpose of the collection/generation/processing	First, electro-chemical signal responses from the breath samples of healthy and CRC patients will be collected in order to establish a database. Then, a pattern recognition method for CRC screening and early detection will be established using AI-modules for VOC's biomarkers (ICCS will develop the AI-modules for VOC's biomarkers pattern recognition).
Expected Size	2 MB per patient (total: several GB, depending on the number of subjects)
Format	The data from the breath sampling device will be saved as JSON files
Metadata	Yes.
Will the data be made available for re-use	Yes (the AI modules for VOC's biomarkers pattern recognition).

How will the security of the data be ensured?	Data anonymization. Data encryption. Secure backups. Access control and management.

Partner	**UMINHO**
Partner's role	Leading partner (as far as ONCO-CTC is concerned)
Solution/Tool/Task	**T3.2 Liquid Biopsy Non-invasive Screening from Blood, Urine and Faecal Screening (ONCO-CTC)** Is not dependant / linked with other technologies in the project, though ONCO-CRISP can interact with ONCTO-CTC.
Data collected/produced/processed	In this task, UMINHO will generate different types of data/research outputs. The data will be numerical (e.g., number of CTCs isolated, size of cells, etc) and high-resolution images of the microfluidic devices (e.g., multicompartment, fibre net microarchitecture, channels, etc), cells (e.g., shape of cells, Live/dead, type of cells, subcellular structures, etc), and antibodies, etc.
Purpose of the collection/generation/processing	To develop and validate novel biological collection/screening methods, and fabricated PoC devices/kits for isolation/detection of CTC's and possibly EVs, and diagnostics of CRC.
Expected Size	Hundreds of KB to hundreds of GBs
Format	The expected data (in different formats) to be generated will be temporarily stored in several pieces of equipment (e.g., microplate readers, microscopes, flow cytometers, etc). The data will be extracted by authorized users for processing.. The use of local repository (https://labs.3bs.uminho.pt/welcome.php) will be also considered for storage of data as part of elaboration of QMS reports at UMINHO. The datasets can be made available to the rest of the project partners upon request.
Metadata	Yes.

Will the data be made available for re-use	Yes, in particular through the partner's local repository: https://repositorium.sdum.uminho.pt. However, when sharing specific data, the need to protect IP and to strike a balance between science openness and confidentiality, will be taken into account.
How will the security of the data be ensured?	The data security will be assured by UMINHO standardized informatic security measure. UMINHO assures the dedication of an expert team of IT managers that will monitor that security measures are implemented and updated during the project execution and post project (5 years).

Partner	**LSMU**
Partner's role	Leading partner (responsible for the HTA) all medical and technical partners will contribute (supporting role)
Solution/Tool/Task	**T5.5 (Health Technology Assessment, Effectiveness and Quality Assurance Trial evaluation)**
Data collected/produced/processed	• **Data collected at the trial site by local trial staff based on study protocols**: Participant's trial ID, demographics, medical history, HRQoL (EuroQol instrument), risk factors, CRC diagnosis evaluation, FIT report, diagnostic solutions reports (ONCO-VOC, ONCO-CRISP, ONCO-NMR, ONCO-CTC), colonoscopy report, histopathology report. • **Data generated after determination of effectiveness of ONCOSCREEN diagnostics**: Life-years (LYs) gained, QALYs, data on sensitivity and specificity of the diagnostic solutions. • **Data collected by the Health Technology Assessment team in collaboration with trial sites**: Costs related to patient treatment, use of healthcare resources, use of FIT, diagnostic solutions (ONCO-VOC, ONCO-CRISP, ONCO-NMR, ONCO-CTC), colonoscopy, histopathology.

Purpose of the collection/generation/processing	The data gathered in T5.5 will allow to analyse the affordability and cost-effectiveness of ONCOSCREEN solutions with the standard of care of European Union countries.
Expected Size	Currently it is not possible to estimate the exact size of the dataset. Based on initial approximation the data size should not exceed 5 GB.
Format	All trial sites will be fully responsible for their data collection and storage. Al trial sites will be asked to provide their data based on the protocol in digital format coded with passwords being transferred via other means of communication. Data needed for this specific task will be stored within LSMU facility and accessed only by the people in charge of carrying out T5.5. Data will be stored in excel format.
Metadata	No metadata will be provided within this task.
Will the data be made available for re-use	As of now, a decision has not been made, but we will bring these questions for the discussion with the project partners. Cost-effectiveness report and the protocol will be made openly available.
How will the security of the data be ensured?	• data will be anonymized by the trial sites in line with the study protocol. • data will be revised to check for any possible mistakes or missing information. In that case, data will be updated in cooperation with the particular trial site. All versions of updated data set will be stored • access to sensitive data will be strictly controlled and limited to authorized personnel only. • regular backups and data recovery processes will be implemented to safeguard against data loss; backups will be performed on a scheduled basis, and data recovery procedures will be in place to quickly restore any lost or corrupted data.

Partner	**URIOJA**
Partner's role	Leading partner (leads the health economics study) all medical partners provide data and support (supporting partners)
Solution/Tool/Task	**T6.1 Health economics, Affordability & Cost-Effectiveness analysis of ONCOSCREEN solutions vs Standard of Care**
Data collected/produced/processed	URIOJA will analyse data that will be generated in T5.5 (Health Technology Assessment, Effectiveness and Quality Assurance Trial evaluation) for the purpose of comparison with the standard of care.
Purpose of the collection/generation/processing	To analyse the affordability and cost/effectiveness of ONCOSCREEN solutions with the standard of care of European Union countries.
Expected Size	Dependant on the size of data generated in T5.5
Format	Dependant on T5.5.
Metadata	No metadata will be provided within this task.
Will the data be made available for re-use	As of now, a decision has not been made, but we will bring these questions for the discussion with the project partners.
How will the security of the data be ensured?	Same as in T5.5. • access to sensitive data will be strictly controlled and limited to authorized personnel only. • regular backups and data recovery processes will be implemented to safeguard against data loss; backups will be performed on a scheduled basis, and data recovery procedures will be in place to quickly restore any lost or corrupted data.

Partner	**EY, AXA, URIOJA, CSIC**
Partner's role	Leading partner: EY and AXA, medical partners and policy makers act as supporting partners
Solution/Tool/Task	T6.2
Data collected/produced/processed	This task will process data set from an earlier project reflecting on annual health reviews of workers (they include presence of CRC and 30+ variables). The data will be processed with statistical and machine learning tools to provide predictive models of presence of CRC, most likely in terms of Bayesian networks and influence diagrams.
Purpose of the collection/generation/processing	To support T6.1, T6.2 and T2.4 and to have a benchmark for the evaluation of the current landscape of financial schemes and provide predictive models of presence of CRC and decision support models for treatment choice.
Expected Size	Ca. 1 GB; The models generated will take up about 50Kb each.
Format	The data will be stored on local repository and then after being processed (with the extraction of specific metrics and measures) will be made available to the rest of the project partners. The original data is a csv file. The models will be (initially) in xslx format (a format for GeNIe) or rmd format (for R)
Metadata	Yes, if considered relevant.
Will the data be made available for re-use	Yes. The models developed will be openly available.
How will the security of the data be ensured?	Data will be deposited in a trusted repository (at least in the ICMAT DataLab github site and probably in Zenodo or other common solution adopted by the consortium.)

Partner	**UzL**

Partner's role	Leading partner
Solution/Tool/Task	**T3.2 Liquid Biopsy Non-invasive Screening from Blood, Urine and Faecal Screening (ONCO-NMR)** Is not dependant on/linked with other technologies in the project.
Data collected/produced/processed	NMR metabolomics data
Purpose of the collection/generation/processing	Examination of a series of metabolic factors (taurine, alanine, valine citrate cyclic and others) in combination with fumarate and succinate
Expected Size	Depending on the sample numbers: Derived metadata -a few MB. Original NMR data 1GB per sample.
Format	Original NMR spectra or XML and PDF files
Will the data be made available for re-use	Yes after a period of 1-2 years beyond project closure considering the finalization of any pending publications
How will the security of the data be ensured?	UzL will ensure that any data produced or collected as part of its tasks in the ONCOSCREEN project is stored securely, adhering to all relevant data protection laws and regulations. Specific measures will include secure storage solutions, limited access to data, and appropriate encryption methods.

Partner	**LSMU**
Partner's role	Leading partner (with the contributions from medical partners and policy makers)
Solution/Tool/Task	**T6.3 ONCOSCREEN Living Guidelines and Clinical Recommendations for CRC wide population screening**

Data collected/produced/processed	Patient data will not be collected in this task. Preliminary and final results of ONCOSCREEN clinical trials will be used in this task
Purpose of the collection/generation/processing	The data (i.e. results of ONCOSCREEN clinical trials) will be used to create clinical guidelines and recommendations for extension of national and EU CRC screening strategies as well as adoption of ONCOSCREEN ideas for other cancer types' screening
Expected Size	Currently not able to estimate the exact size of the dataset. Based on our initial approximation the data size should not exceed 1 GB.
Format	Data that will be necessary for this task (results, conclusions of the clinical trials) will be produced in digital format (MS Word, PDF).
Metadata	No metadata will be provided within this task.
Will the data be made available for re-use	The final documents of this task might be available for re-use, e.g., implementing recommendations for strategies related to other cancer localisations. The data will be made openly available.
How will the security of the data be ensured?	• access to sensitive data will be strictly controlled and limited to authorized personnel only. • regular backups and data recovery processes will be implemented to safeguard against data loss; backups will be performed on a scheduled basis, and data recovery procedures will be in place to quickly restore any lost or corrupted data.

Partner	**MoH-GR**
Partner's role	Leading partner (with the contribution from MoH-LT, URIOJA, EY, AXA, ESDO, ECPT, POLA)
Solution/Tool/Task	**T6.4**

Data collected/produced/processed	For the purpose of this task, a structured, closed-ended questionnaire will be distributed among the consortium partners investigating the legislative provisions/policies regarding CRC at place across EU countries. During the 2 roundtable discussions the data that will be collected will pertain solely to sustainable suggestions for up taking the ONCOSCREEN solutions. To this end, anonymized, yet non-personal data will be collected.
Purpose of the collection/generation/processing	To support the development of the replicable regulatory roadmap for the sustainable uptake of newly established solutions concerning CRC prevention. Despite, the suggestions being formulated based on the legislative/policy gaps identified, the data collected will support MoH-GR suggestions in terms of feasibility, availability, suitability.
Expected Size	Less than 400MB
Format	The data will be collected in a digital format and stored in word and excel files. The two roundtables if recorded will be stored in WMV format and the corresponding transcripts in word files.
Metadata	N/A
Will the data be made available for re-use	N/A
How will the security of the data be ensured?	The data will be safely stored in a personal computer in the MoHGR premises for the remaining of the project's duration and then deleted. The computer is solely accessible by his designated operator following a two-step identification process.

Partner	**MOH-LT**
Partner's role	Supporting (contributing) partner

Solution/Tool/Task	T2.5, T 6.1-T6.4, T7.1 and T7.2
Data collected/produced/processed	Questionnaires
Purpose of the collection/generation/processing	Co-shaping the activities for the ONCOSCREEN wide update in the EU and in the National level from the hospital sites' countries participating in the consortium. Providing input for the living guidelines (T6.3) and providing health economics related data for the LT health system to be used in Tasks T6.1 and T6.2. Supporting the activities of the end-user requirements extraction in T2.5 (WP2) and it is actively involved in the dissemination and communication activities of WP7 in Tasks T7.1 and T7.2 (citizens awareness campaigns).
Expected Size	Less than 400MB
Format	The data will be collected in a digital format and stored in word and excel files (WMV format and the corresponding transcripts in word files).
Metadata	N/A
Will the data be made available for re-use	Yes after a period of 1-2 years beyond project closure considering the finalization of any pending publications
How will the security of the data be ensured?	The data will be safely stored in a personal computer in the MoH-LT premises for the remaining of the project's duration and then deleted. The computer is solely accessible by his designated operator following a two-step identification process.

Partner	**ESDO**
Partner's role	Supporting partner
Solution/Tool/Task	**T6.1, T6.3, T6.4 Workshops**

Data collected/produced/processed	ESDO will organise a workshop / symposium whereby they collect and prepare findings and content which is relevant for the ONCOSCREEN project.
Purpose of the collection/generation/processing	The information and findings which we gather from the workshop participants will contribute to the content of the project.
Expected Size	Max. 1 GB
Format	Data will be collected and stored digitally in word, excel, pdf and pptx files
Metadata	N/A
Will the data be made available for re-use	Data will be available during the project's time frame.
How will the security of the data be ensured?	ESDO collects and stores data by means of data processing equipment (EDP) to fulfil its statutory purposes and tasks in accordance with the provisions of the GDPR and has implemented appropriate technical and organizational measures to ensure that no unauthorized access to the data made available (and no unlawful processing of the data) takes place.

Partner	**ECPT**
Partner's role	Leading partner
Solution/Tool/Task	T7.2
Data collected/produced/processed	ECPT will be involved in citizen awareness campaigns. It will be responsible for applying an extensive campaign by informing end-users about the CRC disease, the value of early screening in the QoL improvement and communicate them the novelties of ONCOSCREEN's contribution towards that direction. Detailed communication packages will be created and distributed to project's open workshops to all attendees and hospital sites that will be responsible to implement project's clinical trials. These

	packages will include short presentations, leaflets, short videos and tutorials on ONCOSCREEN's technologies, screening methods and technical tools including also project's ongoing results
Purpose of the collection/generation/processing	Citizen awareness campaigns
Expected Size	Data generated will primarily be qualitative and not particularly large in terms of volume, ca. 1GB
Format	This data will likely be stored in formats like CSV, Excel files, or databases, all encrypted. For communication activities, analytics data from websites or social media platforms might also be collected and stored in formats like PNG, JPG, MP4 or WebM.
Metadata	N/A
Will the data be made available for re-use	Yes.
How will the security of the data be ensured?	Any data produced or collected as part of ECPT tasks in the ONCOSCREEN project will be stored securely, adhering to all relevant data protection laws and regulations. Specific measures will include secure storage solutions, limited access to data, and appropriate encryption methods.
Partner	**HSGO**
Partner's role	Supporting partner
Solution/Tool/Task	T2.1, T2.5, WP5, T6.3, T7.1-T7.2
Data collected/produced/processed	End-users' needs and requirements, ONCOSCREEN living guidelines
Purpose of the collection/generation/processing	Extraction of end user needs and requirements, ONCOSCREEN solution effectiveness based on clinical trial results

Expected Size	End-user requirements: the data generated will primarily be qualitative and not particularly large in terms of volume. Rough estimate: ca. 2 GB
Format	Data will likely be collected digitally, through online surveys, digital feedback forms, and digital analytics tools. This data will likely be stored in formats like CSV, Excel files, or databases, all encrypted. For communication activities, analytics data from websites or social media platforms might also be collected and stored in formats like PNG, JPG, MP4 or WebM.
Metadata	N/A
Will the data be made available for re-use	Yes.
How will the security of the data be ensured?	Any data produced or collected as part of HSGO tasks in the ONCOSCREEN project will be stored securely, adhering to all relevant data protection laws and regulations. Specific measures will include secure storage solutions, limited access to data, and appropriate encryption methods.

Partner	**YCE**
Partner's role	Supporting partner
Solution/Tool/Task	T2.5, T7.1-T7.4
Data collected/produced/processed	YCE will gather and produce qualitative data from stakeholder and end-user engagement activities. This data may include feedback on the ONCOSCREEN project and its tools, as well as data regarding dissemination activities such as reach, engagement, and impact metrics. This data may be collected from end-users through the following typical means: workshops, surveys, feedback forms, and digital engagement on social media platforms.
Purpose of the collection/generation/processing	The purpose of data collection in YCE's tasks is to gauge and optimize the impact of the ONCOSCREEN project, align

	development with end-user needs, and effectively communicate and disseminate information about the project.
Expected Size	Due to the nature of YCE's work in tasks T2.5 and T7.1-T7.4, the data generated will primarily be qualitative and not particularly large in terms of volume. However, considering various forms of data such as feedback, reports, and communication/dissemination materials, a rough estimate could be: T2.5 (End-User Requirements and Continuous Adaptation): Approximate data size could be around 2GB. This accounts for responses from end-users in the form of structured data (surveys) and unstructured data (open-ended responses, notes, etc.) T7.1 (Dissemination and communication activities): The data size for this task might be around 10 GB considering it would involve storing website analytics, engagement metrics, multimedia content for dissemination like images, videos, etc. T7.2 (Citizen Awareness Campaign Implementation): If this task involves similar activities to T7.1, such as metrics from outreach activities and content for the campaigns, the data size could be around 8GB. T7.3 (ONCOSCREEN Living Lab for Open and Gendered based innovations): This task might generate significant data, especially if it involves storing user interactions with the Living Lab platform. Assuming there might also be multimedia content and user-generated data, a rough estimate could be 20GB. T7.4 (Impact creation activities with other EU initiatives and projects): Depending on the amount of correspondence and documentation of joint initiatives, the data size could be around 5GB. Please note that these are just rough estimates and actual data sizes may vary depending on the specifics of the tasks and the level of engagement and interaction.

Format	Data will likely be collected digitally, through online surveys, digital feedback forms, and digital analytics tools. This data will likely be stored in formats like CSV, Excel files, or databases, all encrypted. For communication activities, analytics data from websites or social media platforms might also be collected and stored in formats like PNG, JPG, MP4 or WebM.
Metadata	As per the specific role of YCE in tasks T2.5 and T7.1 - T7.4, the organization primarily participates in requirements gathering, stakeholder engagement, dissemination, communication, and impact creation activities. Given this nature of participation, YCE won't be directly involved in data creation or management, and therefore, won't be responsible for the creation of rich metadata for any datasets. The responsibility of metadata creation and management will likely fall to the partners more directly involved in research data collection and handling.
Will the data be made available for re-use	Anonymized or aggregated data might be made available for re-use, subject to privacy and ethical considerations. However, any data that could potentially identify individual end-users will be subject to stricter controls. This is due to privacy considerations and the specific nature of the data, which may not lend itself to broader re-use outside the context of the ONCOSCREEN project. YCE will manage other research outputs by making sure that they are effectively communicated and disseminated to the relevant stakeholders, ensuring maximum visibility and impact. The nature of the research outputs from YCE's tasks - primarily engagement, feedback, and dissemination data - may not lend themselves to re-use in the traditional sense. However, where appropriate, we will ensure that outputs are produced and maintained in a manner that allows for potential re-use according to the FAIR principles.
How will the security of the data be ensured?	Data security is a priority for YCE. We will ensure that any data produced or collected as part of our tasks in the ONCOSCREEN project is stored securely, adhering to all relevant data protection laws and regulations. Specific

	measures will include secure storage solutions, limited access to data, and appropriate encryption methods. However, as our primary tasks in this project do not involve the handling of sensitive personal data, we anticipate that our data security requirements will be relatively straightforward.

Partner	**EXUS**
Partner's role	Leading partner
Solution/Tool/Task	**T4.3 ONCO-RISTE (Cancer patients' risk-based stratification engine)** Linked with ONCO-CAWA and Data Fusion.
Data collected/produced/processed	Retrospective data from participating countries in regards to the profile of CRC patients and existing CRC registries. Data from the EU Cancer Inequalities Registry (for inequalities-based classification and cut-off values for group creation in regards to education level, income and urbanization level). ONCO-RISTE will generate dynamically a risk-based calculation considering all factors, stressors, the results from WP3 Diagnostic Tools, faecal tests, colonoscopy adenoma/polyp classification, tissue biopsy adenoma/polyp the background of individuals and the pool of expert rules registered in the Clinical Knowledge base (T2.3). Other data may derive from the responses of patients in standardised questionnaires that will be used in the clinical trials.
Purpose of the collection/generation/processing	To develop a semi-empirical risk stratification process to automatically identify dependencies and reveal correlations among a variety of features concerning the clustering of citizens/patients into their respective risk-Level classification.
Expected Size	2GB

Format	Semi-empirical rules will be in textural format. Rule weights will have a numerical value. Other retrospective data that will be collected will be textural or numerical format and not images or videos in any case.
Will the data be made available for re-use	The train data of the risk-based stratification engine will be kept as proprietary. Any other data that will derive from the clinical trials will be offered as open through ONCO-BIOBA tool.
How will the security of the data be ensured?	EXUS as system integrator will provide encryption for data-on-transfer for the various data flows of ONCO-RISTE. Furthermore the KGEN tool for privacy preservation that will be developed by TLBG will be utilised in order to ensure the protection of citizen's identity and personal data. Overall, a multilevel security by design approach will be followed, based on the ISO:27001 standard, as EXUS is already certified and thus in compliance.

Partner	**FIRALIS**
Partner's role	Leading partner
Solution/Tool/Task	**2.3 Clinical Knowledge Base**
Data collected/produced/processed	Bibliographical data regarding the CRC biomarkers and CRC prevalence
Purpose of the collection/generation/processing	Creation of a clinical knowledge base for healthcare professionals, that will utilise existing open datasets from European Cancer Registries, scientific literature review focusing on Whole Genome analysis for large cancer populations and corresponding cancer stratification and genomic cancer prevalence (including the latest findings in Polygenic Risk Scores), pan-cancer analysis of whole genomes, and cancer experts' opinion and publications and based on these, conduct an evidence-based study in order to identify the CRC biomarkers under examination throughout the duration of project.

Expected Size	Yet to be estimated as the Clinical Knowledge Base has not been created yet.
Format	CVS (Excel spreadsheet)
Metadata	Yes.
Will the data be made available for re-use	Yes.
How will the security of the data be ensured?	Clinical data generated by clinical sites and entered into the eCRF will be securely stored by the eCRF provider (contractor). The contractor will be compliant to health data security and storage standards. Firalis assures the dedication of an expert manager that will monitor that security measures are implemented and updated during the project execution. The access to the eCRF data will be limited to authorized users only and will be accessed via secured internal network.

Partner	**ROSENBAUM**
Partner's role	Supporting partner to clinical trials
Solution/Tool/Task	**WP5 Clinical trials design implementation and validation**
Data collected/produced/processed	During the WP together with Firalis we conduct the Clinical trial, Rosenbaum Consulting under management of Firalis will be responsible for the Source Data Verification of the research hospitals in Slovakia, Czech Republic, Hungary, Greece, and Bulgaria. The involved research hospitals must prepare paper-based source document for each patient visit according to the protocol. During the study the research hospital must transfer the data from the paper-based source document to the Electronic Data Capture system which will be set by Firalis. This transfer is conducted by the research sites. Rosenbaum Consulting will be responsible for visiting the sites on a periodical basis, and for verifying the accuracy of the data entered in the Electronic Data Capture system with the data entered

	in the paper-based source documentation to ensure consistency and accuracy of the data.
Purpose of the collection/generation/processing	Rosenbaum Consulting will be having an audit like role for the research sites to ensure safety and efficacy of the Protocol.
Expected Size	At this stage the site of the expected Dataset is very hard to determine however it will not exceed 20 GB.
Format	The Monitoring Visit reports will be written in Templates in Word and after completion transferred to PDF. Having at the end both Word and PDF files stored on the Project Cloud.
Metadata	No
Will the data be made available for re-use	The produced data will be used only for the ONCOSCREEN project
How will the security of the data be ensured?	As for the Electronic Data Capture the safety is ensured via single user secured access. This is facilitated by the EDC administrator. As for the Patient documentation prepared on site must be secured and administered by the research staff at each clinic. The paper source document of the patients must be stored in metal lockable cupboard in the restricted area available only to the research members involved in the Clinical Trial. Once the Clinical trial is over, the source documents has to be stored for the period of minimum 15 years in the hospital archive, for audit related purposes. Additionally, as per the ICH-GCP (International Conference of Harmonization-Good Clinical Practice) all the data created during the clinical trial stage must be prepared in the blinded manner. This means that all the participants in the Clinical Trial must have hidden names and their Personal Identification Numbers. When entering Clinical Trial the patient received a subject Identification Number i.e. 1000-001 and all the procedures during the study participation will be registered in a blinded manner in the Electronic Data Capture where we will then review to ensure safety and quality of the clinical trial.

Partner	**VITO, ICCS**
Partner's role	Leading partner: VITO, ICCS – Supporting partner (contributes to the development of the intelligent analytics dashboard)
Solution/Tool/Task	**4.6 ONCO-EVIDA Evidence-based Decision Analytics Dashboard** ONCO-EVIDA will need retrospective data from other diagnostic tools like ONCO-RISTE
Data collected/produced/processed	• questionnaire responses • Existing MUG CRC bio-bank • Belgian cancer registry data (morbidity, mortality, prevalence, incidence) • Screening data (FIT) Flanders (Centre for Cancer Tracking – CKO, Flanders) • the E-HIS Flemish online dashboard (co-developed by VITO for the Flemish Agency of Care and Health) for the testing of the platform along with • third-party data coming from open databases and/or the European Space Agency (ESA) Copernicus Sentinel 5P data previously analysed by ICCS • Data on socioeconomic status (SES) • Economic data such as the CRC-related information for financial evaluation • Data on air pollution • Behavioural data, such as consumption data (data on use of certain products e.g. low bran/high bran; fruits; vegetables, meat and dairy) in neighbourhoods All above data will with collected at the highest geospatial level of detail possible without hampering GDPR criteria.
Purpose of the collection/generation/processing	The intelligent analytics dashboard will generate a variety of output formats including graphs, tables, and reports, to

	provide policy makers with a comprehensive view of the data and help them make informed decisions related to CRC screening and environmental health management. Graphs and visualizations can help policy makers to quickly identify patterns and trends in the data, which can inform decision making. Tables and reports can provide detailed information on specific aspects of the data, enabling policy makers to delve deeper into the data and make more informed decisions. Outcomes and potential risk factors will be mapped at geospatial levels with ONCO-EVIDA as to allow policymakers to review potential links.
Expected Size	The expected size of the outputs generated by the intelligent analytics dashboard will depend on several factors, such as the amount of data being analysed, the complexity of the analyses being performed, and the specific output format being used (e.g., graphs, tables, reports). It is estimated that the volume will oscillate between 1 and 5GB
Format	CVS and excel files
Will the data be made available for re-use	VITO is a data processor as we do not directly collect data but we are secondary user of already collected data hence, it is not in our power to make the data openly available. However, the results of the data analysis (the system developed) will be made available to be tested for public use with registration with an academic or governmental account.

Partner	**TLBG**
Partner's role	Leading partner
Solution/Tool/Task	**T4.2 Privacy-preserving (data anonymization)**
Data collected/produced/processed	In this task no data is collected. In this tasks, starting from a dataset not anonymized, an anonymized dataset is generated using an anonymization technique based on

	the concept of k-anonymity. To achieve this, TLBG utilizes its proprietary anonymization tool called KGen. KGen is specifically designed to anonymize large datasets by employing a genetic algorithm. The tool iteratively generates anonymized versions of the dataset by applying various transformations and evaluations based on its fitness function. These transformations may include generalization, suppression, pseudonymization, or other anonymization methods. The genetic algorithm then evolves and refines the anonymized versions over multiple generations to optimize the balance between data utility and privacy preservation.
Purpose of the collection/generation/processing	The purpose is to anonymize a given dataset. By anonymizing the dataset, any personally identifiable information that could potentially lead to the identification of individual is removed or obscured.
Expected Size	The expected size of the dataset cannot be determined a priori, as it depends on the specific dataset we receive for anonymization. However, the general principle is that the size of the anonymized dataset should be similar to the size of the non-anonymized dataset provided as input.
Format	In our approach, we aim to build a microservice architecture where data collection and storage are not part of the process. Instead, our focus is on providing endpoints or APIs to anonymize a given non-anonymized dataset without persisting the data. The data anonymization process will be performed on-the-fly as a service. When a client makes a request to anonymize a dataset, the microservice will accept the dataset (in JSON or CSV format) as input through the API endpoint. The anonymization process will then be executed in memory without storing any data persistently. The microservice will process the dataset according to the defined anonymization techniques and requirements. This may involve applying generalization, suppression, pseudonymization, or other anonymization methods based on the provided metadata or predefined rules. The resulting anonymized dataset will be generated as a response to the client's request. Once the anonymization

	process is completed and the anonymized dataset is delivered to the client, the data will not be stored within the microservice or any external storage. This ensures that the data remains secure and private, as it is not retained beyond the immediate scope of the anonymization process.
Metadata	We do not provide metadata as part of the anonymization process. Instead, we require metadata defining the attributes in the dataset and specifying which data should be anonymized or excluded from the anonymization process to be provided to us as an input.
Will the data be made available for re-use	Given the nature of our involvement in the project, we cannot determine a priori whether the produced data can be re-used. The reusability of the data depends on the permissions and agreements in place regarding the original dataset. If the original dataset has permissions or licenses that allow for re-use, it is likely that the anonymized dataset can also be re-used. However, it is essential to review and adhere to the terms and conditions associated with the original dataset to ensure compliance with any restrictions or limitations.
How will the security of the data be ensured?	We do not offer backup services for the anonymized dataset generated as we do not store them. This decision is based on privacy considerations and limitations of our infrastructure.

2.6 Data collection/generation purposes in relation to the project's objectives

ONCOSCREEN invests on prevention, early detection and imaging technologies of colorectal cancer blending a set of novel non-invasive or minimally invasive technologies that will target the stages before the onset of CRC cancer, and its early phases. The ONCOSCREEN project aims to gather and analyse research data that can contribute to the:

- the development of novel, non-invasive, easy-to-use, low-cost, CRC screening technologies of high sensitivity and specificity for early polyp indication, detection, classification and personalized risk status stratification (**ONCO-VOC, ONCO-CTC, ONCO-NMR, ONCO- CRISPR**)
- Provision of a multi-tier, CRC risk-based stratification methodology on target population groups, based on genetic prevalence, socio economic status, environmental stressors,

behavioural factors, gender aspects, educational background, urbanization inequalities both between and within EU countries (**ONCO-RISTE**)

- Enhancement of the existing methodologies for precise CRC detection from clinicians and evidence-based decision making from Policy makers, through AI-based integrated diagnostics and analytics (**ONCO-EVIDA**)
- Designing and evaluating a novel multi-tier CRC screening program for target populations through lab validation and clinical trials in EU countries and regions
- Responding to the questions regarding the existence of statistically significant inequalities across Europe and secondly on whether new CRC diagnostic tests can be optimized for the examined groups (e.g. examining on whether different biomarkers are suitable for different risk population groups). To achieve this goal, ONCOSCREEN will analyse the potential linkages of dietary metabolites at procarcinogenic level, environmental stressors (including carcinogenic substances), gene mutations and socio-economic status as causative factors of CRC.

The above objectives will be achieved by deploying clinical trials in 10 clinical sites (by participating medical health care providers), targeting different EU regions and populations, in order to retrieve new correlations reflecting the differences within and between countries and regions.

ONCOSCREEN will compare and benchmark the proposed methods against standard-of-care methods detecting and screening colon cancer. The comparison will be made not-only on the effectiveness on the methods on colon cancer detection and screening but also in terms of financial viability for patients and healthcare systems with an ultimate goal to be adopted by National/EU healthcare.

2.7 Expected size of the data

For the detailed size of the respective data collected or generated by the respective solutions and tasks within ONCOSCREEN please refer to tables in Section 2.5 above.

2.8 Data utility

Where possible all data contributing and /or validating to project research findings, that do not constitute personal data, will be made available via online repositories in accordance with the limitation set out in this DMP so as to make them reusable by other researchers. In particular, the ONCOSCREEN data may be useful for:

- researchers/academia/scientific community (working on similar projects, who could reuse the ONCOSCREEN data)
- citizens/patients and their families (by providing low-cost/affordable, accurate, fast, personalized and non-invasive CRC screening and early detection methods)

- healthcare professionals, public health services in EU (providing them with new, fast personalized screening methods that are accurate and non-invasive as well as AI assisted diagnostic tools that will allow for screening and early detection of CRC in larger groups of people)
- policy makers (by allowing them to adopt organized screening programs for CRC, increasing the number of people that will be examined for CRC using the ONCOSCREEN solutions)
- medical device manufacturers (who may be interested in investing in ONCOSCREEN diagnostics methods)
- cancer organizations research community
- AI industry

3 FAIR Data

This section describes how the Project will adhere to the principles of data FAIRIFICATION. An effort will be made for providing anonymized clinical data of CRC stages along with behavioural, socio-economic and environmental data across participating (in the clinical trials) countries to reveal potentially health inequalities and currently unknown linkages with parameters like socio-economic status level, environmental pollution, living environment.

3.1 Making data findable, including provisions for metadata

Will data be identified by a persistent identifier?

Yes (by Digital Object Identifiers (DOIs)).

Will rich metadata be provided to allow discovery? What metadata will be created? What disciplinary or general standards will be followed? In case metadata standards do not exist in your discipline, please outline what type of metadata will be created and how?

ONCOSCREEN approach is to create a meta-data ontology that will enable the publication of data across different open repositories including European Cancer Registries provided from the European Network of Cancer Registries (ENCR).

A specific metadata template will be defined, in order to describe, discover and trace existing data collected by ONCOSCREEN and the data that will be generated by it over the next few years.

For each dataset we plan to create the following metadata (i.e., they are all included in the default Zenodo metadata options):

- **Dataset:** title, description, authors, publication date,
- **Persistent identifier**: a DOI is issued in any submission in Zenodo
- **Authors**: name, affiliations, ORCID ID (if available)
- **Funding:** Horizon Europe, grant (name, acronym and number)
- **Access rights and licensing terms**: open access with creative Common Public Domain Dedication (CC 0) or equivalent
- **Link to related publications or other research outputs**: persistent identifiers of related publications and other research outputs
- **Keywords:** we will provide search keywords to optimize the possibility of discovery and re-use
- **Embargo period**: if any.

Will search keywords be provided in the metadata to optimize the possibility for discovery and then potential re-use?

Yes, keywords will be used to increase the possibility of discovery/reuse. These will be as e.g. CRC, colorectal cancer, cancer screening, cancer diagnosis.

Will metadata be offered in such a way that it can be harvested and indexed?

Yes. By following relevant standards (for classification and structure) and making the metadata accessible online, the metadata can be harvested and indexed. ONCOSCREEN will follow best practices suggested by the Research Data Alliance. An example of such an approach could be the BEACON-2 API, which can be adapted to the data of the ONCOSCREEN project.

3.2 Making data openly accessible

All the knowledge collected during the ONCOSCREEN project will be open to access when not covered by copyright or included in patents. Data, protocols, models, strategies, tools and other results, all publications, presentations, training materials, deliverables as well as the interim, periodic and milestone reports will be uploaded and made available on the ONCOSCREEN open website, which will also be maintained for a period of 5 years after the end of the project. Before publicly releasing any data, the consortium will carefully consider and address any related privacy and ethical issues using appropriate aggregation and anonymization techniques.

ORE (Open Research Europe) platform is going to be used by the Consortium in order to publish open access materials to fellow researchers.

Furthermore, ONCOSCREEN embraces Open Access publishing following the guidelines presented by the European Commission. The project results will be published mainly at open access scientific journals, following the Open Access **Gold method**, due to the high impact associated with certain journals. It is anticipated that our researchers and CRC experts will occasionally also follow the Open Access **Green method** in the case of conferences and workshop contributions. Finally, to promote and actively support reproducible research, the ONCOSCREEN team will use open research collaboration platforms for publishing the open software code to allow researchers to reproduce online the results presented in the associated publications.

External researchers will have unrestricted access, via the Zenodo service. ONCOSCREEN will publish a subset of the data within the Zenodo repository service, providing easy access to research results via an innovative viewing option, integration with existing online services, using persistent identifiers, Digital Object Identifiers (DOIs) (as already indicated above). Integration with OpenAIRE corpus infrastructure will be supported. For the datasets that will be made publicly available, a respective Web page will be created on the ONCOSCREEN site that will provide a description of the dataset, along with a download URL.

3.2.1 Repository

Will the data be deposited in a trusted repository?

ONCOSCREEN will act in the systematic promotion of knowledge sharing. A series of actions have been planned to comply with European Commission's Open Innovation 2.0 strategy. In particular, the project's findings will be ascertained by registering the respective results, before actual publication, in pre-print form, in various open and trusted research data repositories as well as in our own ONCO-BIOBA. The pre-publications will be stored in medRxiv/arXiv.

Have you explored appropriate arrangements with the identified repository where your data will be deposited?

Yes. The use of Zenodo (by either uploading or downloading of data) denotes agreement with the terms of use as described at the Zenodo website (i.e. https://about.zenodo.org/terms/) .

Does the repository ensure that the data is assigned an identifier? Will the repository resolve the identifier to a digital object?

Yes. A DOI is issued to every published record on Zenodo.

3.2.2 Data

Will all data be made openly available? If certain datasets cannot be shared (or need to be shared under restricted access conditions), explain why, clearly separating legal and contractual reasons from intentional restrictions. Note that in multi-beneficiary projects it is also possible for specific beneficiaries to keep their data closed if opening their data goes against their legitimate interests or other constraints as per the Grant Agreement.

All major data datasets generated as a result of the project research activities (and NOT constituting personal data) will be made available for the purpose of making them reusable by all identified Project stakeholders.

Since a lot of the data within ONCOSCREEN shall be very sensitive (patients' health data) they cannot be made publicly available. Such data must either be fully anonymized or omitted before making it available. Personal data, managed within ONCOSCREEN will be anonymized (and thus will not constitute personal data) through T4.2 and stored in a form which does not permit identification of users. Before publicly releasing any data, the consortium will carefully consider and address any related privacy and ethical issues using appropriate aggregation and anonymization techniques.

If an embargo is applied to give time to publish or seek protection of the intellectual property (e.g. patents), specify why and how long this will apply, bearing in mind that research data should be made available as soon as possible.

In principle, data will be made available as soon as possible, if legally permitted and after an embargo period due to IP or publication considerations (i.e. at least 1,5 years). The embargo period will be lifted as soon as the interests of partners are protected with respect to IP and publications, allowing data to be made publicly available will be made available as soon as is

possible (for patent related data - once the patent filing is made, the data will be made available in a publication and/or presentation). No access restrictions are foreseen for data that can be made available. Data with restrictions due to legal reasons (e.g., medical data) will not be made available.

Will the data be accessible through a free and standardized access protocol?

Yes. Through the http(s), or FTP protocol. For sensitive personal data a contact protocol will be used, i.e. an email and telephone number of a contact person capable of providing access to the data, will be specified. Contact protocol will be captured in the metadata.

If there are restrictions on use, how will access be provided to the data, both during and after the end of the project?

Data denoted as protected can be shared out of the consortium, as long as interested parties a priori request access from the consortium, explaining how these datasets will be used (e.g. for research purposes).

How will the identity of the person accessing the data be ascertained?

Security-by-design approaches will be followed and by sure any data derived by ONCOSCREEN that will be offered as FAIR/Open will be privacy preserved and will not contain personal information that can lead back to the identity of a particular person. No access restrictions are foreseen for data that can be made available. Data with restrictions due to legal reasons (e.g., medical data) will not be made available and hence there is no need to ascertain the identity of the person accessing the data.

Is there a need for a data access committee (e.g. to evaluate/approve access requests to personal/sensitive data)?

ONCOSCREEN will process large amounts of personal data of sensitive nature (health data). Protection of this data is of utmost importance to the ONCOSCREEN. As such, ONCOSCREEN will appoint a DAC which make sure that the shared data do not contain any personally identifiable information, and that data will be used within the scope of broad consent provided by data subjects participating in the ONCOSCREEN studies. The DAC will be composed of a representative of EXUS (project coordinator), TIMELEX (a legal support partner to the consortium), TLBG (partner responsible for privacy preservation), MUG (FAIR leader), ICCS (as the Technical Manager, ensuring a secure by design architecture) and SERVTECH (as providing expressive descriptions and structure for the retrospective data).

3.2.3 Metadata

Will metadata be made openly available and licenced under a public domain dedication CC0, as per the Grant Agreement? If not, please clarify why. Will metadata contain information to enable the user to access the data?

Yes. Where provided, metadata will be made openly available and licensed under the CC0 license.

How long will the data remain available and findable? Will metadata be guaranteed to remain available after data is no longer available?

Research data will remain available and findable for a period of at least 5 years after the project ends. A sub-part of metadata derived from research institution work will remain publicly available beyond the period of 5 years. Metadata derived from SME's work that may be essential to lead to commercial products will be kept as proprietary until the product's launch. A decision at a later stage will be made and any potential changes will be documented in the periodic report. No original directly, identifiable data will be retained in the context of ONCOSCREEN after the project completion.

Will documentation (or reference about any software) be needed to access or read the data? Will it be possible to include the relevant software (e.g. in open source code)?

If any documentation be needed to access or read the data, the relevant software will be included in open source code.

3.3 Making data interoperable

What data and metadata vocabularies, standards, formats or methodologies will you follow to make your data interoperable to allow data exchange and re-use within and across disciplines? Will you follow community-endorsed interoperability best practices? Which ones?

Complying with national and international standards and protocols around the exchange of data, ONCOSCREEN will create an integrated framework, focusing on open platforms and open data standards, such as the CDISC standards. Also, the HL7 International's FHIR (Fast Healthcare Interoperability Resource) data standard will be considered to achieve data homogenisation. Support will be offered for electronic health data exchange, based on the CEN/ISO 13606. ONCOSCREEN core ontological models will be fully aligned with existing standards and repositories (OpenAIRE, European Open Science Cloud EOSC, etc.), performing the necessary harmonization and schema mapping. ONCOSCREEN will try to use standard vocabularies for all data types to allow their inter-disciplinary interoperability.

The step-by-step FAIRification workflow proposed by Jacobsen et al. will be followed. Gaps for data curation, validation, de-identification, versioning, and indexing will be addressed based on Sinaci et al. and Holub et al. suggestions.

In case it is unavoidable that you use uncommon or generate project specific ontologies or vocabularies, will you provide mappings to more commonly used ontologies? Will you openly publish the generated ontologies or vocabularies to allow reusing, refining or extending them?

If uncommon ontologies or vocabularies are used, the project will follow a methodological approach that includes seeking to reuse existing ontologies and providing mappings to commonly used ones. Any generated ontologies or vocabularies will be openly published to facilitate their reuse, refinement, and extension.

Will your data include qualified references to other data (e.g. other data from your project, or datasets from previous research)?

Yes. The following other data from previous research will be referred to by the consortium partners:

- **ONCO-VOC**: TECHNION will utilize D5.8 and D6.1 from H2020 SniffPhone for device architecture and data analysis methods modifying them for CRC screening and early detection, combined with the GCMS VOCs identification.
- **ONCO-CRISPR:** The D6.1, D6.2 and D6.3 of the H2020 STAMINA Project (EXUS, CCASSSURED) will be utilized for the protocol described, biomarkers design and the tools CRISP-Cas Dipstick SHERLOCK, DETECTR.
- **ONCO-NMR**: UzL will utilize the results of H2020 projects AMMODIT and BiomarCaRE for diagnostic risks assessment models and biomarker identification methodology that will be utilized in ONCO-NMR and the H2020 PanCareLIFE, PanCareSurPass for cancer related early indications
- **ONCO-CTC**: UMINHO will extend the application of its microfluidic devices and tumour-on-ships developed in H2020 WIDESPREAD (D1.3) & ERA FoReCaST (D1.4) for the isolation and detection of CTCs for CRC diagnosis.
- **ONCO-AICO**: ICCS will utilize the general-purpose AI framework developed in H2020 AI@EDGE for reusable, secure, and trustworthy AI for programmable pipelines for the creation, utilization, and adaptation of the secure, reusable, and trustworthy AI/ML models. Similarly, the explainable AI engine of ONCO-AICO will benefit from H2020 COALA trustworthy digital assistant for prescriptive quality analytics, AI novel explanation engine, addressing AI ethics during design, deployment, of new solutions.
- **ONCO-AITI**: MUG will extend the algorithms developed for the histopathological image analysis from the H2020s HEAP (WP7) & BIGPICTURE (WP5).
- **ONCO-RISTE**: EXUS & CERTH will extend its cancer stratification algorithms from the H2020 ONCORELIEF (D4.3, D4.4), for the CRC classification of citizens and patients into homogenous groups based on their risk level, through human-in-the-loop verification from clinicians and risk factor correlation, fusing behavioural, genetic, environmental and other heterogeneous information and QoL assessment instruments through PROMs/PREMs.
- **ONCO-CAWA**: ISPINT will utilize the work done in the frame of H2020-RE-SAMPLE (D5.2) to facilitate RealWorld Data collection (PROMs/PREMs, behavioural data). Furthermore, the recommendation engine of the app will capitalize upon H2020-iHelp (D5.6) and risk models of H2020-INFINITECH Project (D7.1).
- **ONCO-CLIDE & ONCO-EVIDA**: The EU Human Biomonitoring Dashboard (D10.9) of H2020 project HBM4EU will be utilized by VITO. ICCS will augment its portfolio of integrated Copernicus Services and Data Sources being developed in the H2020 EIFFEL and H2020 DIONE by extending the defined toolbox for environmental impact analysis (D2.2, DIONE) to health-related applications and by including Sentinel5P data and Copernicus

Atmosphere Monitoring Service to the integration, augmentation and analysis of time series datasets (D4.1, D4.2, DIONE)

- **ONCO-BIOBA**: MUG will utilize the standardised ontologies and promote synergies with the H2020 projects: EOSC-Life (WP6 Provenance Management & FAIR data management), CY-BIOBANK (WP3, FAIRtoolbox, WP4, Biobank Sample Management), ADOPT BBMRI-ERIC (WP 2 Cohort data & WP 3 IT gateway).

Linkages and interdependencies between various ONCOSCREEN tools and data sets have been indicated in tables in section 2.5 above.

3.4 Increased data re-use

How will you provide documentation needed to validate data analysis and facilitate data re-use (e.g. readme files with information on methodology, codebooks, data cleaning, analyses, variable definitions, units of measurement, etc.)?

A wiki page providing details for data collection, usage, and proper use of the tool will be provided within the ONCOSCREEN website.

Will your data be made freely available in the public domain to permit the widest re-use possible? Will your data be licensed using standard reuse licenses, in line with the obligations set out in the Grant Agreement?

Yes. ONCOSCREEN will distribute the shareable data by adopting licenses that allow re-use of the data. For the majority of freely available data the CC-BY 4.0 will be provided (i.e. the data will be made available, unless otherwise stated, under the Creative Commons **CC-BY 4.0 license** that allows users to reuse the data with proper attribution to the project ONCOSCREEN and the European Commission).

Data will be made openly available immediately at the time of publication of public reports and scientific papers. Data will be given full citation from official project publications and web sites and they will be made available through institutional or public data repositories compliant with OpenAIRE requirements.

Will the data produced in the project be useable by third parties, in particular after the end of the project?

Yes. The data generated in ONCOSCREEN may be usable by:

- Other researchers/academia
- citizens/patients
- clinicians/healthcare professionals
- policy makers
- cancer organizations research community
- ICT-AI industry

- medical device manufacturers

Will the provenance of the data be thoroughly documented using the appropriate standards?

Yes. The generated data will be well-documented and they will have clear provenance information through a wiki file that will be created for ensuring that the data can be correctly interpreted and re-analysed by others. The wiki will include a) a short description of data included b) definitions of column headings and row labels, data codes and measurement units; c) any data processing steps that may affect interpretation of results; d) a description of what associated datasets are stored elsewhere; e) contact information.

Describe all relevant data quality assurance processes.

MUG has experience and tools that enable formalized robust data quality assessments also supporting federated settings. This work will be undertaken as part of T3.4. ONCOSCREEN will seek for synergies contributing to the consolidation of a European Health Data Space. We will build on results of relevant EU initiatives and projects, including those providing models for linking clinical data and samples to CRC research on prevention and early detection on initiatives for cancer, such as, the Knowledge Centre on Cancer the UNCAN.eu platform, European Cancer Information System (ECIS) with the European Network of Cancer Registries (ENCR), the European Reference Networks (ERNs), the Innovative Partnership for Action Against Cancer (iPAAC) Joint Action, the European Commission Initiative on Colorectal Cancer (ECCIC). Also, we will consider already established ESFRI infrastructures and relevant ESFRI cluster projects. MUG as part of the BBMRI project will assist to set up relevant synergies and collaborations.

ONCOSCREEN will liaise with EU-wide and national/regional CRC screening registries and with the European Reference Networks dealing with cancer and other related health science initiatives (EURACAN40, EuroBloodNet, Genturis, EOSC-Life) and integrate data governance procedures in project's Data Management Plan to allow quality assurance and consolidation. For this purpose, the JRC – ENCR Cancer Registries Data Quality Check Software (v1.8.1) will be used for data validity.

4 Management of other research outputs

Digital (e.g. software, workflows, protocols, models, etc.) or Physical (e.g. new materials, antibodies, reagents, samples, etc.).

- Software and protocols
- Urinal, blood and breath samples

How other research outputs will be managed and shared, or made available for re-use, in line with the FAIR principles

Details on how other research output will be managed and shared will be provided in further iterations of this deliverable.

5 Allocation of resources

What will the costs be for making data or other research outputs FAIR in your project (e.g. direct and indirect costs related to storage, archiving, re-use, security, etc.)? How will these be covered?

Costs for open access fees have already been allocated when drafting the project budget at the proposal stage. These costs are included in the overall budget. There are no costs associated with data preservation in institutional servers or the Zenodo service. There are no other relevant costs. Any unforeseen data curation and preservation costs will be covered through the budgets of the involved beneficiaries' assurance (e.g. partners will be in charge of the long-term storage of the data they collected in the context of ONCOSCREEN for at least five years post project, covering any associated costs).

Who will be responsible for data management in your project?

As defined by the Grant Agreement, the entity responsible for data management is Timelex, supported by ALL partners who collect, generate or make data available.

How will long term preservation be ensured? Discuss the necessary resources to accomplish this (costs and potential value, who decides and how, what data will be kept and for how long)?

The costs for long-term preservation of the data have been taken into account by the consortium and will be covered through the own funding of involved partners.

6 Data Security

What provisions are or will be in place for data security (including data recovery as well as secure storage/archiving and transfer of sensitive data)?

The consortium subscribes to the following guidelines in order to ensure the security and confidentiality of data:

- Federated learning: We are currently investigating various approaches to ensure privacy while enabling collaborative analysis and model training. Consistently with the ENISA[3] guidelines for private-by-design computing over sensitive data, one of the mechanisms we are currently exploring is federated artificial intelligence (AI). Based on its assumptions and already available technical opportunities[4], using federated AI can be a valuable approach for the ONCOSCREEN project. Federated AI enables collaboration and model training across multiple nodes or organizations without the need to share raw data, thereby addressing privacy concerns and data protection regulations. By leveraging federated AI techniques, the project can tap into the collective knowledge and resources of the partners while respecting data privacy and security by design. This approach aligns well with the goal of ONCOSCREEN to leverage diverse datasets from different sources while ensuring the confidentiality and privacy of sensitive data. By exploring federated AI, the project can potentially unlock new opportunities for collaborative analysis, model training, and knowledge sharing while maintaining data privacy and security.
- Anonymisation: Personal data managed within ONCOSCREEN will be anonymized through T4.2 and stored in a form which does not permit identification of users (the personal data is pseudonymised upon collection, before it is shared with the consortium). The pseudonymized data is stored in secured cloud spaces and in a form that does not permit the identification of users, thus ensuring compliance with privacy regulations and GDPR.[5]
- Access control and management: access to research data within the ONCOSCREEN project is strictly controlled. All individuals seeking access to research data must sign a data-sharing agreement that outlines their responsibilities regarding data privacy and security; the form is already defined and available upon request and complies with the EU regulations for data integrity; access management controls are implemented to limit access to partners' personnel involved in the research, subject to adequate confidentiality obligation;

[3] https://www.enisa.europa.eu/publications/privacy-and-data-protection-by-design

[4] Isichella, M., Lax, G. & Russo, A. (2022). Partially-federated learning: A new approach to achieving privacy and effectiveness. *Information Sciences*, . doi: https://doi.org/10.1016/j.ins.2022.10.082

[5] Tamburri, D. A. (2020). Design principles for the General Data Protection Regulation (GDPR): A formal concept analysis and its evaluation.. Inf. Syst., 91, 101469.

- Authorisation and authentication procedure for personnel accessing the data: the ONCOSCREEN project recognized the need for robust authorization and authentication procedures to control data access and ensure accountability; the project will define access control policies and mechanisms based on the needs and specifications of patients and user scenarios; healthcare professionals, administrators, and other personnel accessing the data will be granted access based on their roles and responsibilities, following the defined authorization and authentication procedures.
- Encryption: where appropriate, personal data will be encrypted for storage; SSL/TSL will be used for secure data exchanges to ensure that data is encrypted during transmission.
- Transfer of data outside the Economic European Area (EEA) will only occur when appropriate safeguards are in place. These safeguards may include an adequacy decision by the European Commission[6] or the use of Standard Contractual Clauses ensuring that data protection standards essentially equivalent to those in the EEA are upheld.
- Compliance and review: The ONCOSCREEN project is committed to continuous improvement and compliance with data security regulations. Regular reviews of this policy will be conducted to align with evolving best practices and legal requirements.
- Training and awareness: All personnel involved in the ONCOSCREEN project will receive appropriate training and awareness programs to ensure they understand their responsibilities regarding data security, privacy, and confidentiality.

Will the data be safely stored in trusted repositories for long term preservation and curation?

Retrospective data

ONCOSCREEN Retrospective data, collected by SERVTECH in T2.4 will be stored in the Data Lake's repository.

Prospective data

Through the data fusion tool that will be exploited within ONCOSCREEN a number of other data, including patient measurements, patients' demographics, patients' behavioural data and data gathered from other ONCOSCREEN tools (the so called "prospective data") will be collected. This variety of prospective data will be fused with data related to genetic data and/or diagnostic images and produce as output a repository of various types of data related to cancer risk and diagnosis (ONCOSCREEN Data Lake). For privacy reasons, the prospective data will not be stored in a central database. Rather it will be accessed and processed in the data sources where it belongs. ONCOSCREEN takes the federated data management approach, to provide a unified view of data from multiple edge data sources respecting the autonomy of data sources without compromising data privacy and security.

[6] Isreal Adequacy Decision

To preserve privacy for medical applications the edge data sources (nodes) will not be interconnected to one another but rather to a logically centralised control authority called virtual data lake (ONCOSCREEN Data Lake) which comes equipped with appropriate privacy/security mechanisms, e.g., encryption and role-based access.

The virtual data lake will include enriched schema and meta-data descriptions, associations between meta-data descriptions and references (links) to the edge data sources (nodes) to which the meta-data refer.

Data will not migrate to the virtual data lake but rather will be accessed and processed locally at the relevant nodes, which will retain complete autonomy.

Objective is to submit a query (user or tool initiated request) at the level of the virtual data lake which will decomposed & distributed over the edge nodes with the partial answers aggregated and refined centrally & stored at the virtual data lake - an approach which is very similar and conforms to the federated AI execution model.

7 Legal and ethical aspects

7.1 General

The ethical aspects of the ONCOSCREEN Project will be assessed under WP1 (T1.5) which sets out the legal and ethical requirements that the project must comply with. A data manager (Timelex) will be in charge of the supervision and coordination of data-related aspects, will be responsible for ensuring consortium's data governance processes and compliance with policy and/or regulatory obligations.

Protection of the patients' fundamental rights, health, dignity, privacy and personal data, and overall research integrity, constitute important principles to which the ONCOSCREEN project will adhere.

There are three core legal and ethical issues to be dealt with within ONCOSCREEN:

- access to and use of personal data of patients (i.e. health data), to train the AI and ML toolkit
- involvement of real patients, healthcare professionals and policy makers in the conduct of the validation clinical trial;
- the design and use of artificial intelligence.

To address them within the ONCOSCREEN project, a broad range of ethics management tools will be applied including:

- the conduct of a legal and ethical support task to ensure the legal and ethical compliance throughout the course of the project;
- the application of a legal and ethics assessment framework to systematically and rigorously identify legal and ethical challenges.
- state of the art compliance measures including (but not limited to): ensuring legal basis for the use of patient data in full compliance with the requirements of the GDPR and applicable national laws, layered information notices to ensure transparency (unless an exemption applies), data segregation and data pseudonymization/anonymization (where possible) to minimize unlawful access and re-use of data, adherence to privacy-by-design and privacy-by-default principles, continued supervision of the project by a qualified data protection officer(s) and legal and ethical support Partner (TLX).

7.2 Ethical and GDPR compliance

Given that the clinical trials will be conducted in medical health care providers' centres, the necessary ethics approvals will be obtained. Participants will be provided with patients' informed consent forms and information sheets and will be able to withdraw from participation at any given time. The patient selection may not be discriminatory or result in an unfair treatment. Patients

will also - prior to the collection of their personal data - be informed of the reason, the purpose and the consequences of the processing of their personal data. They will be provided with tools allowing them to access their personal data and manage it (as required under the GDPR and other applicable laws). Personal data of the patients will be kept only for a specific period of time; patients will be aware of that time or the criteria used to determine the storage duration. The deliverables resulting from research activities will not include any personal data. No original, directly identifiable data will be retained in the context of ONCOSCREEN after the project completion.

All activities of the ONCOSCREEN project will follow:

- legal requirements, derived principally from the GDPR and from the recent proposal for a regulation laying down harmonised rules on artificial intelligence (EU AI Act), complemented by general product safety regulation and provisions regarding medical devices and Electronic Health Records rules;
- ethics requirements, derived through the application of the principle of Responsible Innovation, on the basis mainly of the European Charter of Fundamental Rights, European Code of Conduct for Research Integrity, the WMA Helsinki (medical research involving human subjects) and Taipei Declaration (research on health databases big data and biobanks), the EU guidelines on ethics in artificial intelligence; and the Ethics guidelines for trustworthy AI.

The project will ensure lawful, fair and transparent data processing based on a suitable/lawful processing ground. Whenever possible and relevant, processing of personal data will be based on consent, but other legal basis will be selected when more appropriate; where possible, the data subjects will have the possibility to exercise their right to object to the processing.

Processing of sensitive personal data may additionally fall under national restrictions, which will be taken into account, where applicable. If required, given the context of processing and specific legislation, the involved data subjects shall be informed of the processing in a clear and transparent manner.

7.3 DPIA

Since ONCOSCREEN project involves processing of health related data, a Data Protection Impact Assessment will be required. Any updates to the DPIA will be included in the periodic reports and the final version of this DMP.

7.4 Use of AI

The consortium commits to meet the key requirements of trustworthy AI in the implementation of the AI systems. In more specific terms, the project will be designed taking into 7 key requirements of trustworthy AI systems:

- Human Agency and Oversight;
- Technical Robustness and Safety;
- Privacy and Data Governance
- Transparency
- Diversity,
- Non-discrimination and Fairness
- Societal and Environmental Well-being
- Accountability).

Assessment List for Trustworthy Artificial Intelligence (ALTAI)[7] tool will be used. AI HLEG guidelines will be followed in regards to avoid biases in regards to race, origin, colour, ethnicity, sexual orientation etc.

ONCOSCREEN will support the close collaboration of humans and AI models, allowing humans to communicate with intelligent systems by responding and assessing the AI-generated recommendations, thus providing feedback to the system's decisions. In this respect, it is imperative to protect patients' health and dignity, by informing them about their interactions with AI. Proper oversight mechanisms must and will be ensured; a human contact person will be available to explain the underlying logic and intended functioning of the AI. The clinicians will remain in control at all times and may not defer to the AI models. The AI models will only serve as a tool in the process and will not replace the decision of a qualified healthcare specialist. The design of the models will ensure to limit bias and will not lead to discriminatory outcomes. The AI used will respect fundamental human rights and freedoms. The applications developed within the project will additionally be monitored for compliance with the pending EU Artificial Intelligence Act.

7.5 Transfers outside the EEA

The activities will be performed mostly in UE, though the Project will also involve data transfers to and from Israel (the Israel adequacy decision form 2011 will be used to justify such transfers).

7.6 Data sharing agreement

The consortium will draft and implement data sharing framework (e.g. a joint controllership agreement) to ensure responsibility and oversight of any personal data processed by ONCOSCREEN partners engaged in personal data processing.

7.7 Collection of human tissue and cells

[7] https://futurium.ec.europa.eu/en/european-ai-alliance/pages/welcome-altai-portal

Within ONCOSCREEN human blood, urine and stool samples will be collected and processed. Directive 2004/23/EC sets the standards of quality and safety for the donation, procurement, testing, processing, preservation, storage and distribution of human tissues and cells. A key feature of the banking of human biomaterials for research in the field of regenerative medicine is the collection of associated information and data such as technical details regarding cells and tissue samples, personal information about sample donors, and research datasets generated from the use of human bioresources. Collection of human biological samples, associated to personal health information, which will be used for biomedical research will take place in ONCOSCREEN project and therefore the protection of the privacy of individual donors and research participants is of crucial importance. ONCOSCREEN project recognizes that all activities and procedures involving the handling of cells and tissues is subject to specific rules (in particular, concerning donor selection/protection; accreditation/designation/authorisation/licensing of tissue establishments and tissue and cell preparation processes; quality management of cells and tissues; procurement, processing, labelling, packaging, distribution, traceability, and imports and exports of cells and tissues from and to third countries).

7.8 Intellectual property

Issues of intellectual property rights and ownership are governed by the provisions of the Consortium Agreement and the Grant agreement, as signed by all project partners. Furthermore, D7.4 ONCOSCREEN Exploitation and IPR Management which will further detail the IP strategy to be followed by the consortium.

8 Other Issues

Will ONCOSCREEN make use of other national/funder/sectorial/departmental procedures for data management? If yes, which ones (please list and briefly describe them)?

In the context of ONCOSCREEN project, UMINHO shall utilize a well-established DMP to govern the handling of research data and other research outputs, which will be shared with partners. This DMP adheres to the principles of findability, accessibility, interoperability, and reusability (FAIR), as follows:

- **Findability**: selected data will be shared through both the website and social media. Data repository that provides a DOI upon deposition will be selected, ideally one that is recognized within the relevant discipline or research community. Access may be restricted prior to filing patents or if a company needs to retain proprietary know-how;
- **Accessibility:** data and research outputs will be made publicly available via sharing or publication as soon as possible, except in cases where IP sensitivity precludes such sharing. Furthermore, all data that underlies publications will be shared upon paper publication;
- **Interoperability**: International System of Units and standard programs such as Microsoft Office Package and GraphPad will be used. High-resolution images files will be, when needed, converted to standard formats (e.g., jpg., png.) to ensure compatibility across all platforms;
- **Reusability**: Raw data will be made publicly available. However, UMINHO will also take into account the need to protect IP and strike a balance between science openness and confidentiality when sharing specific data.

9 Joint Data Management Strategies of Cluster Projects

9.1 "Prevention and early detection" cluster

Prevention, screening and early detection is the most cost-efficient and long-term cancer control strategy. It is known that 40% of cancers could be prevented, but a more personalized understanding of the disease is needed as well as improvements in the existing prevention programmes and general health literacy in Europe and across the globe.

In this context and within the framework of Horizon Europe, the European Commission launched the Mission Cancer to undertake concrete actions with the ambition of delivering tangible results by 2030. The initiative is supported by the establishment of clusters dedicated to the main objectives of the Mission, namely:

1. Prevention,
2. Optimisation of diagnostics and treatment,
3. Support the quality of life of cancer patients, survivors, and their caregivers,
4. Equitable access to all the aforementioned areas,
5. Understand, as the basis of the four previous actions.

The "Prevention, including screening" cluster combines the efforts of 7 projects financed by the Horizon Europe programme HORIZON-MISS-2021-CANCER-02-01: Develop new methods and technologies for cancer screening and early detection. The programme aims at ensuring equitable access to diagnosis and treatments, through the development of new methods and technologies for screening and early detection to allow for less invasive treatments, increase chances of survival and improve the quality of life.

These projects are the following:

- **LUCIA –** Understanding Lung Cancer Related Risk Factors and their Impact
- **DIOPTRA** – Early Dynamic Screening for Colorectal Cancer via Novel Protein Biomarkers Reflecting Biological Initiation Mechanisms
- **MAMMOSCREEN** – Innovative and safe microwave-based technology to make breast cancer screening more accurate, inclusive and female friendly
- **PANCAID** – PANcreatic Cancer Initial Detection via liquid biopsy
- **SANGUINE** – Early detection and screening of haematological malignancies
- **THERMOBREAST** – An innovative non-contact and harmless screening modality set to change the course of breast cancer detection and patient monitoring
- **ONCOSCREEN** – A European "shield" against colorectal cancer based on novel, more precise and affordable risk-based screening methods, and viable policy pathways.

The main goal of the "Prevention and early detection" cluster (hereafter simply referred to as Prevention cluster) is therefore to support the EU Cancer Mission, create added value, establish a policy feedback loop and increase the impact of the EU funding.

The projects in the Prevention cluster fully adopt the European Commission's views on encouraging inter-project collaborations, and thus will act in the systematic promotion of knowledge sharing. Particularly, the Cluster members must be aligned in performing joint tasks and activities such as the synergistic collaboration for the respective Data Management Plans production. The latter include the drafting of this common chapter. The leader of the task is ONCOSCREEN.

The projects within the cluster work on:

1. the integration of retrospective information from European registries, cohort studies and biobanks (including from clinical partners of the projects) on different types of cancer with prospective data to complement missing features. (ONCOSCREEN, PANCAID, DIOPTRA, LUCIA)
2. prospective data for the clinical validation of new technologies for cancer screening (MammoScreen, ThermoBreast, DIOPTRA, ONCOSCREEN, LUCIA and SANGUINE)
3. AI based or AI enabled analysis and screening methods.

Specifically, the projects will process inter alia:

- Clinical/medical data (age, gender, race, ethnicity, medical history, treatments and disease evaluations from EHRs),
- Behavioural data (e.g. exercise/physical activity, dietary patterns)
- Exposomics incl. environmental, sociodemographic and lifestyle data (e.g. air pollution, chemicals, climate, socioeconomic status, oxidative stress)
- Genomic data (measured genotypes, sequence data, gene expression, DNA methylation)
- Medical images (colonoscopy WSI, mammogram, dynamic thermal and microwaves images)
- Tissue, blood, urine, and stool samples (for the purpose of diagnostics based on MS and NMR metabolomics, VOCs from breath biopsy, microfluidic assay for CTCs, blood protein biomarkers).

9.2 FAIR data management

The data used and generated in the Prevention cluster projects can be useful beyond these projects, most notably to healthcare professionals and cancer researchers wishing to understand risk factors and to diagnose cancer at earlier stages. As such, the data generated within one of the projects can be of substantial use to the other projects from the cluster as well .(of course to the extent permitted by confidentiality and intellectual property related provisions of the respective Consortium Agreements).

The purpose of this common chapter is to find common practices to share the information in pan-European research infrastructures, such as the European biobanking platform (BBMRI-ERIC) or the future UNCAN.eu platform, a federated cancer data hub platform currently under development. This is a particularly critical point, as at the present time patient health data networks in Europe show a high level of heterogeneity in terms of involvement of EU Member States as well as the types and interoperability of collected data, organisation and governance of data storage, security or the possibility to use this data for research purposes.

As such, the cluster projects are committed to manage their data in accordance with the FAIR principles and in full compliance with all the applicable European and national legislation. A close collaboration between the projects will be established in this regard, so as to address commonalities on data standards, data validation, the best practices regarding data privacy (pseudonymization or anonymization techniques), data storage and data exchange protocols.

The projects plan on implementing individual measures to make the data **findable, accessible, interoperable and re-usable.**

But in general, the projects consider possibilities for joint exploitation of data within the cluster. The following possibilities are being examined:

- sharing data during the projects,
- sharing risk scores and models towards the end of the projects,
- publishing a common paper,
- implementing the results in healthcare policies and screening programmes.

In order to structure and accelerate collaboration in the above-mentioned areas, the following actions have been taken:

- Creation of a data management task force (including a representative from each of the projects, ONCOSCREEN being the coordinator of the joint cluster efforts) which will act as a working group to discuss and agree on the common aspects related to the management of the data generated in each of the projects (standards, validation protocols, privacy, storage, etc.) as well as on how to foster data exchange between the cluster projects. This task force will also monitor and update the DMPs throughout the duration of the cluster projects.
- Organization of regular virtual meetings of the data management task force; during the first 6 months the meetings have occurred bi-weekly and weekly to discuss basic commonalities of the cluster projects to be addressed in the initial versions of the respective projects' DMPs; they will continue to be held for the remainder of the projects' duration, every quarter. The collaboration will be also conducted via email correspondence (as needed).

Due to the fact that all of the cluster projects are in their early stages, all of the commonalities related to data management will be addressed in further detail in future iterations of this deliverable.

10 Next Steps

ONCOSCREEN

Based on the information received, and in consultation with the partners in order to define strategic orientation when necessary, the DMP will be maintained internally throughout the duration of the Project, and any changes and supplementations will be documented in a table included in periodic reports on project management. All updates will be included in the second iteration of DMP deliverable D1.4 - Data Management Plan (final version).

Considering that this deliverable is due at M6 and hardly any datasets have been generated thus far, it is beyond certain that a number of aspects outlined in the present version of this deliverable will have to be refined or adjusted in the second iteration of this deliverable.

Joint Data Management Strategies of Cluster Projects

As far as the joint data management strategies for cluster projects are concerned, the cluster data management task force will continue to meet regularly to:

A) agree on the common approach to the DM framework of the cluster projects, to the extent that sharing such a common approach is possible considering the character and nature of the projects
B) to discuss any updates and/or recommendations for improvement.

During the cluster projects' collaboration, the data management task force will further explore the possible shared commonalities and will provide more detailed versions of the common aspects of data management in the Prevention cluster in the upcoming versions of the DMPs.

Conclusion

The purpose of this Deliverable 1.6 - Data Management Plan (DMP) is to set out the first data management plan for the ONCOSCREEN project. It has been drafted based on an initial assessment of the Description of Action (DoA) and the partners' early stage contributions. The current version presents the initial strategy in regards to data management for the ONCOSCREEN Action. However, as many aspects and elements are either under development or are still yet to be defined by the consortium, changes may need to be made. This version of the DMP outlines the current understating about the handling of the research data collected and/or generated within ONCOSCREEN and it will act as a living document, incorporating updates (considering also common updates from the clustered projects). The final updates will be duly documented and submitted in the second iteration of this deliverable in M48.