ONCOSCREEN

D4.1 ONCOSCREEN Co-DESIGNED SYSTEM ARCHITECTURE (FIRST VERSION)

31 January 2024

Funded by the European Union

Submission date: M13

Due date: M13

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	D4.1: ONCOSCREEN Co-Designed System Architecture (First Version)		
Work Package	WP4 – ONCOSCREEN Intelligent Platform & Tools for Citizens, Clinicians & Policy Makers		
Type	R – Document, report	**Dissemination Level**	PU - Public
Lead Beneficiary	ICCS		
Contributions	ICCS, EXUS, TECHNION, UMINHO, SERVTECH, CCASSURED, VITO, CERTH, TLBG, UzL, MUG		
Authors	Ioannis Gallos, Dimitra Dionysiou (ICCS), Eleftheria Karataraki (EXUS), Simon Van den bergh (VITO), Jos Bessems (VITO), Gidi Shani (TECHNION), Joaquim Miguel Oliveira (UMINHO), Ulrich Günther (UzL), Mike Papazoglou (SERVTECH), Rogier Louwen (CCASSURED), Thanassis Mavropoulos (CERTH), Athina Tsanousa (CERTH), Daniel De Pascale (Tilburg), Paul Torke (MUG), Aristodemos Pnevmatikakis, George Labropoulos, George Dafoulas (iSprint), Šarūnas Narbutas (POLA LT), Katie Rizvi (YCE)		
Co-authors	N/A		
Reviewers	Pavlos Kosmides (CATALINK), Anaxagoras Fotopoulos (EXUS)		

DISCLAIMER

COPYRIGHT MESSAGE

DOCUMENT HISTORY

Version	Date	Changes	Contributor(s)	Comments
V0.1	19.12.2024	First draft: ToC Released. Partners assignments defined.	ICCS	Proposed changes were discussed with the Project Coordinator
V0.2	15.01.2024	Second draft: Sections updated based on input by partners.	ICCS, TECHNION, UMINHO, SERVTECH, CCASSURED, VITO, CERTH, TLBG, UzL, MUG, EXUS, iSprint	Initial input was included by 15.01.2024
V0.3	23.01.2024	Third Draft Initial Reviewers' comments Addressed	ICCS	Requests were sent to partners and received input was included.
V0.4	26.01.2024	Pre-final version: peer-review comments addressed	ICCS	Addressed comments from peer reviewers.
V1.0	31.01.2024	Final version submitted to EC	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
AI	Artificial Intelligence
AUC	Area Under Curve
BA	Breath Analyzer
CA	Consortium Agreement
Cas	CRISPR associated protein
cDSS	clinical Decision Support System
CI/CD	Continuous Integration and Delivery
CNN	Convolutional Neural Networks
CRC	Colorectal Cancer
CRISPR	Clustered Regularly Interspaced Short Palindromic Repeats
CTC	Circulating Tumour Cells
D	Deliverable
DHKM	Data Homogenization and Knowledge Model tool
EMA	European Medicines Agency
EMT	Epithelial-Mesenchymal Transition
EpCAM	Epithelial Cell Adhesion Molecule
ETL	Extract, Transform & Load
EV	Extracellular Vesicles
FAIR data	data which meet principles of Findability, Accessibility, Interoperability and Reusability
FHIR	Fast Health Interoperability Resources
FIT	Faecal Immunochemical Test

GA	Grant Agreement
GC-MS	Gas Chromatography-Mass Spectrometry
GDPR	General Data Protection Regulation
GelMA	Gelatin Methacrylate
GMP	Good Manufacturing Practices
GNP	Gold Nano Particles
GradCAM	Gradient-weighted Class Activation Mapping
HCC	HepatoCellular Carcinoma
HTTPS	Hyper Text Transfer Protocol Secure
IPR	Intellectual Property Rights
IVD	In Vitro Diagnostics
IVMDR	In Vitro Diagnostic Medical Devices Regulation
JCA	Joint Controllership Agreement
KPI	Key Performance Indicator
LASSO	Least Absolute Shrinkage and Selection Operator
LIME	Local Interpretable Model-agnostic Explanations
LIT	Laboratory Integration Test
LIT	Laboratory Integration Test
M	Month
MDD	Medical Device Directive
MDR	Medical Device Regulation
NMR	Nuclear Magnetic Resonance (spectroscopy)
ONCO-AICO	Training platform for junior colonoscopists/ nurses to enhance their diagnostic ability

ONCO-AITI	Training platform for enhancing the ability of pathologists to classify histopathological CRC images
ONCO-BIOBA	The data catalogue of ONCOSCREEN
ONCO-CAWA	The companion mobile app of ONCOSCREEN
ONCO-CLIDE	The clinical decision support system of ONCOSCREEN
ONCO-CRISPR	ONCOSCREEN's CRC diagnostic tool based on CRISPR
ONCO-CTC	ONCOSCREEN's CRC diagnostic tool based on CTC
ONCO-NMR	ONCOSCREEN's CRC diagnostic tool based on NMR
ONCO-VOC	ONCOSCREEN's CRC diagnostic tool based on VOCs
PEGD	PolyEthylene Glycol Diacrylate
PPT	Privacy Preservation Tool
RBAC	Role Based Access Control
ROI	Region Of Interest
SHAP	SHapley Additive exPlanations
SNOMED	Systematized Nomenclature of Medicine Clinical Terms
SOP	Standard Operating Procedure
SSL	Secure Sockets Layer
T	Task
VOC	Volatile Organic Compounds
WP	Work Package
WSI	Whole Slide Images

Executive Summary

The first version of D4.1 includes information on the Co-Designed System Architecture as well as the end user and technical requirements of ONCOSCREEN within the first year of the project (M1-M12). Setting the user requirements is a continuous and evolving process and this document will set the basis for further updates and refinements in later stages of development. More informed and final form of this deliverable will be made available on the M29 project in D4.2

This document provides a thorough description of the Co-Design methodology adopted by the project and presents the conceptual description and physical architecture of its subsequent tools. Provisions for security and compliance to standards are presented along with a concrete version of the overall architecture of ONCOSCREEN which has been derived as a result of thorough discussions and continuous interactions between partners. A first consolidated list of end user and technical requirements has been already achieved and is exhibited for each tool. Risks and mitigation measures are reported right after in an effort to always plan ahead of time and be prepared for any setback we might face.

The last part focuses on the next steps that is soon to be followed to finalize the System Architecture and re-iterate requirements to always be in accordance with the needs of the users.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

1 Introduction

This Deliverable focuses on the Co-Designed System Architecture of the project ONCOSCREEN, and the technical characteristics/ specifications required to address the needs set by the end users.

The report starts with the description of the Co-Designed methodology adopted by the project, ultimately aiming to the elicitation of the functional and non-functional requirements for each component. Setting the user requirements is considered as a continuous evolving process. This procedure reflects user needs as the outcome of several interactions and user workshops within the realms of the project. For example, the end users (e.g., clinicians, patients/ citizens, policy makers) are prompted to actively think about tools they would use in every day clinical practice (for clinicians), their routine screening (patients/ citizens) and the policy making process (for policy makers). These efforts focus on how the contemporary screening practices can be more effective and efficient, ultimately aiming to raise the awareness of citizens around screening for CRC, revolutionizing screening on the everyday clinical praxis and help to form a policy making plan to control the evolution of CRC incidences.

The third section demonstrates a first draft of the overall architecture which includes a hybrid data management model along with the data flows and pathways between all the subsequent modules. Next, it introduces the tools that constitute ONCOSCREEN presenting their conceptual description, the potential end users and functionalities. This part is comprised of dedicated subsections each one of them named after the corresponding tool. Reading these subsections the reader should acquire a basic understanding of what the module does, who it concerns (Patients/ Citizens, Clinicians, Policy Makers), its physical architecture, interconnection with other tools and its provisions for security and compliance when applicable.

The next section focuses on how the information flows within the ONCOSCREEN platform and integration and deployment aspects are comprehensively descripted. A separate section is dedicated for the "Security by design" methodology that governs all aspects of the project since sensitive medical data will be at the heart of ONCOSCREEN. With retrospective data planned to be utilized from several Hospitals and clinical trials in two Phases, security of information is a top priority and is incorporated from the stages of design. The project embraces the adoption of a zero-trust framework through secure protocols in communication, federated management and data anonymization ensuring data safety and security. Similar emphasis is also given to the compliance to standards and thus the sixth section discusses the standards upon which the physical centralized repository of ONCOSCREEN is based on.

The seventh section of the document is focused on the end user and technical requirements. The section starts with a brief history of events around the interaction of the users with technical partners stating in detail the rounds of interaction and the iterative process towards the elicitation of requirements. Comprehensive lists of requirements (end user/ technical requirements) for each tool are presented, stating the identifier, the priority, the type of each

requirement (functional/ non-functional, for the case of technical requirements), the related user requirements (only for the list of technical requirements) along with some dedicated notes for the status of the particular requirement. It is worth mentioning that since the process is ongoing, some of the requirements are still under consideration/ development and for some backend components, no user requirements are available for the moment. This is mainly because some other components need to progress first before obtaining feedback from the users.

The last part concerns the potential risks and mitigation measures and how ONCOSCREEN plans to react in these cases. In particular, we write down potential obstacles along the way such as the absence of data due to legal constraints and offer indicative workarounds in a manner of being proactive and prepared. Right after, the next steps of the project are discussed concerning future rounds of the Co-Designed methodology, technological advances and the roadmap to safely go to D4.2 and the final version of the "Co-Designed System Architecture" by M29 of the project.

1.1 Role of Task 4.1 within the ONCOSCREEN

As part of the WP4 "Intelligent Platform & Tools for Citizens, Clinicians & Policy Makers", Task 4.1 "Functional, non-functional requirements and Compliant to standards & Secure-by-design System Architecture" is focused on exploiting the expertise of the End Users (Clinicians, Patients and Policy Makers) through interactive methods (including User Workshops) towards the elicitation of functional and non-functional requirements (e.g. security, reliability etc.) of the entirety of ONCOSCREEN Toolkit. This procedure is very closely connected with the conceptualization and development of the overall System architecture which must be secure and compliant to standards by design. T4.1 serves as a foundational step in guiding the subsequent development cycles of all ONCOSCREEN solutions always ensuring the seamless integration of users and their needs into the development process. Next, the translation of user requirements into (precise) technical functional and non-functional requirements during the development stages ultimately leads to the formulation of the overarching architecture of the project. Through these efforts, the task contributes significantly to the project's success by establishing a solid framework/ architecture that aligns (by design) with both user expectations and the contemporary medical technology standards (e.g., the provisions for security and the compliance to standards).

1.2 D4.1 objectives

ONCOSCREEN DoA requirements	Deliverable addressing the requirements	Brief description

WP4: ONCOSCREEN Intelligent Platform & Tools for Citizens, Clinicians & Policy Makers	D4.1: Co-Designed System Architecture (First Version)	T4.1 entails the translation of user requirements into technical functional and non-functional requirements, along with the formulation of key performance indicators (KPIs) for guiding the development cycles of all ONCOSCREEN solutions. Additionally, the task encompasses the design and structuring of the system's architecture in a way that is secure and compliant by design.

Table 1 Description of Action: Task 4.1

1.3 Relationship with other deliverables and tasks

This deliverable receives input from all technical partners. The present document (D4.1) constitutes the first version of the Co-Designed System Architecture and will be submitted in month thirteen (M13) of the Project. As the process is ongoing and many tools are under development the document will evolve and gain more in formation in its final version (D4.2) which will be submitted in M29 of the project. While this document discusses the overall architecture and system requirements, it includes a brief conceptual and technical description of all the subsequent tools. For more details regarding the diagnostic tools the reader should refer to the corresponding sections in D3.1 "ONCOSCREEN CRC Diagnostics" (first version, the final version is to be delivered M29), while for further information concerning the Platforms for Citizens, Clinicians & Policy Makers should redirect to D4.3 "Integrated Intelligent Platform for Citizens, Clinicians & Policy Makers" (first version, the final version is to be delivered M29). Finally, in regard to the methodology of harmonisation of data among other cluster projects and data management more information can be sought from the dedicated working groups as described in D1.3.

Table 2 Linkages between D4.1 and other ONCOSCREEN deliverables

Deliverable	Description of the deliverable	Link to D4.1
D4.2	Co-Designed System Architecture (Final Version)	D4.1 and D4.2 are the First and Final version respectively
D4.3	Integrated Intelligent Platform for Citizens,	While D4.1 provides a brief description and the end user and technical requirements of the intelligent platforms for Citizens, Clinicians and

	Clinicians & Policy Makers (First Version)	Policy Makers, a more in-depth description can be found in D4.3
D3.1	ONCOSCREEN CRC Diagnostics	While D4.1 provides a brief description and the end user and technical requirements of the diagnostic tools, a more targeted and thorough description can be found in D3.1
D1.3	Data Management Plan (First Version)	D1.3 discusses the Harmonization of Data between joint Cluster Projects while D4.1 discusses the harmonization of retrospective datasets that will be used in ONCOSCREEN
D2.1	ONCOSCREEN Clinical Knowledge Base	D4.1 provides a brief description of the Data Homogenization and Knowledge Model tool prioritizing architecture, technical specifications, security and compliance with standards. D2.1 includes an in-depth analysis of the homogenization procedure and also a detailed description of the CRC Knowledge Model in ONCOSCREEN.

2 Co-design Methodology

ONCOSCREEN adhered to a co-design methodology with end users, recognizing the paramount importance of involving clinicians, citizens, and policy makers in the development process. This collaborative approach ensured that the perspectives, needs, and insights of the end users were integrated seamlessly into the project's design and development phases. Setting the user requirements in ONCOSCREEN is viewed as an ongoing and dynamic process shaping an iterative process that spans until month 29 (M29) of the Project. The co-design methodology encompasses various interactive methods to promote collaboration between partners.

A series of nine rounds of end-user requirement co-creation unfolded, underscoring a comprehensive and iterative approach. The project's inception at the ONOSCREEN kick-off meeting on January 13, 2023, saw partners engaging in face-to-face discussions, culminating in a unanimous agreement on a collaborative co-creation methodology. Developed by POLA LT and ICCS, this methodology provided a structured framework for end-user cohort partners to articulate requirements through an online spreadsheet on ONCOSCREEN SharePoint. The technical partners offered thorough descriptions for each tool (e.g. conceptual description, functionalities, actors etc.), receiving valuable feedback from end-users until March 24, 2023. Subsequently, the end-users scrutinized these technological descriptions and roles, expressing their commitment to co-create requirements until April 14, 2023. The collaborative spirit continued with 12 virtual meetings between technical partners and end-users from April 19 to 21, 2023, fostering valuable input documented in spreadsheet files hosted on the dedicated SharePoint of the project. Further refinement ensued as end-users prioritized requirements by order of importance, engaging in prioritization exercises until April 27. Technical partners, in turn, provided feedback on rankings and proposed lists of requirements to be developed during virtual meetings from May 3 to 5, 2023. Ensuring alignment in the co-creation process, alignment teleconferences conducted by POLA LT and YCE on May 10-12, 2023, served as a platform where technical partners presented their perspectives on end-user requirements, elucidating timelines for developing prototypes.

Progress in collecting end-user requirements was reported by POLA (from the End-Users' side and ICCS (from the technical partners' side) to the Management Board meetings. Following a thorough assessment of the already gathered information, the project advanced to the 1st LIT (Laboratory Integration Test) exercise on September 22, 2023. Co-organized by ICCS and POLA, the technical partners demonstrated some first progress on the development of their tools trying to satisfy some indicative end-user requirements. End-users actively participated, providing valuable feedback and listing additional requirements until October 11, 2023. Later on, the Management Board resolved to conduct the 2nd LIT exercise on December 5, 2023. Again, technical partners demonstrated their progress always in accordance with the latest feedback they got from the end-users. Additional clarifications were sought bilaterally from both the end-users and the technical partners. An agreement was reached to commence the next co-creation step after end-users had been provided with prototypes of ONCOSCREEN tools.

The insights gained from these interactions resulted in the end user requirements and their translation to technical requirements as outlined in Section **7**.

An additional workshop is now scheduled for the 1st plenary meeting of the project scheduled for February 1, 2024, in Paris. This workshop will facilitate in-depth discussions, aiming to update, refine and/ or create additional requirements ultimately aiming to fulfill the projects' goals. Ultimately, the iterative process will continue until M29 of the project. As the process is ongoing and the tools become more mature, new feedback rounds from end users will be initiated, thus affecting the original architecture and enabling them to exert an even bigger influence on the proposed system.

3 Conceptual description of Tools

3.1 ONCOSCREEN overall architecture

The overall architecture of ONCOSCREEN features a hybrid data management model that combines a federated database with a centralized repository. This hybrid approach is designed to leverage the strengths of both structures, providing a flexible and efficient solution for data management. The federated database allows for distributed data access and processing, enabling seamless collaboration and resource optimization, while the centralized repository ensures a robust and standardized FHIR compliant storage environment (Bender & Sartipi, 2013). ONCOSCREEN is carefully aligned with established standards (e.g., SNOMED CT) in medical

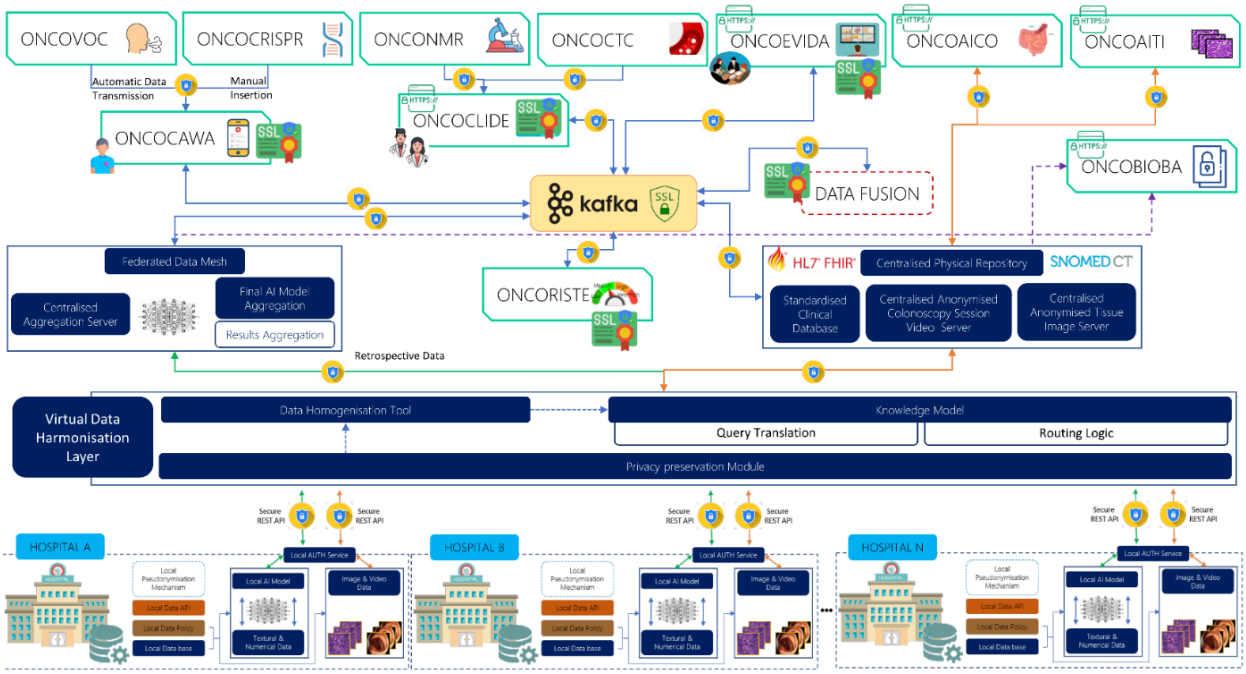

Figure 1 Overall Architecture of ONCOSCREEN depicting the main dataflows of all participating tools in the ecosystem.

technology (Donnelly, 2006), ensuring adherence to industry best practices and interoperability. SNOMED provides a common clinical terminology for consistent medical communication, while FHIR's modern interoperability standards facilitate efficient electronic health information exchange, aligning with the EU's goals for a harmonized and collaborative healthcare landscape. By complying with these standards, the system facilitates seamless integration with other medical technologies and promotes compatibility across diverse healthcare environments.

In Figure **1** we present a schema of the Overall high-level architecture of the project as conceptualized in this first Version of System Architecture. This graphical representation includes all tools comprising ONCOSCREEN (e.g., The diagnostic tools, the clinical decision support system, the training platforms, the policy making support tools and the companion mobile app) along with data flows through the entire framework. More information and in-depth explanations of these tools and their functionalities can be sought on D3.1 and D4.3. Though the

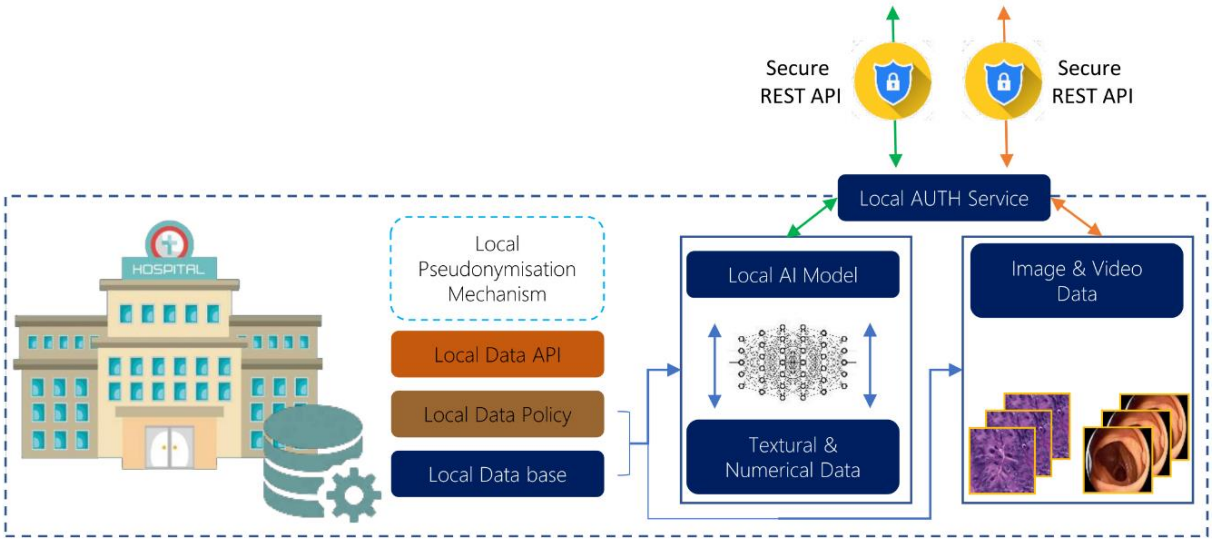

Figure 2 A view of the data flows from the Hospital to the project. This figure is a detailed graphical representation ***of the data flows in the Hospitals also shown in Figure 1***

architecture is still subject to change, the main roadmap of the project along with the provisions for security (e.g. https protocol for web apps, secure APIs, federated data management, SSL protocol), compliance to standards (e.g., FHIR, SNOMED CT), the main orchestrator (e.g., Apache Kafka) and the data pathways are encapsulated within this sketch. Arrows in orange denote the transfer of image/ video data types with (pseudo)anonymised privacy preserved metadata, while in green colour (retrospective data) and the dotted shape depict the transfer of metadata only. A closer dedicated look from the hospital's side is also presented in Figure **2**. This is to showcase how the data flows from the edge (e.g., a hospital) and enters the ONCOSCREEN ecosystem.

3.2 ONCOSCREEN modules

This subsection introduces the tools that constitute the ONCOSCREEN approach for fighting CRC. From conceptual description to their functionalities and potential end users, it aims to give a basic understanding of the role of each module in the project. The toolkit is comprised of clinical diagnostic tools (ONCO-VOC, ONCO-CTC, ONCO-CRISPR, ONCO-NMR), a companion mobile application which serves as the frontend of ONCOSCSREEN (ONCO-CAWA), two training/assisting platforms (ONCO-AITI, ONCO-AICO), a personalized risk stratification engine (ONCO-RISTE), the data lake and fusion engine tool (which comprises of the main

centralized repository of the project and the data fusion engine, a Data Homogenization and Knowledge Model (DHKM) module which will operate as a channel through which different databases with varying structures will interoperate, Privacy Preservation Tool (PPT), a data catalogue (ONCO-BIOBA) featuring the total body of AI-ready and reusable meta-data cohorts of the project, a clinical decision support system (ONCO-CLIDE) and finally a tool for supporting the policy makers (ONCO-EVIDA).

3.2.1 ONCO-VOC

3.2.1.1 Introduction/Conceptual description

Volatile organic compounds (VOC) analysis devices available today in the market, are mainly based on various derivatives of mass spectrometry They are fairly large, inaccurate, and are affected by humidity and other confounding factors. Low- cost miniaturized devices, such as Cyranose-320, Mint, Vivatmo Pro and Vivatmo Me, are either based on the detection of a single biomarker in breath or, they do not show sufficient sensitivity and specificity for a wide variety of VOCs. These past attempts were shown to be inefficient for the screening or even pre-screening of the general population for the development of CRC. Investigation of volatile organic compounds (VOCs) is a novel technique for early detection and investigation of various diseases, including CRC and is based on the emerging sector of Volatolomics. VOCs can be recognized from blood, urine, stool, and breath.

Hypothesis: The VOCs are associated with the pathogenesis of the disease and by investigation of exhaled breath (Altomare et al., 2013), CRC-related VOC patterns can be identified and utilized for screen of asymptomatic general population to achieve early detection. Initially a Gas Chromatography-Mass Spectrometry (GC-MS) instrument will be used for analysing retrospective data of VOC patterns (gas biomarkers) from healthy and CRC-diagnosed patients (establishing a VOC signal database). In this way we will characterize the indicative VOCs associated with CRC. In parallel exhaled samples will be analysed by the ONCO-VOC analyser.

Proposed Technological Solution: a short-exhaled breath sample (2sec.) will be evaluated in less than 2 minutes by the ONCO-VOC breath analyzer that includes 48 gold nano particle (GNP)sensors, in an array format. The reactive pattern of the sensor array to the breath sample will be analysed and evaluated. Analysis of the electro-chemical signal responses with pattern recognition algorithms, will lead to a similarity index to CRC-linked patterns. This should serve as an early indication for the potential development of CRC.

3.2.1.2 Statistical analysis and AI considerations

As for the GCMS analysis, the focus will shift towards identifying statistically significant differences in VOC abundance between patients and healthy controls. These disparities will serve as pivotal biomarkers or features for constructing diverse candidate classification models. While this analysis may lack immediate portability for rapid lab diagnosis, its interpretability

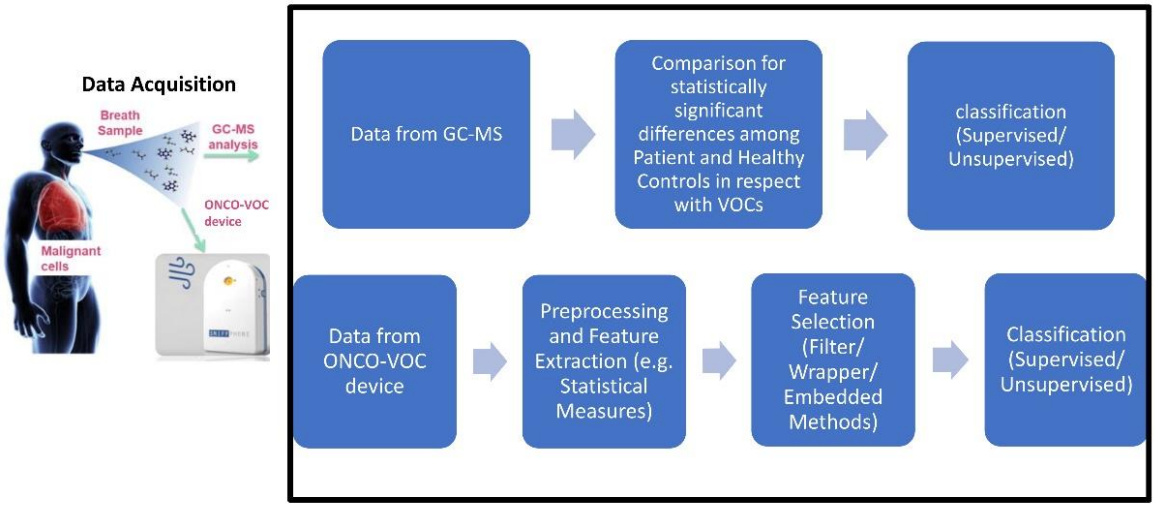

Figure 3 A graphical representation of the envisioned ONCO VOC analysis

offers valuable insights into specific VOCs integral to Colorectal Cancer (CRC) manifestation.

As for the electrochemical signals directly acquired from ONCO VOC device, the analysis is more of black-box nature, since the interpretability at this point is limited. Instead, the analysis will weigh more on the side of sophisticated dimensionality reduction and pattern recognition methods, such as Manifold learning, to facilitate visualization and generate meaningful features. The final strategy formulation hinges upon several factors including parameter count, sample size, data quality (e.g., intra and inter-subject variability), and aligning with the project's end-user needs. An outlined indicative strategy includes a phased approach starting with linear baseline models like logistic regression for binary classification. Subsequently, it progresses to more intricate models such as Least Absolute Shrinkage and Selection Operator (LASSO) regression (Tibshirani, 1996) for feature selection, followed by established machine learning models like random forests and boosting algorithms. Finally, if the sample size permits, exploration of deep learning models like Convolutional Neural Networks (CNNs) will be attempted. Each step in this progression will evaluate the relative benefits of complexity against explainability, aligning model assessments with prevailing scientific consensus, often utilizing the Area Under Curve (AUC) metric (accounting for both sensitivity and specificity). In Figure **3,** a schematic depiction outlines the envisioned ONCO-VOC pipeline of the forthcoming analysis.

3.2.1.3 Functionalities and actors

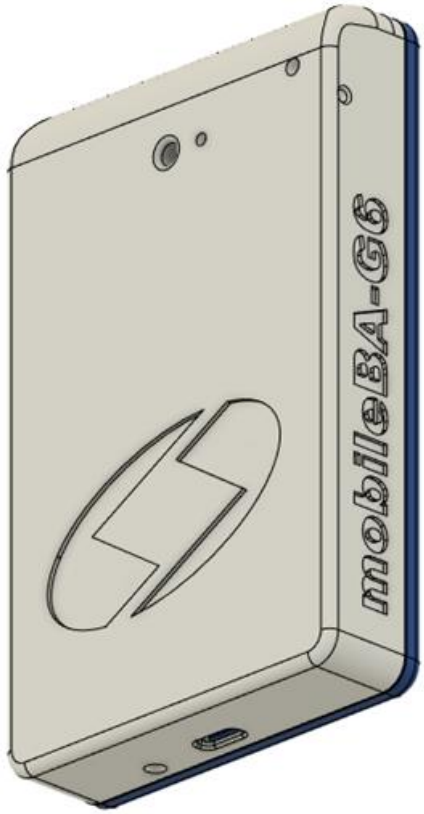

Figure 4 Rear view of the Breath Analyzer (BA-G6)

Potential end-users are clinicians. ONCO-VOC is expected to be used at healthcare centers to collect the breath sample and the resulting VOC reactive pattern. The results will also feed the personalized risk stratification engine (ONCO-RISTE).

3.2.1.4 Physical architecture

Technically, the Breath Analyser (BA-G6) devices (Figure **4**), "ONCO-VOC", consists of an array of 48 molecularly modified gold nanoparticles sensor chip, auxiliary sensors (humidity, temperature), a microfluidic system including a microchamber, valves and a micro pump, heating units to avoid the condensation of water vapor in the fluidic system, a USB PC connector, and an operating software. In the future this device can be developed to harbour Bluetooth Low-Energy (BLE) and a SD card for data storage without the need of a phone or other communicating devices. A short instruction of the correct exhaling procedure will be supplied by the guiding nurses and the rest is all integrated into the electronic unit. The subject will fill out a short questionnaire, the device will acquire room air for reference and immediately, take the breath sample from the subject. The sample should linger in the sampling chamber inside the device for 60 seconds. The sample analysis will start right away and initial results should be displayed via the dedicated software, on the PC. This existing prototype is still experimental and therefore the data would be obtained for future analysis and model development.

3.2.1.5 ONCO-VOC's provisions for Security

The analysis of the breath sample is completely non-invasive. No physical contact is made between the subject and the breath analyser. Moreover, all data transferred from the BA to the server and back will be via secured communication means (eg., kafka SSL).

3.2.1.6 ONCO-VOC's provisions for Compliance to Standards

The BA is still a prototype. Yet even this primary design completely meets CE standards and ensures adherence to European GMP requirements. In addition, as this product matures, all parts and modules will be reinspected thoroughly to reiterate complete adherence to European GMP.

3.2.2 ONCO-CRISPR

3.2.2.1 Introduction/Conceptual description

IN CRC there are only some minimal invasive diagnostic tests available such as the FIT test that screens for occult blood in faecal samples. Clinically relevant genetic biomarkers in colorectal adenomas, colorectal carcinomas in tissue and non- or minimal invasive obtained patient samples, are currently only addressed by mainly the PCR test. Molecular tests such as the PCR test can increase sensitivity and specificity in the detection of (early) CRC, but most of these tests do not reach the clinical setting due to poor performance and thus fail during the clinical validation phases. CRISPR-Cas technologies might provide a solution for this poor performance due to increased sensitivity and specificity. Specifically, the ONCO-CRISPR technology from CCassured is applicable in analysing genetic (early) CRC specific biomarkers in liquid biopsies such as blood, urine and stool samples. Moreover, when this technology is successfully validated, it might even come equipment-free such as seen with a pregnancy or the well-known SARS2 test.

Hypothesis: In CRC there is an unmet need for early detection with tools that are minimal or non-invasive of low cost and harbor more than >95% sensitive and specific, which we can address with different CRISPR-Cas technologies.

Proposed technical solution: The ONCO-CRISPR tool a point of care test will address this unmet need by delivering an increased sensitivity and specificity, affordable/user-friendly tools, equipment free and will deliver to the end-users the prerequisite of the ASSURED criteria (Otoo & Schlappi, 2022). Indeed, more importantly with this tool we can focus on a more reduced invasive procedure like liquid biopsies such as faeces, urine, finger prick or a small blood draw (2-5ml) for obtaining a correct diagnosis.

Moreover the established technology for cancer detection by using small microRNAs that originate from the discovered human CRISPR (van Riet et al., 2022) will be further explored beyond tissue samples, because the latter requires invasive methods. With different partners within the ONCOSCREEN project their presence is therefore explored in liquid biopsies, which would allow their detection with the ONCO-CRISPR tool if their microRNA expression values are black and white, thereby differentiating CRC from healthy controls. Unfortunately, based on our established knowledge (van Riet et al., 2022) their activity is mostly visible in expression value variations, for example in cancer a factor 10 downregulated in comparison to healthy controls. This requires a different detection approach in which we will explore the usage of colorimetrics in combination with the ONCO-CRISPR tool. The development of such a new ONCO-CRISPR tool method will be our target in 2024, next to the clinical validation of our ONCO-CRISPR dipstick prototype four different variants have already been developed in M1-M13. Two ONCO-CRISPR dipstick prototypes enable the detection of CRC specific mutations and two that enable the detection of CRC specific methylation patterns.

3.2.2.2 Functionalities

The ONCO-CRISPR tool can be used on faeces material and liquid biopsies (blood or urine). Our human CRISPR microRNA biomarkers can also be tested as RNA probes on CRC tissue biopsies for pathological examination using RNAscope. Both the ONCO-CRISPR tool and the

human CRISPR microRNA biomarkers allow testing for the presence or absence of colorectal adenomas or colorectal carcinomas. The ONCO-CRISPR tool will enable the detection early-stage colorectal cancers. For the human CRISPR microRNAs their highly sensitivity >0.95 specificity >0.95 is established in tissue but is investigated in liquid biopsies in ONCOSCREEN.

The obtained results are easy to interpret/understand and harbor a good cost/benefit ratio (ideal end user price <25euros per examination). The ONCO-CRISPR tool will require in the end no complex material or laboratory assays to execute the test and interpreted the result. The plan is to test at the point of care and in later stages we expect them to be bought by the EU-citizens themselves in which the tool is equipment free and can be stored at room temperature.

Currently, the ONCO-CRISPR tool can be easily stored under standard conditions (room temperature, 4°C or -20°C). The biomarkers will be preferably obtained from non-invasive samples such as stools, urine and in later stages we cannot exclude the introduction of a minimal invasive finger prick.

The end-users will be the clinicians and laboratory personnel for now, but in the end the EU citizens could become the end-users, in case the tool is/becomes equipment free. Policymakers can use the obtained data blinded for further analyses and policymaking purposes. In ONCOSCREEN, ONCO-CRISPR will interact with ONCO-CLIDE, results can be visualized in ONCO-CAWA and RNA probes detecting human CRISPR activity in tissue can be developed and applied with ONCO-AITI.

3.2.2.3 Physical architecture

The tool consists of a detection dipstick module or a colorimetric readout system module, a reporter, a Cas protein (nuclease, DNA/RNA cutting enzyme) and a guiding RNA, primers, isothermal amplification enzymes and input material, such as faeces, urine or a blood droplet, from which RNA or DNA could be obtained so that the biomarker of interest can be detected (Figure **5**).

The tool requires at the moment minimal work in a laboratory to isolate RNA or DNA, but in the future, we envision having a tool developed that can be used equipment free, such as seen with a pregnancy or SARS-CoV-2 based test.

A movie generated to visualize the usage of the dipstick tool can be seen via this link[1].

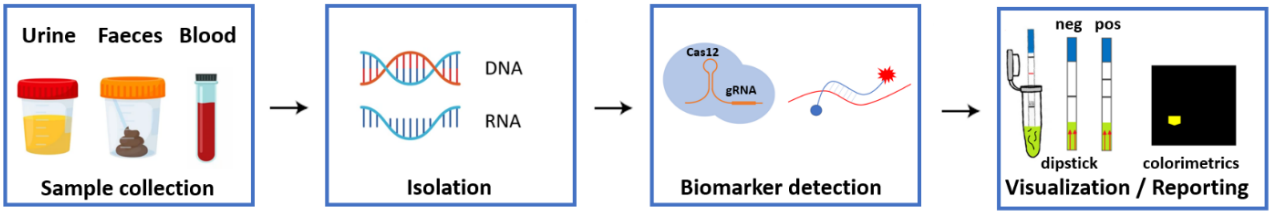

Figure 5 Module components and procedure for ONCO-CRISPR

[1] https://www.youtube.com/watch?v=YyxoSsM13zQ

3.2.2.4 ONCO CRISPR's provisions for Security

For liquid biopsies the procedure is non-invasive, which will not lead to security risks, correct usage will be further defined in a Standard operating procedure. Concerning a finger prick or tissue biopsies, risks of a breach in security are always present that can lead to health problems, such as bleeding, infections, damage to the colon and will require additional attention in the standard operating procedure when ONCO-CRISPR will also include the development of such diagnostic tests. Concerning the tissue retrieval during colonoscopy such standardized protocols are already available. Moreover, in later stages when the diagnostic results are obtained a breach in data security could lead to an unwanted release of the diagnostic results of a patient. Also, when patients, in the end, will test themselves, leakage of diagnostic data might occur, which requires the investment in (bio-)informatic security measures.

3.2.2.5 ONCO-CRISPR's provisions for Compliance to Standards

The ONCO-CRISPR tools are still used as a prototype (TRL4) as stated above. The first clinical validation tests will be initiated in 2024. When proven accurate and able to pass the clinical validation phase (TRL6), it will comply with the various inspections that it will undergo that will ensure to meet the requirement of a commercial product. All parts and modules will be thoroughly inspected to adhere to European GMP and CE marking of diagnostic tests will be requested. CE marking is required for all in vitro diagnostic (IVD) devices sold in Europe. CE marking indicates that an IVD device complies with the European In-Vitro Diagnostic Devices Directive (IVDD 98/79/EC) and that the device may be legally commercialized in the EU. Indeed, in the European Union (EU) the ONCO-CRISPR tools must undergo as any other diagnostic device a conformity assessment to demonstrate they meet legal requirements to ensure they are safe and perform as intended (TRL7-9). This process is regulated at an EU Member State level, but the European Medicines Agency (EMA) is involved in the regulatory process.

3.2.3 ONCO-CTC

3.2.3.1 Introduction/Conceptual description

Circulating tumour cells (CTCs) as epithelial cancer cells can move, migrate and invade blood vessels after epithelial-mesenchymal transition (EMT) and have been characterized as the main causal factor of tumour metastasis mediation. Compared to other cancer biomarkers, CTCs contain molecular and biological information about the tumour as a whole, supporting single cell analysis and ongoing changes in tumours at all stages. At the same time, phenotypic and molecular characteristics of CTCs, can reveal the mechanism of pathogenesis and metastasis of CRC and identify specific mutations in target genes (Jiang et al., 2021). In contrary with the conventional theory that the metastatic dissemination of cancer cells represents the final stage of a deteriorating process, it has been found that CTCs often disseminate at the early stages during the process of tumorigenesis, invading distant organs and eventually developing into

overt metastatic lesions (Rachel et al., 2023). Therefore, the detection of CTCs in the circulation may be proved a feasible way to improve the early diagnosis and treatment of patients with CRC prior to metastasis. However, CTCs are rarely found in blood, at levels typically >1 in a billion cells, and they are fragile. Therefore, the enrichment of CTCs with purity and high recovery is a great challenge. Various assays for identification of CTCs are not yet standardized confusing the scientific community. The CellSearch system, which identifies the EpCAM, is the only standardized system approved by the US Food and Drug Administration for CTC detection in patients with metastatic CRC. However, carcinoma cells that have passed through a partial or complete EMT process are no longer detectable by epithelial phenotype (EpCAM) in peripheral blood. Recent studies reported that a significant portion of CTCs are EpCAM negative showcasing significant limitations of the CellSearch. In addition, the detection methods vary across laboratories and the optimal cut-off value thresholds for CTCs remains unclear. There is a need for large clinical studies with various cut-off values using to assess the prognostic usefulness of CTCs. Although CTCs are vastly described to be present in patients with advanced tumours, little is known about the CTC presence in the blood of individuals with adenomas, some publications reported to find CTCs in patients with adenomas (Tsai WS, 2018). Therefore, studies should be conducted with the devices from ONCO-CTC, to evaluate the ability to detect CTCs in adenoma patients.".

Hypothesis: CTCs can be found in blood as independent cells, and clusters of both CTCs alone and cells aggregates comprising of neutrophils, platelets, and CTCs. Thus, blood-based analysis of CTCs could therefore function as a "liquid biopsy," allowing repeated sampling.

Proposed Technological Solution: UMINHO will use a well-established photolithography technology and will explore the different designs for microfluidics fabrication under an ISO 5

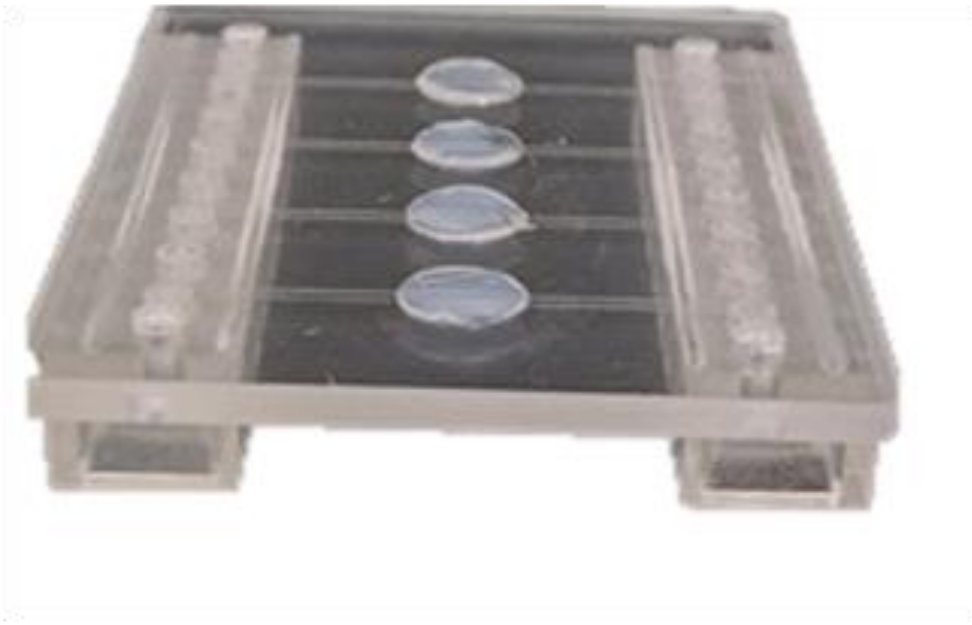

Figure 6 PDMS chip used to develop the 1st generation ONCO-CTC

facility, for focusing on the detection of isolated CTCs that are usually formed at twice the rate of CTC clusters.

In terms of physical properties, CRC cancer CTCs have a diameter less than 10 μm and a cross-sectional area of 40–65 μm^2, similar to white blood cells. The proposed chip platform is based on size exclusion isolation and detection of CTCs, the so-called ONCO-CTC. ONCO-CTC will have also a strong IP background based on the patent Enzymatically Crosslinked Silk Fibroin Hydrogel Microfluidic Platform (WO/2021/038507, 2019). The final technology (can be based on a 1st and 2nd generation prototype) will be optimised by testing different (bio)materials and is expected to be part of a cost-effective CRC diagnostic chips/test kit(s). ONCO-CTC 1st generation is expected to cost less than 35 euros, but other advanced tools and kits (2nd generation) may be more expensive.

The proposed technology could also be used to isolate extracellular vesicles (EVs), which could a way to use this technology in terms of screening since EVs are more prone to be present in early stages of the disease. Although further studies need be conducted. Thus, ONCO-CTC needs to be optimized, i.e. engineering the type of meshes and pore size for isolation of EVs.

3.2.3.2 Functionalities and actors

Innovation and impact: It will provide qualitative and quantitative information across the cancer continuum (initiation, progression, metastasis and relapse), not only providing indication/detection insights, but also classification capabilities. It is expected to provide one sub-component to be integrated in the 2nd generation prototype that can allow personalised testing of anti-cancer therapies.Potential end-users are clinicians and colonoscopists.

3.2.3.3 Physical architecture

The ONCO-CTC 1st generation prototype is comprised by the different sub-components which includes a filtration component for biological entities isolation made of eletrospun fibre meshes. This component is then assembled into a traditional PDMS chip (Figure **6**).

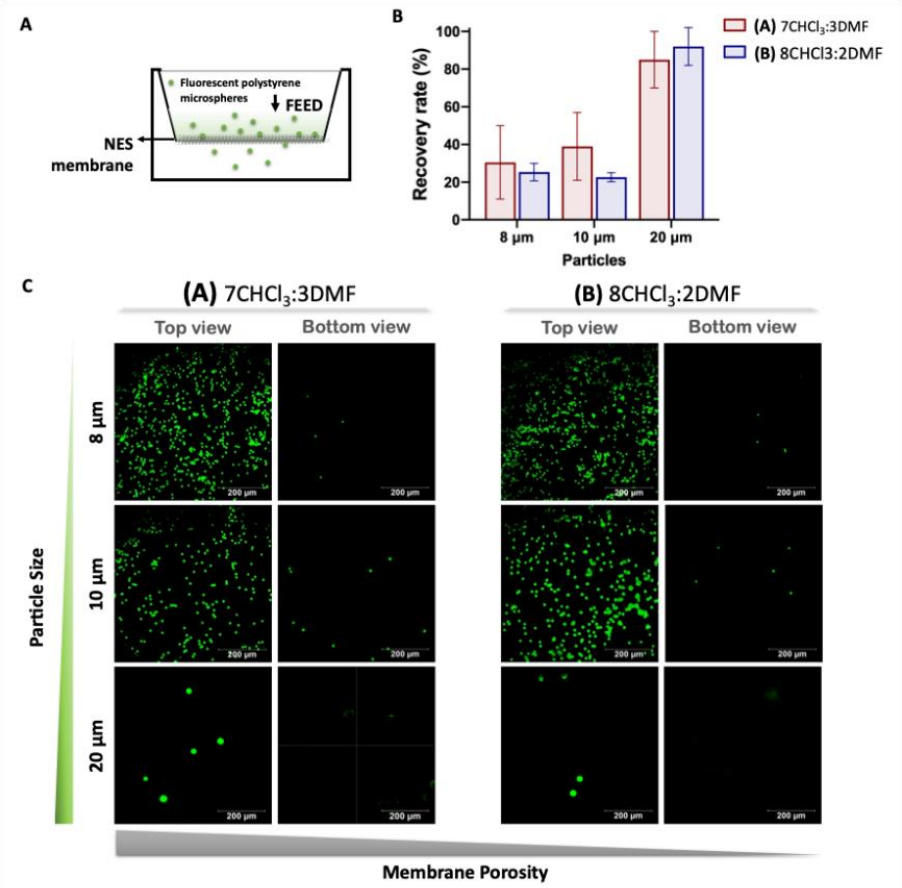

Figure 7 Evaluation of electrospun fiber meshes efficiency for separation of microparticles entities. A) Schematic cross-section view of the structure and particles distribution on EFN in vitro evaluation model in static conditions; (B) The recovery yields obtained in the presence of fluorescent polystyrene microspheres with the diameters of 8, 10, and 20 µm, mimic the behaviours of red blood cells, white blood cells, and CTCs, respectively. (C) Tracking the purification (retention/filtration) process on top and bottom of EFN membranes using fluorescent microscopy.

The design and fabrication of the fibre meshes are optimized by providing a size-exclusion of circulating tumour cells (CTC) and possibly extracellular vesicles (EV's). The first proof of concept studies using microparticles (at a conc. of 200 µm/mL) of different size demonstrated the efficiency of the developed meshes for size-dependent separation under static conditions (Figure **7**).

The cell retention studies were also carried out in static conditions, in vitro (Figure **8**). For this study, HCT-116 and THP-1 cell types were used. Several problems were found with THP 1 cells, including: i) Instability of THP-1 cells; ii) Loss of phenotypic characteristics; iii) Increased permeability; and iv) Non-specific staining. Thus, the ONCO-CTC validation using patient blood should be further carried out. It is noteworthy that UMINHO has already developed an SOP for processing patient blood, enrichment, and CTC's isolation.

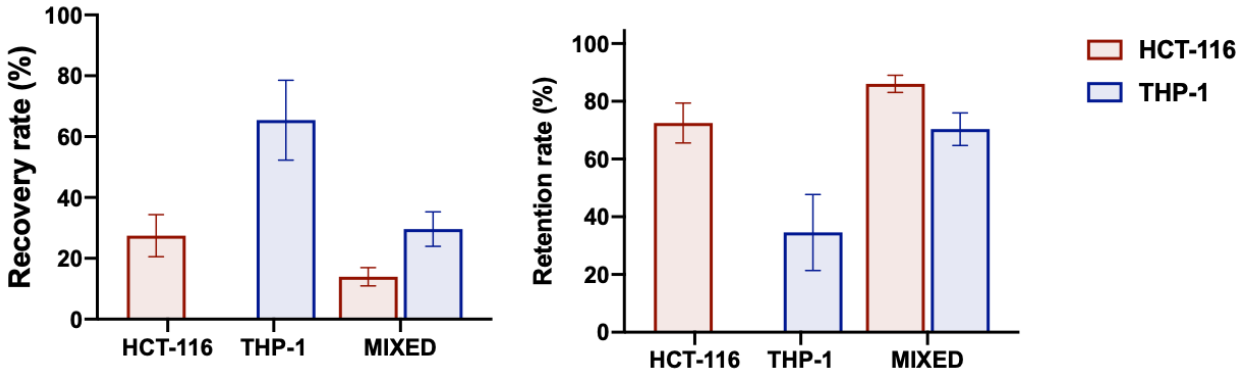

Figure 8. Count cell in Fluorescence microscope by hemocytometer. HCT 116 cells were stained orange with Celltracker

UMINHO also started designing and fabricating an important sub-component of the 2nd generation ONCO-CTC aiming to be used in diagnostics and personalised medicine approaches. This aim consisted of making use of engineering principles for the development of an in vitro tissue model. In vitro models are crucial for studying the intricate interplay between the intestinal epithelium and microbiota in a controlled environment. This work focuses on developing a 3D in

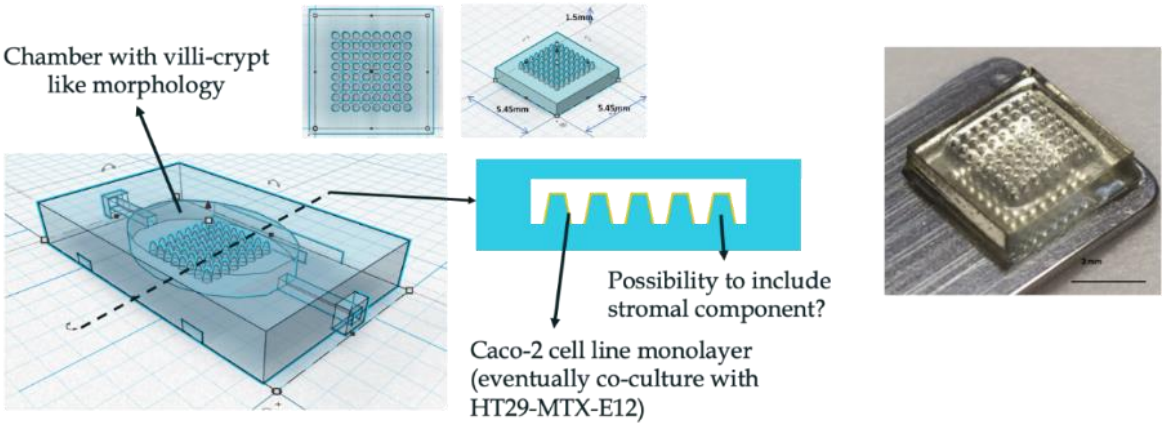

Figure 9 Engineered villi-crypt scaffold-on-chip mimicking the intestinal epithelium fabricated by digital light processing 3D printing

vitro model using DLP-3D printing to mimic the complex intestinal crypt-villi system within a perfusion microfluidic device (Figure **9**).

A blend of home-produced gelatin methacrylate (GelMA) and low molecular weight polyethylene glycol diacrylate (PEGDA), dissolved in DMEM, was employed. The production parameters such as GelMA methacrylation, pre-hydrogel solution composition (polymers, photo-initiator and photo-absorber concentrations), layer thickness and exposure time, and power intensity were optimized. Different formulations were analysed for mechanical properties using

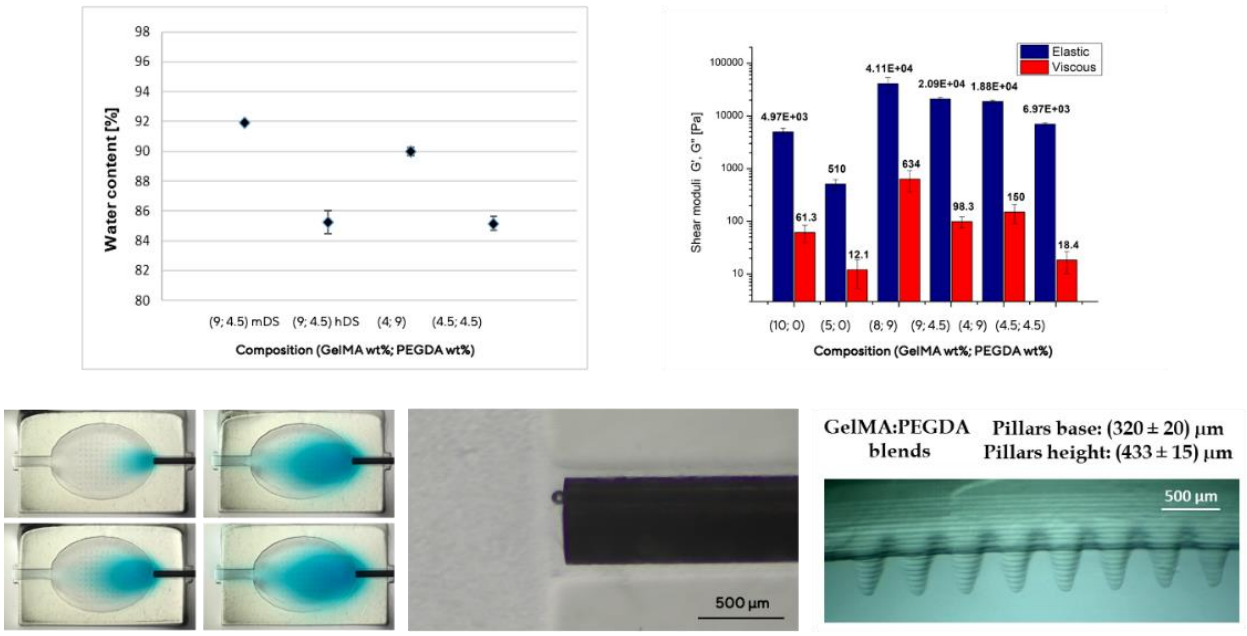

Figure 8 Optimization and characterization of the perfusable villi-crypt 3D in vitro model

a rheometer, to find the optimal balance between stiffness mimicry and printing fidelity and were evaluated for cytocompatibility using Caco-2 cells (Figure **10**). The desired architectures were achieved, having parabolic villi (433 ± 15 µm height; 320 ± 20 µm base), and cylindrical crypts (141 ± 6 µm height; 134 ± 10 µm diameter). GelMA with high degree of substitution was successfully produced, achieving high printing fidelity, long-term stability and cytocompatibility. The chosen GelMA/PEGDA compositions 9.0/4.5 (w%/w%) and 4.5/4.5 (w%/w%) displayed a shear elastic modulus of 21.9 ± 1.3 kPa and 7.0 ± 0.4 kPa, respectively. The closed devices were finally produced with compliant squared channels of 500 µm side, and their perfusability was assessed over a 0 – 100 µL/minute range of flow rates. Both the compositions revealed good cell attachment and proliferation of Caco-2 cells. The Resazurin reduction assay performed at 24 hours revealed increased metabolic activity for the (9; 4.5) composition. This result was also associated with a higher level of cell adhesion.

Overall, we successfully developed a closed microfluidic device, comprising the desired internal villi-crypt like 3D morphology, with tunable mechanical properties and able to sustain the attachment and growth of the Caco-2 cell line opening up the possibility for developing an integrated perusable component in the 2nd generation prototype of the ONCO-CTC (2nd generation prototype) can allow envision the realization of both diagnosis and personalised testing of drugs/therapies.

3.2.3.4 ONCO CTC's provisions for Security

The ONCO-CTC prototype will be produced in an ISO 5 facility following the best practices. The instructions and safety protocols for the final user can be defined later in the project for both 1st generation and 2nd generation ONCO-CTC tool. The training workflow (preparation and procedures) will be defined for each trainee (cell culture technician, lab

technician, clinicians, and colonoscopists) and other end-users by providing demo video(s) and standard operating procedure (SOP's).

3.2.3.5 ONCO-CTC's provisions for Compliance to Standards

The ONCO-CTC prototype and validation studies are being carried out following the standardized in vitro protocols for pre-clinical validation that are currently well-established at UMINHO. The future efficiency studies using patient liquid biopsies will follow the already defined SOP which comprises the use of the current gold standard methods (e.g. CellSearch), i.e. It will be compared to the current standardized protocols for patient blood enrichment and CTC's isolation.

3.2.4 ONCO-NMR

3.2.4.1 Conceptual description

NMR metabolomics provides approx. 40 metabolic parameters, ca. 100 detailed lipoprotein parameters (subclasses regarding triglyceride and cholesterol content of different particle sizes) and (in a recent version developed by UzL) a range of glycoprotein parameters from blood serum or plasma samples. There are several reports showing the potential of metabolomics for colorectal cancer (reviewed in Salmerón et al., 2022), and a recent publication showing metabolite and lipoprotein responses in metastatic CRC patients undergoing liver resection (Constantini et al., 2023). NMR is a highly robust method with excellent reproducibility and minimal variation between measurements. It has been used to study cancer samples in many studies and has been shown to provide good signatures for cancer and cancer progression.

The ONCO-NMR protocol for metabolites and lipoproteins is sold as IVD for research, but not for clinical tests. In the US an NMR protocol for lipoproteins has gained FDA approval and is used commercially by Labcorp. Recent advances show that the glycoprotein signature derived by NMR from blood samples provide a powerful signature for a number of diseases, including cancer, as evidenced for hepatocellular carcinoma. This is based on specific fucoses bound to acute phase proteins. We expect a similar signature for colorectal cancers. From above cited publications we know that there is a pronounced metabolic signature in CRC, according to Constantini et al. 2022 we also expect a lipoprotein signature. The glycoprotein analysis is a new development, but it is well known that glycosylation responds sensitively to cancers.

3.2.5 ONCO-AICO

3.2.5.1 Introduction/Conceptual description

ONCO-AICO is a web platform with a training focus, specifically to assist/ train junior colonoscopists or nurse practitioners in their diagnosis, by employing the support provided by AI algorithms to suggest potential areas of interest in colonoscopy videos for adenoma or carcinoma classification. The platform allows users to view videos, select areas of interest, and

classify them into pre-defined categories as declared by consortium's experts in the field. It then auto-suggests potential areas of interest and provides confidence levels, classification types, and other metadata information. Among those, ONCO-AICO will provide explainable characteristics of the prediction model in order to further and support trainees in terms of understanding the functionality (the outcome) of the AI algorithms on how they gain insight into the provided decision. The platform calculates the correctly classified information and provides users with a score, which can be tracked and improved upon. The platform's objectives include assisting in the training of junior colonoscopists and acting as an automated annotation tool for colonoscopy images enhancing the decision support of the experts/end users.

3.2.5.2 Functionalities and actors

The envisioned end-users of the platform are either junior colonoscopists or nurse practitioners (trainee) and also senior expert colonoscopists (trainer). It features:

- A video player component that can handle different video formats and provide standard playback controls such as play, pause, and seek.
- A user management service that can authenticate users, manage user roles and permissions, and provide appropriate access to different parts of the system based on user type.
- A scoring service that can evaluate trainees based on their submitted selections and calculate their score, providing feedback and guidance as needed.
- Integration of Explainable AI algorithms like Gradient-weighted Class Activation Mapping (GradCAM) (Selvaraju et al., 2017), SHapley Additive exPlanations (SHAP) (Lundberg & Lee, 2017), and Local Interpretable Model-agnostic Explanations (LIME) (Ribeiro et al., 2016), that can interpret (based on heatmaps of significances) the results of AI algorithms in an agnostic way to enhance traceability and trustworthiness of results.
- A metadata generation service that can generate a JSON file containing information on the selected dataset, training parameters, selected algorithm, and other related information for each session, ensuring traceability of calculations and trustworthiness of results.
- A service for trainers to evaluate asynchronously ambiguous images submitted by trainees, annotating them and improving the evaluation service to update its dataset on every new training session.
- User interface controls that are user-friendly, easy to use and can provide options for video manipulation, profile, track record of scores, and other relevant features.
- A feedback and improvement mechanism that can collect feedback from trainers and trainees to enhance the usability and effectiveness of the platform, while also ensuring user privacy and data security.

Trainee workflow:

1. Log in: The trainee logs in to the platform using their account credentials.
2. Select a dataset: The trainee selects a dataset to work on, choosing from the available retrospective datasets such as HyperKvasir, LDPolyVideo, CVC-Colon, PIBAdb, KUMC, and PolypsSet.
3. Choose an algorithm: The trainee selects an open-source adenoma/polyp detection algorithm from a drop-down menu provided on the front-end.
4. View colonoscopy videos: The trainee views the colonoscopy videos provided, using user-friendly controls to manipulate the video playback as needed.
5. Select areas of interest: As the video plays, the trainee selects areas of interest using the platform's annotation tools, indicating potential adenomas or polyps.
6. View AI suggestions: The platform automatically suggests potential areas of interest based on the selected algorithm and displays the results alongside the trainee's annotations.
7. View the explainable oriented features of the AI suggestions as they provided by users' request
7. Classify areas of interest: The trainee classifies the areas of interest as adenomas or polyps based on their own judgement, using pre-selected categories provided by the platform.
8. Submit selections: The trainee submits their selections for review and scoring by the platform.
9. Review score: The trainee can review their score and track their progress over time, with the platform automatically evaluating their performance based on the accuracy of their selections.
10. Continue training: The trainee can continue to work through the available datasets and refine their skills with the assistance of the ONCO-AICO platform.

Trainer workflow:

1. Log in: The trainer logs in to the platform using their account credentials.
2. Access trainee data: The trainer can access data on trainees, including their scores and progress, as well as their selected datasets and algorithms.
3. Review ambiguous selections: The trainer can review any ambiguous selections made by trainees, annotating them, and improving the evaluation service to update its dataset on every new training session. Provided xAI features can be used for decision support.

4. Evaluate trainee performance: The trainer can evaluate trainee performance asynchronously based on their submitted selections and score them, accordingly, providing feedback and guidance as needed.
5. Monitor progress: The trainer can monitor trainee progress over time, tracking improvements in accuracy and identifying areas for improvement.
6. Evaluate and improve the platform: The trainer can provide feedback on the platform's functionality and suggest improvements to enhance its usability and effectiveness in training junior colonoscopists.

3.2.5.3 Physical architecture

This section cannot be fully covered on the deliverable D4.1 as the tool is under development. Based on the latest version of Architecture, ONCO-AICO will be connected with ONCO-AITI, the centralized repository and the privacy preservation tool.

3.2.5.4 ONCO AICO's provisions for Security

Security wise ONCO-AICO platform follows the consortium standards. It requires authentication to gain access in the platform and features three roles (trainee, trainer and administrator). All the security related dependencies are aggregated by the Spring Security framework. Using this framework provides comprehensive support for authentication, authorization, and protection against common exploits. Additionally, ONCO-AICO will enhance user privacy and security by implementing HTTPS (Hypertext Transfer Protocol Secure), encrypting all data exchanged between users and the ONCO-AICO server. This uses SSL/TLS protocols to establish a secure connection, preventing unauthorized access and ensuring a trusted environment for user interactions.

3.2.5.5 ONCO AICO's provisions for Compliance to Standards

This section cannot be covered on the deliverable D4.1 as the tool is under development and the provisions for compliance to standards is yet to be determined.

3.2.6 ONCO-AITI

3.2.6.1 Introduction/Conceptual description

In routine clinical pathological diagnosis, histopathological examination of specimens (e.g. haematoxylin and eosin (H&E) stained glass slides) is conventionally done under light microscopy (Gurcan et. al.; Janowczyk & Anant, 2016). A biopsy consists out of several hundred glands. Whole slide images (WSIs) are the digitised counterparts of extracted biopsy (glass slides) obtained via specialised scanning devices, and they are considered to be comparable to microscopy for primary diagnosis. During the diagnostic process, a pathologist has to identify the dysplastic glands, where a single dysplastic gland can change the complete diagnose! Therefore, a system that can handle efficiently Big Data WSIs information, reducing time in annotation and histopathological image examination, is e definite need.

The aim of this tool is to utilize scalable AI algorithms for supporting junior pathologists during the learning process, to become an expert. This support should not be understood as clinical decision support system, but as a training tool to train the diagnosis on CRC histopathological images. Therefore, the tool aims on junior pathologists, to offer additional information during the training/learning process of WSIs.

3.2.6.2 Functionalities and actors

The end users are junior pathologists who are assisted during the learning process of diagnosis of CRC histopathological WSIs with the help of a viewer-based learning tool. With the help of AI algorithms, the software will offer the user suggestions about possible interesting regions (e.g., as a heat map) that can have a decisive influence on the decision of the diagnosis. In addition, experienced pathologists are observed when viewing WSIs in order to obtain information about regions in the image that are important for the diagnosis. This information should also be made available to the trainee in the training tool as helpful suggestions for regions of interest. With the help of these suggestions, junior pathologists can be supported during their training time to learn the diagnosis of CRC histopathological images.

3.2.6.3 Physical architecture

ONCO-AITI is a learning tool in which not only the person being trained interacts with the tool, but also experienced experts. The basic principle is that a trainee can use the tool to improve the assessment of WSIs through training. As an aid, the tool not only provides visualization of the WSIs but also helpful additional information that can be displayed on request. These can consist of pointing out interesting regions in a WSI that could be important for an assessment of the WSI. This identification of regions of interest [ROIs] can be done in different ways and can refer to different sources of information. Possible representations could be, for example, heatmaps or view boxes that show the amount of magnification and the corresponding coordinates in the image a trainee can see interesting information. The source of such information can be the results of AI algorithms or the knowledge from reviews observed from experienced experts. The illustration below (Figure **9**) shows this concept in a very simplified form.

The tool itself is planned as web-based software into which users can log in via a user management system. Since the aim is to avoid that personal data can be exchanged between partners through the tool, it is also possible to consider that the software could be installed as a

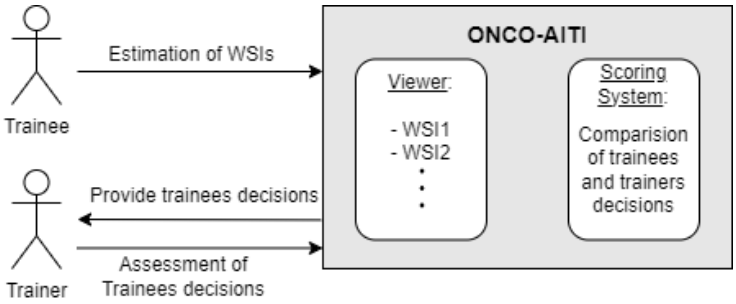

Figure 9 A schematic diagram of ONCO-AITI

stand-alone on a computer. This would allow users to optionally use in-house WSIs for training without this data having to be stored on a specific server.

3.2.6.4 ONCO AITI's provisions for Security

Data listed and described in this tool must comply with Austrian and EU-wide laws regarding the protection of personal rights. An example of this is compliance with the EU-wide GDPR (General Data Protection Regulation). All data is taken from retrospectively existing data sets. The AI algorithm is trained with the help of in-house data sets. However, this patient-related data is not later made available to the user. Rather, the user will be offered possible open-source data sets in the tool or will have the opportunity to use the tool on their own data sets. All data is taken from retrospectively existing data sets. The AI algorithm is trained with the help of in-house data sets. However, this patient-related data is not later made available to the end-user. Rather, the user will be offered anonymized, open-source data sets in the tool or will have the opportunity to use the tool as an offline version in order to apply their own data sets. In any case, an exchange of patient-related data between different users or partners is prevented.

3.2.6.5 ONCO AITI's provisions for Compliance to Standards

Since ONCO-AITI is a tool that could process patient-related data, we will adhere to domestic and European standards in this regard. An example of this is compliance with the EU-wide GPDR. The Medical University of Graz has its own compliance requirements that must be observed and adhered to when developing ONCO-AITI. In principle, as little data as possible, but as much data as necessary should be used. When working with patients' data related, even when pseudonymized data, every action is previously described in an ethics application and approved by an ethics committee. In general, the in-house data used to train the AI algorithm is not made available to any other partner. Rather, the goal is for the tool to be able to process data that all partners can choose independently without having to share it with other partners. In order to comply with the Joint Controllership Agreement (JCA) and the rules on personal data, the tool is developed in such a way that the tool developers do not have direct access to personal data of other partners. In the case of ONCO-AITI, there are no plans to exchange patient-related data with other partners.

3.2.7 ONCO-BIOBA

3.2.7.1 Introduction/Conceptual description

The cost of unFAIR "data is estimated to be €10.2 billion/year across sectors in the EU68. FAIRification of legacy data is not sustainable, given that data scientists spend an estimated 80% of their time on data curation and 20% on doing their job (i.e., analytics) (DG for Research and Innovation, 2010). The availability of high-quality data has direct financial benefits: a reduction in costly data duplication, increased effectiveness and accuracy of decision making, and increased quality of life and better health of citizens through preventive measures. In order to promote pioneering research in the field of integrated diagnostics across Europe, there is a

definite need for having for a given citizen/patient multiple, high-quality, standardized, privacy-preserved, open data. The role of ONCO-BIOBA is to provide an overview of the total body of AI-ready and reusable data cohorts and describe them in a federated setting, which will be available for the scientific community of ONCOSCREEN. In order to implement this role, a data catalogue will be developed that lists and describes the planned and existing data sets and cohorts. In principle, each partner can have a collection of samples created by them entered, which can be available for later research under defined conditions. This collection is described by uniform metadata, which can be made visible in ONCO-BIOBA. The description of such collections is provided in a harmonized form, which adheres to the FAIR principles. With the help of such a catalogue, interested parties such as research institutes, organizations and industry can obtain an overview of the existing databanks and all relevant information for contacting them.

3.2.7.2 Functionalities and actors

ONCO-BIOBA will provide rich and privacy-preserved metadata, as well as FAIR integrated-data to the scientific community. This data will describe available datasets and cohorts and will mainly be processed by the end-users GERCOR, TIMELEX and UFC, but can also be made accessible to other partners in the project.

3.2.7.3 Physical architecture

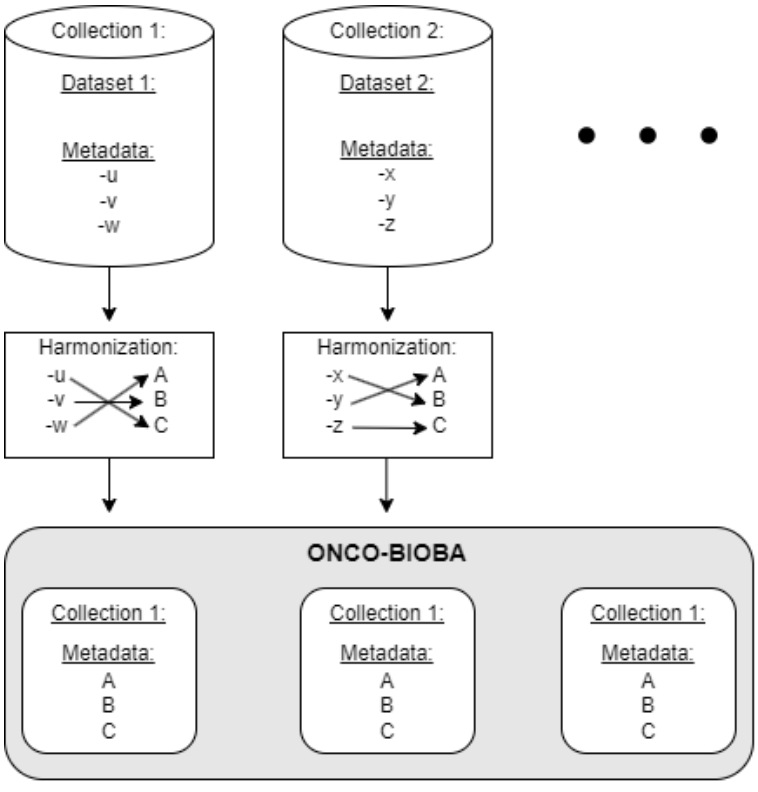

Figure 10 A graphical representation of ONCO-BIOBA's architecture

In ONCO-BIOBA various collections are described via metadata. These collections can be of diverse nature and differ greatly from one another. A provider can contain several collections, and many collections from different providers can arise over the course of the project. As diverse as the providers and their collections will be, the description of these collections can also be as diverse. In ONCO-BIOBA there will be a harmonization of the metadata for the description for standardization and clarity, to ensure the description of such a collection in a uniform and comparable way. In the following figure (Figure **10**), this harmonization is indicated. It is illustrated that a collection consists of the collected samples summarized in a dataset and the description of these samples, as well as the description of the entire collection. However, since the collections in ONCO-BIOBA are only described at an aggregated level, this tool no longer contains any data records or physical,

as well as digital datasets and no conclusions can be drawn about personal data. The key concept is the description of collections at an aggregate level.

3.2.7.4 ONCO BIOBA's provisions to Security

Data listed in this catalogue must comply with Austrian and EU-wide laws regarding the protection of personal rights. An example of this is compliance with the EU-wide GPDR (General Data Protection Regulation). It is important to note that, this tool only describes collections and cohorts at an aggregated level and therefore does not process any personal data.

In-house security aspects are guaranteed by the IT department of the Medical University of Graz as well as through access restrictions to the buildings (key cards) and access restrictions to the computers (password and admin rules).

3.2.7.5 ONCO BIOBA's provisions for Compliance to Standards

For the uniform description of collections on an aggregated level, ONCO-BIOBA will refer to appropriate and assessed standards. The aim is to unify and harmonize this description. Since the use of suitable standards will be determined in consultation with the users during the course of the project, it would be too early to list established standards at this point. However, the MIABIS standard (Merino-Martinez et al., 2016), which deals with the minimum information for describing biobanks, can be named as an example. We plan to adapt this standard for the goals in ONCO-BIOBA on the description of collections and to adapt it to the needs of the partners in the project.

3.2.8 ONCO-CAWA

3.2.8.1 Introduction/Conceptual description

ONCO-CAWA is a personalised mobile app for subjects' (patients or citizens) self-assessment, monitoring and citizen education/awareness of population. It is the subjects front-end facilitating monitoring, by enabling data collection at the subject end. Subjects use the mobile app to provide self-assessments in the form of questionnaires answered. Widgets offer access to data from other devices, as needed. ONCO-CAWA is providing two-way information flow, as subjects are also informed on their data and status, and they go through dialogues, increasing their education and awareness.

ONCO-CAWA is used to visualize information created by other ONCOSCREEN tools. More specifically the risk stratification engine ONCO-RISTE (getting risk assessments) and the ONCO-VOC (getting data from the VOC biomarkers pattern recognition software). Whether all such information will be made available to ONCO-CAWA via direct integrations to the tools, or following a centralized approach where ONCOSCREEN platform collects and redistributes information as needed, is still unclear.

The ONCO-CAWA mobile app is communicating with a backend system provided by Innovation Sprint for managing users and data. This backend system provides an API for the

distribution of data to the ONCOSCREEN systems. The API is extended to cover the data types of ONCOSCREEN and is hence part of ONCO-CAWA. The rest of the ONCOSCREEN systems access the data via the API. From their viewpoint, ONCO-CAWA is a type of composite sensor, providing its diverse range of measurements and self-reports.

ONCO-CAWA will require modifications to the Healthentia app, both in terms of appearance, but also in terms of supported measurements, devices, widgets, questionnaires and dialogues.

3.2.8.2 Functionalities and actors

The following functionalities are supported by ONCO-CAWA:

- Registration of subjects in the prospective study of ONCOSCREEN.
- Connection of activity trackers.
- Enabling of widgets for different type of data collected (e.g., physical activity, liquid consumption, nutrition, blood pressure, SPO2, etc.)
- Reception and answering of questionnaires.
- Ingestion of ONCO-RISTE and ONCO-VOC data.
- Visualization of all collected data about self.
- Running dialogues for information and awareness.
- Delivering collected data to ONCOSCREEN data lake.

The end-users are envisioned to be the patients/citizens.

3.2.8.3 Physical architecture

ONCO-CAWA is implementing the subjects' companion app for the ONCOSCREEN study participants. It consists of two modules: the backend and the mobile frontend. The backend is a collection of API endpoints and the associated services that:

- Manage studies and the investigators conducting them.
- Manage subjects and the data collected from them.
- Collect data from devices integrated via APIs.
- Support the operation of the mobile app.
- Offer the collected data to authorized 3rd parties.

It is implemented in C# using the .NET framework.

The mobile frontend implements the study participants' mobile companion app, offering:

- Consent management system.
- Subject's profile management.
- Data collection: Manual input in widgets, question/answering system, measurements via the SDKs of devices.
- Subject's information: Graphs and other data representations in the different widgets.
- Subject's education/coaching: A notification and a dialogue delivery system.

It is implemented as a React Native application, from which both the Android and iOS systems are generated.

3.2.8.4 ONCO-CAWA's provisions for Security

ONCO-CAWA is built on top of Innovation Sprint's main product, Healthentia. Healthentia is already a Medical Device, certified as Class I under EU's Medical Device Directive (MDD). Although there are no longer new MDD certifications, the validity of the existing certifications is extended for a period that covers the duration of ONCOSCREEN for all Class I medical devices that have already applied for a certification under the new Medical Device Regulation (MDR). This is indeed the case for Healthentia.

All functionalities that are involved in data collection, management and distribution are included in the certification. Only subject education/coaching is not already certified, but it is covered in the new application under MDR. Both MDD and MDR include the necessary security standards, with which Healthentia and thus ONCO-CAWA comply (see next section).

More specifically, ONCO-CAWA's authorization system is based on the Role Based Access Control (RBAC) component that authorize user's actions according to the role that the user is assigned. Roles in ONCO-CAWA define a group of rights/permissions that can be assigned to a user. The rights/permissions of the system indicate the operations that a user is allowed to perform while interacting with the platform as well as they define what the user can see on different screens of the application. API access is also controlled by the RBAC component, thus securing the interconnections of the system.

3.2.8.5 ONCO-CAWA's provisions for Compliance to Standards

ONCO-CAWA is built on top of a product of Innovation Sprint that is a certified medical device, thus as a software system (and the processes followed by Innovation Sprint in building ONCO-CAWA) complies with the requirements of the international standards ISO9001:2015, ISO 13485: 2016 and ISO 27001: 2013, and the following applicable regulatory requirements:

- 21CFR820 and related US laws
- Medical Device regulation 2017/745 EU, as amended
- Guideline for Good Clinical Practice (GCP) E6(R2)
- ISO 9001:2015 ISO 9001:2015 Quality management systems
- ISO 27001:2013 Information security management systems
- BS 10012 Personal Information Management (GDPR)
- ISO 13485:2016 Medical devices - Quality management systems —Requirements for regulatory purposes
- ISO 14971:2019 Medical devices - Application of risk management to medical devices
- IEC 62366-1:2015+A1:2016 Application of usability engineering to medical devices MEDDEV 2.4/1 Classification of Medical Devices
- IEC 62304+A1:2015 Medical device software - Software life cycle processes
- EN 1041: 2008/A1: 2013 Information supplied by the manufacturer

- EN ISO 15223-1: 2016 Medical devices - Symbols to be used with medical device labels, labelling and information to be supplied - Part 1: General requirements

3.2.9 Knowledge model and data harmonisation tool

3.2.9.1 Introduction/Conceptual description

Knowledge Model

Harmonizing disparate relational database schemas involves creating a common understanding of the data structures and relationships across different databases that contain retrospective CRC data. This can be achieved using database specific technical metadata and a unifying knowledge model. A knowledge model for federating medical relational databases should encompass the necessary information to understand the content, structure, relationships, policies, and constraints within the distributed databases. This knowledge model serves as a foundational resource for coordinating queries, ensuring data privacy, and maintaining compliance with regulatory standards.

The knowledge model (Figure **11**) conveys a typical disease profile for colorectal cancer and includes a comprehensive set of data that provides a detailed description of the disease, its characteristics, and its impact on the patient. This profile is essential for diagnosis, treatment planning, and monitoring and is an essential part of the Knowledge Model and described briefly below. More details about the disease profile can be found in deliverable D2.1.

- Patient Demographics: typical attributes include Age, Gender, Ethnicity, and Country of Residence, and Personal medical history
- Genetic & Hereditary CRC Syndromes: typical attributes include Family history of hereditary CRC syndromes, and Polyposis Syndromes.
- Lifestyle and Behavioural Data: typical attributes include Diet (fibre intake, consumption of red and processed meats, etc), Physical activity, Obesity, Smoking, Alcohol consumption, Inflammatory Bowel Diseases, Screening and Prevention.
- Environmental Data: typical attributes include Exposure to Carcinogens (air pollution, hazardous chemicals, radiation), Geographic Data, Occupational Factors (location and occupation of an individual)
- Psychological Well-being: typical attributes include Emotional distress, Anxiety level, Depression Level, Stress Level, Coping Mechanisms, Social support.
- Clinical Data: typical attributes include symptoms such as rectal bleeding, changes in bowel habits, abdominal pain, weight loss, iron deficiency anaemia, fatigue, Duration of symptoms, Medical Conditions and Medications, Biopsy and pathology reports, Tumour staging and Surgical reports.

- Laboratory Data: typical attributes include quantitative and qualitative information about various biomarkers and substances related to colorectal cancer such as Tumour Markers, Blood Chemistry and Haematology Tests, Genetic and Molecular Tests, Tissue DNA Analysis and Diagnostic Data.
- CRC Medications and Treatment Data: typical attributes include data related to cancer treatments, such as chemotherapy regimens, radiation therapy plans, targeted therapies, and immunotherapies. They also include information about medications, including dosages and administration schedules.
- CRC Risk: typical attributes include CRC Risk Assessment, Genetic Counselling and Testing, Screening and Surveillance Recommendations, Lifestyle and Behavioural Interventions.

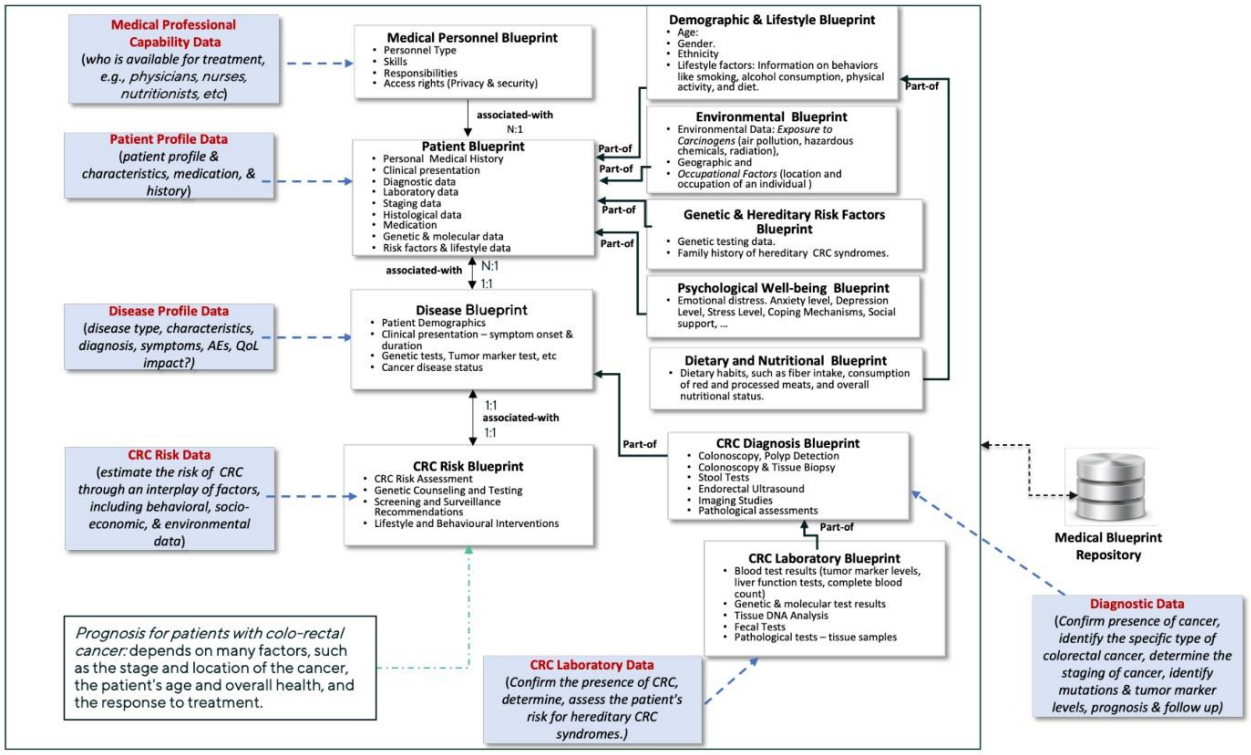

Figure 11 illustrates the CRC Digital knowledge model and its associated knowledge parts (called medical blueprint frames or simply blueprints). The frames in Figure 13 collectively provide a comprehensive understanding of the patient's colorectal cancer disease profile, and is crucial for early detection and effective management of colorectal cancer. This is facilitated by traversing the interconnected blueprint types. This helps healthcare professionals and tools understand the content and context of the data, which is crucial for accurate analysis and decision-making.

Figure 12 illustrates the CRC digital knowledge model using a more formal representation scheme that employs a UML structure diagram. The "…" in the diagram reflects the extensibility feature of the blueprints model, which means that the model can be modularly extended to incorporate and interlink new concepts and relationships throughout the agile and iterative design and implementation of the CRC digital knowledge model. For example, as shown in Figure 12, "CancerMedicalHistory" is modelled as a sub-class of "MedicalHistory" super-class; more disease classes that has an "IsA" relationships with "MedicalHistory" can be easily defined and incorporated in the CRC digital knowledge model, such as "CardioVascularMedicalHistory".

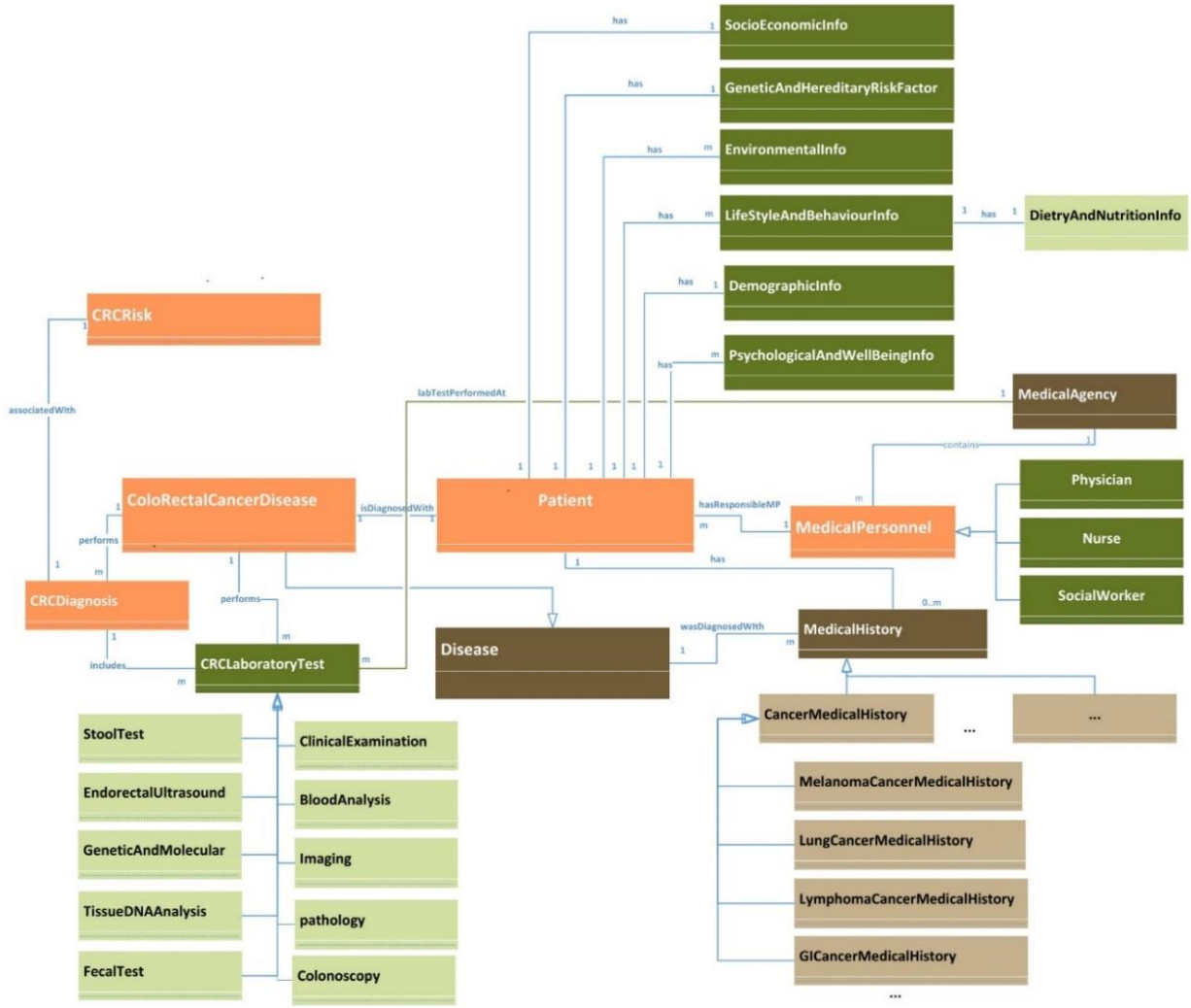

Figure 12 represents the informal knowledge model in Figure-13 using a UML structure diagram.

Data Harmonization/ Homogenization

To integrate disparate database schemas database technical meta-data and the unifying knowledge model are used by creating a layer of abstraction that allows different databases with varying structures to work together seamlessly. Such technical meta-data includes information about tables, columns, indexes, constraints, and any other relevant details in the disparate databases and is associated with the relevant blueprint entities in the unifying knowledge model so that they can be discovered in their respective databases.

3.2.9.2 Functionalities

Through the data lake and fusion engine tool that will be developed within ONCOSCREEN, several data, including patient measurements, patients' demographics, patients' behavioural data and data gathered from other ONCOSCREEN tools, will be collected. This variety of retrospective data will be fused with data related to genetic data or laboratory/diagnostic data to help estimate the risk of CRC through an interplay of factors, including behavioural, socio-economic, and environmental data, and diagnose patients with colorectal cancer based on many factors, such as the stage and location of the cancer, a patient's age and overall health, and the response to treatment. All such information is contained in the knowledge model.

Harmonizing and eventually federating retrospective data from relational databases using a hybrid federated/data-mesh approach involves decentralizing ownership, implementing domain-oriented data products, and fostering a collaborative culture across different medical domains. The ONCOSCREEN data-sharing architecture is designed as a federation of domain-oriented data products. Each data product is owned and managed by a specific medical domain, e.g., demographic, dietary, and nutritional data sources. These data products expose standardized interfaces for other domain-specific data sources to access and use their data. This includes:

1. **Defining Data Products**: This entails breaking down the retrospective data into domain-oriented data products. Each data product should encapsulate the data, logic, and storage related to a specific medical domain. For example, we can consider creating separate data products for Patient Demographic Data, Clinical History and Physical Examination Genetic and Hereditary Risk Factors, Lifestyle and Behavioural Factors, Dietary and Nutritional Factors etc.
2. **Decentralizing Data Ownership**: assign ownership of each data product to the respective business domain. This decentralization promotes autonomy and accountability, aligning with the data-mesh principles.
3. **Data Product APIs**: Develop APIs for each data product to provide standardized and controlled access to the retrospective data. These APIs serve as the interface for other domains or applications to interact with the data products.

The steps to homogenize/harmonize disparate relational database schemas include:

1. **Understand Existing Schemas**: analyse the schemas of the disparate relational databases to understand their content, structures, entities, relationships, and data types.
2. **Metadata Extraction**: extract data content metadata from each retrospective database.
3. **Define a Unifying Knowledge Model**: design a common unifying knowledge that represents the data content, characteristics and essential entities in a diversity of retrospective databases. This model will serve as the basis of unifying disparate retrospective database schemas.

4. **Create a Knowledge Model Repository**: establish a metadata repository or catalogue to store and manage the extracted knowledge model metadata. This repository will serve as a central source of information for the unifying knowledge model entities.
5. **Incremental Harmonization/Homogenization**: an incremental approach to homogenize data, allows for the phased integration of different databases. This approach can be more manageable and less disruptive to ongoing operations.

3.2.9.3 Physical architecture

Decomposing and dispatching queries to relational databases based on metadata and a unifying knowledge model involves understanding the structure of the databases, utilizing metadata to guide query decomposition, and optimizing the distribution of queries. This is shown in Figure **13**. This figure illustrates ***a unifying virtual retrospective data layer*** which combines all diverse, distributed data sources, and enables logically centralized access, combination, and provision all retrospective data to meet requirements through the use of the common knowledge model. The building blocks and steps in this architecture which follows a hybrid federated data mesh approach are as follows:

1. **Define the Database Structure and Relationships Meta-data Model**: create a model that represents the structure and relationships of the databases to be federated on the basis of data products in 3.2.9.2. This model should provide a unified view of the data and serve as a basis for query decomposition.
2. **Metadata Extraction**: Extract metadata from each relational database. This includes information about tables, columns, indexes, constraints, and any other relevant details. Populate the database structure and relationships meta-data model with this metadata.
3. **Query Decomposition Rules**: Develop query decomposition rules based on the database structure and relationships model. Define how queries against this unifying model can be decomposed into subqueries that can be executed against individual databases.
4. **Routing Logic**: Develop routing logic to dispatch subqueries to the appropriate relational databases. Use the metadata in the to det database structure and relationships model ermine which databases contain the relevant data for each part of the query.

5. **Federated Query Processing Engine**: Implement a federated query engine that can understand and process queries against the knowledge model and database structure and relationships model. This engine should use the query decomposition rules and

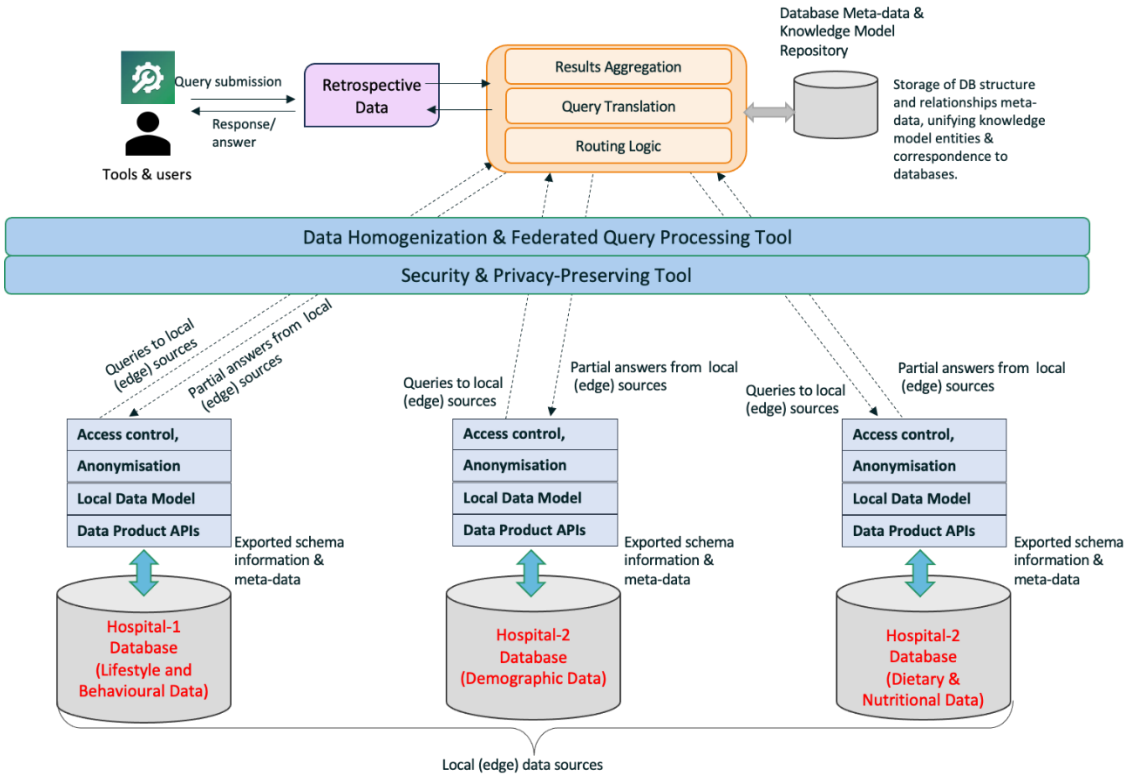

Figure 13 A Graphical representation of the Hybrid Federated Data-Mesh Management Approach.

routing logic to efficiently execute subqueries across the relevant databases.

6. **Query Mapping**: Develop mechanisms for mapping queries written against the knowledge model into queries that are compatible with the underlying databases. This may involve translating SQL syntax using the specific details of the Database Structure & Relationships Meta-Data Model.
7. **Security and Privacy Preserving Tool**: Integrate security measures to ensure that query dispatching adheres to access controls and security policies. Enforce access controls based on the user's permissions and the security requirements of each database.

3.2.9.4 Data flows

Below we present an activity diagram for the CRC Data Homogenization and Federated Processing tool with a general representation of the flow of activities, data, and interactions within the system.

Figure **14** represents a UML activity diagram that captures the behaviour of the ONCOSCREEN data homogenization and federated querying tools, and its interactions with ONCOSCREEN tools & users and (edge) local data sources (as architecturally described in Figure **13** of section 3.2.9.3). An activity diagram is an important behavioural diagram in UML to describe dynamic

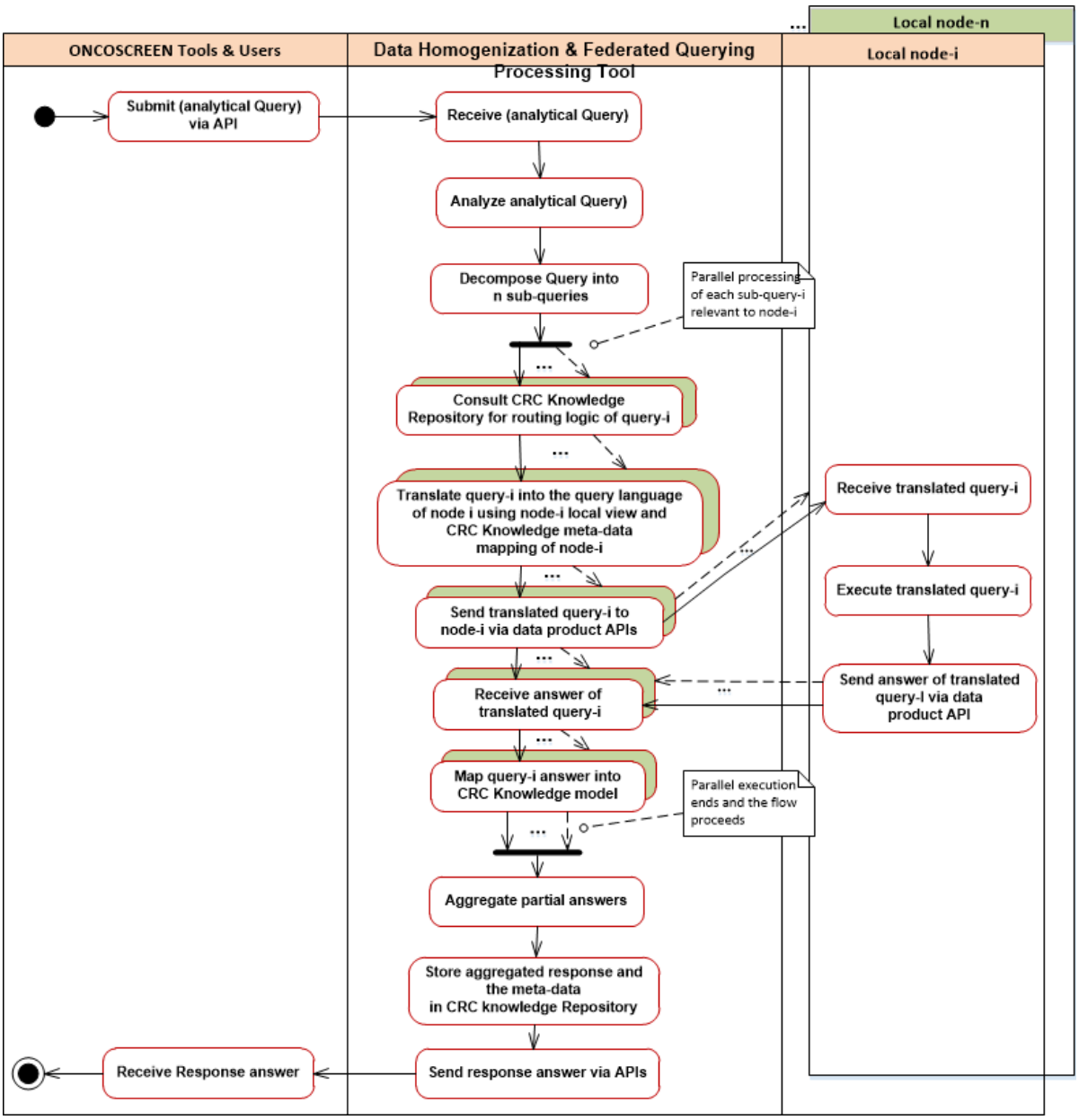

Figure 14 Activity diagram capturing the behaviour, data flow and interactions of the CRC Data Homogenization and Federated processing tool

aspects of the system. Activity diagrams are graphical representations of workflows of stepwise activities and actions with support for choice, iteration and concurrency.

3.2.9.5 Security by Design Approach

Security and Access Control measures need to be implemented on top of the knowledge model to ensure that data access is controlled and follows the security policies of each source database. Access controls based on user roles and permissions need to be also applied. Access control policies need to be implemented within the knowledge model. We specify who can access what data and under what conditions. Consider roles, permissions, and contextual factors to tailor access controls.

For privacy reasons, the prospective data will not be stored in a central database as it is the case with the traditional Extract, transform, and load (ETL) approach. Rather it will be accessed and processed in the data sources where it belongs.

Privacy: Should include

- **Access Controls**: Technical database-specific metadata can specify access controls and permissions for each data element. This ensures that only authorized individuals or systems can access sensitive information, helping to protect patient privacy in healthcare data, for example.
- **Data Anonymization**: Technical database-specific metadata can include information about data anonymization techniques and policies. It can specify which data elements require anonymization and provide guidelines for protecting individuals' identities while still allowing meaningful analysis.
- **Data Sensitivity Labels**: Metadata can tag data elements with sensitivity labels, indicating whether the data contains personal, sensitive, or confidential information. This labelling helps enforce privacy policies and ensures that data is handled appropriately.

Security: Should include

- **Encryption**: Metadata can describe the encryption mechanisms used to protect data during transfer and storage. This information is critical in ensuring that data remains secure throughout its lifecycle.
- **Audit and Compliance**: Metadata can store audit information, including logs of who accessed the data, when, and for what purpose. This supports security auditing and compliance with data protection regulations.

3.2.9.6 Compliance by Design to Standards

To develop the Knowledge model and achieve seamless data integration standards that help ensure that data from different sources and systems can be exchanged, shared, and used effectively. These include:

SNOMED CT (Systematized Nomenclature of Medicine - Clinical Terms): is a comprehensive clinical terminology standard that provides a common language for healthcare data (Donnelly, 2006). It is used to encode clinical concepts, making it easier to standardize and exchange medical data.

Fast Healthcare Interoperability Resources (FHIR): is designed to facilitate the exchange of healthcare information in a structured and standardized format. FHIR is designed to be highly flexible and extensible (Bender & Sartipi, 2013), allowing healthcare organizations and systems to create custom profiles and extensions for specific use cases, including colorectal cancer management. By adhering to FHIR standards, the knowledge model can better exchange and integrate data related to colorectal cancer in several ways including patient records, observations and laboratory results; clinical documents that include colonoscopy reports,

pathology reports, and other clinical notes related to the diagnosis and management of colorectal cancer; medications and treatment plans, and genomic data.

3.2.10 Privacy preservation tool

3.2.10.1 Conceptual description

KGEN is a meta-heuristic approach designed for anonymizing big datasets. The core functionality of KGEN lies in its application of a genetic algorithm to anonymize datasets, providing a solution for privacy-preserving data management. To meet privacy standards, KGEN uses k-anonymity criteria to assess the solutions it generates. This ensures that the resulting dataset is GDPR compliance, guaranteeing privacy.

One key aspect of KGEN is its integration within an API framework. This choice makes KGEN more accessible, enabling users to seamlessly incorporate its anonymization capabilities into their existing workflows. By leveraging this API platform, KGEN facilitates user interaction with the anonymization process in a straightforward manner.

Initiating from a dataset and an accompanying metadata file detailing attribute type, KGEN autonomously determines the anonymization rules to be applied. The output is an anonymized dataset, providing users with a secure and privacy-aware version of their original data.

3.2.10.2 Functionalities

KGEN's functionalities are tailored to meet the diverse needs of end-users seeking to anonymize datasets, particularly in CSV or JSON format.

KGEN empowers end-users to anonymize their datasets through a user-friendly API. This functionality ensures that integration into diverse workflows is seamless and straightforward. Users can interact with KGEN's anonymization capabilities in real-time, facilitating the dynamic incorporation of privacy measures into their data management processes.

3.2.10.3 Physical Architecture

KGEN is structured to efficiently process user inputs, implement anonymization strategies, and deliver anonymized datasets. The architecture (see Figure **15**) consists of several key components:

1. **User Input:** the user provides input data and metadata to the platform through the API, specifying the dataset and describing the attributes.
2. **Generalization Criteria Generation:** KGEN selects anonymization strategies based on the provided metadata.
3. **Solution Encoding:** The dataset is encoded to create a lattice of potential anonymization solutions. This step involves representing the dataset in a structured format that allows for easy manipulation and processing.

4. **Genetic Algorithm Solution Process:** KGEN employs a genetic algorithm to explore and optimize potential anonymization solutions. This involves iteratively refining the anonymization process to find the most effective solution.
5. **Solution Decoding:** The selected solution is decoded to create a JSON file containing the anonymized dataset. This involves translating the encoded solution back into a usable dataset format.
6. **Output to User:** The anonymized dataset, now in JSON format, is sent back to the end user through the API, completing the anonymization process.

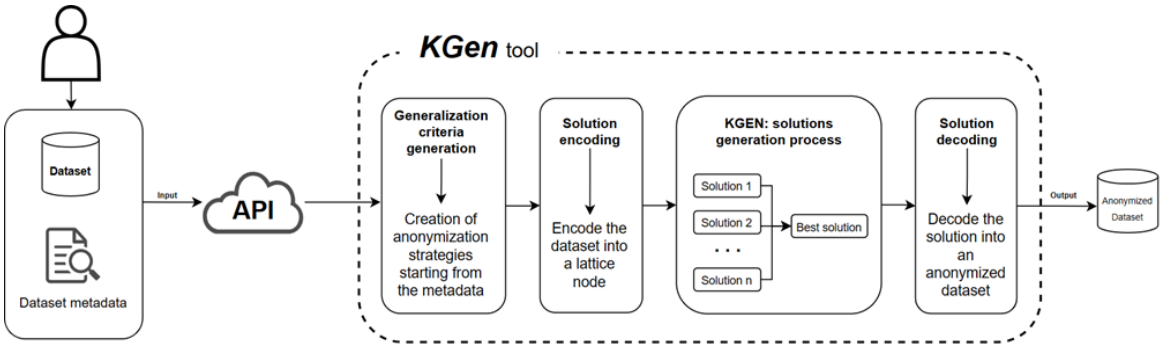

Figure 15 Physical architecture of the envisioned Privacy Preservation Tool

3.2.11 Data lake and fusion engine tool

3.2.11.1 Introduction/Conceptual description

This tool comprises of two separate modules, namely the data lake and the fusion engine. The multi-source data fusion refers to the process of integrating and analysing data from various sources to generate insights and identify patterns related to colon cancer risk and detection. The multi-source data fusion system aims to address several needs, including efficient data ingestion from multiple sources in various formats, storage and processing of structured, semi-structured, and unstructured data, large-scale data processing and analysis to generate insights and identify patterns, robust security measures to protect sensitive data, data integration from disparate sources, and archiving of data that is no longer needed while preserving its integrity and accessibility.

The system interacts with several other tools, such as the ONCO-RISTE engine, where it provides the source data for the analysis. ONCO-RISTE then uses clustering algorithms to identify dependencies and reveal correlations among various features related to colon cancer risk. In addition, it interacts with the clinical Decision Support System (cDSS) that is responsible for facilitating the work of medical experts and clinicians during their day-to-day activities when screening and treating citizens and patients for CRC.

The multi-source data fusion system aims to address several needs, including efficient data ingestion from multiple sources in various formats, storage and processing of structured, semi-structured, and unstructured data, large-scale data processing and analysis to generate insights and identify patterns, robust security measures to protect sensitive data, data

integration from disparate sources, and archiving of data that is no longer needed while preserving its integrity and accessibility.

In summary, data fusion refers to the combination of data from open sources mostly, with the scope to discover underlying factors for CRC, while data lake refers to a database that will be developed after M13.

3.2.11.2 Functionalities and actors

The envisioned functionalities of the multi-source data fusion are the following:

- Efficient ingestion of data from multiple sources in various formats and structures.
- Storage of a variety of data types, including structured, semi-structured, and unstructured data (data lake).
- Research open databases and identify possible underlying factors that may affect CRC.
- Robust security measures should be in place to protect sensitive data from unauthorized access or misuse.
- Support for data integration from disparate sources, including APIs, databases, and file systems in the data lake.
- Mechanism for archiving data that is no longer needed while still preserving its integrity and accessibility.
- Fusion of heterogeneous information from third-party sources.

3.2.11.3 Physical architecture

During this stage of the project, only the predictive and analytic aspect of the data fusion task has been developed. Data fusion has merged datasets from various open sources, to investigate the effect of environmental, lifestyle (nutrition, smoking, alcohol habits), comorbidities (obesity) and socioeconomic factors on CRC (Figure **16**).

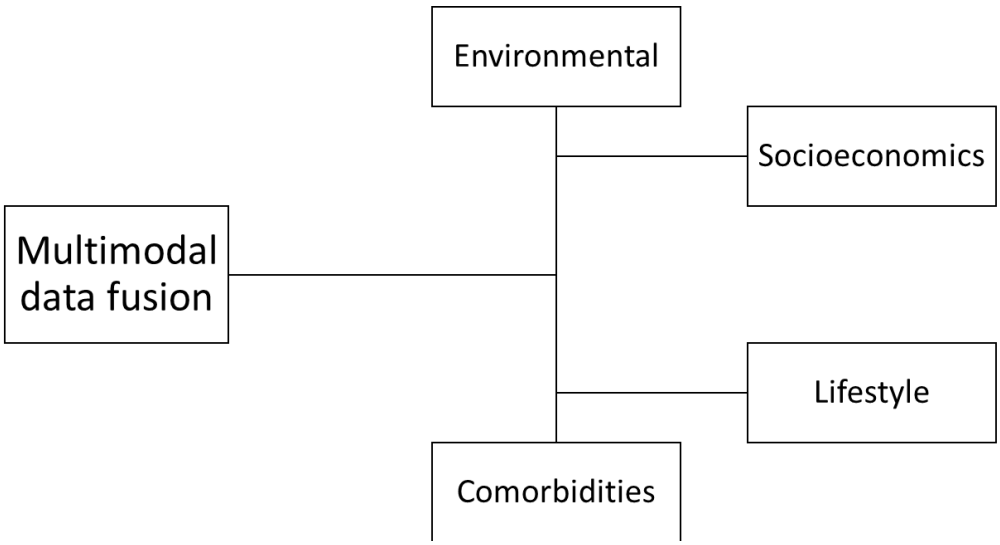

Figure 16 Types of risk factors from the merged dataset that are studied for their relation to CRC

3.2.11.4 Data flows

The data lake stands as a centralized repository between the virtual data harmonization layer and the rest of the ONCOSCREEN tools (see Figure **1**). The data lake and the virtual data harmonization layer are connected via REST API in order to receive anonymized clinical data, anonymized colonoscopy session video data and anonymized tissue image data. These data are then forwarded to the appropriate ONCOSCREEN tools in order to be analysed.

The data lake is connected via secure REST API with ONCO-AICO and ONCO-AITI to exchange colonoscopy session videos and tissue images respectively. The analysis results are also sent back to the data lake through the same REST API protocol. Also, the data lake is connected via REST API to ONCO-BIOBA.

The data lake is also connected via Kafka SSL with the rest of the ONCOSCREEN tools to exchange necessary data for the performed analyses. The data fusion tool is also connected via Kafka SSL with the rest of the ONCOSCREEN tools.

3.2.11.5 Security by Design Approach

The data lake introduced by the multimodal fusion tool will be a centralized repository dealing with sensitive user data. It is important to introduce access control and security measures to ensure safe exchange of sensitive information along the different ONCOSCREEN modules. Secure Sockets Layer (SSL) encryption protocol will be adopted for ensuring encrypted links between server and clients. Also, data anonymization techniques and access control to data through access restrictions will be introduced to deal with privacy issues.

3.2.11.6 Compliance by Design to Standards

It is important for the data lake to follow the SNOMED CT and the FHIR protocols to ensure that sensitive data from different sources and systems can be exchanged, shared, and used effectively. The SNOMED CT standard is used as a common language for healthcare data while the FHIR protocol is adopted to ensure the exchange of healthcare information in a structured and standardized format. The protocols are described in detail at the 3.2.9.6 subsection.

3.2.12 ONCO-RISTE

3.2.12.1 Introduction/Conceptual description

There are different and diverse population groups that have exposure to workplace carcinogens, live in air/water polluted areas, have a family history in cancer, genetic mutations in genes related to cancer pathways, un-healthy lifestyle behaviours, people that avoid examination etc. In addition, population groups come from different socio-economic status, educational backgrounds, different (and difficult-to-reach sometimes) geographical areas. EU is really interested to understand on whether there are such inequalities that play a role. An accurate cancer prevalence grouping based on evidence the above factors and novel risk stratification algorithm that could accurately estimate the risk of emerging CRC in those groups.

ONCO-RISTE is a semi-empirical stratification engine that identify dependencies and reveal correlations among a variety of features concerning the clustering of citizen/patients and their risk level classification. To do that, it utilizes different AI architectures to handle uncertainty and probabilistic empirical rules from clinical experts.

ONCO-RISTE is connected to the Data Lake and Data Fusion to collect heterogeneous information regarding behavioural, socio-economic and environmental factors; ONCO-CAWA for citizen/patient data; and to ONCO-CLIDE to provide the output of the stratification engine. This output will be a risk-level on a 5-level scale.

3.2.12.2 Functionalities

ONCO-RISTE will:

- Provide the risk level to three visualization tools such as ONCO-CAWA, ONCO-CLIDE and ONCO-EVIDA.
- Allow clinicians to provide input and modify the existing rules used for the engine to calculate the risk level.
- Update the risk level based on the input received. This could lead to the risk level to be decreased/increased based on changes in another tool, i.e. ONCO-CAWA.
- Be integrated with the aforementioned tools and the data fusion engine.
- Use the KGEN tool to preserve and secure all the patient data.

3.2.12.3 Physical architecture

ONCO-RISTE is one of the tools of ONCOSCREEN project. The interconnection between this tool and the rest of them is shown in Figure **19**.

ONCO-RISTE is connected via REST API with ONCO-CAWA tool, which provides information about the citizen/patient via questionnaires. Those questionnaires provide information

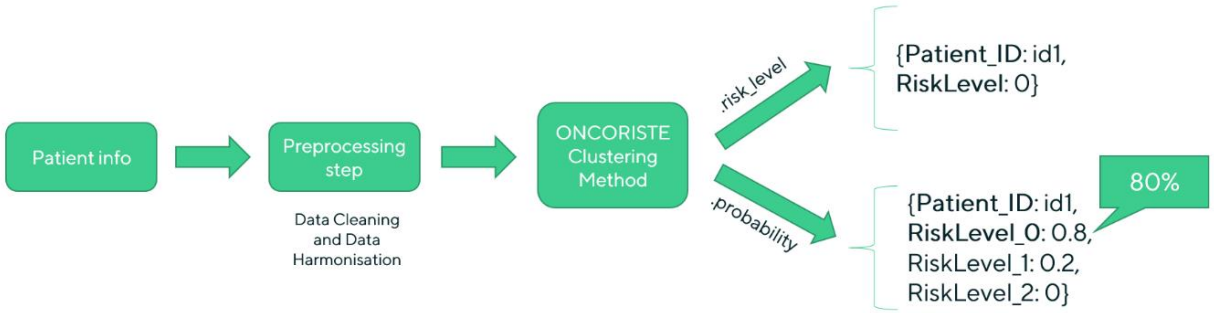

Figure 17 ONCO-RISTE fuzzy module workflow

regarding personal (gender, age, work status, etc), behavioural (smoking, alcohol consumption), and clinical factors among others, that is received and formatted by ONCO-RISTE backend to feed with it the fuzzy approach and calculate the risk level.

ONCO-RISTE is also connected via Kafka SSL with ONCO-CLIDE and Data Fusion tools. The output of the tool is sent to ONCO-CLIDE to be seen and checked by clinicians with all the

information regarding the patient. About Data Fusion, environmental information and demographic factors are added to the patient info in order to improve the quality and add more complete information to the one gathered to classify correctly the risk level.

How this classification works is shown in Figure **18**. Given the patient info along the rest of the information given by the Data Fusion, a pre-processing step is carried out to clean and harmonized the data before feeding the fuzzy clustering method. There are two different outputs: the first one provides only the risk level along with the patient id; the second one provides, again, the patient id, and also the probability of that patient to belong to different risk levels according to the provided information. This way, ONCO-RISTE is taking advantage of the

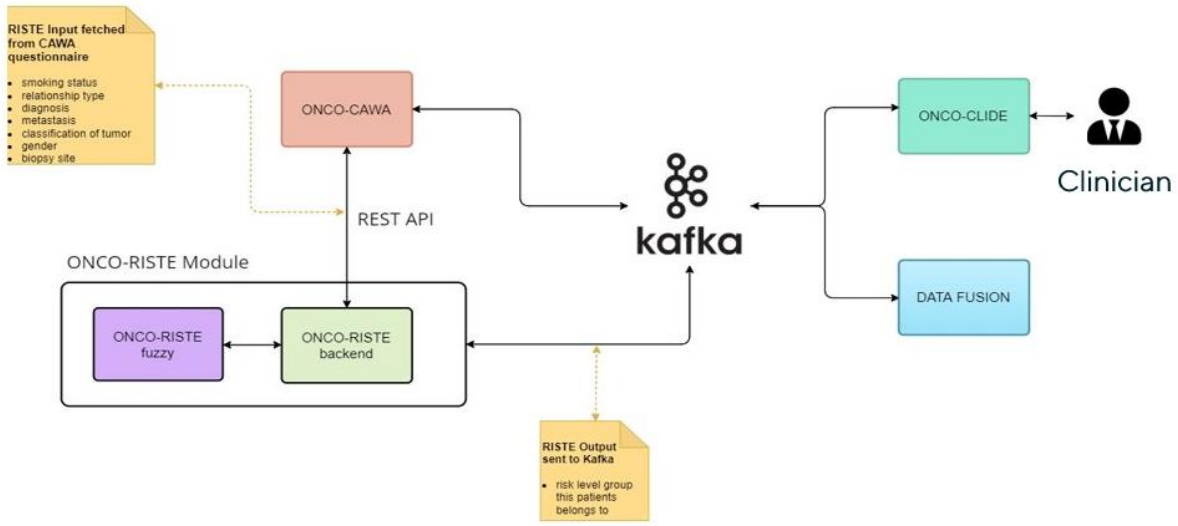

Figure 18 Conceptual diagram of ONCO-RISTE connections with other tools

fuzzy characteristics of the approach, providing also more information to the clinicians in order to decide the best way to proceed with the current information. The risk level showed by the first option is the highest probability risk calculated by the second one.

3.2.12.4 ONCO RISTE's provisions for Security

Regarding security, the open patient data that were used for the algorithm training and the calculation of the risk level in the different approaches is anonymized. ONCO-RISTE for the communication of the different data flows between ONCOSCREEN tools utilises the Kafka SSL, that will be used across the project towards a zero-trust framework ensuring security by design.

3.2.12.5 ONCO RISTE'S provisions for Compliance to Standards

ONCO-RISTE utilises standardised communication protocols with advanced security like SSL, while it is planned (as it will be depicted in the second iteration of the deliverable) that its data will follow the SNOMED and FHIR standards.

3.2.13 ONCO-CLIDE

3.2.13.1 Introduction/Conceptual description

ONCO-CLIDE (ONCOSCREEN Clinical Decisions Support System for CRC Integrated Diagnosis) represents an advancement in the field of clinical decision support systems (cDSS), specifically designed for colorectal cancer (CRC) diagnosis. Distinguishing itself from existing cDSS tools that focus primarily on treatment, ONCO-CLIDE aims to provide more precise and timely integrated diagnoses. It achieves this by incorporating a wide array of diagnostic data, including liquid and breath biopsies, stool tests, histopathological images, and colonoscopies, tailored to each patient's unique background (data from WP3). This approach contributes to current medical practice by combining diverse diagnostic information into integrated algorithms, thus enhancing the accuracy and comprehensiveness of CRC diagnosis.

3.2.13.2 Functionalities and actors

The envisioned functionalities of the tool are the following:

1. Web Interface:

- User-friendly dashboard.
- Integration of diagnostic data.
- Secure login for patient confidentiality.
- Intuitive navigation with access to various modules.

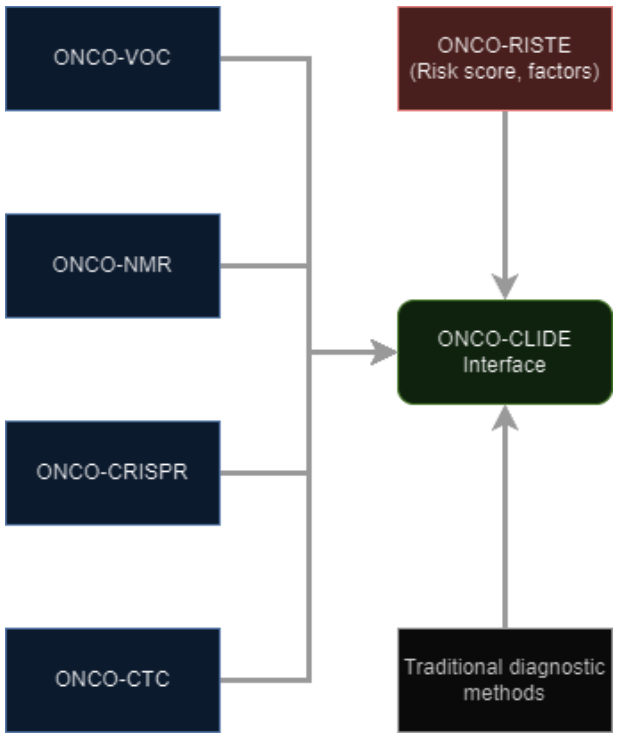

Figure 19 ONCO-CLIDE's interface and the connection with other tools

2. Suggestion and Classification Models:

- Machine learning algorithms for diagnostic suggestions.
- Prediction of cancer stages.

3. Virtual Tumour Board:

- Platform for medical professionals to discuss patient cases.
- Collaborative approach for cancer care.
- User-friendly interface for ease of access and regular use.

During this phase of the project, our efforts have been concentrated on developing the diagnostics frontend and the central backend. This stage has been crucial in laying the foundational structure for the cDSS. The other functionalities (2,3) are still in the planning stages since they need to be developed later in the project.

The envisioned end users of the tool are CRC clinicians and the laboratory personnel, who will be responsible for entering results related to ONCO-NMR and ONCO-CTC.

3.2.13.3 Physical architecture

ONCO-CLIDE can be thought of as being divided into three parts/modules. The first one is the web interface that displays the data from the various project sources as shown in the below diagram. The second one is the suggestion models that will help during the diagnostic process and the classification model that will output the stage of cancer. Finally, the third part is the virtual tumour board where doctors will be able to communicate on a specific case. The above are still in a prototype stage and changes might be made during the course of the project.

A Kafka server provided by EXUS manages the data streams from the ONCO diagnostic tools. It organizes the data into topics for systematic processing and delivery. The server prepares the data to be compatible with the ONCO-CLIDE Interface, ensuring it is ready for presentation.

The ONCO-CLIDE Interface (see Figure **19**), at the receiving end of this pipeline, subscribes to these Kafka topics. It retrieves the processed diagnostic data and risk assessments, synthesizing them into a coherent, interactive dashboard. The interface is designed to not only present this information clearly but also to allow for the incorporation of additional data from traditional diagnostic methods, thereby painting a complete picture of the patient's diagnostic landscape.

The data flows from the traditional diagnostic methods is not yet finalized. It is likely to be differentiated depending on each clinical partner.

3.2.13.4 ONCO CLIDE's provisions for Security

All communications with the ONCO-CLIDE platform will be encrypted. Regarding the medical data gathered in the project due to their sensitivity the idea is that no database will be used for the tool itself, but they will be fetched from a central ONCOSCREEN database as required.

3.2.13.5 ONCO CLIDE's provisions for Compliance to Standards

This section cannot be covered on the deliverable D4.1 as the tool is under development and the provisions for compliance to standards is yet to be determined.

3.2.14 ONCO-EVIDA

3.2.14.1 Introduction/Conceptual description

The intelligent analytics dashboard of ONCO-EVIDA is designed to address the needs of policy makers in health care as well as in public health and the environment & health sector by providing them with evidence-based recommendations for decision-making related to (1) effectiveness of CRC screening campaigns and to help focus them, e.g. on vulnerable groups or low participation rate groups (to increase the participation rate) and (2) (potential) environmental CRC risk factors and environment & health management (for example, to lower

exposure to (indirect) risk factors such as air pollution). The dashboard integrates data from multiple sources, including diagnostic tools (CRC screening data), retrospective data, open databases, socio-economic data, and air quality data, to provide a comprehensive view of the data and enable policy makers to make informed decisions. It will also be able to map out the queried risk factors on a local and/or regional level. Consulting these maps enables policy makers to target CRC screening or screening promotion more efficiently.

Furthermore, the tool is envisioned to have an intuitive and interactive design that enables end-users to explore and analyse data easily. It should provide real-time processing and display of data to enable policy makers to react to changes in data quickly. Users will be able to customize the layout and design of the dashboard to suit their specific needs and preferences.

Access control and user authentication will be implemented to secure underlying data and prevent unauthorized access. The dashboard will allow users to filter and stratify data based on various criteria to enable them to focus on relevant data. It will also enable users to aggregate data based on different metrics and dimensions to identify trends and patterns in the data.

The intelligent analytics dashboard will provide alerts and notifications based on predefined criteria to notify policy makers of important changes in the data. It will also allow users to export data and visualizations for further analysis or sharing with others.

The policy makers in the environmental health part of public health will use ONCO-EVIDA to find potential risk factors by mapping factors like pollution or socio-economic status onto CRC screening and prevalence data. Once risk factors have been identified, ONCO-EVIDA can be used to support environment and health policy makers to reduce those risk factors.

3.2.14.2 Functionalities and actors

The end users for ONCO-EVIDA will be policymakers such as officials of the various Ministries of (Public) Health of EU member states. They can expect the following functionalities of the tool (see Figure **20**):

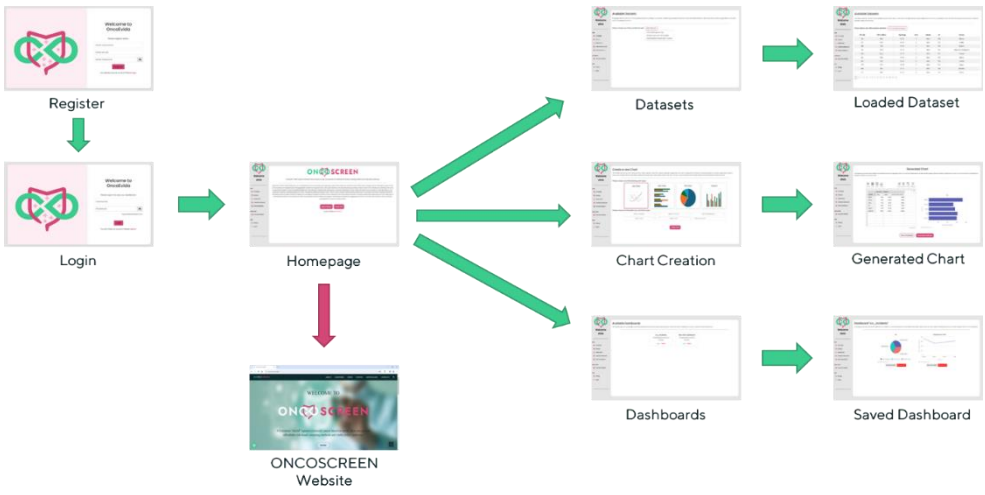

Figure 20 Provisional layout of ONCO-EVIDA web application

- Data integration from multiple sources

- Real-time data processing and display
- Intuitive and interactive visualizations
- Customizable layout and design
- Access control and user authentication
- Data filtering based on various criteria
- Data aggregation based on different metrics and dimensions
- Alerts and notifications based on predefined criteria
- Exporting data and visualizations for further analysis or sharing with others
- Integration with machine learning algorithms for advanced analysis and insights

3.2.14.3 Physical architecture

ONCO-EVIDA is envisioned to be a web application. Below, we present a graphical representation of its functionalities and components.

3.2.14.4 ONCO EVIDA's provisions for Security

As a web application, ONCO-EVIDA will implement a secure communication protocol by incorporating the HTTPS (Hypertext Transfer Protocol Secure) protocol. This ensures that all data transmitted between users and the ONCO-EVIDA server is encrypted and secure. HTTPS employs SSL/TLS (Secure Sockets Layer/Transport Layer Security) protocols to provide a secure connection, preventing unauthorized access and safeguarding sensitive information exchanged during user interactions with the application. This security measure enhances user privacy, protects against potential cyber threats, and establishes a trustable environment for users interacting with ONCO-EVIDA.

Creating a password-secured user account is necessary to access ONCO-EVIDA's features. While the dashboards created with the tool can be saved and exported, the underlying data cannot. All the security-related dependencies are aggregated by the Spring Security framework. Using this framework provides comprehensive support for authentication, authorization, and protection against common exploits.

3.2.14.5 ONCO EVIDA's provisions for Compliance to Standards

This section cannot be covered on the deliverable D4.1 as the tool is under development and the provisions for compliance to standards is yet to be determined.

4 Data Flows and Integration

Data flows are a key element to understand how information moves within ONCOSCREEN platform. This section breaks down how data travels from one point to another, explaining the steps and changes it undergoes, as depicted in Figure (**21**).

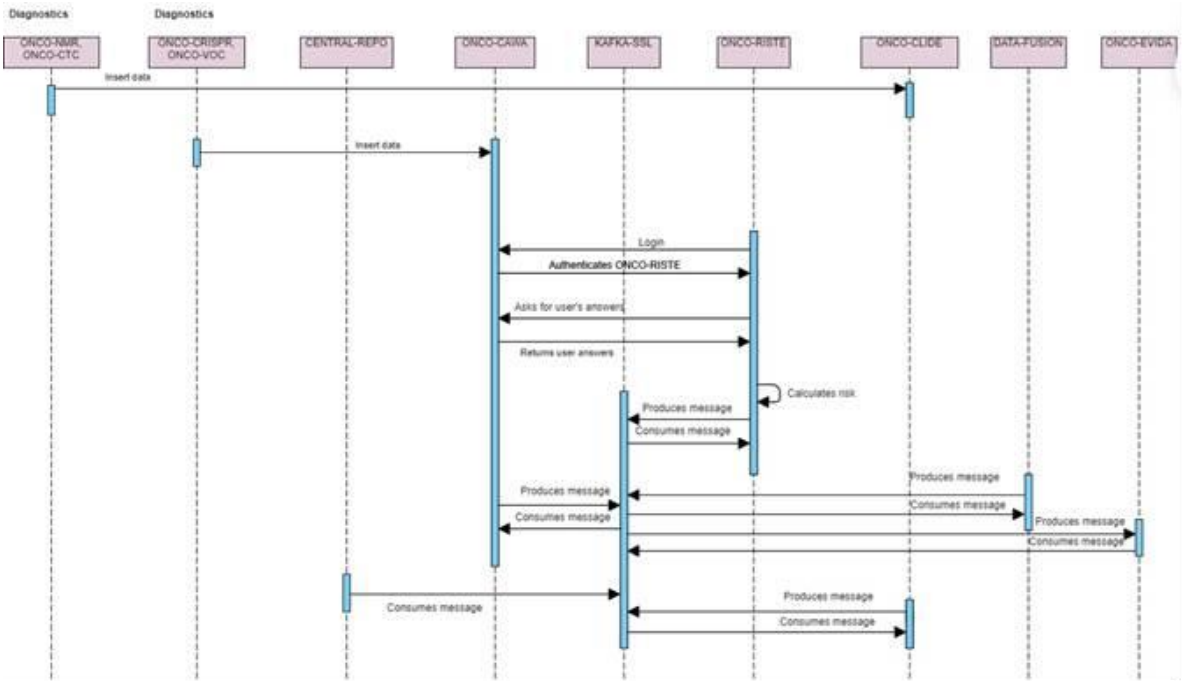

Figure 21 Data flows within ONCOSCREEN ecosystem

4.1 Integration

To facilitate the integration of all tools, EXUS deployed Kafka SSL. Apache Kafka[2] is a distributed streaming platform designed for building real-time data pipelines and streaming applications. Developed by the Apache Software Foundation, Kafka provides a scalable, fault-tolerant, and high-throughput framework for handling data streams. It allows seamless integration between various systems, enabling the efficient ingestion, storage, processing, and retrieval of data in a real-time or near-real-time manner. Kafka's architecture is based on a publish-subscribe model, where producers publish data to topics, and consumers subscribe to those topics to process the information. Widely adopted in enterprises for its reliability and performance, Kafka plays a crucial role in scenarios such as event sourcing, log aggregation, and building reactive applications. Kafka SSL, or Secure Sockets Layer, is a security protocol implemented in Apache Kafka to establish a secure communication channel between Kafka brokers and clients. It ensures the confidentiality and integrity of data exchanged within the Kafka ecosystem. By encrypting the data during transmission, Kafka SSL safeguards against unauthorized access and potential

[2] https://kafka.apache.org

eavesdropping. This security feature utilizes cryptographic certificates, including public and private key pairs, to authenticate and secure the communication between producers, consumers, and Kafka brokers. Overall, Kafka SSL is a fundamental component for enhancing the security posture of Kafka clusters, making it a reliable choice for organizations prioritizing data protection in their distributed streaming environments.

4.1.1 Kafka SSL deployment

In our search for a sophisticated and efficient infrastructure deployment model, we have designed and implemented a robust system utilizing Terraform for infrastructure provisioning, Ansible[3] for configuration management, and GitHub Actions for automated deployment. These technologies form an end-to-end pipeline that ensures consistency, repeatability, and reliability in our infrastructure.

Terraform for Infrastructure Provisioning

Terraform[4] is an Infrastructure as Code (IaC) tool that allows us to define and provision infrastructure in a declarative manner. Terraform scripts describe the desired infrastructure state, including server specifications, networking configurations, and any necessary resources. With a simple command, Terraform provisions and manages the infrastructure, ensuring consistency across environments.

Ansible for Configuration Management

Ansible, the chosen configuration management tool, plays an important role in maintaining and configuring our server. Utilizing YAML-based playbooks, Ansible succinctly describes the desired state of our systems, encompassing everything from package installations to intricate application setups. Seamless connectivity to provisioned servers ensures Ansible's ability to enforce configurations consistently. This approach not only enhances system manageability but also establishes a comprehensive audit trail, vital for compliance and troubleshooting.

GitHub Actions: Streamlined CI/CD for Automated Deployment

GitHub Actions serves as the bases in our automated deployment process, providing a continuous integration and delivery (CI/CD) framework integrated seamlessly into the GitHub repository. The workflows respond to specific triggers, such as code pushes or pull requests. This orchestrates a precise sequence of operations wherein Terraform provisions or updates infrastructure, Ansible configures servers, and our services are deployed—all without human intervention. The integration of GitHub Actions makes the development pipeline agile and reliable.

Summarizing, the steps we followed to deploy Kafka SSL (see also Figure **22**) were the above:

[3] https://www.ansible.com

[4] https://www.terraform.io

1. Developer commits the code to the GitHub repository.
2. GitHub Actions automatically triggers workflows in response to defined events.
3. Terraform provisions or updates infrastructure, adhering to defined specifications.
4. Ansible configures servers based on playbooks
5. Automated deployment of service

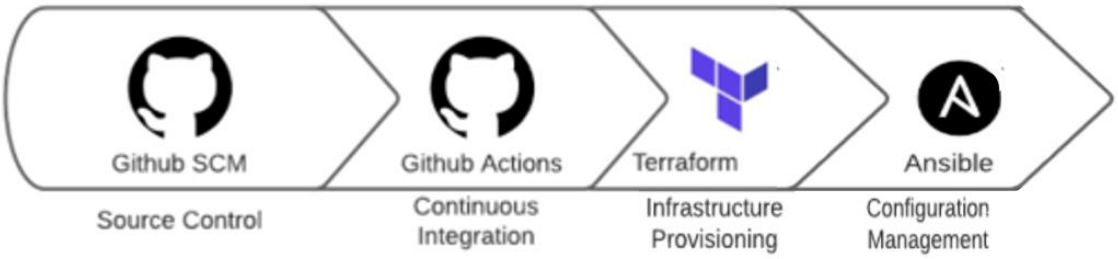

Figure 22 Steps for the deployment of Kafka SSL

5 Security by design

ONCOSCREEN has been designed to incorporate a comprehensive set of security provisions to safeguard its operations ultimately pursuing a zero-trust framework. Such a framework ensures a comprehensive and meticulous approach to data protection with its main concept being the "never trust, always verify" approach, which means that users and devices should not be trusted by default, even if they are connected to a permissioned network. Below we discuss the main security features that ONCOSCREEN incorporates towards the safe transfer of data within its framework.

All web applications use an HTTPS (i.e., ONCO-EVIDA, ONCO-CLIDE, ONCO-AICO, ONCO-AITI, ONCO-BIOBA) that ensures secure communication between users and the platform, encrypting data transmission to prevent unauthorized access. HTTPS employs SSL/TLS (Secure Sockets Layer/Transport Layer Security) protocols to provide a secure connection, preventing unauthorized access and safeguarding sensitive information exchanged during user interactions with the application. The inclusion of authenticator protocols adds an additional layer of security, verifying the identity of users to prevent unauthorized access and enhance overall system integrity.

The overall architecture of the project is envisioned to be hybrid, involving a federated database (transferring meta data only) and a centralised repository. The adoption of a federated database is a key security measure, allowing aggregation, analysis, and AI model training to be conducted within the data source (e.g., the data never leaves its source). This decentralized approach minimizes the need for centralized data transfers, reducing the risk of data exposure and enhancing privacy and security. On top of that, a dedicated privacy preservation tool will operate in conjunction with the federated database module for further privacy preservation.

Finally, the implementation of the Kafka SSL protocol enhances the security of data streaming within ONCOSCREEN. This security feature ensures secure communication and data exchange in real-time, preventing potential vulnerabilities during the streaming process. It utilizes cryptographic certificates, including public and private key pairs, to authenticate and secure the communication between producers, consumers, and Kafka brokers. Overall, Kafka SSL is a fundamental component for enhancing the security posture of Kafka clusters, making it a reliable choice for organizations prioritizing data protection in their distributed streaming environments. Additionally, the use of secure REST APIs (Application Programming Interfaces) further fortifies data interactions by incorporating security measures at the interface level.

6 Compliance to standards by design

ONCOSCREEN has been strategically designed to address compliance requirements, particularly by integrating the Fast Healthcare Interoperability Resources (FHIR) standards and Systematized Nomenclature of Medicine Clinical Terms (SNOMED CT) into its centralized repository and federated database (Donnelly, 2006). The adoption of FHIR standards (Bender & Sartipi, 2013) ensures a standardized and interoperable approach to health data exchange, facilitating seamless communication and integration with other healthcare systems and applications. This interoperability enhances data sharing, collaboration, and the overall efficiency of the ONCOSCREEN platform within the broader healthcare ecosystem. Incorporating SNOMED CT into the centralized repository further enhances the system's ability to represent and encode clinical information consistently. SNOMED CT, as a comprehensive clinical terminology, provides a standardized way of capturing and sharing clinical data, promoting clarity and consistency in medical records and communications. By aligning with these industry-standard frameworks, ONCOSCREEN not only ensures compliance with established norms in health informatics but also positions itself to effectively contribute to the evolving landscape of healthcare technologies. This approach enhances the platform's versatility, adaptability, and its ability to provide valuable insights and support in the complex domain of CRC research and patient care.

Finally, all ONCOSCREEN diagnostic tools adhere to the IVMDR (In Vitro Diagnostic Medical Devices Regulation) and CE (Conformité Européene) standards, ensuring compliance with the regulatory requirements set by the European Union for in vitro diagnostic medical devices

7 End-user and technical requirements

7.1 End User Requirements

End-User Requirements and Continuous Adaptation is an iterative process that will continue until M29 of ONCOSCREEN project. 28 partners, representing 4 stakeholder groups are involved in this task: (1) technical partners (creators and developers of the ONOSCREEN tools); (2) clinical end-users, involving associations of healthcare professionals, healthcare facilities, and academic institutions; (3) policy making end-users, involving National Ministries of Health, consultancy companies, and think-tanks; (4) patient end-users, involving patient organisations.

End-user requirement process is led by POLA LT, with a substantial coordination effort made by the YCE team. The co-creation process was cautiously designed not to require the collection of any personal data.

Between M1-M12 there were a total of 9 rounds of end-user requirement co-creation, namely:

(1) During the ONOSCREEN kick-off meeting on 13 January 2023 the first co-design workshops with end-user requirements and technical requirements took place. During face-to-face discussion, ONCOSCREEN partners agreed on a joint co-creation methodology developed by POLA LT and ICCS and partners in the end-user cohort have started to create end-user requirement by completing online spreadsheet file on ONCOSCREEN SharePoint, hosted by EXUS on a secure server.
(2) After kick-off technical partners have prepared technological description for each ONCOSCREEN tool, end-users have provided feedback / clarifications until 24 March 2023 (Step 1).
(3) All end-user partners have been asked by POLA LT to review technological descriptions and their roles and were able to express their interest in co-creating end-user requirements beyond the pre-assigned technologies until 14 April 2023 (Step 2).
(4) A series of end-user teleconferences (12 virtual meetings in total) took place on 19-21 April 2023 between technical partners and end-users, that were coordinated by POLA LT and YCE. During and after the meeting end-users have been asked to provide input to spreadsheet files that were co-created by POLA LT and ICCS and hosted on ONCOSCREEN SharePoint by EXUS (Step 3).
(5) End-users were asked to rank the requirements in the order of importance that by completing the prioritisation exercise in the assigned spreadsheet files until 27 April 2023 (Step 4).
(6) Technical partners have provided feedback on the ranking of end-user requirements and proposed list of end-user requirements to be developed during the series of virtual meetings, coordinated by ICSS and POLA LT during 3-5 May 2023 (Step 5).
(7) Series of alignment teleconferences were coordinated by POLA LT and YCE between 10-12 May 2023, where technical partners have presented their view on end-user

requirements, end-users were able to clarify some end-user requirements and the timeline for developing prototypes have been agreed (Step 6).

The iterative process, referenced in the points 2-7 is further detailed in the scheme below (see Figure **23**).

(8) During the ONCOSCREEN Management Board meetings progress of collecting end-user requirements has been reported by POLA LT and the progress of technical partners work has been reported by ICCS. After concluding that sufficient information has been collected, the Management Board has decided to organize LIT1 (Laboratory Integration Test) exercise. LIT1 was co-organized by ICCS and POLA LT and took place on 22 September 2023. During the meeting all technical partners presented the progress update on the technical tools they are developing for all ONCOSCREEN tasks, end-users were able to provide feedback and list additional end-user requirements until 11 October 2023, coordinated by POLA LT.

(9) Management Board has noted the substantial progress for development of ONCOSCREEN tools and has decided to organize LIT2 exercise on 5 December 2023. LIT2 was co-organized by ICCS and POLA LT. Technical partners have reported on their progress in developing ONCOSCREEN tools and have asked for additional clarifications from end-users regarding certain requirements. It was agreed that the next co-creation step will commence after end-users will be provided with the prototypes of ONCOSCREEN tools.

7.2 Translation of End User to technical Requirements

End-user requirements were finally translated to functional and non-functional requirements through bi-weekly technical steering teleconferences between the technical partners always with the aid of the coordinator of end users and POLA LT. In addition to this meeting, many ad-hoc meetings took place (e.g. between the corresponding technical partner and the responsible end users) with the goal of refining and/ or shaping specific technical solutions tailored to facilitate specific end user requirements. Functional requirements are critical, essential tasks for each system component aligned based on user needs. In contrast, non-functional requirements specify criteria that can be used to judge the operation of a system, rather than specific behaviours. In other words, while functional requirements specify what the system does, non-functional requirements define how the system operates. Finally, there are also general/horizontal requirements including aspects like security, compliance and reliability that can be seen as overarching requirements of the system. These requirements play a pivotal role since the project involves sensitive medical data and ensuring confidentiality and integrity is crucial, guarding against unauthorized access and cyber threats. Adhering to established standards not only meets regulatory requirements but also fosters trust among stakeholders.

Prioritizing these requirements underscores the project's commitment to ethical practices, legal compliance, and the success of this critical European Commission initiative in advancing medical technology and patient care.

7.3 Structure of the End User and Technical Requirements

Below we present the current list of User and Technical Requirements for each tool. In the cases of Data Lake and Fusion Engine Tool, Privacy Preservation Tool (PPT) and the Data

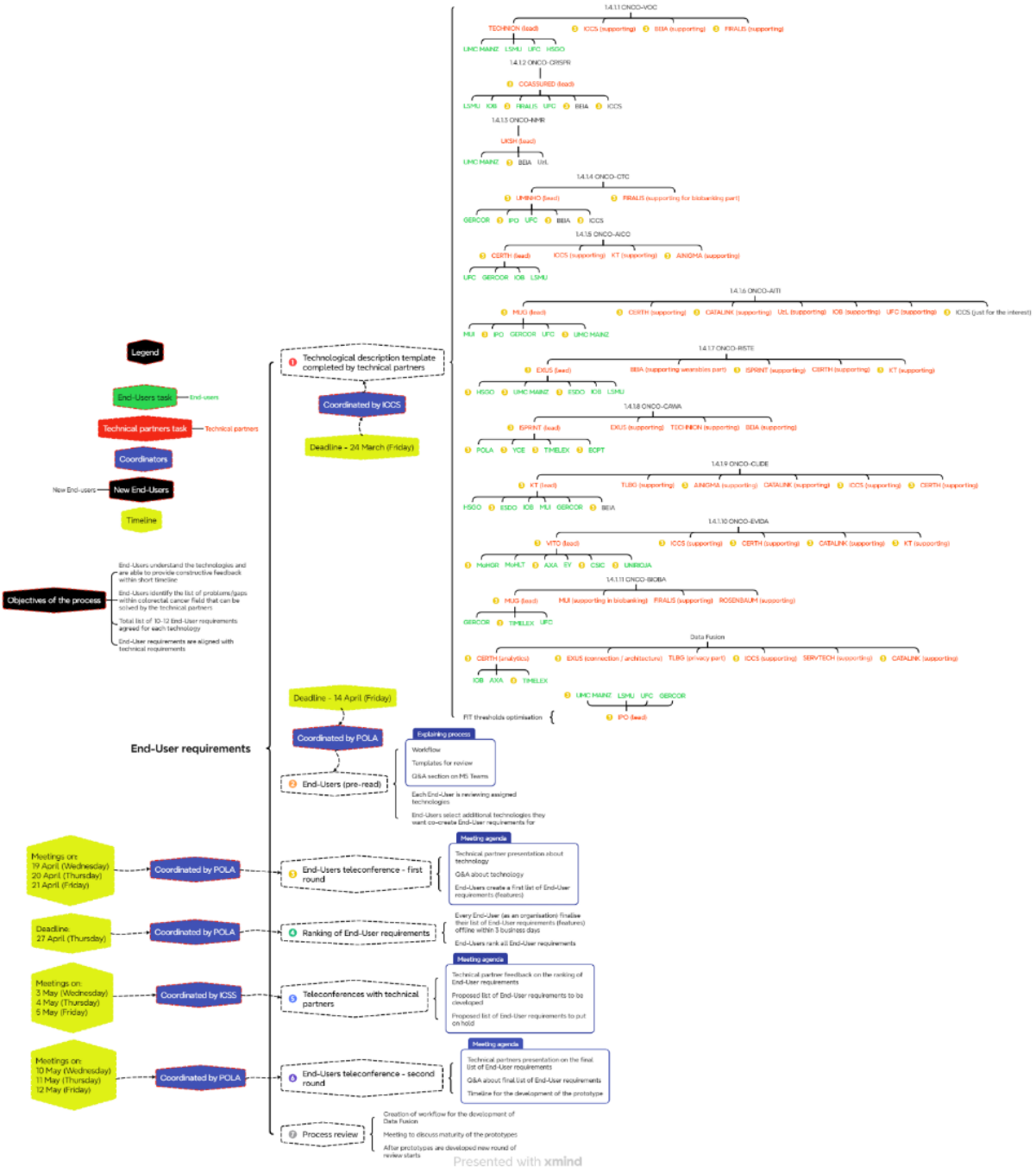

Figure 23 A diagram depicting activities towards the elicitation of functional and non functional requirements in ONCOSCREEN

Homogenization and Knowledge Model tool (DHKM), no end user requirements were collected. This is because these tools have not yet entered discussions with the end users. This is natural, as these components work on top of the outputs of other tools that need to progress first.

The structure for the list of requirements is as follows:

ID column is the identifier of the specified user/ technical requirement. Type is predefined as Functional (F) and Non-Functional (NF). Description column includes actual description of each requirement. The column Related URs, applies only on the list of technical requirements and it points the corresponding user requirement that is satisfied with the incorporation of this specific technical requirement. Priority column characterizes the priority a specific requirement has been set from either the end user or technical partners. A requirement may be of High (e.g., mandatory), Medium or Low priority. The Notes column serve as a disclaimer for the evolutionary nature of some user and/ or technical requirements (e.g., a requirement may be under construction or more feedback is needed, to be defined further etc.)

7.4 ONCOVOC

7.4.1 User Requirements

Table 3 Table of User Requirements for the ONCO-VOC tool

ID	Description	Priority	Notes
VOC_UR1	non-invasive	High	n/a
VOC_UR2	low cost (<10eur)	Medium	n/a
VOC_UR3	integration of data results	High	n/a
VOC_UR4	easy to explain and to use	Medium	n/a
VOC_UR5	low storage cost	High	n/a
VOC_UR6	high/improved sensitivity and specificity (>95%)	Medium	n/a
VOC_UR7	after integration: evidence- based decision making and recommendation for patient	High	n/a
VOC_UR8	personalized risk assessment	High	n/a

VOC_UR9	easy to integrate in the clinical praxis/routine	High	n/a
VOC_UR10	user-friendly communication of results to health clinic practitioner and patient	High	n/a
VOC_UR11	easy to integrate in everyday practice	High	n/a
VOC_UR12	easy to get equipment (availability)	Medium	n/a
VOC_UR13	quick result (without limiting the precision of diagnosis)	High	n/a
VOC_UR14	defined sensitivity specificity to distinguish benign vs malignant vs dysplasia	Medium	n/a
VOC_UR15	stratification of dysplasia grade	Medium	n/a
VOC_UR16	molecular categorization /profiling of tumour (so that it might guide therapeutic decision & medication selection	Medium	n/a
VOC_UR17	should be able to be performed in medical office, so that in case of a positive result, a physician can guide the patient for the next steps	High	n/a

7.4.2 ONCO-VOC Technical Requirements

Table 4 Table of Technical Requirements for the ONCO-VOC tool

ID	Type	Description	Related to UR(s)	Priority	Notes
VOC_TR1	NF	The tool should be cheap.	VOC_UR2	High	n/a
VOC_TR2	NF	The tool should be easily explained and easy to use.	VOC_UR4	High	n/a
VOC_TR3	NF	The tool's data results should be integrated with other components of ONCOSCREEN.	VOC_UR3, VOC_UR7	High	n/a
VOC_TR4	NF	The cost of storage of samples should be low.	VOC_UR5	High	n/a
VOC_TR5	F	The tool should utilize statistical Machine Learning methods to discriminate between healthy and CRC patient.	VOC_UR6-8	High	n/a
VOC_TR6	NF	The tool should be easy to integrate in the clinical "everyday" praxis.	VOC_UR1, VOC_UR9-12, VOC_UR17	High	n/a
VOC_TR7	NF	The tool should be able to deliver a quick result.	VOC_UR13	High	n/a
VOC_TR8	F	The tool should distinguish between benign/ malignant / dysplasia along with stratification of dysplasia grade	VOC_UR14-16	Low	n/a

VOC_TR9	F	The tool should be able to result in a risk percentage of having the disease, thus enabling it to account for uncertainty.	VOC_UR7-8	High	n/a

7.5 ONCO-CRISPR

7.5.1 ONCO-CRISPR User Requirements

Table 5 Table of User Requirements for the ONCO-CRIPSR tool

ID	Description	Priority	Notes
CRISPR_UR1	testing at points of care	Medium	n/a
CRISPR _UR2	high sensitivity and specificity (>95%)	High	n/a
CRISPR _UR3	easy to interpret/understand the result	High	n/a
CRISPR _UR4	non-invasive	High	n/a
CRISPR _UR5	no complex material needed	Medium	n/a
CRISPR _UR6	easy storage conditions (room temperature or 4°C)	Medium	n/a
CRISPR _UR7	low cost (<10eur)	Medium	n/a
CRISPR _UR8	quick result (few hours)	Medium	n/a
CRISPR _UR9	no special preparation for patients	High	n/a

CRISPR _UR10	validation in 12 months	Medium	n/a

7.5.2 ONCO-CRISPR Technical Requirements

Table 6 Table of Technical Requirements for the ONCO-CRISPR tool

ID	Type	Description	Related to UR(s)	Priority	Notes
CRISPR_TR1	NF	Non or minimally invasive test with ease of use and no special equipment necessary.	CRISPR_UR4	High	n/a
CRISPR _TR2	F	CRISPR protocol of max 2 hours.	CRISPR_UR8	Low	n/a
CRISPR _TR3	NF	High sensitivity and specificity (>95%)	CRISPR_UR2	High	n/a
CRISPR _TR4	NF	Easy Storage conditions (room temperature or 4°C) or -20 °C (normal freezer) for specific compounds	CRISPR_UR6	Medium	n/a
CRISPR _TR5	F	Able to identify CRC-associated single nucleotide polymorphisms (SNPs).	CRISPR_UR2-3, CRISPR_UR10	High	n/a
CRISPR _TR6	F	Able to identify CRC-associated methylation patterns.	CRISPR_UR2-3, CRISPR_UR10	High	n/a

CRISPR _TR7	F	Able to identify CRC-associated hCRISPR RNA expression patterns	CRISPR_UR2-3, CRISPR_UR10	High	n/a
CRISPR _TR8	NF	Able to put a CRISPR-Cas based CRC detection kit into the market for an affordable price	CRISPR_UR1-10	Medium	n/a

7.6 ONCO-CTC

7.6.1 ONCO-CTC User Requirements

Table 7 Table of User Requirements for the ONCO-CTC tool

ID	Description	Priority	Notes
CTC_UR1	is an affordable device	High	1st generation prototype can be made of PDMS
CTC_UR2	Is easy to operate	Medium	A liquid biopsy SOP has been established. The 1st generation prototype can allow staining procedures for Y/N outputs for CTC isolation without the use of a microscope
CTC_UR3	has the capability to detect CTC in early stages of disease	Low	2nd generation prototype may possibly be functionalized for immunoassays

CTC_UR4	has the possibility to detect specific biomarkers in CTCs allowing cells phenotyping	High	2nd generation prototype may possibly be optimized for sorting cells and characterization by FACS
CTC_UR5	no need for special equipment to isolate or identify CTC	Low	1st generation prototype can allow staining procedures for Y/N outputs
CTC_UR6	can recover CTC for downstream analysis (genomics, transcriptomics)	Medium	1st and 2nd generation prototypes may possibly be optimized for sorting cells
CTC_UR7	can recover viable cells suitable for organoid creation and therapy outcome tests	High	2nd generation prototype can include a sub-component for 3D in vitro model of intestine for personalised testing
CTC_UR8	design a multipurpose test with the ability to also isolate other analytes that could be better suitable for screening	High	2nd generation prototype can include a sub-component for 3D in vitro model of intestine for personalised testing
CTC _UR9	CTC are available only for invasive lesions and not appropriate for screening	Low	The prototype fibre meshes can be optimised for EVs isolation and characterization. Alternatively, ONCO-CTC can be combined with ctDNA or VOC or FIT to define the risk of invasive or advanced lesions

CTC _UR10	the test should be combined with CRISPR or VOC or FIT to define the risk of invasive or advanced lesions	Medium	2nd generation prototype may possibly be functionalized for immunoassays
CTC_UR11	acceptance and validation by endoscopist	High	1st generation prototype can allow processing of liquid biopsy sample (blood) for microscopy analysis

7.6.2 ONCO-CTC Technical Requirements

Table 8 Table of Technical Requirements for the ONCO-CTC tool

ID	Type	Description	Related to UR(s)	Priority	Notes
CTC_TR1	F	The tool should be affordable and easy to operate.	CTC_UR1-2	High	Under construction
CTC _TR2	NF	The tool should be able to detect the disease even in early stages.	CTC_UR3, CTC_UR10	Low	To be defined further
CTC _TR3	NF	The tool should detect biomarkers in Circulating Tumour Cells for allowing cells phenotyping and downstream analysis (genomics, transcriptomics).	CTC_UR4, CTC_UR9	Medium	To be defined further

CTC _TR4	F	The tool should be able to operate without special equipment.	CTC_UR5	Medium	Developed by UMINHO
CTC _TR5	NF	The tool should be able to recover viable cells, suitable for organoid creation.	CTC_UR7-8	High	Under validation
CTC _TR6	F	The tool should be validated by endoscopists.	CTC_UR11	Medium	To be defined further

7.7 ONCO-NMR

7.7.1 ONCO-NMR User Requirements

Table 9 Table of User Requirements for the ONCO-NMR tool

ID	Description	Priority	Notes
NMR_UR1	Easy storage conditions	Low	Samples can be stored in NMR tubes
NMR_UR2	no complex material needed	High	Suitable for standard blood serum or plasma

NMR_UR3	No special preparation of patients	High	Standard patient blood samples suffice
NMR_UR4	quick results	Medium	NMR is by nature non-invasive
NMR_UR5	possible in-house evaluation of results	Low	After initial evaluation by expert software on the spectrometers further external evaluation is possible
NMR_UR6	high sensitivity and specificity (>95%)	High	Has been demonstrated for HCC
NMR_UR7	evidence-based decision making	High	The goal is to enable doctors to identify relapses and treatment responses
NMR_UR8	personalized risk assessment	High	Evaluation whether the method is suitable for risk assessment
NMR_UR9	easy to integrate into clinical practice / routine	High	To be demonstrated that the method is suitable for clinical use
NMR_UR10	user-friendly communication of results to health clinic practitioner and patient	High	Requires data integration, e.g. using the ONCO-CLIDE tool (decision support system)
NMR_UR11	Validation in 12 months	Medium	n/a

7.7.2 ONCO-NMR Technical Requirements

Table 10 Table of Technical Requirements for the ONCO-NMR tool

ID	Type	Description	Related to UR(s)	Priority	Notes
NMR _TR1	F	Detection of lipoproteins related to CRC.	NMR_UR7-10	High	n/a
NMR _TR2	F	Detection of glycoproteins related to CRC.	NMR_UR7-10	High	n/a
NMR _TR3	NF	Minimally invasive test with fast measurement protocol of 30 minutes.	NMR_UR2,3	High	n/a
NMR _TR4	NF	The tool should be able to deliver results using samples of either blood serum or plasma.	NMR_UR2,3	High	n/a
NMR_TR5	NF	Easy storage conditions (-20°C)	NMR_UR1	High	n/a
NMR_TR6	F	NMR protocol of max 30min	NMR_UR4	High	n/a
NMR_TR7	NF	Provide a user friendly list of biomarkers as end result	NMR_UR5,10,11		

7.8 ONCO-AICO

7.8.1 ONCO-AICO User Requirements

Table 11Table of User Requirements for the ONCO-AICO tool

ID	Description	Priority	Notes
AICO_UR1	ensure low rate of mistakes (<5%)	High	n/a
AICO_UR2	gain acceptance and validation by endoscopists	High	XAI features will promote the acceptance by the experts
AICO_UR3	facilitate a faster increase of ADR for junior endoscopists during training and enhance learning usability	Medium	XAI features will support the learning /training process by providing intuitive insights to the AI based decisions provided to the trainees.
AICO_UR4	provide an easy-to-use interface and design	Medium	n/a
AICO_UR5	achieve good differentiation between subtypes of adenomas/ polyps	High	n/a
AICO_UR6	allow easy integration into everyday practice	Medium	n/a
AICO_UR7	provide an understandable summary of findings	Medium	n/a
AICO_UR8	make it easy to reclassify findings for experts	Low	n/a

AICO_UR9	highlight areas for pathologists to review	Low	n/a
AICO_UR10	ensure the software shows why and how the decision are made, contributing to the learning process	Medium	XAI features will support the learning /training process by providing intuitive insights to the AI based decisions provided to the trainees.

7.8.2 ONCO-AICO Technical Requirements

Table 12 Table of Technical Requirements for the ONCO-AICO tool

ID	Type	Description	Related to UR(s)	Priority	Notes
AICO_TR1	F	AICO_TR1 (F): Create user-friendly and easy-to-use interface controls for video manipulation, profile, track record of score etc.	AICO_UR4, AICO_UR6	High	n/a
AICO_TR2	F	AICO_TR2 (F): Develop a video player component that supports multiple formats and implements standard playback controls (play, pause, seek).	AICO_UR4, AICO_UR6	High	n/a
AICO_TR3	F	AICO_TR3 (F): Enable trainees to annotate regions of interest effectively, supporting features like bounding box and classification options.	AICO_UR9	High	n/a

AICO_TR4	F	AICO_TR4 (F): Develop user management service with authentication, role, and access control.	-	High	n/a
AICO_TR5	F	AICO_TR5 (F): Develop a scoring service for trainee evaluations and provide feedback and guidance based on their scores.	AICO_UR6-7	High	n/a
AICO_TR6	F	AICO_TR6 (F): Integrate Explainable AI algorithms and model-agnostic explanation methods to improve understanding and transparency.	AICO_UR2-3, AICO_UR10	Medium	The xAI aspect of the tool will be integrated with the n/a e AI models and output on the ONCO-AICO platform
AICO_TR7	F	AICO_TR7 (F): Develop a metadata generation service to include session information, dataset details, training parameters, and algorithm information.	AICO_UR7	High	n/a
AICO_TR8	F	AICO_TR8 (F): Enable asynchronous evaluation for trainers of non-correct selections to effectively handle ambiguous images.	AICO_UR9	High	n/a
AICO_TR9	F	AICO_TR9 (F): Establish a feedback and improvement mechanism that ensures privacy and	-	High	n/a

		data security while collecting feedback from trainers and trainees.			
AICO_TR10	F	AICO_TR10 (F): The tool should enable AI-based illustration of Regions of Interest (ROIs) and also offer the option of show/ hide ROIs for improved visualization.	AICO_UR10	High	n/a
AICO_TR11	F	AICO_TR11 (F): Support and facilitate learning of decisions by providing ground truth annotations	AICO_UR8-9	High	n/a
AICO_TR12	F	AICO_TR12 (F): Provide the ability to display explainable artificial intelligence (xAI)	AICO_UR2-3, AICO_UR10	Medium	Both intrinsic and extrinsic approaches will be employed in order to enhance the explainability provided to meet UR. Intuitive explanations will be in terms of heatmaps.

7.9 ONCO-AITI

7.9.1 ONCO-AITI User Requirements

Table 13 Table of User Requirements for the ONCO-AITI tool

ID	Description	Priority	Notes
AITI_UR1	highlight the areas for pathologist to review	High	n/a
AITI_UR2	similar or increased accuracy compared to standard of care	High	n/a
AITI_UR3	time saving - faster results /or diagnostics	High	n/a
AITI_UR4	software should show why and how the decision is made	High	n/a
AITI_UR5	classification between physiological and dysplastic glands	Medium	This UR do not directly match the goal of AITI described in the GA, but can be seen as an option or suggestion for improvement that further enhances the tool
AITI_UR6	improved illustration / visualization for young doctors and nurses	High	n/a

AITI_UR7	accessibility for a vast majority of clinics/hospitals	High	n/a

7.9.2 ONCO-AITI Technical Requirements

Table 14 Table of Technical Requirements for the ONCO-AITI tool

ID	Type	Description	Related to UR(s)	Priority	Notes
AITI_TR1	F	The tool should enable the illustration of ROIs (Regions of interest)	AITI_UR1	High	n/a
AITI_TR2	NF	Suggestions and assessments should have similar or better accuracy than decisions made by pathologists without this tool	AITI_UR2	High	n/a
AITI_TR3	NF	the tool should support and/or facilitate the learning of pathological decisions and thus accelerate the learning time from junior to expert	AITI_UR3	High	n/a
AITI_TR4	F	the tool should provide the ability to display xAI (explainable artificial intelligence)	AITI_UR4	High	n/a
AITI_TR5	F	The tool should enable the discrimination between physiological and dysplastic glands	AITI_UR5	Medium	n/a

AITI_TR6	F	The tool could be developed as a viewer for WSIs. The tool can offer a choice to show or hide the ROIs to improve and facilitate visualization depending on the situation	AITI_UR6	High	n/a
AITI_TR7	NF	Access to the tool should be enabled as open-source software	AITI_UR7	Medium	n/a

7.10 ONCO-BIOBA

7.10.1 ONCO-BIOBA User Requirements

Table 15 Table of User Requirements for the ONCO-BIOBA tool

ID	Description	Priority	Notes
BIOBA_UR1	would like to have an overview of available collections and cohorts in the project	High	n/a
BIOBA _UR2	Need of information about collections/cohorts, which kind of the samples/data they contain and the provider	High	n/a
BIOBA _UR3	Wish to get access to the data or the providers	High	n/a
BIOBA _UR4	There should be an option to provide own cohorts to the catalogue, if wished.	High	n/a

BIOBA _UR5	uniform description of the collections and cohorts	High	n/a
BIOBA _UR6	Somehow an overview about what kind of samples/data is in a collection or cohort	High	n/a

7.10.2 ONCO-BIOBA Technical Requirements

Table 16 Table of Technical Requirements for the ONCO-BIOBA tool

ID	Type	Description	Related to UR(s)	Priority	Notes
BIOBA_TR1	F	User friendly overview of listed collections, cohorts, provider	BIOBA_UR1	High	n/a
BIOBA _TR2	F	The content of the collections/cohorts should be descripted, as well as the corresponding provider	BIOBA_UR2, BIOBA_UR6	High	n/a
BIOBA _TR3	F	The tool should provide contact information to enable the access to the provider	BIOBA_UR3	High	n/a
BIOBA _TR4	F	The user should be able to inform the provider of BIOBA about an own cohort or collection and to obtain information, what is necessary to include information about the cohort in the list of the catalogue	BIOBA_UR4	Medium	n/a

BIOBA _TR5	NF	Certain metadata, which describes collections/cohorts and their provider has to be uniform/harmonized	BIOBA_UR5	High	n/a

7.11 ONCO-CAWA

7.11.1 ONCO-CAWA User Requirements

Table 17 Table of User Requirements for the ONCO-CAWA tool

ID	Description	Priority	Notes
CAWA_UR1	as an ONCOSCREEN app user, I expect to have access to the devices necessary for the needed measurements (C42)	High	n/a
CAWA_UR2	as an ONCOSCREEN app user, I expect the app translated to my native language (D40)	High	n/a
CAWA_UR3	as an ONCOSCREEN app user, I expect to be able to consent to the use of the app (D40)	High	n/a
CAWA_UR4	as an ONCOSCREEN app user, I expect to consent to the processing of my data by ONCOSCREEN partners (link with Data Fusion) (D40)	High	n/a
CAWA_UR5	as an ONCOSCREEN app user, I expect to read a privacy policy, explaining to me why my data is being collected, who it is going to be shared with, how it is protected, what are my rights in this regard, before I give my consent (D40)	High	n/a

CAWA_UR6	as an ONCOSCREEN app user, I need to be certain that my data is securely protected from unauthorized access, loss, destruction, modification (app will be with CE mark, it will also be stored in Data Fusion) (D40)	High	n/a
CAWA_UR7	as an ONCOSCREEN app user, I expect to read EULA/terms of use (D41)	Medium	n/a
CAWA_UR8	as an ONCOSCREEN app user, I expect that all information provided to me is easy to comprehend, written in plain language that I understand (native language) - link with ONCO-CLIDE (generating recommendations, what to say, to whom to say it and when to say it)	High	n/a
CAWA_UR9	as an ONCOSCREEN app user, I expect it available both for Android and iOS (B41)	High	n/a
CAWA_UR10	as an ONCOSCREEN app user, I expect to have access to the devices necessary for the needed measurements (C42)	High	n/a
CAWA_UR11	as an ONCOSCREEN app user, I expect it to have an intuitive UI (B44)	High	n/a
CAWA_UR12	as an ONCOSCREEN app user, I expect its UI to adapt to the expectations of my age group / personality (messages in different tones to judge the effectiveness (B45)	Medium	n/a
CAWA_UR13	as an ONCOSCREEN app user, I expect it to offer gamification features (B46)	Medium	n/a
CAWA_UR14	as an ONCOSCREEN app user, I expect it to show data from wearable devices (e.g., from apple kit, Fitbit, Garmin, polars etc.) (B47)	Medium	n/a
CAWA_UR15	as an ONCOSCREEN app user, I expect it to be using validated questionnaires for capturing PROM/PREM (B48) - related to WP5 (protocol design)	High	n/a
CAWA_UR16	as an ONCOSCREEN app user, I expect it to visualize data from diagnostic devices (B49)	High	n/a

CAWA_UR17	as an ONCOSCREEN app user, I expect it to show CRC prevalence analysis with at least comparison from EU Member States and if possible, within different regions of same Member State (C43)	Medium	n/a
CAWA_UR18	as an ONCOSCREEN app user, I expect the app to allow me to understand the factors leading to the decision of the risk stratification algorithm (C44)	High	n/a
CAWA_UR19	as an ONCOSCREEN app user, I expect the app to present to me my risk target group (C46)	High	n/a

7.11.2 ONCO-CAWA Technical Requirements

Table 18 Table of Technical Requirements for the ONCO-CAWA tool

ID	Type	Description	Related to UR(s)	Priority	Notes
CAWA_TR1	F	CAWA will be written in React Native, to support both iOS and Android devices from a single code base	CAWA_UR1, CAWA_UR9	High	n/a
CAWA_TR2	NF	The app will use minimum text. All messages will be translated to the necessary languages	CAWA_UR2	High	n/a
CAWA_TR3	F	CAWA will be integrated with devices for automatic data entry via the manufacturers' SDKs	CAWA_UR10	-	Not clarified yet which devices will be used

CAWA_TR4	F	CAWA will be integrated with devices for automatic data entry via the manufacturers' APIs	CAWA_UR10	-	Not clarified yet which devices will be used
CAWA_TR5	F	CAWA will be accepting manual input for devices that do not offer any radio communications	n/a	High	n/a
CAWA_TR6	NF	CAWA will be offered with an intuitive UI comprising single-purpose widgets	CAWA_UR11	High	n/a
CAWA_TR7	NF	Both the main screen and the widgets of CAWA will have minimum text, based mainly on icons	CAWA_UR11	High	n/a
CAWA_TR8	NF	App branding: CAWA is adopted from the Healthentia mobile app, sporting the logo and colour theme of the ONCOSCREEN project	n/a	High	n/a
CAWA_TR9	NF	CAWA UI will adapt to age group expectations	CAWA_UR12	Low	n/a
CAWA_TR10	F	Gamification features (goals & rewards) will be promoting CAWA usage	CAWA_UR13	Medium	n/a
CAWA_TR11	F	CAWA will be contacting the users with the recommendations generated by CLIDE. These should be easy to understand, trying different approaches for the same end-goal, attempting to match users' personality	CAWA_UR19, CAWA_UR8	High	n/a

CAWA_TR12	F	CAWA will be providing optimum timing of messages to users, based on their app usage history	CAWA_UR12	High	n/a
CAWA_TR13	F	CAWA will be able to forward questionnaires to users, preferably validated ones	CAWA_UR15	High	n/a
CAWA_TR14	F	CAWA will be offering proper visualizations for the information collected from the screening tests	CAWA_UR16	High	n/a
CAWA_TR15	F	CAWA will be offering proper visualizations for the information collected from activity tracking	CAWA_UR14	High	n/a
CAWA_TR16	F	CAWA should be able to get and offer users the information generated by RISTE in a comprehensive way	CAWA_UR18	High	n/a
CAWA_TR17	NF	CAWA should be offering the users the terms of use, EULA, privacy policy & consents, recording their acceptance	CAWA_UR7, CAWA_UR3-4	High	n/a
CAWA_TR18	NF	CAWA is built upon the Healthentia mobile app. Healthentia is a CE-marked medical device, soon to get Class IIa certification, that covers all data security concerns	CAWA_UR5-6	High	n/a
CAWA_TR19	F	CAWA is offering the collected information to the data fusion via the Healthentia API	CAWA_UR4	High	n/a

7.12 Data Homogenization and Knowledge Model tool

7.12.1 DHKM User Requirements

Table 19 Table of User Requirements for the DHKM tool

ID	Description	Priority
DHKM_UR1	The Knowledge model should be designed with user-friendly interfaces to ensure accessibility for healthcare practitioners with varying levels of technical expertise.	High
DHKM_UR2	The Knowledge model should convey CRC specific knowledge for effectively use and support users & tools to interpret the information provided by the model.	High
DHKM_UR3	The knowledge model and data homogenization tool should provide standard APIs that promote interoperability and the exchange of information between different systems.	High
DHKM_UR4	The knowledge model should be extendable so that it accommodates additional information regarding CRC characteristics and patient conditions.	High

7.12.2 DHKM Technical Requirements

Table 20 Table of Technical Requirements for the DHKM tool

ID	Type	Description	Related to UR(s)	Priority	Notes
DHKM_TR1	F	The tool should connect to relational database schemas for retrospective data	DHKM_UR2	High	n/a
DHKM_TR2	F	Enrich relational schemas with meta-data	DHKM_UR1	High	n/a

DHKM_TR3	F	Develop Knowledge Model for CRC	DHKM_UR4	High	n/a
DHKM_TR4	F	Describe the relational data sources using the CRC Knowledge Model	DHKM_UR1	High	n/a
DHKM_TR5	F	Create repository for the CRC Model descriptions for retrospective data	n/a	High	n/a
DHKM_TR6	F	Map the relational data source schemas to the constructs of the CRC Knowledge Model	DHKM_UR2	High	n/a
DHKM_TR7	F	Connect the CRC Knowledge Model to the ONCORISTE tool	DHKM_UR3	Medium	n/a

7.13 Privacy Preservation Tool

7.13.1 PPT User Requirements

This tool has not received requirements from the end users yet.

7.13.2 PPT Technical Requirements

Table 21Table of Technical Requirements for the DAT tool

ID	Type	Description	Related to UR(s)	Priority	Notes
DAT_TR1	F	The tool should automatically select the best anonymization algorithm based on the dataset received in input.	n/a	Medium	n/a

DAT_TR2	F	The tool should allow the integration of custom anonymization strategies based on the end-user dataset metadata.	n/a	High	n/a
DAT_TR3	F	Creation of a medical anonymization strategy.	n/a	Medium	n/a
DAT_TR4	F	The anonymization output will be temporarily saved (for one week) to facilitate pagination, enabling the end-user to retrieve the information later or save it to another database. Afterward, it will be removed from the anonymization platform.	n/a	High	n/a

7.14 Data lake and fusion engine tool

7.14.1 User Requirements

This tool has not received requirements from the end users yet.

7.14.2 Technical Requirements

Table of Technical Requirements

Table 22 Table of Technical Requirements for the Data Lake and Fusion Engine tool

ID	Type	Description	Related to UR(s)	Priority	Notes
FUSION_TR1	F	The system should be able to efficiently ingest data from multiple sources in various formats and structures.	n/a	High	To be integrated with the

					ONCO_RISTE tool
FUSION_TR2	F	The system should utilize a repository capable of efficiently handling structured data.	n/a	High	n/a
FUSION_TR3	F	The system should be able to handle large-scale data processing and analysis to generate insights and identify correlations and patterns.	n/a	High	n/a
FUSION_TR4	F	The repository should have robust security measures in place to protect sensitive data from unauthorized access or misuse.	n/a	High	n/a
FUSION_TR5	F	The system should support data integration from disparate sources, including APIs, databases, and file systems.	n/a	High	n/a
FUSION_TR6	F	The system should have a mechanism for archiving data that is no longer needed, while still preserving its integrity and accessibility.	n/a	High	n/a
FUSION_TR7	F	The system should be able to fuse large-scale, heterogeneous, graph-structured data sources and recognize multidimensional patterns.	n/a	High	n/a
FUSION_TR8	F	The system should support data deletion to comply with GDPR legislation.	n/a	High	n/a
FUSION_TR9	F	The system should have a predefined naming convention for its data.	n/a	High	n/a
FUSION_TR10	F	The system should be adaptable to handle potential data source changes.	n/a	High	n/a
FUSION_TR11	F	Insights from the system should be easily exportable via retrieval calls.	n/a	High	n/a

7.15 ONCO-RISTE

7.15.1 ONCO-RISTE User Requirements

Table 23 Table of User Requirements for the ONCO-RISTE tool

ID	Description	Priority	Notes
RISTE_UR1	the tool should be able to classify the risk in a 5-level stratification	High	n/a
RISTE_UR2	as user, I expect the tool to give me the time distribution of the risk based on my risk level (i.e., short vs long-term)	Medium	n/a
RISTE_UR3	as an HCP user/patient, I expect to know the importance of the factors in the given risk level.	High	n/a
RISTE_UR4	tool should dynamically change the output if the patient changes the factors given.	High	n/a
RISTE_UR5	tool must show the rules to the medical partners and give them the option to change them.	High	n/a
RISTE_UR6	Tool must communicate the high-level information of risk levels with demographic data	High	n/a

7.15.2 ONCO-RISTE Technical Requirements

Table 24 Table of Technical Requirements for the ONCO-RISTE tool

ID	Type	Description	Related to UR(s)	Priority	Notes
RISTE_TR1	F	ONCORISTE will be integrated with ONCO-CAWA	RISTE_UR1, RISTE_UR3-4	High	n/a

RISTE_TR2	F	ONCORISTE will receive the semi-empirical rules from Retrospective CRC screening Data Collection & Analysis	RISTE_UR1	Low	n/a
RISTE_TR3	F	ONCORISTE will be integrated with data fusion service	RISTE_UR6	Medium	n/a
RISTE_TR4	F	ONCORISTE will implement a classification process based on ML algorithms	RISTE_UR1-3	High	n/a
RISTE_TR5	F	ONCORISTE output will be a risk-based score for each individual	RISTE_UR1, RISTE_UR3	High	n/a
RISTE_TR6	F	ONCORISTE will be integrated with all the previous tools via the virtual data lake and not directly	- n/a	Low	n/a
RISTE_TR7	F	An aggregator service will be implemented in order to collect all the data and transform them to an ONCORISTE input	RISTE_UR1, RISTE_UR3, RISTE_UR6	Medium	n/a
RISTE_TR8	F	The ONCORISTE output will be saved to a dedicated database	- n/a	Medium	n/a
RISTE_TR9	NF	Authentication and password management will be implemented	- n/a	Medium	n/a
RISTE_TR10	F	ONCORISTE will integrate with ONCO-CLIDE	RISTE_UR4-5	High	n/a

RISTE_TR11	NF	All components will communicate via a secure network	- n/a	High	n/a
RISTE_TR12	NF	All components will communicate with secure communication protocols	- n/a	High	n/a

7.16 ONCO-CLIDE

7.16.1 ONCO-CLIDE User Requirements

Table 25 Table of User Requirements for the ONCO-CLIDE tool

ID	Description	Priority	Notes
CLIDE_UR1	to have precise diagnosis of presence or absence of malignancy	Medium	This will be implemented on the classification component.
CLIDE_UR2	to come with a positive/negative predictive value (high sensitivity and specificity)	Medium	This will depend on the selected classification algorithm.
CLIDE_UR3	to be able to detect precancerous lesions and degree of dysplasia	Low	This is not really connected to the scope of the predictive models.
CLIDE_UR4	to be able to distinguish the stage of malignancy as well the molecular subtype / mutational profile of the tumour.	Medium	The stage will be determined by the relevant model. The others are probably not possible but might be considered depending on enough historical data.

7.16.2 ONCO-CLIDE Technical Requirements

Table 26 Table of Technical Requirements for the ONCO-CLIDE tool

ID	Type	Description	Related to UR(s)	Priority	Notes
CLIDE_TR1	F	Integration with the various diagnostic solutions proposed in this project via the virtual data lake	n/a	High	Some first steps have been done on this.
CLIDE_TR2	F	Integration/communication with the clinical centre's systems to receive data from colonoscopies, CT scans, biopsies etc. via the virtual data lake	n/a	High	n/a
CLIDE_TR3	F	The system should be able to consider the individual's risk level provided from ONCO-RISTE when formulating recommendations.	n/a	High	n/a
CLIDE_TR4	F	The system must include a Virtual Tumour board that facilitates communication between doctors. This will be included in the web interface.	n/a	High	This probably needs to be done at a later stage.
CLIDE_TR5	F	The system should be able to perform correlation-based classification of citizens/cases across various adenoma-carcinoma sequence levels (0, I, II, III, IV) based on the European CRC guidelines	CLIDE_ UR1-4	High	This probably needs to be done at a later stage when data is available.

CLIDE_TR6	F	The system should generate a set of recommendations based on the diagnostic results, classification level (staging) and individual risk level.	CLIDE_UR1-4	High	This probably needs to be done at a later stage when data is available
CLIDE_TR7	F	Interaction between doctors and ONCO-CLIDE needs to happen through a dedicated web interface	n/a	High	This needs to be split into multiple requirements through an iterative progress with the medical partners.
CLIDE_TR8	F	There needs to be an authentication system for the web platform with different roles.	n/a	High	This has been implemented for some user types.

7.17 ONCO-EVIDA

7.17.1 ONCO-EVIDA User Requirements

Table 27 Table of User Requirements for the ONCO-EVIDA tool

ID	Description	Priority	Notes
EVIDA_UR1	include uncertainty and sensitivity aspects; Include explanations of the results. Use explainable and interpretable models. Sensitivity	EVIDA_UR1a: Medium	EVIDA_UR1a: Include uncertainty and sensitivity aspects. As the dashboard

	analysist tools; Easy access to underlying data to do focused analyses; Avoid biases : As the dashboard might portray aggregated data and visuals based on them, it is important to low as much as possible uncertainty and biases; Any policy suggestions made by the tool is important to be as precise and targeted as possible while corresponding to specific population group at risk; Use a coherent approach to communicate forecasts and risks. Avoid standard risk matrices. Use colours maybe but how you get the numbers behind them are important	EVIDA_UR1b: High EVIDA_UR1c: High EVIDA_UR1d: High	might portray aggregated data and visuals based on them, it is important to lower as much as possible uncertainty and biases. EVIDA_UR1b: Easy access to underlying data to do focused analyses. EVIDA_UR1c: Include explanations of the results. Use explainable and interpretable models. EVIDA_UR1d: Any policy suggestions made by the tool is important to be as precise and targeted as possible while corresponding to specific population group at risk.
EVIDA_UR2	Bayesian methodology to aggregate information	Low	Needs a more concrete explanation of how user envisions this, then priority could be increased.
EVIDA_UR3	mapping areas with populations per risk classification; population level analytics (e.g. ability to compare cohorts among themselves; or populations within a cohort with similar populations in other cohorts)	High (Map of areas where risk factors are more prominent but can only be done when the risk	We can easily use open-source data but including risk classification is dependent on the progress of ONCO-RISTE.

		factors first have been established)	
EVIDA_UR4	filtering of data; interactive dashboard; Automatic indicator of associations; visualization of risks by 'cause'/'risk factor'; Visualization of risk	High	A filtering function will be provided for each one of the data related tables and for the graphs that will be present on the dashboard.
EVIDA_UR5	language switching	High	The dashboard will be multi-language and support the corresponding language of each pilot.
EVIDA_UR6	possibility to provide suggestions to the developers and bug signalling in the pilot phase.	High	A contact section will be available where the partners will be able to submit their feedback and/or report bugs.
EVIDA_UR7	the tool should make it possible to identify (on maps) areas with populations classified according to willingness to participate in early detection programs; The tool should contain information at least at three geographic-political levels (national, subnational, local)	EVIDA_UR7a: High EVIDA_UR7b: High (Subnational & Local mapping: Medium)	EVIDA_UR7a: The tool should make it possible to identify (on maps) areas with populations classified according to willingness to participate in early detection programs. EVIDA_UR7b: The tool should contain information at least at three geographic-political levels (national, subnational, local).

			Maps below country level are currently not supported on the platform. "Willingness to participate" not measurable in some countries (i.e. countries that don't actively send out invitations).
EVIDA_UR8	automatic indicator of current policies; automatic inclusion of current policies	Low	Needs a more concrete explanation of how user envisions this, then priority could be increased.
EVIDA_UR9	option to see deviation of results from a predefined standard (real life evidence vs theoretical standard). have thresholds that determine severity of deviation	Low	n/a
EVIDA_UR10	multicriteria decision making algorithm	Medium	n/a
EVIDA_UR11	secure the system to avoid misuse and leakage	High	A token-based authentication system will be in place and only authorized users will be able to retrieve/store/see information on the dashboard.
EVIDA_UR12	access rights/agreements established: since it will support multiple actors across different entities access rights need to be suitable for each one's competencies; not impeding them from acquiring/sharing critical information, without compromising data security	High	n/a

EVIDA_UR13	retrospective information (3 years back); possibility to visualize information on dynamic timeline (to observe changes over time of specific phenomena)	High	n/a
EVIDA_UR14	present the impact of the policies vs the current status quo; possibility of simulating alternative policies; ability to test "what-ifs" scenarios and model the impact on forecasts (e.g. show hypothesis of how implementing certain policies could lead to inflection in the forecasted CRC mortality…); suggest policies and provide a bunch of them ordered according e.g. to their expected utilities or similar. Such policies should be personalised to the policymaker, patient, tool user, doctor…; the tool should incorporate (according to data availability) the costs of activities related to the early detection of colorectal cancer	EVIDA_UR14a: Medium EVIDA_UR14b: High EVIDA_UR14c: High	EVIDA_UR14a: Present the impact of the policies vs the current status quo; Possibility of simulating alternative policies; Ability to test "what-ifs" scenarios and model the impact on forecasts (e.g. show hypothesis of how implementing certain policies could lead to inflection in the forecasted CRC mortality…) EVIDA_UR14b: Suggest policies and provide a bunch of them ordered according e.g. to their expected utilities or similar. Such policies should be personalised to the policymaker, patient, tool user, doctor…; EVIDA_UR14c: The tool should incorporate (according to data availability) the costs of activities related to the early detection of colorectal cancer.

			The idea is that policymakers can fiddle around with variables to see what the impact of changing one modifiable exposure (e.g. increase screening) would be on CRC outcomes on one hand and health care costs on the other hand.
EVIDA_UR15	translate forecasts of CRC risks into deaths, QUALY's, etc.	Low	n/a
EVIDA_UR16	to have a possible real-time scenario of: a) prevalence; b) Factor affecting disease	Medium	Often cancer-related data takes a while to compile process so aggregated data of the current calendar year ("real-time") is usually not available, but a recent approximation should be.
EVIDA_UR17	data Virtualization section: a specific area where researchers and practitioners can access to data and run their own study mixing and combining different sources	Low	Would require ONCO-EVIDA to have a lot of data processing and statistical functionality, otherwise the distinction with other URs (1b, 4, 7 and 10) is unclear.
EVIDA_UR18	ability to potentially host payment data	Medium	Suggested by MoHGR "Payment data" to be interpreted as data of screening costs, likely obtainable from health insurers.

			Important for the cost/benefit analyses that policymakers want to be able to do in ONCO-EVIDA.
EVIDA_UR19	mechanism for continuous update regarding changes in legislation (is linked to user rights to change legislation)	Low	Suggested by MoHGR
EVIDA_UR20	proposed policy interventions to be extracted as legislative suggestions so that they have a more long-term orientation/character.	Low	Suggested by MoHGR. Seems like legal expertise would be required to properly implement this UR.

7.17.2 ONCO-EVIDA Technical Requirements

Table 28 Table of Technical Requirements for the ONCO-EVIDA tool

ID	Type	Description	Related to UR(s)	Priority	Notes
EVIDA_TR1	F	The tool should be able to account for bias and uncertainty in the data.	EVIDA_UR1a	Medium	n/a
EVIDA _TR2	F	The tool should provide access to the underlying data for subsequent focused analysis.	EVIDA_UR1b	High	n/a
EVIDA _TR3	F	The tool should be able to provide explanations of results.	EVIDA_UR1c	High	n/a
EVIDA _TR4	F	The tool should make policy suggestions based on population risk score.	EVIDA_UR1d	High	n/a

EVIDA _TR5	F	The tool should be able to use Bayesian methodology to aggregate information.	EVIDA_UR2	Low	n/a
EVIDA _TR6	F	The tool should be able to deliver interactive dashboard enabling the filtering of data and calculate associations, visualize risk and the cause of it.	EVIDA_UR4	High	n/a
EVIDA _TR7	F	The tool should support different languages.	EVIDA_UR5	High	n/a
EVIDA _TR8	F	The tool should support alarm warning in case of suspected bugs during the pilot phases.	EVIDA_UR6	High	n/a
EVIDA _TR9	F	The tool should be able to identify on maps (spatially) populations willing to participate in the early screening programs.	EVIDA_UR7a	High	n/a
EVIDA _TR10	F	The tool should provide information at least on 3 geographic political levels (i.e., national, subnational, local).	EVIDA_UR7b	High / Medium	n/a
EVIDA _TR11	F	The tool should be able to include and take into account the current policies.	EVIDA_UR8	Low	n/a
EVIDA _TR12	F	The tool should be able to make decisions based on multiple criteria.	EVIDA_UR10	Medium	n/a
EVIDA _TR13	NF	The tool should be secure in order to avoid misuse, data leakage etc.	EVIDA_UR11	High	n/a
EVIDA _TR14	NF	The tool should be designed in an adaptable manner to enable multiple actors to make use of it in a secure and safe way.	EVIDA_UR12	High	n/a

EVIDA _TR15	F	The tool should be able to utilize retrospective data in order to show the dynamics of specific phenomena).	EVIDA_UR13	High	n/a
EVIDA _TR16	F	The tool should be able to compare among different policies with a possibility of simulating a reference policy on a "what-if" scenario. This functionality should also account for the cost.	EVIDA_UR14 EVIDA_UR15	Medium Low	n/a
EVIDA _TR17	F	The tool should be able to host a time-wise scenario on prevalence and different factors affecting the disease.	EVIDA_UR16	Medium	n/a
EVIDA _TR18	F	The tool should be able to adapt on changes in legislation.	EVIDA_UR19	Low	n/a
EVIDA _TR19	F	The tool should be able to host payment data.	EVIDA_UR18	Low	n/a
EVIDA_TR20	F	The tool should be able to map areas with populations according to their risk level.	EVIDA_UR3	High	n/a
EVIDA_TR21	F	The tool should have the option to see deviation of results from a predefined standard.	EVIDA_UR9	Low	n/a
EVIDA_TR22	F	The tool should have a specific area where researchers and practitioners can access the data and combine different resources to run their own study.	EVIDA_UR17	Low	n/a
EVIDA_TR23	F	The tool's proposed policy interventions are to be extracted as legislative suggestions so that they have a more long-term character.	EVIDA_UR20	Low	n/a

8 Risks and Mitigation Measures

ONCOSCREEN is not without its share of potential data-related risks, particularly when legal constraints may lead to the absence of essential data. In such cases, a potential mitigation strategy involves the utilization of synthetic data to ensure continuity in development and testing processes. Even though this workaround introduces a new set of challenges, we are prepared to develop (if needed) new modules for the creation and meticulous validation of synthetic data (e.g. by using available open-source data) that accurately represents the complexities of a real-world scenario.

Another potential risk is the end-user dissatisfaction and potential solution cancellation. Misconception on the use, the role and functionalities of a particular tool can lead to further miscommunications during the stages of development and ultimately the rejection of the final product by the end users despite the effort made. To address this, the management of the project particularly emphasizes in the early and continuous engagement with end users, seeking their input and feedback throughout the development lifecycle. This proactive approach not only minimizes the likelihood of unforeseen user issues but also fosters a collaborative environment where potential concerns can be identified and addressed in a timely manner, enhancing the overall success and acceptance of the project. The detailed list technical risks for each tool is contained in D4.3 and D3.1.

9 Next Steps

Moving forward, the ONCOSCREEN project will continue its System Co-design approach, fostering additional rounds of interaction between end users and technical partners starting from our 1st Plenary meeting on M14. A collective decision has been made to initiate the next co-creation step once end users have been provided with prototypes of ONCOSCREEN tools, creating an invaluable feedback loop. This phase will involve a careful re-iteration and refinement of existing requirements, paving the way for more advanced developments, particularly in anticipation of Phase A, covering Data Collection and Pre-evaluation, and the onset of the clinical trials. The forthcoming 3rd Laboratory Integration Test (LIT) on M28 will draw upon insights from the previous LITs (LIT1 and LIT2) and leverage retrospective data from Task 2.4, soon to be made available. By that time, the aim is to achieve comprehensive system integration. Results of these efforts will be demonstrated in Deliverable D4.2, representing the final version of the Co-Designed System Architecture, scheduled for release on M29 of the project.

10 Conclusion

This deliverable presents the first draft of the Overall Architecture of ONCOSCREEN and introduces the tools that comprise the project. ONCOSCREEN adheres to a Co design methodology with the end users which in turn, provide requirements for the technical partners. The purpose of this procedure is twofold. First, to acquire a proper understanding of end users' needs and how the technical partners can help them be more efficient in their field of work and second, how the latter can translate these requirements into technical specifications of the system. This evolving process of setting the user requirements will be sought until M29. The iteration of user requirements listed on this deliverable will set the basis for further refinements and upcoming technical developments in later phases of the project.

With a combination of advanced security technologies and enhanced compliance practices, the project ambitions to maintain the relevant standards of data security and regulatory adherence in handling sensitive medical information. ONCOSCREEN prioritizes robust security and compliance measures to safeguard sensitive information. Employing a zero-trust framework, ensures a comprehensive and diligent approach to data protection. The utilization of Kafka SSL establishes a secure and encrypted communication channel, complemented by secure APIs that carefully regulate access to sensitive information. Federated virtual databases enhance the project's ability to analyse data and/or train AI models within the source securely, maintaining data integrity and minimizing exposure. Finally, the projects provisions for compliance adheres to SNOMED CT and FHIR standards on the physical centralized repository, reinforcing interoperability and standardization.

Ahead of the forthcoming 3rd LIT on M28, the project aims for a comprehensive integration, drawing on insights from previous tests, leveraging upcoming retrospective data and the latest technical developments. This will be the last test prior to D4.2 (on M29 of the project) that will finalize the current first Draft of the Overall Architecture of the system and its end user and technical requirements.

11 References

Altomare, D. F., Di Lena, M., Porcelli, F., Trizio, L., Travaglio, E., Tutino, M., Dragonieri, S., Memeo V. & de Gennaro, G. (2013). Exhaled volatile organic compounds identify patients with colorectal cancer. *Journal of British Surgery*, 100, 144-150.

Bender, D., & Sartipi, K. (2013). HL7 FHIR: An Agile and restful approach to healthcare information exchange. *Proceedings of the 26th IEEE International Symposium on Computer-Based Medical Systems*, 326-331. https://doi.org/10.1109/cbms.2013.6627810

Costantini, S., Di Gennaro, E., Capone, F., De Stefano, A., Nasti, G., Vitagliano, C., Setola, S. V., Tatangelo, F., Delrio, P., Izzo, F., Avallone, A., & Budillon, A. (2023). Plasma metabolomics, lipidomics and cytokinomics profiling predict disease recurrence in metastatic colorectal cancer patients undergoing liver resection. *Frontiers in Oncology*, 12, 1110104. https://doi.org/10.3389/fonc.2022.1110104

Donnelly K. (2006). SNOMED-CT: The advanced terminology and coding system for eHealth. *Studies in health technology and informatics*, 121, 279–290.

Gurcan, M. N., Boucheron, L. E., Can, A., Madabhushi, A., Rajpoot, N. M., & Yener, B. (2009). Histopathological image analysis: A review. *IEEE reviews in biomedical engineering*, 2, 147-171.

Janowczyk, A., & Madabhushi, A. (2016). Deep learning for digital pathology image analysis: A comprehensive tutorial with selected use cases. *Journal of pathology informatics*, 7(1), 29.

Jiang, M., Jin, S., Han, J., Li, T., Shi, J., Zhong, Q., Li W.,Tang W., Huang Q. & Zong, H. (2021). Detection and clinical significance of circulating tumour cells in colorectal cancer. *Biomarker research*, 9, 1-20.

Lundberg, S. M. & Lee, S.I. (2017). A unified approach to interpreting model predictions. In Proc. 31st Int. Conf. *Neural Information Processing Systems*, 30, 4768–4777

Lawrence, R., Watters, M., Davies, C. R., Pantel, K., & Lu, Y. J. (2023). Circulating tumour cells for early detection of clinically relevant cancer. *Nature Reviews Clinical Oncology*, 1-14.

Merino-Martinez, R., Norlin, L., van Enckevort, D., Anton, G., Schuffenhauer, S., Silander, K., Mook, L., Holub, P., Bild, R., Swertz, M. & Litton, J. E. (2016). Toward global biobank integration by implementation of the minimum information about biobank data sharing (MIABIS 2.0 Core). *Biopreservation and biobanking*, 14(4), 298-306.

Otoo, J. A., & Schlappi, T. S. (2022). REASSURED Multiplex Diagnostics: A critical review and forecast. *Biosensors*, 12(2), 124

Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). " Why should i trust you?" Explaining the predictions of any classifier. *In Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*,1135-1144.

Salmerón, A. M., Tristán, A. I., Abreu, A. C., & Fernández, I. (2022). Serum Colorectal Cancer Biomarkers Unraveled by NMR Metabolomics: Past, Present, and Future. *Analytical Chemistry*, 94(1), 417–430. https://doi.org/10.1021/acs.analchem.1c04360

Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., & Batra, D. (2017). Grad-cam: Visual explanations from deep networks via gradient-based localization. *In Proceedings of the IEEE international conference on computer vision*, 618-626.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. Journal of the *Royal Statistical Society Series B: Statistical Methodology*, 58, 267-288.

Tsai, W. S., Nimgaonkar, A., Segurado, O., Chang, Y., Hsieh, B., Shao, H. J., ... & Mei, R. (2018). Prospective clinical study of circulating tumor cells for colorectal cancer screening. *Journal of Clinical Oncology*, 36(4_suppl), 556.

van Riet, J., Saha, C., Strepis, N., Brouwer, R. W., Martens-Uzunova, E. S., van de Geer, W. S., ... & Louwen, R. (2022). CRISPRs in the human genome are differentially expressed between malignant and normal adjacent to tumor tissue. *Communications Biology*, 5(1), 338.

DG for research and Innovation (EC), PwC EU Services, "Cost-benefit analysis for FAIR research data Cost of not having FAIR research data", European Commision, 2010