D1.1: Data Management Plan

Revision: v.1.0

Work package	WP1
Task	Task 1.1
Due date	30/06/2023
Submission date	30/06/2023
Deliverable lead	UOI
Version	1.0
Authors	Christos Androutsos (UOI), Vasileios Pezoulas (UOI), D.I. Fotiadis (UOI)
Reviewers	Stavros Miloulis (ICCS)
Abstract	This document presents the DMP of the DIOPTRA project, outlining the strategies and approaches that will be implemented to ensure efficient handling of the data collected and generated during the project on colorectal cancer screening. The document provides a detailed description of the data that will be collected per Work Package (WP), in addition to the mechanisms for data collection, documentation, metadata generation, and data evaluation. Furthermore, the storage and backup mechanisms, data preservation, sharing, and access methods, as well as the DMP's compliance with ethical and legal standards, are described.
Keywords	Data Management Plan, FAIR, GDPR

Document Revision History

Version	Date	Description of change	List of contributor(s)
V0.1	20/04/2023	ToC and 1st edit	Christos Androutsos (UOI), Vasileios Pezoulas (UOI)
V0.2	17/05/2023	Review and 2nd edit	Christos Androutsos (UOI), Vasileios

Grant Agreement No.: 101096649
Call: HORIZON-MISS-2021-CANCER-02
Topic: HORIZON-MISS-2021-CANCER-02-01
Type of action: HORIZON-RIA

			Pezoulas (UOI)
V0.3	24/05/2023	Added executive summary, introduction,	Christos Androutsos (UOI), Vasileios Pezoulas (UOI)
V0.4	01/06/2023	Added content on the DIOPTRA datasets and FAIR sections	Christos Androutsos (UOI), Vasileios Pezoulas (UOI)
V0.5	13/06/2023	Added content in the data introduction and description sections	Christos Androutsos (UOI), Vasileios Pezoulas (UOI)
V0.6	15/06/2023	Comments on the content of deliverable and added content in data description and storage and backup sections.	Stavros Miloulis (ICCS)
	16/06/2023	Comments on the content of deliverable and added content in data introduction and description sections.	Zheshen Jiang (CHUL)
	16/06/2023	Added content in the legal and ethical aspects section	Niamh Christina Gleeson (ARTHUR)
V0.7	19/06/2023	Comments on the content of deliverable and added content in the data description section	Christos Fotis (PAO)
V0.8	19/06/2023	Updated data description, data storage and backup and data security sections	Grigoris Antonopoulos (INTRA), Filopoimin Lykokanellos (INTRA), Michalis Vourtzoumis (INTRA)
V0.9	20/06/2023	Review comments and updated content in all sections.	Christos Androutsos (UOI), Vasileios Pezoulas (UOI)
V1.0	20/06/2023	Prepared final consolidated version	Christos Androutsos (UOI), Vasileios Pezoulas (UOI), Dimitrios I. Fotiadis (UOI)
V1.0	30/06/2023	Deliverable approved by the PC and submitted	Stavros Miloulis (ICCS)

Disclaimer

Funded by the European Union

Funded by the European Union (DIOPTRA, 101096649). Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or the Health and Digital Executive Agency. Neither the European Union nor the granting authority can be held responsible for them.

Copyright notice

Project funded by the European Commission in the Horizon Europe Programme		
Nature of the deliverable:	**DMP—Data Management Plan**	
Dissemination Level		
PU	*Public, fully open, e.g. web (Deliverables flagged as public will be automatically published in CORDIS project's page)*	✔
SEN	*Sensitive, limited under the conditions of the Grant Agreement*	
Classified R-UE/ EU-R	*EU RESTRICTED under the Commission Decision No2015/ 444*	
Classified C-UE/ EU-C	*EU CONFIDENTIAL under the Commission Decision No2015/ 444*	
Classified S-UE/ EU-S	*EU SECRET under the Commission Decision No2015/ 444*	

* R: Document, report (excluding the periodic and final reports)

DEM: Demonstrator, pilot, prototype, plan designs

DEC: Websites, patents filing, press & media actions, videos, etc.

DATA: Data sets, microdata, etc.

DMP: Data management plan

ETHICS: Deliverables related to ethics issues.

SECURITY: Deliverables related to security issues

OTHER: Software, technical diagram, algorithms, models, etc.

Executive Summary

D1.1 corresponds to the first version of the Data Management Plan (DMP) for the DIOPTRA project, which outlines the actions to be taken and the processes to be followed to manage the data collected and generated during the project's lifetime. It provides a summary of:

a) the data that will be collected, processed and generated throughout the lifecycle of the project.
b) the mechanisms that will be followed for the collection, management and analysis of the data.
c) the methodologies and standards that will be followed.
d) the conditions under which the data will be shared/made open and the procedure for achieving this.
e) the storage and the back-up processes that will be adopted.
f) the compliance of the DIOPTRA project with findable, accessible, interoperable and re-usable (FAIR) principles.
g) the legal and ethical considerations that will be taken into account.

The deliverable is organised as follows: Initially, the primary scope of this deliverable, the objectives of the data management plan and its relationship to other related deliverables are described. In the Data Introduction section, information regarding the major DIOPTRA data types is provided. The following are the DIOPTRA FAIR principles. The following section, which corresponds to the Data description, provides a detailed description of the data that will be collected per Work Package (WP), as well as the mechanisms for data collection, documentation, metadata generation and data assessment. Additionally, the data storage and backup procedures are also presented in this section. Finally, the risks and proposed mitigation measures, responsibilities and resources, along with the legal and ethical aspects related, among other, also to the General Data Protection Regulation (GDPR) and its application to the DIOPTRA project, are described.

Funded by the European Union

Compliance of D1.1 with DoA

DoA	D1.1 Data Management Plan
"UOI and ICCS will also produce the Data Management Plan (DMP), corresponding to the overall work."	Along with ICCS, UOI presented the DMP's table of contents and added content to the entire document. When revising and finalising the document, the remaining participants' feedback was taken into account.
"DMP will be the blueprint required to make European research data and infrastructures Findable, Accessible, Interoperable and Reusable according to the FAIR Principles."	Section 4 outlines the strategies and approaches for aligning the project with the Findable, Accessible, Interoperable, and Reusable (FAIR) data principles, ensuring appropriate data management.
"Considering that the patients' data are sensitive personal data, the relevant legal framework for the patients' rights is the Data Protection Directive on the protection of individuals with regard to the processing of personal data and on the free movement of such and the General Data Protection Regulation (EU) 2016/67984 will be addressed."	The legal and ethical aspects related, among other, also to the General Data Protection Regulation (GDPR) regulation and its application to the DIOPTRA project, are described extensively in the Section 6.
"Data curation will be based in open format exploitation, while FAIRness will be assessed and enforced in line with DIOPTRA Data Management Plan (DMP)"	Data curation techniques will be applied on the retrospective and prospective EHR data. This data will be stored in a centralised platform uploaded by each clinical partner. Anonymisation tools will be utilised offering an extra layer of privacy protection to the already pseudonymised data before the uploading. On ensuring overall interoperability and addressing heterogeneity and lack of shared semantics across sources, DIOPTRA will leverage widely adopted ontologies and standards, developing extensions to model relevant knowledge in the domains of the project for which no standards exist.
"DIOPTRA consortium will provide a Data Management Plan (DMP, WP1) addressing ownership, management and dissemination of all data and metadata (including guidelines & digital tools) throughout the project and after	The HL7 FHIR (Fast Healthcare Interoperability Resources) 4 standard is to be utilised for the specification of the semantically harmonised information in the data model, while other ontologies can be used for semantic enrichment of the data. DIOPTRA repository will have an

Funded by the European Union

completion, ensuring Findability, Accessibility, Interoperability and Re-Usability based on FAIR data principles in compliance with guidelines and policies of the European Commission."	interoperable and extensible design that will enable accommodation of additional (clinical and real world) datasets, targeting potential benefits well after the end of the project. In terms of the DIOPTRA models, the exact strategy for sharing models will depend on the exploitation strategy to be decided (an on-going process). If an open strategy is concluded, DIOPTRA models will consider open interchangeable formats such as OpenML, ONNX (Open Neural Network Exchange Format) or PMML (Predictive Model Markup Language). A public repository such as GitHub will be examined to facilitate reuse. As far as scientific publications are concerned, DIOPTRA may be oriented towards the Zenodo solution. Zenodo (an OpenAIRE and CERN collaboration) allows researchers to deposit publications research data by adopting the DataCite Metadata Schema (schema.datacite.org).
"Safety concerns remain due to cybersecurity threats in relation to connectivity, data sharing, etc."	Data security is a crucial issue in DIOPTRA and described in the current version of the DMP. All relevant data protection and security standards that will be employed are described in the data security section.

Table of Contents

Funded by the European Union

Funded by the European Union

List of Figures

Funded by the European Union

List of Tables

Funded by the European Union

Abbreviations

Abbreviation	Explanation
CRC	Colorectal Cancer
EHR	Electronic Health Records
DMP	Data Management Plan
DoA	Description of Action
FAIR	Findable, Accessible, Interoperable, Re-usable
GDPR	General Data Protection Regulation
ORDP	Open Research Data Pilot
WP	Work Package
N/A	Not Applicable
EU	European Union
ONNX	Open Neural Network Exchange Format
PMML	Predictive Model Markup Language
DOI	Digital Object Identifier
BCP	Basecamp project
AI	Artificial Intelligence
KPI	Key Performance Indicator
VM	Virtual Machine
DCC	Data Control Committee

Funded by the European Union

1 Introduction

This document presents the first version of the DMP for the DIOPTRA project. It provides an overview of the datasets that will be collected and generated as part of the project, as well as how these datasets will be made accessible. All projects participating in the ORDP of Horizon 2020 and Horizon Europe are mandated to have a DMP, unless they opt out by keeping some or all generated research data confidential.

Since it is anticipated that the DMP will mature throughout the project, a more developed version of the plan will be included in a respective deliverable at a later stage (M36). The DMP will be updated as significant changes occur, including (but not limited to): (i) new data, (ii) changes in consortium policies (e.g., innovation potential, decision to file for a patent), and (iii) changes in consortium composition and external factors (e.g., new consortium members joining or old members leaving).

The DMP of DIOPTRA will realise the data management regarding two types of data: on the one hand the utilisation of the research data that will be generated and collected within the context of the project, and on the other hand the dissemination of the scientific results generated from the project. It will be developed with regard to the processing of personal data on the free movement of such as dictated under the General Data Protection Regulation EU 2016/67984.

Despite the fact that DIOPTRA is not part of the Open Research Data Pilot (ORD pilot), a Data Management Plan was included in the work plan as a valuable guide for the project's data-related operations. The report adheres to the European Commission's published "Guidelines on FAIR Data Management in Horizon 2020"[1] and "EC guidelines on Data Management in Horizon Europe"[2]. Given that most of the related activities and tasks are ongoing and have not yet produced their final results, the data management plan of the project may be revised in the future.

The deliverable is structured taking into account the Horizon 2020 FAIR Data Management Plan[3] and the Horizon Europe Data Management Plan[4].

1.1 Purpose of the DIOPTRA Data Management Plan

The data management process of the DIOPTRA project consists of the steps depicted in Figure 1. The DMP of the DIOPTRA aims to describe each one of these steps and more specifically:

- the data to be collected, processed and generated.
- the methodology of data collection, processing and analysis.
- the data security and storage activities and the related standards that will be followed.
- the strategies for data to be findable, accessible, interoperable and re-usable (FAIR).
- the ethical and legal issues monitoring.
- the risks and contingency plan from risk management.

[1] http://ec.europa.eu/research/participants/data/ref/h2020/grants_manual/hi/oa_pilot/h2020-hi-oa-data-mgt_en.pdf
[2] https://open-research-europe.ec.europa.eu/for-authors/data-guidelines
[3] http://ec.europa.eu/research/participants/docs/h2020-funding-guide/cross-cutting-issues/open-access-data-management/data-management_en.htm
[4] https://enspire.science/wp-content/uploads/2021/09/Horizon-Europe-Data-Management-Plan-Template.pdf

Funded by the European Union

Figure 1: DIOPTRA Risk Management process

Throughout the project's lifetime, three studies will be conducted: a retrospective, a biomarker discovery, and a prospective study. The retrospective and biomarker discovery studies will produce the homogenised DIOPTRA dataset. This dataset will serve as the foundation for analysis and model training. The processed data are the outcomes of the Risk Factor Analysis, the Biomarker discovery and the AI modelling. In addition, the mobile application will be used to monitor and generate specific data for the prospective study.

Funded by the European Union

2 Inter-project Collaboration

2.1 "Prevention, including screening" cluster

Prevention, screening and early detection is the most cost-efficient and long-term cancer control strategy. It is known that 40% of cancers could be prevented, but a more personalised understanding of the disease is needed as well as improvements in the existing prevention programmes and general health literacy in Europe and across the globe. In this context and within the framework of Horizon Europe, the European Commission launched the Mission Cancer to undertake concrete actions with the ambition of delivering tangible results by 2030. The initiative is supported by the establishment of clusters dedicated to the main objectives of the Mission, namely:

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

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

These projects are the following:

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

Funded by the European Union

The main goal of the "Prevention and early detection" cluster (hereafter simply referred to as Prevention cluster) is therefore to support the EU Cancer Mission, create added value, establish a policy feedback loop and increase the impact of the EU funding.
The projects in the Prevention cluster fully adopt the European Commission's views on encouraging inter-project collaborations, and thus will act in the systematic promotion of knowledge sharing. Particularly, the Cluster members must be aligned in performing joint tasks and activities such as the synergistic collaboration for the respective Data Management Plans production. The latter include the drafting of this common chapter. The leader of the task is OncoScreen.

The projects within the cluster work on:

1. the integration of retrospective information from European registries, cohort studies and biobanks (including from clinical partners of the projects) on different types of cancer with prospective data to complement missing features. (OncoScreen, PANCAID, DIOPTRA, LUCIA)

2. prospective data for the clinical validation of new technologies for cancer screening (MammoScreen, ThermoBreast, DIOPTRA, ONCOSCREEN, LUCIA and SANGUINE)

3. AI based or AI enabled analysis and screening methods

Specifically, the projects will process inter alia:

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

2.2 FAIR data management

The data used and generated in the Prevention cluster projects can be useful beyond these projects, most notably to healthcare professionals and cancer researchers wishing to understand risk factors and to diagnose cancer at earlier stages. As such, the data generated within one of the projects can be of substantial use to the other projects from the cluster as well (of course to the extent permitted by confidentiality and intellectual property related provisions of the respective Consortium Agreements).
The purpose of this common chapter is to find common practices to share the information in pan-European research infrastructures, such as the European biobanking platform (BBMRI-ERIC) or the future UNCAN.eu platform, a federated cancer data hub platform currently under development. This is a particularly critical point, as at the present time patient health data

Funded by the European Union

networks in Europe show a high level of heterogeneity in terms of involvement of EU Member States as well as the types and interoperability of collected data, organisation and governance of data storage, security or the possibility to use this data for research purposes.
As such, the cluster projects are committed to manage their data in accordance with the FAIR principles and in full compliance with all the applicable European and national legislation. A close collaboration between the projects will be established in this regard, so as to address commonalities on data standards, data validation, the best practices regarding data privacy (pseudonymisation or anonymisation techniques), data storage and data exchange protocols.
The projects plan on implementing individual measures to make the data **findable, accessible, interoperable and re-usable.**
But in general, the projects consider possibilities for joint exploitation of data within the cluster. The following possibilities are being examined:

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

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

- Organisation of regular cluster meetings on the topic of data management
- Creation of a data management task force (including a representative from each of the projects, ThermoBreast being the coordinator of the joint cluster efforts) which will act as a working group to discuss and agree on the common aspects related to the management of the data generated in each of the projects (standards, validation protocols, privacy, storage, etc.) as well as on how to foster data exchange between the cluster projects. This task force will also monitor and update the DMPs throughout the duration of the cluster projects.

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

2.3 Structure of the Document

In an overall view, section 3 provides an overview of the data types that will be utilised throughout the project's lifetime. The alignment with the FAIR principles is provided in section 4. The 5th section presents a description of the DIOPTRA data. In this section, a summary of the datasets that will be utilised during the DIOPTRA project is provided, along with the data acquisition, processing, and analysis methodology that will be employed. This section also provides information regarding the data storage infrastructure. The operations for legal and ethical compliance are described in section 6 and the risks that may arise during the project in section 7. Finally, the responsibilities and resources are presented in section 8 and the conclusion of the deliverable in section 9.

Funded by the European Union

3 Data Introduction

During the DIOPTRA project, data will be collected from each WP of the project, each one facilitating a specific aim.

- The data collected/extracted within WP1 contain information related to the project management and coordination.
- The data collected/extracted within WP2 contain information related to the project requirements, conceptual architecture and standards that will be followed for the design and development of DIOPTRA.
- The data collected/extracted within WP3 contain information related to the retrospective data that will be provided by clinical partners.
- The data collected/extracted/generated within WP4 & WP5 contain information related to the analysis of DIOPTRA biological samples and dataset.
- The data collected/extracted within WP6 contain information related to the prospective data collected on clinical sites, as part of DIOPTRA system evaluation, training, validation and expandability survey.
- The data collected/extracted within WP7 contain information related to the project dissemination and exploitation.

3.1 Data Types

The data that will be collected during the **WP1** will be:

- Partners contact list
- Financial statements
- Risk log
- Managerial documents

The data that will be collected during the **WP2** will be:

- Requirements list
- Standards for the design and development

The data that will be collected during the **WP3** will be:

- Participant's retrospective data (EHR)
  - **Demographic data**
  - **Behavioural data (e.g. lifestyle and diet)**
  - **Medical history information**
  - **Family history information**
  - **Symptoms**
  - **Diagnosis**

Funded by the European Union

- Biological Samples
  - **Blood and tissue samples**

The data that will be collected during the WP4 & WP5 will be:

- Obtained results from the analysis of the biological samples and data

The data that will be collected during the **WP6** will be:

- Participant's prospective data with follow-up
  - **Demographic data**
  - **Behavioural data (e.g. lifestyle and diet)**
  - **Medical history information**
  - **Colonoscopy procedure data**
  - **Family history information**
  - **Symptoms**
  - **Diagnosis**
- Biological Samples
  - **Blood samples**
  - **Protein biomarker measurements from blood sample analysis**
- Feedback from clinical partners (observations, notes)

The data that will be collected during the **WP7** will be:

- Communication KPIs
- Dissemination materials (website, brochures, posters, newsletter, press releases, video etc.)
- Exploitation plan
- Contact details of linked initiatives
- IPR data

A detailed description of the datasets in terms of[5]:

- What is the purpose of the data collection/generation and its relation to the objectives of the project?
- What types and formats of data will the project generate/collect?
- Will you re-use any existing data and how?

[5] http://ec.europa.eu/research/participants/docs/h2020-funding-guide/cross-cutting-issues/open-access-data-management/data-management_en.htm

- What is the origin of the data?
- What is the expected size of the data?
- To whom might it be useful ('data utility')?

is presented in Section 5.

4 Alignment to the Findable, Accessible, Interoperable, Re-usable (FAIR) data principles

The purpose of this Data Management Plan is to outline the strategies and approaches for aligning the project with the Findable, Accessible, Interoperable, and Reusable (FAIR) data principles (Figure 2). The FAIR data principles aim to maximise the value and impact of research data by promoting its discoverability, accessibility, interoperability, and reusability. This section describes the strategies and approaches that will be implemented to align the project with these principles, considering that the data from the project are not in the ORD pilot and that open access will be limited to scientific publications.

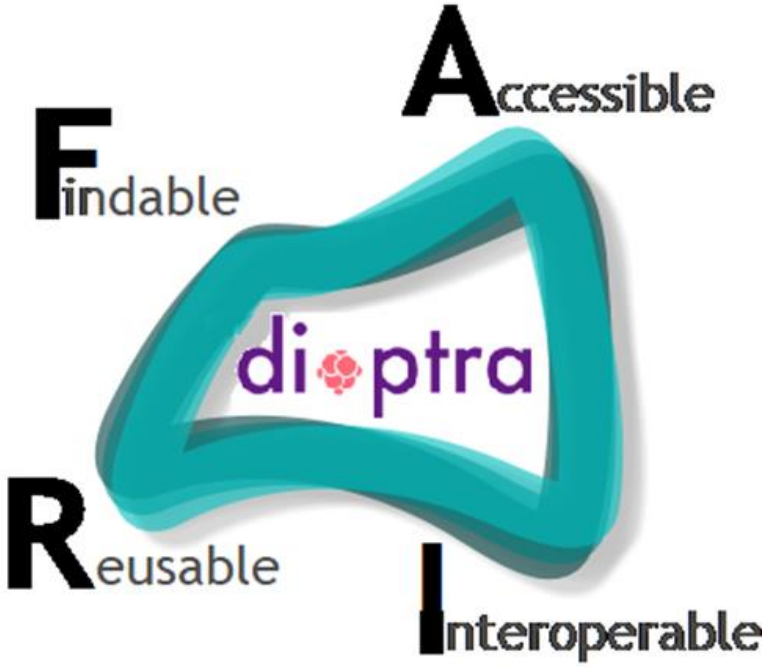

Figure 2: DIOPTRA FAIR data principles.

The DIOPTRA consortium is the end-responsible for the Data Management Plan and makes the research data findable, accessible, interoperable and reusable (FAIR), towards ensuring appropriate data management. FAIR data ensure knowledge discovery, innovation, data extension innovation and reuses. As far as patient data is concerned, due to limitations imposed by the study protocols (for retrospective and prospective data) and of the data sharing agreements of between the data providers and the data processor, the DIOPTRA consortium is considering that data sharing and reuse of these data is not possible. In this sense, the DIOPTRA consortium is considering opting out of the Open Research Data pilot (ORD pilot).

On ensuring overall interoperability and addressing heterogeneity and lack of shared semantics across sources, DIOPTRA will leverage widely adopted ontologies and standards, developing extensions to model relevant knowledge in the domains of the project for which no standards exist. The HL7 FHIR (Fast Healthcare Interoperability Resources) 4 standard is to be utilised for the specification of the semantically harmonised information in the data model,

Funded by the European Union

while other ontologies can be used for semantic enrichment of the data. FHIR specification, which is a standard for exchanging healthcare information electronically, is implemented on top of HL7 standard. DIOPTRA repository will have an interoperable and extensible design that will enable accommodation of additional (clinical and real world) datasets, targeting potential benefits well after the end of the project.
In terms of the DIOPTRA models, the exact strategy for sharing models will depend on the exploitation strategy to be decided (an on-going process). If an open strategy is concluded, DIOPTRA models will consider open interchangeable formats such as OpenML, ONNX (Open Neural Network Exchange Format) or PMML (Predictive Model Markup Language). A public repository such as GitHub will be examined to facilitate reuse.
As far as scientific publications are concerned, DIOPTRA may be oriented towards the Zenodo[6] solution. Zenodo (an OpenAIRE and CERN collaboration) allows researchers to deposit publications research data by adopting the DataCite Metadata Schema (schema.datacite.org). To support the FAIR principles using Zenodo, the following practices will be followed in DIOPTRA (as reported in https://about.zenodo.org/principles/).

4.1 Making data findable, including provisions for metadata

The first step of the FAIRification process for reusing the data is to easily find them inside the large data pools. Thus, both metadata and data should be easily recognisable by both humans and machines.

Table 1: Findable data

Findable	❖ A Digital Object Identifier (DOI) is assigned to every published record on Zenodo enabling researchers to easily locate and reference them. ❖ Zenodo's metadata is compliant with DataCite's Metadata Schema minimum and recommended terms, with a few additional enrichments. ❖ The DOI is a top-level and a mandatory field in the metadata of each record. ❖ Metadata of each record is indexed and searchable in Zenodo's search engine immediately after publishing. ❖ Metadata of each record is sent to DataCite servers during DOI registration and indexed there.

4.2 Making data accessible

This principle refers to easy access of the data from the user. While open access to the data is not feasible, the datasets remain accessible within the boundaries of relevant ethical and legal frameworks.

[6] http://www.zenodo.com/

Table 2: Accessible data

Accessible	❖ Metadata for individual records as well as record collections are harvestable using the OAI-PMH protocol by the record identifier and the collection name. ❖ Metadata is also retrievable through the public REST API. ❖ OAI-PMH and REST are open, free and universal protocols for information retrieval on the web. ❖ Metadata are publicly accessible and licensed under public domain. No authorisation is ever necessary to retrieve it. ❖ Data and metadata will be retained for the lifetime of the repository. ❖ This is currently the lifetime of the host laboratory CERN, which currently has an experimental programme defined for the next 20 years at least. ❖ Metadata are stored in high-availability database servers at CERN, which are separate from the data itself.

4.3 Making data interoperable

To promote interoperability, the data will be stored in standard, widely used formats that are compatible with existing tools and resources. Although the data may not be openly available, detailed information will be provided on the data formats and structures in the scientific publications and accompanying documentation. This will enable researchers to understand the data structure and facilitate integration with other datasets through well-defined mappings and transformations. By adhering to established standards and providing clear guidelines, interoperability is promoted and encourages collaboration and data integration across different research domains.

Table 3: Interoperable data

Interoperable	❖ Zenodo uses JSON Schema as internal representation of metadata and offers export to other popular formats such as Dublin Core or MARCXML. ❖ For certain terms Zenodo refers to open, external vocabularies, e.g.: license (Open Definition), funders (FundRef) and grants (OpenAIRE). ❖ Each referenced external piece of metadata is qualified by a resolvable URL.

4.4 Increase data re-use (through clarifying licenses)

The highest goal of the FAIRification process is to optimise the reuse of data. Although direct access to the datasets may be restricted, comprehensive documentation, including data dictionaries, data collection methodologies, and data processing workflows, will be provided alongside the datasets to ensure their understandability and enable reproducibility. This documentation will aid researchers in understanding the data. Furthermore, clear data usage

Funded by the European Union

policies and licensing information will be established to define the terms of data reuse, fostering transparency and trust among potential data users. By clearly defining the conditions for data reuse, responsible and ethical data usage is promoted. Finally, data preservation strategies, including long-term archiving and digital preservation techniques will be employed to ensure the dataset's long-term accessibility and reusability beyond the project's lifespan.

Table 4: Reusable data

Reusable	❖ Each record contains a minimum of DataCite's mandatory terms, with optionally additional DataCite recommended terms and Zenodo's enrichments. ❖ License is one of the mandatory terms in Zenodo's metadata and is referring to an Open Definition license. ❖ Data downloaded by the users is subject to the license specified in the metadata by the uploader. ❖ All data and metadata uploaded is traceable to a registered Zenodo user. ❖ Metadata can optionally describe the original authors of the published work. ❖ Zenodo is not a domain-specific repository, yet through compliance with DataCite's Metadata Schema, metadata meets one of the broadest cross-domain standards available.

Funded by the European Union

5 Data Description

5.1 Data Summary

The data management plan (DMP) of the DIOPTRA project will focus on the following categories of clinical data:

- Participants retrospective data
  - **Demographic data**
  - **Behavioural data**
  - **Medical history information**
  - **Family history information**
  - **Symptoms**
  - **Diagnosis**
  - **Biological samples (tissue & blood)**
    - Gene expression data generated from the analysis of tissue samples
    - Protein expression data generated from the analysis of blood samples
  - **Biological samples analysis data**
- Participants prospective data with follow-up:
  - **Demographic data**
  - **Medical history information**
  - **Family history information**
  - **Symptoms**
  - **Behavioural data**
  - **Diagnosis**
  - **Biological samples (blood samples)**
    - protein expression data generated from the analysis of blood samples

All these types of data and the implementation of the project will be based on four groups.

- Group I: Healthy
  - **No findings after a colonoscopy or presence of hyperplastic polyps**
- Group II: Non - Advanced Adenomas

Funded by the European Union

- Group III: Advanced Adenomas
- Group IV: Colorectal Cancer (CRC) patients
  - **CRC stage I, II, III**
  - **CRC diagnosis with or without surgery, without neo-adjuvant chemotherapy and/or radiotherapy treatment.**

5.1.1 Retrospective data (EHR)

One of the main achievements targeted by the DIOPTRA project is to interlink data from 8 highly heterogeneous local, regional, national and international cohorts into a dataset that will be utilised for predicting the early onset of the Colorectal Cancer (CRC). The retrospective data will be provided by eight clinical sites, partners of the DIOPTRA consortium:

- BLOKS ZDRAVNI I SOTSIALNI GRIZHI EOOD (BLOCKS), patients' data coming from the pilot site in Bulgaria.
- CENTRE HOSPITALIER UNIVERSITAIRE DE LIEGE (CHUL), patients' data coming from the pilot site in Belgium.
- REGION MIDTJYLLAND (RM-RRH), patients' data coming from the pilot site in Denmark.
- Univerzitetni klinicni center Maribor (UKCM), patients' data coming from the pilot site in Slovenia.
- FUNDACION BURGOS POR LA INVESTIGACION DE LA SALUD (BURGOS), patients' data coming from the pilot site in Spain.
- ETHNIKO KAI KAPODISTRIAKO PANEPISTIMIO ATHINON (NKUA), patients' data coming from the pilot site in Greece.
- LINAC-PET SCAN OPCO LIMITED (GOC), patients' data coming from the pilot site in Cyprus.
- GENIKO ANTIKARKINIKO OGKOLOGIKO NOSOKOMEIO ATHINON O AGIOS SAVVAS (AGSAVVAS), patients' data coming from the pilot site in Greece.

The combined dataset will contain variables from EHR of participants, including demographics, medical and family history, symptoms, diagnosis etc. The dataset will consist of well-structured and properly annotated variables, with defined value-range and units. The clinical variables are based on terminology standards in health such as SNOMED or LOINC with corresponding codes.

5.1.2 Retrospective data (Biological samples)

For the biomarker discovery study, biological samples (blood & tissue) will include cryopreserved serum, plasma and tissue samples along with the clinical information of these samples and will be utilised for protein biomarkers screening. The clinical information has already been processed exclusively in a pseudonymised form using clear identifiers or area-specific personal identifiers in line with the requirement of European Data Protection Law. The dataset produced will be integrated into retrospective dataset. The outcome will be a homogenised dataset that will be processed and further analysed.

Funded by the European Union

5.1.3 Prospective data

The prospective data of DIOPTRA will be similar to the retrospective data. During the prospective study all the above-mentioned clinical partners will participate in acquiring all variables needed during the evaluation and validation of the DIOPTRA system. The prospective dataset will include biological samples, variables defined in DIOPTRA clinical protocol (D6.2) and data from questionnaires and DIOPTRA mobile application. During prospective study, variable list will be submitted to change due to refinement of the DIOPTRA system and will be documented in updated DIOPTRA clinical protocol (D6.3).

5.1.4 Other types of data

Throughout the duration of the DIOPTRA also some other types of data will be collected:

- Scientific publications
- DIOPTRA mobile application

5.1.4.1 Scientific publications

In this category data concerning publications resulting from the scientific and technical work performed within the project lifecycle and presented in peer review journals and conferences will be collected, along with the source files of the publications (journal papers, conferences papers, and posters). The information about the specific category of data is presented in the deliverable D7.1 Dissemination and exploitation plan (will be delivered on M6).

5.1.4.2 DIOPTRA Mobile Application

The mobile application will be utilised during the prospective study. The purpose of the mobile application will be to collect structured risk factors from recruited participants. The behavioural questionnaire containing data that will be used as input for the risk factor module to emphasise the risks associated with the early onset of CRC. The user will receive assisted suggestions for a healthy lifestyle based on the module's output. The behavioural modification suggestions will be according to medical guidelines and validated by clinicians.

5.2 Purpose of the DIOPTRA data collection

The purpose of data collection/generation in this project is to gather comprehensive information on colorectal cancer and its risk factors, as well as to identify potential biomarkers for its early screening. The collected data will be instrumental in achieving the overall objectives of the project, which are to develop effective strategies for early detection and prevention of colorectal cancer. Firstly, retrospective and prospective data obtained from EHRs will be analysed. These records contain valuable clinical information, such as demographic, medical and family history and symptoms related to Colorectal Cancer (CRC) among others. By analysing these data, the aim is to identify significant risk factors associated with colorectal cancer development. This information will help us to understand the demographic, genetic, lifestyle, and environmental factors that contribute to the disease's

early onset and progression. Additionally, the project involves biomarker discovery using blood and tissue samples. The purpose of generating these samples is to extract biomarkers, specifically protein features, which can potentially serve as indicators or predictors of early onset colorectal cancer. The biomarker discovery process involves the analysis of these biological samples, to identify specific protein patterns or signatures that correlate with the presence or progression of colorectal cancer. The identification of reliable biomarkers will aid in the development of accurate and efficient early screening tests for CRC.

The data collection and generation activities are directly aligned with the project's objectives. By analysing the retrospective and prospective data from EHRs, the aim is to identify and validate risk factors associated with colorectal cancer. This knowledge will enhance the understanding of the disease's aetiology and enable the development of targeted prevention strategies. Identifying high-risk populations and risk factor associations will contribute to the project's goal of reducing the incidence and mortality rates of colorectal cancer through early intervention and personalised approaches. Similarly, the generation of biomarker data from blood and tissue samples aligns with the objective of developing early screening methods. By extracting and analysing protein features, the aim is to discover robust biomarkers that can serve as reliable indicators of colorectal cancer. Such biomarkers have the potential to revolutionise the early detection process, allowing for timely intervention and improved patient outcomes. The integration of biomarker discovery with risk factor analysis will enhance the overall effectiveness of the project's strategies for colorectal cancer prevention and early screening.

In summary, the purpose of data collection/generation in this project is to acquire comprehensive information on colorectal cancer risk factors and extract biomarkers for early screening. These activities directly support the project's objectives of understanding disease aetiology, developing targeted prevention strategies, and advancing early detection methods. The data collected/generated will serve as the foundation for evidence-based decision-making and the development of innovative approaches to combat colorectal cancer.

5.3 Methodology for data collection

Table 5 presents a first estimation of the methodology that will be followed for the collection of each one of the DIOPTRA datasets.

Table 5: First estimation of the datasets collection methodology

Datasets of WP1	**The datasets of WP1 contain information related to the project management and coordination.**	
DS1.1	Partners contact list	The data were collected during the initiation phase of the consortium, recorded in an excel file stored in the project's private repository. It will be updated once a modification of the project team appears.
DS1.2	Financial statements	The financial statements information will be collected in M18, M36 and M48 of the project following the template provided by the coordinator.

Funded by the European Union

		They will be stored in the project's private repository and, they will be included in the interim periodic reported deliverables that will be uploaded to the ECAS portal.
DS1.3	Risk log	The Risk Log for the project will be prepared using the PM2 Risk Log template. It will be stored in the project's private repository, in the corresponding Basecamp project (BCP) (Task 1.4: Risk management & Quality assurance).
DS1.4	Managerial documents	Managerial documents include the interim periodic progress report documents, the final report of the project, the deliverables, the milestones, the minutes of the conference call and physical meetings, the WP monthly reports etc. The documents are collected as it is planned in the DoA as far as it concerns the reports, deliverables and milestones, the minutes are collected one week after the end of the conference call and physical meetings, while WP progress reports are prepared by the WP leader and are collected by the coordinator at the end of each month. The managerial documents are stored in the project's private repository, they will be uploaded in the ECAS portal (deliverables, milestones, reports) and they will be available in the site of the project, in case the dissemination level of the documents allow it.
Datasets of WP2	The datasets of WP2 contain information related to the project requirements, conceptual architecture and standards that will be followed for the design and development of DIOPTRA.	
DS2.1	Requirements	For the collection of the DIOPTRA requirements a list will be created by subgroups within the consortium for examining perspectives of clinical partners, researchers, technical partners and regulatory bodies. The list of the requirements, along with detailed information will be provided in a form of a document and it will be available to the consortium partners through private repository and corresponding BCP.
DS2.2	Standards for the design and development	These are the necessary and adequate standard for the requirements collection and hardware/software elicitation. The standards will be determined during the WP2 and will be updated during the development phase of the project. Technical partners of the DIOPTRA consortium will be responsible for this action.
Datasets of WP3	The datasets of WP3 contain information related to the retrospective data that will be provided by the clinical partners.	
DS3.1	EHR data	The Electronic Health Records (EHR) data will be collected from 8 clinical sites.

Funded by the European Union

DS3.2	Biological samples	The biological samples (tissue & blood) will be transferred from the DIOPTRA clinical partners to PAO for analysis.
DS3.3	Variable list	Definition of a list of variables that will be collected and will aid in gathering same data from the clinical partners.
DS3.4	Interface	Clinical interface will be developed, where all the anonymised/pseudonymised data will be uploaded by the clinical partners.
DS3.5	Data curation	The curation techniques will be achieved by analysing all the collected retrospective data for missing values, outliers etc.
DS3.6	Data storage and management	Use appropriateness of the required infrastructure and protocols for data storage and management (e.g., exchange and analysis). Endpoints will be developed for data retrieval.
Datasets of WP4		The datasets of WP4 contain information related to the associated analysis of the data.
DS4.1	Risk Factor Analysis	The curated EHR data will be retrieved from the clinical interface and utilised for the risk factor analysis, where features on the effect of risk factors will be examined and will be used as input to the AI modelling phase of WP5.
DS4.2	Biomarker Discovery	Serum/Plasma samples from individuals belonging to the four study groups will be analysed with the Olink Explore platform to measure the expression of 3000 proteins. Additionally, tissue samples will be analysed with an NGS pipeline to measure gene expression (~20000 genes). The resulting datasets will be analysed with computational methods to produce the protein biomarker hits.
Datasets of WP5		The datasets of WP5 contain information related to the associated analysis of the data.
DS5.1	AI modelling	The examined risk factors along with the extracted proteins will be the input for the AI models, which will propose a small subset of proteins based on a cost-effective screening protocol.
DS5.2	Mobile Application	Based on the outputs of the WP2 and WP4 the mobile application will collect structured risk factors for providing AI assisted suggestions to the user.
Datasets of WP6		The datasets of the WP6 contain information related to the prospective data that will be collected in the clinical sites, as part of DIOPTRA system evaluation, validation and expendability survey.
DS6.1	EHR data	The Electronic Health Records (EHR) data will be collected from 8 clinical sites.
DS6.2	Questionnaires	Data from behavioural and colonoscopy procedural questionnaires will be collected during prospective study at baseline and one year follow-up.

Funded by the European Union

DS6.3	Variable list	Definition of the variable list which contains the clinical information (variables) that will be collected.
DS6.4	Biological samples	All clinical sites will provide blood samples during prospective study at baseline and one year follow-up. Blood samples will be analysed with the xMAP platform (Luminex) to measure the expression of protein biomarkers that were discovered in WP4.
Datasets of WP7	The datasets of WP7 contain information related to the project dissemination and exploitation.	
DS7.1	Communication KPIs	The list of communication KPIs will be completed by the dissemination manager in close cooperation with the partners.
DS7.2	Dissemination and exploitation plan	Through the whole lifecycle of the project a list of the dissemination, communication, and exploitation activities along with the content of the dissemination materials will be collected and stored in the private document repository of the project while it will be uploaded on the website and the social media of the project.

5.4 DIOPTRA datasets

The convention followed for naming the project datasets, it should be noted that the name of each dataset comprises:

1. A prefix "DS" indicating a dataset.
2. Its unique identification number depending on the WP the dataset comes from, e.g., "DS1" for datasets coming from WP1, “DS2” for datasets coming from WP2 etc.
3. A serial number restarting at 1 for each WP indicating the sub-dataset comes from the specific WP: "DS1.1", “DS1.2” etc.
4. A short name indicative of its content and purpose. e.g., “DS1.1_managerial documents”.

5.4.1 Summary of the datasets

Tables 6-12 present a short description of the content of the DIOPTRA datasets.

Table 6: Datasets of WP1.

Datasets of WP1	The datasets of WP1 contain information related to the project management and coordination.	
DS1.1	Partners contact list	This dataset contains the detailed consortium contact information.
DS1.2	Financial statements	This dataset contains the financial statement log file describing the financial statement reports along with a

		small description.
DS1.3	Risk log	This dataset contains the identified risks from the beginning of the project accompanied with the mitigation plans.
DS1.4	Managerial documents	This dataset contains a list of the managerial documents that will be prepared within the lifecycle of the project.

Table 7: Datasets of WP2.

Datasets of WP2	The datasets of WP2 contain information related to the project requirements, conceptual architecture and standards that will be followed for the design and development of DIOPTRA.
DS2.1	Requirements	This dataset contains the list of the identified requirements.
DS2.2	Standards for the design and development	This dataset contains the list of all necessary standards for the requirements collection and hardware/software elicitation.

Table 8: Datasets of WP3.

Datasets of WP3	The datasets of WP3 contain information related to the retrospective data that will be provided by the clinical partners.
DS3 Retrospective data	
DS3.1	EHR data	The EHR data will be collected from 8 clinical sites and will include: ✓ Demographics ✓ Medical history ✓ Family history ✓ Symptoms, clinical and anatomo-pathological diagnoses ✓ Clinical biology ✓ Status of medication
DS3.2	Biological samples	Detailed inventory of biological samples that will be used for biomarker discovery. The inventory will include detailed information of the blood and tissue samples (e.g. storage variables, remaining volume, aliquots, etc.) along with the diagnostic variables.
DS3.3	Variable list	Definition of a list of all variables that will be collected during retrospective study.

DS3.4	Interface	Clinical interface, where all the anonymised/pseudonymised data will be uploaded by the clinical partners and will be stored.
DS3.5	Data curation	The curation techniques will be achieved by analysing all the collected retrospective data for missing values, outliers etc.

Table 9: Datasets of WP4.

Datasets of WP4	The datasets of WP4 contain information related to the associated analysis of the data.	
DS4.1	Risk Factor Analysis	The curated EHR data will be utilised for the risk factor analysis, where features on the effect of risk factors will be examined and will be used as input to the AI modelling phase of WP5.
DS4.2	Biomarker Discovery	The biomarker discovery dataset will contain: • The protein expression measurements. • The gene expression measurements. • The identified biomarker hits.

Table 10: Datasets of WP5.

Datasets of WP5	The datasets of WP5 contain information related to the associated analysis of the data.	
DS5.1	AI modelling	The examined risk factors along with the extracted proteins will be the input for the AI models, which will propose a small subset of proteins based on a cost-effective screening protocol.
DS5.2	Mobile Application	Based on the outputs of the WP2 and WP4 the mobile application will collect structured risk factors for providing AI assisted suggestions to the user.

Table 11: Datasets of WP6.

Datasets of WP6	The datasets of the WP6 contain information related to the prospective data that will be collected in the clinical sites, as part of DIOPTRA system evaluation, validation and expendability survey.	
DS6 Prospective data		
DS6.1	EHR data	The EHR data will be collected during prospective study. Variables collected will subject to changes depending on the refinement of DIOPTRA system.

Funded by the European Union

DS6.2	Questionnaires	Questionnaires for gathering risk factors and family history will be provided during the prospective study.
DS6.3	Variable list	Definition of a list of all variables that will be collected during prospective study.
DS6.4	Biological samples	All clinical sites will provide biological samples for the clinical study. The dataset of biological samples will contain a detailed inventory of the collected samples. Additionally, the dataset will contain the measurements of protein biomarkers.

Table 12: Datasets of WP7.

Datasets of WP7	The datasets of WP7 contain information related to the project dissemination and exploitation.
DS7.1	Communication KPIs	The list of communication KPIs will be completed by the dissemination manager in close cooperation with the partners.
DS7.2	Dissemination and exploitation plan	Through the whole lifecycle of the project a list of the dissemination, communication, and exploitation activities along with the content of the dissemination materials will be collected and stored in the private document repository of the project, while it will be uploaded on the website and the social media of the project.

5.4.2 Datasets description

Tables 13 – 30 provide detailed information for each one of the datasets in terms of: i) generic description, ii) origin of data, iii) nature and scale of data, iv) to whom the dataset could be useful, v) related scientific publications, vi) indicative existing similar data sets, vii) partners activities and responsibilities, viii) standards and metadata, ix) data exploitation and sharing, x) archiving and preservation.

Table 13: DS1.1_ Partners Contact List description.

Data identification: *DS1.1_ Partners Contact List*
Generic description
The contact details of the persons representing each partner organisation in the DIOPTRA project and participating in each WP and task. Contact details include telephone number, skype name and email address.

Origin of data	
The data was collected at the beginning of the project and they will be updated once a change in the personnel of each organisation (DIOPTRA partner) takes place.	
Nature and scale of data	
Spreadsheet data.	
To whom the dataset could be useful	
All partners	
Related scientific publication(s)	
N/A	
Indicative existing similar data sets (including possibilities for integration and reuse)	
N/A	
Partners activities and responsibilities	
Partner owner of the data	DIOPTRA consortium
Partner in charge of the data analysis	DIOPTRA consortium
Partner in charge of the data storage	ICCS
Related WP(s) and task(s)	All
Standards and metadata	
Standards, format, estimated volume of data	Spreadsheet data.
Data exploitation and sharing	
Data access policy/ Dissemination level: confidential (only for members of the Consortium and the Commission Services) or Public	Confidential (only for members of the Consortium and the Commission Services)
Data sharing, re-use, distribution, publication (How?)	Shall be limited only to be carried out between the Project Consortium

	members and the European Commission's Services.
Personal data protection: are they personal data? If so, have you gained (written) consent from data subjects to collect this information?	Personal data is involved. All the partners agreed on this during the kick off meeting of the project.
Access Procedures	None within the project consortium
Embargo periods (if any)	None
Archiving and preservation (including storage and backup)	
Data storage (including backup): where? For how long?	On project private file repository. Shall be maintained and backed up for a period of 4 years until the end of the project.
Indicative associated costs for data archiving and preservation	N/A
Indicative plan for covering the above costs	N/A

Table 14: DS1.2_ Financial statements description.

Data identification: *DS1.2_ Financial statements*
Generic description:
The financial information of each partner of the consortium will be included in the DS1.2 dataset. They will include information about the personnel cost, the justification of travel, the justification of Equipment, the justification of other goods and services, the justification of sub-contracting, the justification of linked third parties and the Justification of contributions linked third parties.
Origin of data:
The information will be provided by each partner to the coordinator along with the interim progress report on M18, M36 and M48.
Nature and scale of data:

Spreadsheet data.	
To whom the dataset could be useful:	
All partners, European Commission.	
Related scientific publication(s)	
N/A	
Indicative existing similar data sets (including possibilities for integration and reuse):	
N/A	
Partners activities and responsibilities	
Partner owner of the data	DIOPTRA consortium
Partner in charge of the data analysis	ICCS
Partner in charge of the data storage	ICCS
Related WP(s) and task(s)	WP1
Standards and metadata	
Standards, format, estimated volume of data	Excel format files.
Data exploitation and sharing	
Data access policy/ Dissemination level: confidential (only for members of the Consortium and the Commission Services) or Public	Confidential (only for members of the Consortium and the Commission Services)
Data sharing, re-use, distribution, publication (How?)	Shall be limited only to be carried out between the Project Consortium members and the European Commission's Services.
Personal data protection: are they personal data?	No personal data.

If so, have you gained (written) consent from data subjects to collect this information?	
Access Procedures	None within the project consortium
Embargo periods (if any)	None
Archiving and preservation (including storage and backup)	
Data storage (including backup): where? For how long?	On project private file repository and European Commission portal. Shall be maintained and backed up for a period of 4 years until the end of the project.
Indicative associated costs for data archiving and preservation	N/A
Indicative plan for covering the above costs	N/A

Table 15: *DS1.3_RiskLog description.*

Data identification: *DS1.3_RiskLog*
Generic description:
A description of the risk, its causes, the kinds of problems that it could result in (potential effects), and risk dependencies.
Origin of data:
Several potential risks have been identified with direct or indirect impact on DIOPTRA project. Risks are grouped into three categories: a) general and administrative, b) technical and scientific, c) exploitation and dissemination. Each partner will estimate and evaluate the associated risks and the respective controls and will monitor the effectiveness of the controls.
Nature and scale of data:
The *Risk Log* for the project is using PM^2 *Risk Log* template and no changes have been done to the structure, fields or values.
To whom the dataset could be useful:

Executive Board and Project Core	
Related scientific publication(s)	
N/A	
Indicative existing similar data sets (including possibilities for integration and reuse):	
N/A	
Partners activities and responsibilities	
Partner owner of the data	CMA
Partner in charge of the data analysis	CMA
Partner in charge of the data storage	CMA
Related WP(s) and task(s)	All WPs
Standards and metadata	
Standards, format, estimated volume of data	PM^2 *Risk Log* template
Data exploitation and sharing	
Data access policy/ Dissemination level: confidential (only for members of the Consortium and the Commission Services) or Public	Confidential (only for members of the Consortium and the Commission Services)
Data sharing, re-use, distribution, publication (How?)	Shall be limited only to be carried out between the Project Consortium members and the European Commission's Services.
Personal data protection: are they personal data? If so, have you gained (written) consent from data subjects to collect this information?	No personal data is involved.
Access Procedures	None within the project consortium

Embargo periods (if any)	None
Archiving and preservation (including storage and backup)	
Data storage (including backup): where? For how long?	On CMA Servers. Shall be maintained and backed up for a period of 4 years until the end of the project.
Indicative associated costs for data archiving and preservation	N/A (part of company's regular backup)
Indicative plan for covering the above costs	N/A

Table 16: DS1.4_ Managerial documents description.

Data identification: *DS1.4_ Managerial documents*	
Generic description:	
This dataset includes information related to the project management and coordination.	
Origin of data:	
The information will be collected throughout the whole lifecycle of the project.	
Nature and scale of data:	
Documents in excel, word and pdf format.	
To whom the dataset could be useful:	
All partners, European Commission.	
Related scientific publication(s)	
N/A	
Indicative existing similar data sets (including possibilities for integration and reuse):	
N/A	
Partners activities and responsibilities	
Partner owner of the data	DIOPTRA consortium

Partner in charge of the data analysis	ICCS
Partner in charge of the data storage	ICCS
Related WP(s) and task(s)	WP1
Standards and metadata	
Standards, format, estimated volume of data	Excel format files. Word format files. PDF format files Volume ~1GB
Data exploitation and sharing	
Data access policy/ Dissemination level: confidential (only for members of the Consortium and the Commission Services) or Public	Confidential (only for members of the Consortium and the Commission Services)
Data sharing, re-use, distribution, publication (How?)	Shall be limited only to be carried out between the Project Consortium members and the European Commission's Services.
Personal data protection: are they personal data? If so, have you gained (written) consent from data subjects to collect this information?	No personal data.
Access Procedures	None within the project consortium
Embargo periods (if any)	None
Archiving and preservation (including storage and backup)	
Data storage (including backup): where? For how long?	On project private file repository and European Commission portal. Shall be maintained and backed up for a period of 4 years until the end of the project.

Funded by the European Union

Indicative associated costs for data archiving and preservation	N/A
Indicative plan for covering the above costs	N/A

Table 17: DS2.1_ Requirements description.

Data identification: *DS2.1_ Requirements*	
Generic description:	
The DS2.1 describes the requirements that will be collected and analysed in order to specify the necessary architecture of the DIOPTRA.	
Origin of data:	
The data will be derived from specific needs of each partner of the consortium.	
Nature and scale of data:	
This dataset will be included in a spreadsheet where the requirements are described.	
To whom the dataset could be useful:	
DIOPTRA consortium	
Related scientific publication(s)	
N/A	
Indicative existing similar data sets (including possibilities for integration and reuse):	
List of standards (DS2.2_Standards for the design and development).	
Partners activities and responsibilities	
Partner owner of the data	ICCS
Partner in charge of the data analysis	ICCS, AGSAVVAS, NKUA, TERAGLOBUS, UOI, PAO, HOPE, DHCE, INTRA, TCR
Partner in charge of the data storage	ICCS
Related WP(s) and task(s)	WP2, WP3, WP4, WP5, WP6
Standards and metadata	

Standards, format, estimated volume of data	Spreadsheet table
Data exploitation and sharing	
Data access policy/ Dissemination level: confidential (only for members of the Consortium and the Commission Services) or Public	Confidential (only for members of the Consortium and the Commission Services)
Data sharing, re-use, distribution, publication (How?)	Shall be limited only to be carried out between the Project Consortium members and the European Commission's Services.
Personal data protection: are they personal data? If so, have you gained (written) consent from data subjects to collect this information?	No personal data is involved.
Access Procedures	None within the project consortium
Embargo periods (if any)	None
Archiving and preservation (including storage and backup)	
Data storage (including backup): where? For how long?	On project file server repository. Shall be maintained and backed up for a period of 4 years until the end of the project.
Indicative associated costs for data archiving and preservation	N/A
Indicative plan for covering the above costs	N/A

Table 18: DS2.2_ Standards for the design and development description.

Data identification: *DS2.1_ Requirements*
Generic description:

The DS2.1 describes the requirements that will be collected and analysed in order to specify the necessary architecture of the DIOPTRA.	
Origin of data:	
The data will be derived from specific needs of each partner of the consortium.	
Nature and scale of data:	
This dataset will be included in a spreadsheet where the requirements are described.	
To whom the dataset could be useful:	
DIOPTRA consortium	
Related scientific publication(s)	
N/A	
Indicative existing similar data sets (including possibilities for integration and reuse):	
List of standards (DS2.2_Standards for the design and development).	
Partners activities and responsibilities	
Partner owner of the data	ICCS
Partner in charge of the data analysis	ICCS, AGSAVVAS, NKUA, TERAGLOBUS, UOI, PAO, HOPE, DHCE, INTRA, TCR
Partner in charge of the data storage	ICCS
Related WP(s) and task(s)	WP2, WP3, WP4, WP5, WP6
Standards and metadata	
Standards, format, estimated volume of data	Spreadsheet table
Data exploitation and sharing	
Data access policy/ Dissemination level: confidential (only for members of the Consortium and the Commission Services) or Public	Confidential (only for members of the Consortium and the Commission Services)

Data sharing, re-use, distribution, publication (How?)	Shall be limited only to be carried out between the Project Consortium members and the European Commission's Services.
Personal data protection: are they personal data? If so, have you gained (written) consent from data subjects to collect this information?	No personal data is involved.
Access Procedures	None within the project consortium
Embargo periods (if any)	None
Archiving and preservation (including storage and backup)	
Data storage (including backup): where? For how long?	On project file server repository. Shall be maintained and backed up for a period of 4 years until the end of the project.
Indicative associated costs for data archiving and preservation	N/A
Indicative plan for covering the above costs	N/A

Table 19: DS3.1_ EHR Data description.

Data identification: DS3.1_ EHR Data
Generic description:
This dataset contains information collected retrospectively from 8 clinical sites (BLOKS, CHUL, RM-RRH, UKCM, BURGOS, NKUA, GOC, and AGSAVVAS). The dataset will include data from 4 groups (healthy, advanced adenomas, non-advanced adenomas and CRC participants). More specifically, the following clinical variables are recorded: • Clinical variables such as demographics, medical and family history and symptoms related to Colorectal Cancer (CRC), among others. • Clinical variables derived from EHR.
Origin of data:

The collection of the data has been performed from the clinical sites during previous years.	
Nature and scale of data:	
The format of each parameter will be defined after the completion of the final variable list, which depends on the variable's availability of each clinical site.	
To whom the dataset could be useful:	
The dataset will be utilised to technically set up the approach that will be followed for the curation, development and training of the models of T3.2 and T4.1.	
Related scientific publication(s)	
Journal and conference scientific publications that will present the developed and validated DIOPTRA system.	
Indicative existing similar data sets (including possibilities for integration and reuse):	
Retrospective data provided by BLOKS, CHUL, RM-RRH, UKCM, BURGOS, NKUA, GOC, AGSAVVAS.	
Partners activities and responsibilities	
Partner owner of the data	BLOKS, CHUL, RM-RRH, UKCM, BURGOS, NKUA, GOC, AGSAVVAS
Partner in charge of the data analysis	UOI, INTRA
Partner in charge of the data storage	ICCS/GRNET
Related WP(s) and task(s)	WP3, WP4, WP5
Standards and metadata	
Standards, format, estimated volume of data	Each parameter value that is recorded has a specific format (Excel, Json).
Data exploitation and sharing	
Data access policy/ Dissemination level: confidential (only for members of the Consortium and the Commission Services) or Public	Confidential (only for members of the Consortium and the Commission Services)

Data sharing, re-use, distribution, publication (How?)	A data processing agreement will be signed between clinical and technical partners.
Personal data protection: are they personal data? If so, have you gained (written) consent from data subjects to collect this information?	Personal data is involved. Anonymisation tools will be utilised offering an extra layer of privacy protection to the already pseudonymised data.
Access Procedures	The data will be accessed by the data providers and data processors.
Embargo periods (if any)	None
Archiving and preservation (including storage and backup)	
Data storage (including backup): where? For how long?	On project centralised platform (clinical interface). Shall be maintained and backed up for a period of 4 years until the end of the project.
Indicative associated costs for data archiving and preservation	N/A
Indicative plan for covering the above costs	N/A

Table 20: DS3.2_ Biological Samples and DS6.3_Biological samples description.

Data identification: DS3.2_ Biological Samples
Generic description:
This dataset contains information collected retrospectively and prospectively from the GRAZ biobank and from the clinical sites of the project. The dataset will contain a detailed inventory of the biological samples that will be collected and analysed for DIOPTRA. The dataset will include data from 4 groups (healthy, advanced adenomas, non-advanced adenomas and CRC participants). More specifically, the following clinical variables are recorded: • Blood and tissue samples along with EHR accompanying these samples.

Origin of data:	
Existing data already collected during previous studies and data collected within the project (based on a related clinical protocol) will be included in the dataset	
Nature and scale of data:	
Structured data and biological material.	
To whom the dataset could be useful:	
The dataset will be utilised to technically set up the approach that will be followed for the extraction of protein biomarkers along with the development and training of the models of T4.2 and T4.3.	
Related scientific publication(s)	
Journal and conference scientific publications that will present the developed and validated DIOPTRA system, as well as specific procedures and outcomes derived from the analysis.	
Indicative existing similar data sets (including possibilities for integration and reuse):	
N/A	
Partners activities and responsibilities	
Partner owner of the data	GRAZ and clinical sites
Partner in charge of the data analysis	PAO
Partner in charge of the data storage	PAO
Related WP(s) and task(s)	WP3, WP4, WP5
Standards and metadata	
Standards, format, estimated volume of data	Each parameter value that is recorded has a specific format (Excel, Json).
Data exploitation and sharing	
Data access policy/ Dissemination level: confidential (only for members of the	Confidential (only for members of the Consortium and the Commission Services)

Consortium and the Commission Services) or Public	
Data sharing, re-use, distribution, publication (How?)	A material transfer agreement will be signed between clinical partners and PAO.
Personal data protection: are they personal data? If so, have you gained (written) consent from data subjects to collect this information?	Personal data is involved.
Access Procedures	The data will be accessed by the data providers and data processors.
Embargo periods (if any)	None
Archiving and preservation (including storage and backup)	
Data storage (including backup): where? For how long?	In the PAO labs for analysis. Shall be maintained and backed up for a period of 4 years until the end of the project.
Indicative associated costs for data archiving and preservation	N/A
Indicative plan for covering the above costs	N/A

Table 21: DS3.3_ Variable list description.

Data identification: DS3.3_ Variable list
Generic description:
This dataset contains the clinical information that will be collected from the 8 clinical sites.
Origin of data:
The clinical information of the data that have been collected from the clinical sites during previous years.

Nature and scale of data:	
Structured data.	
To whom the dataset could be useful:	
The dataset will be utilised to synthesise the retrospective data that will be used for the risk factor analysis, as well as the development and training of the models of T4.1.	
Related scientific publication(s)	
N/A	
Indicative existing similar data sets (including possibilities for integration and reuse):	
Similar to DS3.1 EHR Data	
Partners activities and responsibilities	
Partner owner of the data	BLOKS, CHUL, RM-RRH, UKCM, BURGOS, NKUA, GOC, AGSAVVAS
Partner in charge of the data analysis	UOI
Partner in charge of the data storage	ICCS/GRNET
Related WP(s) and task(s)	WP3, WP4, WP5
Standards and metadata	
Standards, format, estimated volume of data	Each parameter value that is recorded has a specific format (Excel).
Data exploitation and sharing	
Data access policy/ Dissemination level: confidential (only for members of the Consortium and the Commission Services) or Public	Confidential (only for members of the Consortium and the Commission Services)
Data sharing, re-use, distribution, publication (How?)	Shall be limited only to be carried out between the Project Consortium members

Funded by the European Union

Personal data protection: are they personal data? If so, have you gained (written) consent from data subjects to collect this information?	No personal data is involved.
Access Procedures	Clinical and technical partners.
Embargo periods (if any)	None
Archiving and preservation (including storage and backup)	
Data storage (including backup): where? For how long?	On project file server repository. Shall be maintained and backed up for a period of 4 years until the end of the project.
Indicative associated costs for data archiving and preservation	N/A
Indicative plan for covering the above costs	N/A

Table 22: DS3.4_ Interface description.

Data identification: DS3.4_ Interface
Generic description:
This DIOPTRA component will be the centralised platform where all the anonymised/pseudonymised retrospective and prospective data will be uploaded by the clinical partners and will be stored.
Origin of data:
Existing data already collected during previous studies and data collected within the project (based on a related clinical protocol) will be included in the dataset.
Nature and scale of data:
Structured data.
To whom the dataset could be useful:
The dataset will be utilised to technically set up the approach that will be followed for the curation, development and training of the models of T3.2, T4.1, T4.2 and T4.3.

Related scientific publication(s)	
N/A	
Indicative existing similar data sets (including possibilities for integration and reuse):	
N/A	
Partners activities and responsibilities	
Partner owner of the data	Clinical sites and Biobank
Partner in charge of the data analysis	UOI, PAO, INTRA, CSCY
Partner in charge of the data storage	ICCS/GRNET
Related WP(s) and task(s)	WP3, WP4, WP5
Standards and metadata	
Standards, format, estimated volume of data	Each parameter value that is recorded has a specific format (Excel, Json).
Data exploitation and sharing	
Data access policy/ Dissemination level: confidential (only for members of the Consortium and the Commission Services) or Public	Confidential (only for members of the Consortium and the Commission Services)
Data sharing, re-use, distribution, publication (How?)	A material transfer agreement will be signed between clinical and technical partners.
Personal data protection: are they personal data? If so, have you gained (written) consent from data subjects to collect this information?	Personal data is involved. Anonymisation tools will be utilised offering an extra layer of privacy protection to the already pseudonymised data before the uploading of the data.
Access Procedures	The data will be accessed by the data providers and data processors.

Funded by the European Union

Embargo periods (if any)	None
Archiving and preservation (including storage and backup)	
Data storage (including backup): where? For how long?	In the project's centralised platform. Shall be maintained and backed up for a period of 4 years until the end of the project.
Indicative associated costs for data archiving and preservation	N/A
Indicative plan for covering the above costs	N/A

Table 23: DS3.5_ Data Curation description.

Data identification: DS3.5_ Data Curation
Generic description:
The data curation techniques that will be applied to the data, in order to be further processed and analysed.
Origin of data:
The collection of the data has been performed from the clinical sites during previous years.
Nature and scale of data:
Structured data.
To whom the dataset could be useful:
The dataset will be utilised to technically set up the approach that will be followed for the curation of the data (T3.2)
Related scientific publication(s)
N/A
Indicative existing similar data sets (including possibilities for integration and reuse):
N/A
Partners activities and responsibilities

Partner owner of the data	Clinical sites
Partner in charge of the data analysis	INTRA
Partner in charge of the data storage	ICCS/GRNET
Related WP(s) and task(s)	WP3, WP4, WP5
Standards and metadata	
Standards, format, estimated volume of data	Each parameter value that is recorded has a specific format (Excel, Json).
Data exploitation and sharing	
Data access policy/ Dissemination level: confidential (only for members of the Consortium and the Commission Services) or Public	Confidential (only for members of the Consortium and the Commission Services)
Data sharing, re-use, distribution, publication (How?)	A data processing agreement will be signed between clinical and technical partners.
Personal data protection: are they personal data? If so, have you gained (written) consent from data subjects to collect this information?	Personal data is involved.
Access Procedures	The data will be accessed by the data providers and data processors.
Embargo periods (if any)	None
Archiving and preservation (including storage and backup)	
Data storage (including backup): where? For how long?	On the project's centralised platform. Shall be maintained and backed up for a period of 4 years until the end of the project.

Indicative associated costs for data archiving and preservation	N/A
Indicative plan for covering the above costs	N/A

Table 24: DS4.1_ Risk Factor Analysis description.

Data identification: DS4.1_ Risk Factor Analysis	
Generic description:	
This dataset contains information related to all the extracted factors after the analysis of the data.	
Origin of data:	
Existing data already collected during previous studies and data collected within the project (based on a related clinical protocol) will be included in the dataset.	
Nature and scale of data:	
Structured data.	
To whom the dataset could be useful:	
The dataset will be utilised to technically set up the approach that will be followed for the development and training of the models of T4.1.	
Related scientific publication(s)	
Journal and conference scientific publications that will present the developed methodology and the results derived from the analysis.	
Indicative existing similar data sets (including possibilities for integration and reuse):	
N/A	
Partners activities and responsibilities	
Partner owner of the data	Clinical sites
Partner in charge of the data analysis	UOI
Partner in charge of the data storage	ICCS/GRNET

Related WP(s) and task(s)	WP3, WP4, WP5
Standards and metadata	
Standards, format, estimated volume of data	Each parameter value that is recorded has a specific format (Excel, Json).
Data exploitation and sharing	
Data access policy/ Dissemination level: confidential (only for members of the Consortium and the Commission Services) or Public	Confidential (only for members of the Consortium and the Commission Services)
Data sharing, re-use, distribution, publication (How?)	A data processing agreement will be signed between clinical and technical partners.
Personal data protection: are they personal data? If so, have you gained (written) consent from data subjects to collect this information?	Personal data is involved. Informed consent will be collected before the prospective study initiation.
Access Procedures	The data will be accessed by the data providers and data processors.
Embargo periods (if any)	None
Archiving and preservation (including storage and backup)	
Data storage (including backup): where? For how long?	On the project's centralised platform. Shall be maintained and backed up for a period of 4 years until the end of the project.
Indicative associated costs for data archiving and preservation	N/A
Indicative plan for covering the above costs	N/A

Funded by the European Union

Table 25: DS4.2_ Biomarker Discovery description.

Data identification: DS4.2_ Biomarker Discovery	
Generic description:	
This dataset contains information related to the measurements of gene and protein expression of the biological samples of the discovery study.	
Origin of data:	
The dataset will be created using proteomics and transcriptomics analysis methods from the collected biological samples.	
Nature and scale of data:	
Structured data.	
To whom the dataset could be useful:	
The dataset will be utilised to technically set up the approach that will be followed for the extraction of protein biomarker, as well as the development and training of the models of T4.2 and T4.3.	
Related scientific publication(s)	
Journal and conference scientific publications that will present the developed methodology and the results derived from the analysis.	
Indicative existing similar data sets (including possibilities for integration and reuse):	
N/A	
Partners activities and responsibilities	
Partner owner of the data	GRAZ and all clinical sites
Partner in charge of the data analysis	PAO
Partner in charge of the data storage	PAO
Related WP(s) and task(s)	WP3, WP4, WP5
Standards and metadata	

Standards, format, estimated volume of data	Each parameter value that is recorded has a specific format (Excel, Json).
Data exploitation and sharing	
Data access policy/ Dissemination level: confidential (only for members of the Consortium and the Commission Services) or Public	Confidential (only for members of the Consortium and the Commission Services)
Data sharing, re-use, distribution, publication (How?)	A material transfer agreement will be signed between clinical and technical partners.
Personal data protection: are they personal data? If so, have you gained (written) consent from data subjects to collect this information?	Personal data is involved.
Access Procedures	The data will be accessed by the data providers and data processors.
Embargo periods (if any)	None
Archiving and preservation (including storage and backup)	
Data storage (including backup): where? For how long?	On the project's centralised platform and PAO labs. Shall be maintained and backed up for a period of 4 years until the end of the project.
Indicative associated costs for data archiving and preservation	N/A
Indicative plan for covering the above costs	N/A

Table 26: DS5.1_ AI modelling description.

Data identification: DS5.1_ AI modelling
Generic description:

The examined risk factors along with the extracted proteins will be the input for the AI models, which will propose a small subset of proteins on the basis of a cost-effective screening protocol.	
Origin of data:	
Existing data already collected during previous studies and data collected within the project (based on a related clinical protocol) will be included in the dataset.	
Nature and scale of data:	
Structured data.	
To whom the dataset could be useful:	
The dataset will be utilised to technically set up the approach that will be followed for the development and training of the models of T5.1.	
Related scientific publication(s)	
Journal and conference scientific publications that will present the developed methodology and the results derived from the analysis.	
Indicative existing similar data sets (including possibilities for integration and reuse):	
N/A	
Partners activities and responsibilities	
Partner owner of the data	Clinical sites, Biobank
Partner owner of the data derived from the analysis	PAO, UOI
Partner in charge of the data analysis	VILABS, AINIGMA, NOVELCORE
Partner in charge of the data storage	ICCS/GRNET
Related WP(s) and task(s)	WP3, WP4, WP5
Standards and metadata	
Standards, format, estimated volume of data	Each parameter value that is recorded has a specific format (Excel, Json).
Data exploitation and sharing	

Funded by the European Union

Data access policy/ Dissemination level: confidential (only for members of the Consortium and the Commission Services) or Public	Confidential (only for members of the Consortium and the Commission Services)
Data sharing, re-use, distribution, publication (How?)	A data processing agreement will be signed between clinical and technical partners.
Personal data protection: are they personal data? If so, have you gained (written) consent from data subjects to collect this information?	Personal data is involved.
Access Procedures	The data will be accessed by the data providers and data processors.
Embargo periods (if any)	None
Archiving and preservation (including storage and backup)	
Data storage (including backup): where? For how long?	On the project's centralised platform. Shall be maintained and backed up for a period of 4 years until the end of the project.
Indicative associated costs for data archiving and preservation	N/A
Indicative plan for covering the above costs	N/A

Table 27: DS6.1_EHR data

Data identification: DS6.1_ EHR Data
Generic description:
This dataset contains information collected prospectively from 8 clinical sites (BLOKS, CHUL, RM-RRH, UKCM, BURGOS, NKUA, GOC, and AGSAVVAS). The dataset will include data from 4 groups (healthy, advanced adenomas, non-advanced adenomas and CRC participants). More specifically, the following clinical variables are recorded:

• Clinical variables like such as demographics, family history information, symptoms related to Colorectal Cancer (CRC), diagnosis, among others. • Clinical variables derived from EHR.
Origin of data:
The collection of the data will be performed from all the clinical sites during the recruitment.
Nature and scale of data:
Structured data.
To whom the dataset could be useful:
The dataset will be utilised to technically refine and validate the DIOPTRA system.
Related scientific publication(s)
N/A
Indicative existing similar data sets (including possibilities for integration and reuse):
DS3.1 EHR data
Partners activities and responsibilities

Partner owner of the data	Clinical sites
Partner in charge of the data analysis	UOI
Partner in charge of the data storage	ICCS/GRNET
Related WP(s) and task(s)	WP4, WP5, WP6
Standards and metadata	
Standards, format, estimated volume of data	Each parameter value that is recorded has a specific format (Excel).
Data exploitation and sharing	
Data access policy/ Dissemination level: confidential (only for members of the Consortium and the Commission Services) or Public	Confidential (only for members of the Consortium and the Commission Services)

Data sharing, re-use, distribution, publication (How?)	A data processing agreement will be signed between clinical and technical partners.
Personal data protection: are they personal data? If so, have you gained (written) consent from data subjects to collect this information?	Personal data is involved. Informed consent will be collected before the prospective study initiation.
Access Procedures	The data will be accessed by the data providers and data processors.
Embargo periods (if any)	None
Archiving and preservation (including storage and backup)	
Data storage (including backup): where? For how long?	On the project's centralised platform. Shall be maintained and backed up for a period of 4 years until the end of the project.
Indicative associated costs for data archiving and preservation	N/A
Indicative plan for covering the above costs	N/A

Table 28: DS6.2_ Questionnaires description.

Data identification: DS6.2_ Questionnaires
Generic description:
Questionnaires for gathering risk factors, family history and procedural data will be provided during the prospective study.
Origin of data:
The collection of the data will be performed from all the clinical sites during the recruitment.
Nature and scale of data:

Structured data.	
To whom the dataset could be useful:	
The dataset will be utilised to technically refine and validate the DIOPTRA system.	
Related scientific publication(s)	
N/A	
Indicative existing similar data sets (including possibilities for integration and reuse):	
DS3.1 EHR data, DS3.2 Biological samples	
Partners activities and responsibilities	
Partner owner of the data	Clinical sites
Partner in charge of the data analysis	UOI, PAO
Partner in charge of the data storage	ICCS/GRNET
Related WP(s) and task(s)	WP4, WP5, WP6
Standards and metadata	
Standards, format, estimated volume of data	Each parameter value that is recorded has a specific format (Excel).
Data exploitation and sharing	
Data access policy/ Dissemination level: confidential (only for members of the Consortium and the Commission Services) or Public	Confidential (only for members of the Consortium and the Commission Services)
Data sharing, re-use, distribution, publication (How?)	A data processing agreement will be signed between clinical and technical partners.
Personal data protection: are they personal data?	Personal data is involved. Informed consent will be collected before the prospective study initiation.

Funded by the European Union

If so, have you gained (written) consent from data subjects to collect this information?	
Access Procedures	The data will be accessed by the data providers and data processors.
Embargo periods (if any)	None
Archiving and preservation (including storage and backup)	
Data storage (including backup): where? For how long?	On the project's centralised platform. Shall be maintained and backed up for a period of 4 years until the end of the project.
Indicative associated costs for data archiving and preservation	N/A
Indicative plan for covering the above costs	N/A

Table 29: DS6.3_ Variable list description.

Data identification: DS6.3_ Variable list
Generic description:
Definition of the variable list which contains the clinical information (variables) that will be collected.
Origin of data:
The collection of the data will be performed from all the clinical sites during the recruitment.
Nature and scale of data:
Structured data.
To whom the dataset could be useful:
The dataset will be utilised to technically curate the data, refine and validate the DIOPTRA system.

Related scientific publication(s)	
N/A	
Indicative existing similar data sets (including possibilities for integration and reuse):	
N/A	
Partners activities and responsibilities	
Partner owner of the data	Clinical sites
Partner in charge of the data analysis	UOI, PAO, INTRA
Partner in charge of the data storage	ICCS/GRNET
Related WP(s) and task(s)	WP3, WP4, WP5, WP6
Standards and metadata	
Standards, format, estimated volume of data	Each parameter value that is recorded has a specific format (Excel).
Data exploitation and sharing	
Data access policy/ Dissemination level: confidential (only for members of the Consortium and the Commission Services) or Public	Confidential (only for members of the Consortium and the Commission Services)
Data sharing, re-use, distribution, publication (How?)	A data processing agreement will be signed between clinical and technical partners.
Personal data protection: are they personal data? If so, have you gained (written) consent from data subjects to collect this information?	Personal data is involved. Informed consent will be collected before the prospective study initiation.
Access Procedures	The data will be accessed by the data providers and data processors.
Embargo periods (if any)	None

Funded by the European Union

Archiving and preservation (including storage and backup)	
Data storage (including backup): where? For how long?	On the project's centralised platform. Shall be maintained and backed up for a period of 4 years until the end of the project.
Indicative associated costs for data archiving and preservation	N/A
Indicative plan for covering the above costs	N/A

Table 30: DS7.1_ Communication KPIs and DS7.2_ Dissemination and exploitation plan description.

Data identification: **DS7.1_ Communication KPIs, DS7.2_ Dissemination and exploitation plan**
Generic description:
The datasets include information regarding the KPIs, dissemination and exploitation of the DIOPTRA consortium.
Origin of data:
The datasets will be updated during the lifecycle of the project.
Nature and scale of data:
The nature of these dataset can be Excel, Word, pdf documents, while the content of the dissemination materials can be web pages, brochures, flyers, PowerPoint presentations, papers in journals and conferences, videos, images etc.
To whom the dataset could be useful:
All partners
Related scientific publication(s)
Journal and conference publications that will be made during the lifecycle of the project.
Indicative existing similar data sets (including possibilities for integration and reuse):
N/A

Partners activities and responsibilities	
Partner owner of the data	DIOPTRA consortium partners
Partner in charge of the data analysis	MARTEL GMBH
Partner in charge of the data storage	MARTEL GMBH, DCHE, CSI
Related WP(s) and task(s)	All WPs
Standards and metadata	
Standards, format, estimated volume of data	Excel, Word, PowerPoint, PDF Image formats (*.tiff, *.png, *.jpeg etc.) Video formats
Data exploitation and sharing	
Data access policy/ Dissemination level: confidential (only for members of the Consortium and the Commission Services) or Public	Public for the dissemination materials. Exploitation plan and IPR that will be confidential to the consortium partners and the Commission's Services
Data sharing, re-use, distribution, publication (How?)	The dissemination materials can be shared, re-used and distributed following copyright agreements. Exploitation plan and IPR that shall be limited only to be carried out between the Project Consortium members and the Commission's Services.
Personal data protection: are they personal data? If so, have you gained (written) consent from data subjects to collect this information?	No personal data is involved.
Access Procedures	Freely available for the consortium partners. Access procedures policies based on copyright agreements are applied for non-open access publications.

Funded by the European Union

Embargo periods (if any)	None
Archiving and preservation (including storage and backup)	
Data storage (including backup): where? For how long?	Scientific publications are stored in a central repository in the premises of Coordinator. Shall be maintained and backed up for a period of 4 years until the end of the project. Open access publications will be stored in open access repositories.
Indicative associated costs for data archiving and preservation	N/A
Indicative plan for covering the above costs	N/A

5.5 Constraints on data utilisation

Besides the constraints ensuring compliance with regulations such as GDPR for data utilisation, protection and privacy, DIOPTRA system implies in addition some extra constraints in a technical perspective. The prospective and retrospective data that will be stored in the Elasticsearch and processed through the Logstash module of the Data Curation Asset must be tabular (csv or excel format) or in Json format for uploading through the dashboard. As a result, no images are allowed for data ingestion. Moreover, the data format must follow the data model and schema that will be produced in the next months.

5.6 Data storage and backup

For data storage, the infrastructure of GRNET will be used by the project. GRNET S.A. is a public sector technology company in Greece that has been operating since 1998 providing networking, cloud computing, HPC, data management services, and e-Infrastructures to academic and research institutions, educational bodies, and public sector agencies operating under the auspices of the Ministry of Digital Governance. GRNET serves as the coordinator of all e-infrastructures in education and research and contributes to Greece's digital transformation in various sectors. The company develops and maintains advanced computer and networking infrastructures, data centers, and a national fiber optic network.

GRNET is also involved in open science projects, the formulation of the National Artificial Intelligence Strategy, and the Euro QCI (Quantum Communication Infrastructure) are some examples among the others. The company fosters the collaboration with other agencies in the Greek public sector and provides international interconnection through the GÉANT network.

The services that are provided by GRNET are not only used by academic and research communities, but also by students, public hospital personnel, public administration personnel,

business executives, and citizens. The digital transformation in public administration, digital innovation in education, advanced technologies for research, digital capability for health, and accessing culture through modern technologies is a main goal of the company.
GRNET also emphasises its green policy, energy-efficient infrastructures, and its role in the Greek Internet Exchange (GR-IX). The company engages in international partnerships, collaborations, and e-infrastructures, contributing to the development of research and science in Greece and abroad by receiving funding from the Greek State and the European Union.
In the context of the DIOPTRA Horizon project, GRNET will provide us with the following infrastructure and equipment, namely specific virtual machines (VMs):

4. VM1: 8 or 16 cores, 32 GB RAM, 500 GB disk (pref SSD)
5. VM2: 8 cores, 16 GB RAM, 200 GB disk
6. VM3: 8 cores, 16 GB RAM, 200 GB disk
7. VM4: 4 cores, 8 GB RAM, 100 GB disk
8. VM5: 8 cores, 16 GB RAM, 100 GB disk
9. VM6: 4 cores, 8 GB RAM, 100 GB disk

The scientific publications produced by the DIOPTRA project will be:

- retained locally to the premises of each partner
- uploaded to the ECAS participant portal on the defined for the project section
- uploaded to the central repository where all the documents for the project is stored, a repository managed by the coordinator of the project

The AI models that will be developed will be stored:

- internally by each technical partner
- At the DIOPTRA cloud in the provided VMs of the GRNET

The retrospective and prospective data will be stored:

- In the DIOPTRA centralised platform by utilising the ELK STACK. Elasticsearch comes up with Cross-cluster replication (CCR), a way to automatically synchronise indices from the primary cluster to a secondary remote cluster that can serve as backup. If the primary cluster fails, the secondary cluster can take over. Moreover, Elasticsearch provides snapshots as a backup of a running Elasticsearch cluster for data recovery stored in an off-cluster storage location called a snapshot repository.

4.7 Data security

The DIOPTRA centralised platform (ELK STACK) adopts a layered security approach, effectively protecting stored data at various levels (Cluster, index, Document, Field). Its security principles serve as a solid foundation for securely operating Elasticsearch. More precisely:

- Elasticsearch security features provide built-in support for user management and authentication, while also allowing integration with third-party tools and systems. To establish a connection with

the cluster, users must include their credentials with their requests. Two services are available: the token service, which is enabled by default when TLS/SSL is activated for HTTP, and the API key service, which is also enabled by default. Both services can be utilised as credentials to authenticate new requests.

- Authorisation of users is achieved through a role-based access control (RBAC) mechanism. This involves assigning privileges to roles and associating roles with users or groups. For restricting access to documents in search queries and aggregations, an attribute-based control (ABAC) mechanism can be employed. This ensures complete control over user access rights.
- SSL/TLS encryption is implemented to secure node-to-node connections. Certificates and keys for TLS are generated for both the transport and HTTP layers. The TLS settings can be configured in the elasticsearch.yml file.
- IP filtering can be used to restrict connections. This filtering can be applied to application clients, node clients, or transport clients, allowing or denying access based on hosts, domains, or subnets. If a node's IP is blacklisted, the connection is immediately terminated, and no further requests are processed.
- The elastic stack complies with security standards such as FIPS 140-2 and Section 508. It also provides support for meeting GDPR requirements.

Regarding the data uploaded in the Zenodo the security is ensured since, according to the security policy of Zenodo, it provides the following:

- The data centers are located on CERN premises and all physical access is restricted to a limited number of staff with appropriate training and who have been granted access in line with their professional duties.
- The servers are managed according to the CERN Security Baseline for Servers, meaning e.g. remote access to our servers is restricted to Zenodo staff with appropriate training, and the operating system and installed applications are kept updated with latest security patches via the automatic configuration management system Puppet.
- The CERN Security Team runs both host and network-based intrusion detection systems and monitors the traffic flow, pattern and contents into and out of CERN networks in order to detect attacks. All access to zenodo.org happens over HTTPS, except for static documentation pages which are hosted on GitHub Pages.
- Zenodo stores user passwords using strong cryptographic password hashing algorithms (currently PBKDF2+SHA512). Users' access tokens to GitHub and ORCID are stored encrypted and can only be decrypted with the application's secret key.
- Zenodo also employs a suite of techniques to protect your session from being stolen by an attacker when a user is logged in and run vulnerability scans against the application.

6 Legal and Ethical Aspects

6.1 Compliance with the data protection principles in the context of Responsible Research and Innovation

In research settings, the General Data Protection Regulation (GDPR) imposes legal obligations on entities conducting, among other, scientific research and on how they treat research data[7]. In this context, for example, entities are required to provide research subjects with detailed information about what will happen to the personal data that they collect and also requires the organisations processing the data to ensure the data are properly protected, minimised, and destroyed when no longer needed.

In addition to respecting legal obligations, all EU projects are guided by ethical considerations and the values and principles on which the EU is founded and these have been set out in guidance by the European Commission.[8] The Commission guidance emphasises that particular attention is required when research involving special categories of data (formerly known as 'sensitive data'), profiling, automated decision-making, data-mining techniques, big-data analytics and artificial intelligence. It is more likely that a project raises higher ethics risks when it involves the processing of 'special categories' of personal data. While any research project must demonstrate compliance with the GDPR, the fact that research is legally permissible does not necessarily mean that it will be deemed ethical.

In light of the above and considering that DIOPTRA project involves processing of health data, which is a special category of data under the GDPR, and that it involves using the data to test and train AI models, it is a project that requires consideration, also, from an ethical viewpoint. For this reason, the following ethical principles are relevant to the use of data in the DIOPTRA research.

Accountability

Accountability represents a key ethical principle; it entails that someone can be held responsible for their actions. The concept of accountability has a slightly stricter form of responsibility. While responsibility is task-focused concept requiring the one to be responsible for ensuring that a task is completed, accountability goes beyond being responsible for the decisions and actions and expects the relevant actor to explain these decisions and action as well as provide a justification for these actions. Strengthening accountability is one of the highest priorities, as well one of the thorniest issues, in using health data in research. The notion of accountability follows from the fundamental democratic principle that those who exercise political authority should be accountable for their actions. Internationally democratic accountability has been linked increasingly with policies that encourage wider citizen participation in health systems including the implementation of patient and public involvement (PPI) initiatives in decision-making processes, and clinical and research

[7]Article 89 of the Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation)

[8]European Commission, Ethics and Data Protection, 5 July 2021 available at https://ec.europa.eu/info/funding-tenders/opportunities/docs/2021-2027/horizon/guidance/ethics-and-data-protection_he_en.pdf

governance structures.[9] There are various types of potential accountability in research, on the one hand, around achieving more patient-centred outcomes, greater transparency in health research funding processes and more democratic clinical decision-making and, on the other hand, promoting participatory accountability as a form of citizen-based evaluation to complement existing forms of accountability – for example, the public administrative, managerial and professional accountabilities.

In the context of using health data in a multi-phase research project, the issue of accountability is a rather complex one. There are varying and complex accountabilities that come into play in the healthcare context: from the satisfaction of the Hippocratic Oath, summarised as 'first do no harm' (clinical accountability), to showing respect for the integrity and autonomy of patients (ethical accountability), as well as professional, legal and financial accountabilities (including data protection accountability and compliance with GDPR).[10]

In respecting the ethical principles for confidentiality, as well as the legal obligations in the GDPR, and within the various phases of the project, there must be appropriate accountability mechanisms to demonstrate how these various different accountabilities are taken into account. The concept of accountability is about owning and co-owning roles and responsibilities, making things happen, and offering assistance should anything go wrong. Therefore, accountability must not be an afterthought but should be engrained throughout the various technical and organisational measures of the project and appropriately communicated to patients and to the rest of DIOPTRA partners, including clinical partners. To ensure accountability for data processing, mechanisms that facilitate the system's auditability are also fundamental. For example, such systems would require, in principle, the traceability and logging of data sharing activities between the different partners of the project.

Data protection by design and default

The concept of 'data protection by design' in the GDPR requires data controllers to implement appropriate technical and organisational measures to give effect to the GDPR's core data protection principles (Articles 5 and 25 GDPR). This concept is one of the best ways to address ethical concerns with research. In the context of research, measures to achieve data protection by design could include: pseudonymisation or anonymisation of personal data; data minimisation; applied cryptography; and various techniques that enable data subjects to exercise their fundamental rights. Applying the principle of data protection by design in a research project can help mitigate the ethics risks. In addition, where there is a possibility to enhance the level of data protection for research subjects, there measures should be applied by default, rather than making them available only as an optional extra.

Informed Consent

Informed consent is the cornerstone of research ethics and requires that, to the extent relevant, all DIOPTRA partners and particularly clinical partners explain to research participants what the research is about, what their participation in the project will entail and any risks that may be involved. Only after this information has been conveyed to the participants – and they have fully understood it – it is possible to seek and obtain their express

[9] Performing accountability in health research: A socio-spatial framework Aris Komporozos-Athanasiou, Mark Thompson, and Marianna FotakiHuman Relations 2017 71:9, 1264-1287

[10] Ibid.

permission to include them in the research project and for their personal data to be processed (Articles 4(11) and 7 GDPR).

Whenever any person or body collects personal data directly from research participants, they must seek their informed consent by means of a procedure that meets the minimum standards of the GDPR. This requires consent to be given by a clear affirmative act establishing a freely given, specific, informed and unambiguous indication of the subject's agreement to the processing of their personal data.[11] This may take the form of a written statement, which may be collected by electronic means, or an oral statement. Wherever possible, this process should be integrated into a broader informed consent procedure that meets the standards set out in the Commission's Guidance note on informed consent. In particular, for projects involving particularly complex or sensitive data-processing operations, such as collecting and re-using health data, there is a requirement to have a specific informed consent process covering the data-processing component of the research project. There is a requirement to keep records documenting the informed consent procedure, including the information sheets and consent forms provided to research participants, and the acquisition of their consent to data processing. These may be requested by data subjects, funding agencies or data protection supervisory authorities.

For consent to data processing to be 'informed', the data subject must be provided with detailed information about the envisaged data processing in an intelligible and easily accessible form, using clear and plain language. The data subjects must also be made aware if data are to be used for any other purposes, shared with research partners or transferred to organisations outside the EU (see Article 13 GDPR).

As with any research project involving human subjects, if the data processing entails potential risks to the data subjects' rights and freedoms, they must be made aware of these risks during the informed consent procedure. The consent process(es) and the information given to the data subjects should cover all the data-processing activities related to their participation in the research and for the processing of their personal data. From a research ethics perspective, and in accordance with the principles of fair and transparent data processing, if the project intends to use or make their data available for future research projects, it is best practice to obtain their additional, explicit consent to the secondary use of the data. Even if the controller may legally use the same data for the work of another research project without further steps to ensure lawfulness of processing that data, informing the subjects and asking for new consent is consistent with research ethics and the principle of fair processing.[12]

Use of previously collected personal data (secondary use)

The use of personal data collected for one purpose and then used for other research, without the knowledge or consent of the data subject, is potentially a breach of ethical standards. When processing personal data in research without the express consent of the data subjects, there must be a clear recorded explanation about how this data are obtained to justify their use in the project and to ensure that the processing is fair to the data subject. If the collection

[11] See also Article 7 GDPR and Guidelines on consent under Regulation 2016/679, Article 29 Working Party (adopted 28 November 2017).

[12] See case study and example, given by European Commission in Ethics and Data Protection guidance (p.12) citing Handbook on European data protection law: 2018 edition, EU Fundamental Rights Agency, European Court of Human Rights, Council of Europe and European Data Protection Supervisor (2018); http://fra.europa.eu/en/publication/2018/handbook-european-data-protection-law).

or use of data raises specific ethics issues (for example, as regards consent and transparency, privacy and the rights and expectations of the data subjects), there is a requirement to provide a detailed overview of the planned data collection and further processing operations and to explain how the ethics concerns will be mitigated.
If any clinical partner or other partner of an EU project intends to use personal data that were collected from a previous research project, it must ensure that it can provide detailed information regarding the initial data collection, methodology and informed consent procedure and confirm that it has permission from the owner/manager of the dataset(s) to use the data in this project.
Where the planned use of data is predicated on the 'legitimate interests' of the data controller, the nature and purpose of the dataset must be set out in detail, together with the safeguards (e.g. anonymisation or pseudonymisation techniques) that warrant its use in the project.[13]
If the intended data processing is based on national legislation or international regulations authorising the research, or a demonstrable overriding public interest (e.g. public health, social protection) that allows the partner participating in an EU research project to use a particular dataset, this must be made explicit and there must be explicit reference to the relevant Member State or Union law or policy.
In principle, if clinical partners or other project partners are using personal data provided by a third party and the data subjects have not expressly consented to its use in research projects, the partner must, in accordance with the GDPR, inform them that it has acquired the data and what it will be using them for (Article 14 GDPR). The partner must also provide the data subjects with the same basic information about the data processing and their rights as data subjects that they are obliged to provide to people that they are collecting data from directly. However, these requirements do not apply where it is not possible or would involve a disproportionate effort to contact the data subjects. In such cases the partner must implement appropriate safeguards, including technical and organisational measures to ensure respect for the principle of personal data minimisation and protect the subjects' fundamental rights. Crucially, the GDPR requires that pseudonymisation or anonymisation techniques be implemented wherever viable (Article 89 GDPR).

Data security

There are both ethical and legal obligations when collecting personal data to ensure that research participants' information is properly protected. The GDPR requires all data controllers and processors to implement appropriate technical and organisational measures to ensure a level of data security that is commensurate to the risks faced by the data subjects in the event of unauthorised access to, or disclosure, accidental deletion or destruction of, their data (Article 32 GDPR). There should be appropriate technical and organisational measures that will be implemented to protect the personal data processed in the course of the research project, e.g. with reference to the host institution's and research partners' data protection and information security policies. Such measures may include the pseudonymisation and encryption of personal data, and policies and procedures to ensure the

[13] According to the GDPR, '[t]he legitimate interests of a controller, including those of a controller to which the personal data may be disclosed, or of a third party, may provide a legal basis for processing, provided that the interests or the fundamental rights and freedoms of the data subject are not overriding, taking into consideration the reasonable expectations of data subjects based on their relationship with the controller'. See also recital 47 and Article 89 GDPR.

confidentiality, integrity, availability and resilience of processing systems. Where higher-risk processing is envisaged (e.g. and this includes processing health data), this should be taken into account and given an enhanced level of data security by choosing appropriate research methods and data-processing tools that keep communications and data secure from unauthorised access.

6.1.1 Regulatory developments in data protection law

Ongoing developments in EU regulation relevant to data protection will be closely monitored and to the extent relevant and necessary will be addressed in future DIOPTRA deliverables, including under the updated version of the Data Management Plan, deliverable D1.2 due in M36 of the project.
To note, key regulatory developments relevant to DIOPTRA are as follows:

- The **Data Governance Act**[14], which becomes applicable on 24 September 2023, introduces the concept of data altruism. Data altruism is about individuals giving their consent for the processing of their personal data or organisation giving permission to make available non-personal data that they generate, voluntarily and without reward, to be used for objectives of general interest such as healthcare or scientific research purposes. The Data Governance Act aims to create trusted tools that will allow the creation of pools of data of a sufficient size in Europe to allow data analytics and machine learning, including across borders. Therefore, in case the specific needs of DIOPTRA for research data are not addressed otherwise and/or in the occurrence of unforeseen risks relating to the datasets, currently, planned to be made available for the purposes of DIOPTRA, the DIOPTRA project could potentially benefit from data altruism and have access to a vast amount of data from healthcare institutions.
- On 3 May 2022, the European Commission presented a **proposal for a regulation on the European Health Data Space** (EHDS)[15]. The proposed regulation aims to lay down the legal foundation of the EHDS and to ensure a safe and secure exchange, use and reuse of health data across the EU. In that regard, the proposal comprises of EU-wide rules, common standards and practices, infrastructures and a governance framework for processing of a wide range of health data in the EU. The proposed EHDS regulation builds upon the GDPR, the Data Governance Act within the EU legal framework for data.

6.2 Further Ethical Aspects

DIOPTRA aims to use research data collected to train and test Artificial Intelligence (AI) models, with the goal of risk stratification, behavioural monitoring and interventions. This entails machine learning models and various data analytics techniques that will be used within the project. On this basis it seems appropriate to devote a section on the ethics of AI. This section provides an overview of the guiding ethical principles and standards concerning AI and it afterwards examines the steps to ethical AI by design.

[14] Regulation (EU) 2022/868 of the European Parliament and of the Council of 30 May 2022 on European data governance and amending Regulation (EU) 2018/1724 (Data Governance Act), available at https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A32022R0868.
[15] European Commission, Proposal for a Regulation of the European Parliament and of the Council on the European Health Data Space (European Health Data Space Regulation), COM/2022 197 final. Available at https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A52022PC0197.

6.2.1 Ethics Principles for Trustworthy AI

The EU strategy on AI aims to create a standard for trustworthy AI. In 2021, as part of its strategy on AI, the European Commission published Guidelines on Ethics By Design and Ethics of Use Approaches for Artificial Intelligence[16] which is developed from the EU's High-Level Expert Group's Ethics Guidelines for Trustworthy AI published in 2019[17]. In addition, OECD has also set forth ethical principles for trustworthy artificial intelligence in the Recommendation of the Council on Artificial Intelligence in 2019.[18] Although the Commission's Guidelines and the OECD's recommendation are not legally binding and do not offer pieces of advice on legal compliance for AI, they are still considered as two of the most prominent and influential documents on AI in the EU. Therefore, the high-level principles of the ethical AI set forth under these documents are relevant to the development and use of AI in the DIOPTRA project.
The Ethics Guidelines of the EU's High-Level Expert Group lays out a framework of four (4) ethical principles and seven (7) key requirements that trustworthy AI systems are supposed to meet. The ethical principles are respect for human autonomy, prevention of harm, fairness and explicability. The seven (7) key requirements involve (1) human agency and oversight; (2) technical robustness and safety; (3) privacy and data governance; (4) transparency; (5) diversity, non-discrimination and fairness; (6) social and environmental wellbeing; and (7) accountability.

6.2.2 Ethical AI by Design

Building on these ethical principles, the Commission's Guidance on Ethics by Design and Ethics of Use Approaches for Artificial Intelligence of 25 November 2021[19], aims at raising awareness in the scientific community, and especially within beneficiaries of EU research and innovation projects, guides the developers through the incorporation of these ethical considerations into the development process. The objective is to address ethical concerns during the design and development stages, rather than trying to fix them later in the process and thereby to prevent ethical issues from arising in the first place. In that respect, it should be analysed if the AI system may unintentionally or intentionally create any ethics risk by way of, for instance, creating social disadvantages to people either by the AI-based system, or by the way it will be deployed. Afterwards, in order to properly address these identified risks, it is recommended to modify the elements of the AI design and to implement the measures and procedures in the development process in a way that they could mitigate the risks. Ethics by Design approach is described with the five-layer model illustrated in the figure below.

[16] European Commission, Ethics by Design and Ethics of Use Approaches for Artificial Intelligence, Version 1.0, November 2021.
[17] High-Level Expert Group on AI, Ethics Guidelines for Trustworthy AI, 2019.
[18] Organisation for Economic Co-operation and Development, Recommendation of the Council on Artificial Intelligence, OECD/LEGAL/0449, 22 May 2019.
[19] European Commission, Ethics By Design and Ethics of Use Approaches for Artificial Intelligence, 25 November 2021.

Funded by the European Union

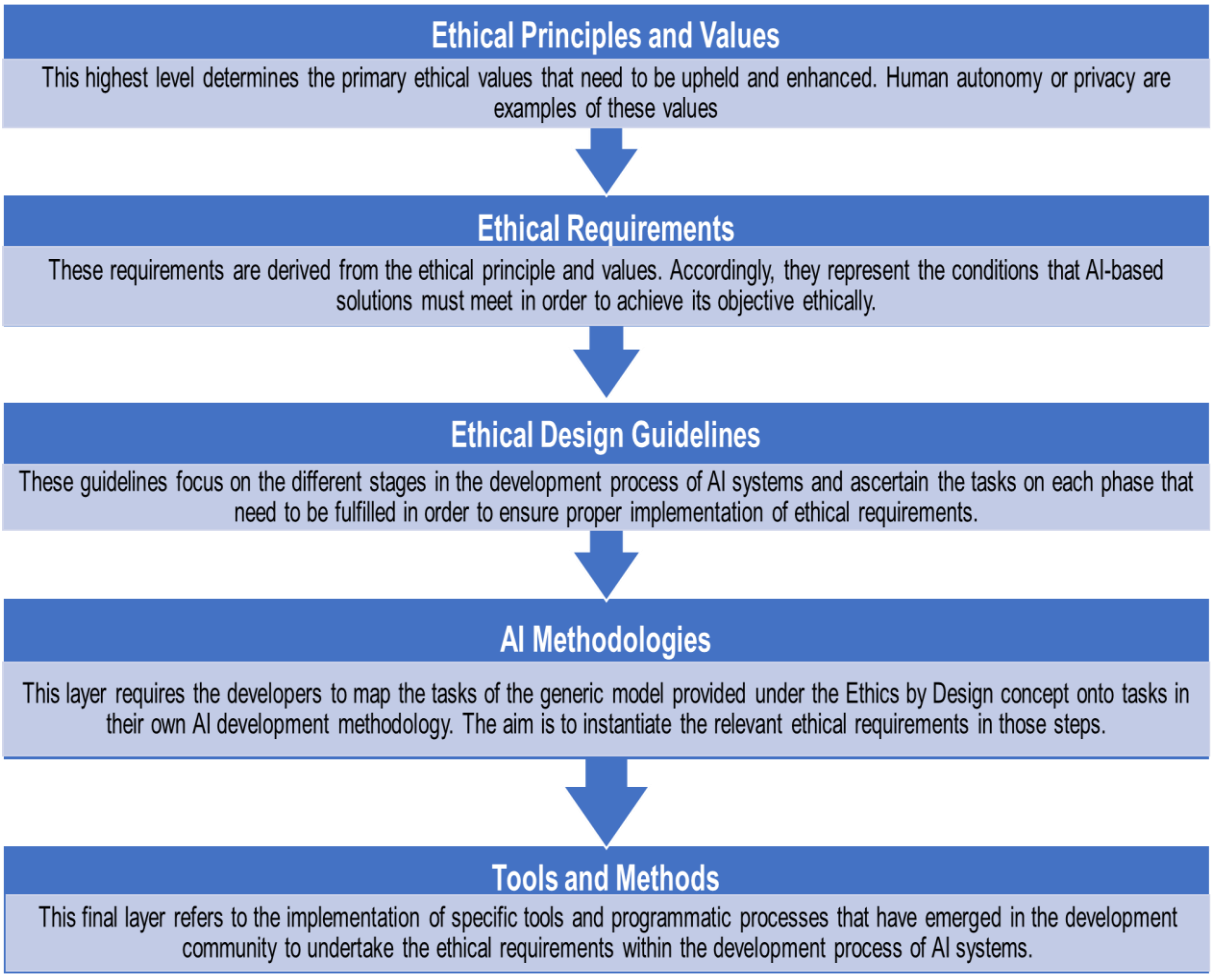

Figure 3: The 5-layer Model of Ethics by Design[20]

A generic model containing six phases had been developed to guide the developers to properly embed Ethics by Design in the development processes. These six phases refer to the specific points in the AI development methodology in which the ethical requirements need to be instantiated as tasks, goals, constraints, and measures to prevent ethical risks arising in the first place. Although the six phases of the generic model under the Ethics by Design approach are illustrated in a list format below, the EU Commission explicitly states that it is not a sequential process, and the developers are advised to tailor these phases into the similar steps in their chosen development methodology.

[20] The content of the figure 1 is largely based on the information provided under the EU Commission's Guidance on Ethics by Design and Ethics of Use Approaches for Artificial Intelligence of 25 November 2021 and under the Deliverable Report of Ethics by Design and Ethics of Use in AI and Robotics published by the EU-funded SIENNA project.

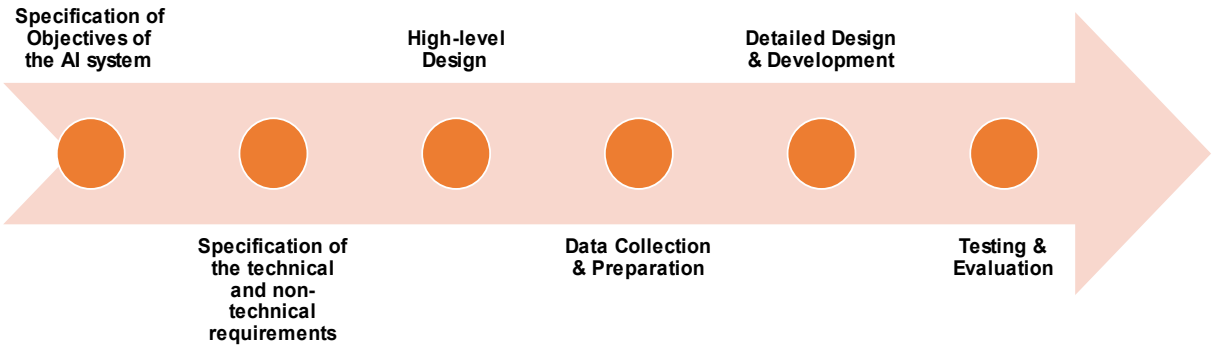

Figure 4: The EU Commission's Generic Model for AI Development[21]

It should be noted that each of these steps includes different tasks which must be undertaken by the developers in order to ensure ethics compliance of their AI systems.[22]

6.2.3 Regulatory developments in AI

Regulatory developments that are relevant to AI use in the DIOPTRA project and could be adopted during the lifetime of the project and, same as it was the case for the developments pertaining to personal data protection, these will be, also, actively monitored for future DIOPTRA deliverables, including under the updated version of the Data Management Plan, deliverable D1.2 due in M36 of the project. The most important is the proposed Artificial Intelligence Act[23], which introduces a risk classification of AI systems, and which is currently going through the legislative process and is likely to be adopted by early autumn 2023.

[21] The six phases of the generic model shown in the Figure-2 is based on the illustration provided under the EU Commission's Guidance on Ethics by Design and Ethics of Use Approaches for Artificial Intelligence of 25 November 2021.
[22] Further detailed discussion and explanation of each specific tasks under the six phases of the generic model can be found in the EU Commission's Guidance on Ethics by Design and Ethics of Use Approaches for Artificial Intelligence of 25 November 2021 as well as in the Deliverable report of Ethics by Design and Ethics of Use in AI and Robotics published by the EU-funded SIENNA project in 2020.
[23] Proposal for a Regulation of The European Parliament and of The Council Laying Down Harmonised Rules on Artificial Intelligence (Artificial Intelligence Act) and Amending Certain Union Legislative Acts COM/2021/206 final.

Funded by the European Union

7 Risks and mitigation measures

Risk management includes identification, assessment and prioritisation of risks, followed by coordinated actions to eliminate or minimise and control the impact of difficulties that can arise in performing the tasks, and of unfortunate events. The next table describes some of the risks that may arise during the lifecycle of the project along with mitigation measures for these specific risks.

Table 31: Risks and proposed mitigation measures.

Risk No#	Description	Work Package No(s)	Proposed Mitigation Measures
1	Retrospective data availability/quality issues due to damaged/unfit samples or limited features	WP3, WP4, WP5	DIOPTRA plans an early start of prospective data collection, compensating for potential retrospective data issues. Moreover, if a partner underperforms with regard to data flow, other partners will be asked to target an increased contribution.
2	Identification of a data breach incident	WP3, WP6	DIOPTRA storage and sharing mechanisms are based on well-established and tested platforms that ensure security assurance. Moreover, a dedicated anonymisation tool has been provided by the technical partners as an extra protection layer that will be applied by the clinical partners on top of their existing pseudonymisation processes before data sharing. Nevertheless, if a data breach occurs, the Project Coordinator, the Clinical Manager, the Technical Manager, the Data Protection Officer, the Quality Assurance & Risk Manager, and the partner responsible for security protocols will be notified. The impact of the event will be assessed and relevant measures will be taken, which will also include the notification of data owners.
3	Clinical sites are unable to provide the estimated prospective data volumes	WP6	Estimated volumes are based on partner capabilities and past evidence. If a partner faces unforeseen difficulties, other members will be asked to increase contribution accordingly.

Funded by the European Union

4	Unpredicted risk during the project	WP1	T1.4 “Scientific & Technical Management” will reassess risks based on project progress, updating the current table and aiming to prevent potential problems.
5	Partner not performing well or leaving the project	WP7, WP1, WP2, WP3, WP5, WP4, WP6	Partners have been selected to fulfil all objectives, ensuring multiple capacities. Accordingly, the coordinating partner will decide whether the predefined work can be conducted by another partner or a replacement will be needed.
6	High data variability among clinical sites hindering harmonisation	WP3, WP4	In case of major discrepancies on data types, format & protocols, DIOPTRA will incorporate into its model all risk factors & indicators, covering differentiations and allowing for null values among datasets. With regard to modality variability (e.g. FIT/colonoscopy), data will be curated separately on a modality-specific level, while sample size asymmetries will be prospectively addressed, adapting requests for new data flows provided by the clinical partners.

8 Responsibilities and Resources

The Data Control Committee (DCC), comprised of three coordinating partners, has been established to assume data controller responsibilities for the project. This committee will be responsible for implementing the DMP, assigning pertinent roles and responsibilities (from data capture and data quality assurance to metadata production and data archiving and sharing), and ensuring that it is reviewed and revised as needed.
The DCC is comprised of the following members:

- **Project coordinator:** INSTITUTE OF COMMUNICATION AND COMPUTER SYSTEMS (ICCS), PIC 999654356, established in Patission Str. 42, ATHINA 10682, Greece, responsible for data protection: *Prof. Dr. Georgios Matsopoulos.*
- **Technical coordinator:** PANEPISTIMIO IOANNINON (UOI), 74670, established in PANEPISTEMIOYPOLE, PANEPISTEMIO IOANNINON, IOANNINA 45110, Greece, PIC 999852818, responsible for data protection: *Prof. Dr. Dimitrios I. Fotiadis.*
- **Clinical coordinator:** CENTRE HOSPITALIER UNIVERSITAIRE DE LIEGE (CHUL), PIC 999495276, established in AVENUE DE L HOPITAL 1, LIEGE 4000, Belgium, responsible for data protection: *Head of Information Systems Management Department Prof. Phillippe Kolh and CHUL's CEO Mr. Marc De Paoli.*

The set-up of data protection framework for the project is led by the DIOPTRA legal partner ARTHUR'S LEGAL BV (ARTHUR).
It should be noted that the list of the persons and entities serving as data controllers and data processors who will be responsible for handling data management issues will be included in the letter to the Data Processing Agreement.

Funded by the European Union

9 Conclusions

This document presented the DMP for the DIOPTRA project, covering the strategies and approaches that will be implemented to ensure effective management of the data collected and generated in the project on early screening of colorectal cancer. The document presents a detailed description of the data that will be collected per Work Package (WP), as well as the mechanisms for data collection, documentation, metadata generation and data assessment. Additionally, the storage and back up mechanisms, data preservation, sharing and access methods along with the ethical and legal compliance of the DMP, are described. Finally, the specific responsibilities regarding the DMP implementation and review (if required) have been described. An updated version of the DMP will be included in respective deliverable at a later stage (M36).

Funded by the European Union