Ref. Ares(2023)8900308 - 29/12/2023

D2.1: DIOPTRA Full Specification Report

Revision: v.1.0

Work package	WP2
Task	Tasks 2.1 & 2.2
Due date	31/12/2023
Submission date	29/12/2023
Deliverable lead	ICCS
Version	1.0
Authors	Stavros Miloulis (ICCS), Ioannis Kouris (ICCS), Zheshen Jiang (CHUL), Christos Fotis (PAO), Nikos Tsolakos (PAO), Leonidas Alexopoulos (PAO), Sophia Stamatatou (PAO), Christos Androutsos (UOI), Stavros Pitoglou (CSCY), Grigoris Antonopoulos (INTRA), Philopoimin Lykokanellos (INTRA), Michalis Vourtzoumis (INTRA), Ioannis Vezakis (TCR), George Domalis (NOVELCORE), Aris Gioutlakis (NOVELCORE), Ioannis Livieris (NOVELCORE), Lefteris Koumakis (STS), Mattia Pirani (I2G)
Reviewers	Simos Symeonidis (AINIGMA), Iliana Korma (CMA)
Abstract	This deliverable presents the requirements related to the development and pilot implementation of DIOPTRA, identified via the requirements' analysis that engaged project partners and external stakeholders. Use cases implicating ecosystem actors are listed, together with the specific components that comprise the technical architecture of DIOPTRA aiming to support the studies of the project. Functional and non-functional requirements are delineated for each component, aiming to ensure the quality of development and implementation.
Keywords	Requirements, Architecture, System Components, Clinical Workflow, Biomarker Analysis

GrantAgreementNo.:101096649
Call:HORIZON-MISS-2021-CANCER-02
Topic: HORIZON-MISS-2021-CANCER-02-01
Type of action: HORIZON-RIA

Document Revision History

Version	Date	Description of change	List of contributor(s)
V0.1	02/11/2023	1st Release of Table of Contents &Contribution Assignments	Stavros Miloulis (ICCS), Ioannis Kouris (ICCS)
V0.2	08/12/2023	1st Consolidated Version	Stavros Miloulis (ICCS), Zheshen Jiang (CHUL), Christos Fotis (PAO), Nikos Tsolakos (PAO), Leonidas Alexopoulos (PAO), Sophia Stamatatou (PAO), Christos Androutsos (UOI), Stavros Pitoglou (CSCY), Grigoris Antonopoulos (INTRA), Philopoimin Lykokanellos (INTRA), Michalis Vourtzoumis (INTRA), George Domalis (NOVELCORE), Aris Gioutlakis (NOVELCORE), Ioannis Livieris (NOVELCORE), Lefteris Koumakis (STS)
V0.3	14/12/2023	2nd Consolidated Version	Stavros Miloulis (ICCS), Christos Androutsos (UOI), Ioannis Vezakis (TCR)
V0.4	19/12/2023	3rd Consolidated Version	Stavros Miloulis (ICCS), Christos Androutsos (UOI), Zheshen Jiang (CHUL) Mattia Pirani (I2G), Ioannis Vezakis (TCR), Grigoris Antonopoulos (INTRA), George Domalis (NOVELCORE)
V0.5	22/12/2023	Minor updates & Incorporation of internal reviewers' comments	Stavros Miloulis (ICCS), Ioannis Kouris (ICCS), George Botis (ICCS), Christos Fotis (PAO), George Domalis (NOVELCORE), Simos Symeonidis (AINIGMA), Iliana Korma (CMA)
V1.0	29/12/2023	Final version approved and submitted by the Project Coordinator	Maria Haritou (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.

©2023-2026 DIOPTRA

COPYRIGHT NOTICE

© 2023-2026 DIOPTRA

Project funded by the European Commission in the Horizon Europe Programme		
Nature of the deliverable:	**R***	
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: Datasets, 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

This deliverable describes the **full set of requirements** for the **technical development** and **operation** of the DIOPTRA workflow, based on the interdisciplinary collaboration of consortium partners during the initial phases of the project. The document initiates from a **requirements' analysis** and the **description of use cases** with regard to both the project research and innovation actions and the envisioned real-world use, translating this information to a **complete technical architecture** featuring **specific components** that are set to serve fulfilment of the use cases. Specifications and operational principles for **non-technical components** are also included, eliciting workflows and information pathways that pose further needs to be addressed via the technical infrastructure.

Based on the above, this document serves as a **reference guide defining the following**:

- Clinical Workflows
- User Groups & Use Cases
- System Components & Overall Architecture
- Functional & Non-functional Requirements for each Component
- Security Compliance Roadmap

The following means have been employed to generate the content of this document:

- Dedicated discussions among technical partners
- Dedicated discussions among clinical partners
- Discussions involving stakeholders with experience in citizen engagement and real-world deployment in the health sector
- Interdisciplinary discussions engaging stakeholders that will be involved in ecosystem assessment, dissemination actions, policy-making and exploitation
- Requirements' elicitation techniques (including task analysis, literature review, surveys, etc.)

Since this document is produced as a **roadmap**, some details may change during actual development & implementation. Moreover, specific points and methodologies are referred within a generic framework and will be further clarified throughout the workplan. Finalised information will be included in the deliverables referring to the final system outcomes and evaluation.

Funded by the European Union

©2023-2026 DIOPTRA

TABLE OF CONTENTS

Funded by the European Union

©2023-2026 DIOPTRA

©2023-2026 DIOPTRA

LIST OF FIGURES

Funded by the European Union

©2023-2026 DIOPTRA

LIST OF TABLES

Funded by the European Union

©2023-2026 DIOPTRA

Funded by the European Union

©2023-2026 DIOPTRA

Funded by the European Union

©2023-2026 DIOPTRA

ABBREVIATIONS

AA	Advanced Adenoma
AI	Artificial Intelligence
API	Application Programming Interface
AT	Anonymisation Tool
BD	Becton Dickinson
BP	Base Pair
BSA	Bovine Serum Albumin
C<#>	DIOPTRA Component No#
CRC	Colorectal Cancer
CSI	Clinical Site Interface
CSV	Comma-Separated Values
CV	Coefficient of Variation
D<#>	Deliverable No#
DoA	Description of Action
DCS	Data Curation and Storage System
DIOPTRA_UC1	DIOPTRA Use Case No#
EHR	Electronic Health Record
ETSI	European Telecommunications Standards Institute
EU	European Union
FR	Functional Requirement
G<#>	DIOPTRA Actor Group No#
GB	Gigabyte
GCLP	Good Clinical Laboratory Practices
GCP	Good Clinical Practice
GDPR	General Data Protection Regulation

Funded by the European Union

©2023-2026 DIOPTRA

GMP	Good Manufacturing Practices
GRNET	National Infrastructures for Research and Technology
HADEA	European Health and Digital Executive Agency
HLM	Health Literacy Module
ICT	Information and Communications Technology
ID	Identifier
IEC	International Electrotechnical Commission
ISMS	Information Security Management System
ISO	International Standardisation Organisation
IT	Information Technology
IVD	In Vitro Diagnostics
JAD	Joint Application Development
JSON	JavaScript Object Notation
JTC	Joint Technical Committee
LLM	Large Language Model
MA	Mobile App
M2M	Machine-to-Machine
ML	Machine Learning
non-AA	Non-Advanced Adenoma
N/A	Not Applicable
NAS	Network Attached Storage
NFR	Non-Functional Requirement
NGS	Next Generation Sequencing
NHS	N-Hydroxysuccinimide
NISD	Network and Information Security Directive
PCR	Polymerase Chain Reaction
PEA	Proximity Extension Assay

Funded by the European Union

©2023-2026 DIOPTRA

PEG	Polyethylene Glycol
PII	Personally Identifiable Information
RAM	Risk Assessment Module
REST	Representational State Transfer
RTBF	Right to be Forgotten
QC	Quality Control
QMS	Quality Management System
RIN	RNA Integrity Number
RNA	Ribonucleic Acid
SAD	Structured Analysis and Design
SAI	Screening AI Module
SC	Sub-Committee
SOP	Standard Operating Procedure
UPS	Uninterruptable Power Supply
VM	Virtual Machine
WP	Work Package

Funded by the European Union

©2023-2026 DIOPTRA

1 INTRODUCTION

As per the **DoA**, this deliverable represents the outcome of WP2 ("DIOPTRA Requirements & Conceptual Architecture") aiming to define and document the following:

- Specific needs to be addressed
- Ecosystem requirements for addressing the identified needs
- Distinct components of the above ecosystem & their functionality
- Specific requirements (e.g. technical, clinical) for design, development & implementation of both the different components and the DIOPTRA system as a whole

As such, the **structure of the main document content** has been organised as follows:

- **Section 2** presents the requirements' analysis conducted within the project.
- **Section 3** describes the use cases of DIOPTRA based on the outcomes of requirements' elicitation.
- **Section 4** presents the full architecture of DIOPTRA, listing specific functional & non-functional requirements for each component, addressing component integration & security compliance guidelines.
- Finally, **Section 5** includes the main conclusions and summarises key details of the deliverable's content and the related work conducted.

Funded by the European Union

©2023-2026 DIOPTRA

2 REQUIREMENTS ANALYSIS

2.1 REQUIREMENTS' ELICITATION

The elicitation of development and implementation requirements of the DIOPTRA ecosystem was conducted from clinical, technical and overall operational perspectives, aiming to support the delivery of the final project outcomes. The corresponding process included the following steps[1], [2]:

1. **Understanding the Application Domain**: Use scenarios within the real world were considered, starting from the target service delivery (i.e. CRC screening) and listing the required workflows. These include a) clinical workflows (e.g. data collection, decision making), b) data analysis workflows (biological & non-biological), and c) technical workflows required to support the use scenarios.

2. **Identification of Stakeholders**: The main stakeholders involved in the development and/or use of DIOPTRA include the following:

   - **Healthcare Staff** (administrative staff, doctors, nursing staff): This group corresponds to the direct testers and end-users of the DIOPTRA system. Healthcare staff can exploit the DIOPTRA components both during development (data collection and management within the clinic, testing on screening population) and within the envisioned real-world application (clinical decision support). In this regard, this group is strongly related to clinical requirements affecting system applicability and implementation details within pilot studies throughout development and testing.

   - **Biomarker Analysis Specialists**: This group is responsible for developing the biological screening test which constitutes a direct outcome of DIOPTRA. The corresponding work is heavily depended on the collection and efficient management of high-quality data, as well as on the application of demanding techniques (leveraging high-cost resources). Therefore, this group is in position to set requirements for the implementation of the whole workflow that initiates from biological data collection (along with related metadata) and results to the delivery of a validated screening tool as well as to real-world application guidelines.

   - **Technical Developers**: The technical developers are mainly responsible for delivering a) the infrastructure required to enable efficient data collection and management (including curation & monitoring by clinicians), b) software modules for data analysis with regard to clinical decision making on CRC screening, and c) the envisioned mobile application that is planned to support behavioural data collection, triggering of behavioural change and CRC knowledge communication. Overall, this group is needed to translate the requirements of clinical & biological analysis workflows to specific components, as well as define the features and requirements for these components from a developer's perspective.

   - **Citizen Engagement Experts**: The engagement of entities with experience in system interaction with citizens is necessary in order to refine workflows and refine the corresponding system components (e.g. mobile app & related material) via the definition of key features that should be included by the developers.

   - **Policy-Makers**: Consideration of policy-makers' input is crucial in order to define potential needs that should be taken into account to maximise the potential of DIOPTRA to penetrate everyday practice.

Funded by the European Union

©2023-2026 DIOPTRA

- **Business Entities**: Organisations related to exploitation and business modelling of DIOPTRA are part of this group. Within the project scope, they are responsible for identifying opportunities for upscale and overall exploitation of DIOPTRA endpoints based on pilot results and research outcomes. Moreover, this group includes potential actors that could be involved in exploitation and marketing activities beyond the project scope.

- **General Population**: This group is not an end-user of DIOPTRA, though it directly benefits from its outcomes. The general population that visits clinical sites to receive screening services would undergo the DIOPTRA-defined workflow, while in a real-world application scenario their clinical pathway would be affected by the results of DIOPTRA-assisted screening.

3. **Elicitation of Requirements**: The following methods were leveraged within the main process of requirements' elicitation:

   - ***Task Analysis***: In order to define the required components of the complete technical infrastructure, a top-down approach was employed to decompose high-level tasks (e.g. clinical data input, storage & management) into sub-tasks (e.g. data encoding, upload, quality control, curation, inspection). This resulted in a detailed delineation of all events, which have been translated into a technical architecture workflow. Similarly, the same process was followed during the clinical workflow analysis, as well as the biomarker analysis workflow. As a result, the elicited requirements described over the next sections of this document aim to fulfil the needs of the identified sub-tasks.

   - ***Domain Analysis***: Existing documentation and related material were studied mainly with regard to a) behavioural questionnaires, b) data management software, c) CRC-related variables of clinical significance, and d) behavioural changes related to CRC risk minimisation. More specifically, the DIOPTRA behavioural questionnaires will be used for baseline assessment and for 1-year follow-up of a specific participants' subgroup. Previously existing material was used towards the final set of questions. Regarding data management software, features and overall operation of existing software solutions were studied for incorporation into the DIOPTRA technical infrastructure (Section 4). Finally, existing literature and online knowledge was studied to collect information on behavioural strategies and significant CRC variables. Concerning the latter, health records of the DIOPTRA clinical sites were also studied in order to a) identify candidate variables to incorporate into the DIOPTRA data model, and b) For all of the above, comments and feedback from consortium specialists (clinicians, developers) were considered in order to refine the results of domain analysis and formulate the requirements that will be incorporated into DIOPTRA.

   - ***Ethnography***: Observation was the ethnography technique that was employed within DIOPTRA, with biomarker analysts performing on-site visits to clinical partners of DIOPTRA to collect information on the colonoscopy process, aiming to incorporate the biological sampling for the purposes of the project to the standard procedure. Sample collection and management requirements were formed based on their conclusions, with relevant information being incorporated both into the current document and in the DIOPTRA clinical protocols. More specifically, the SOP related to Sample Collection & Management was incorporated as an Annex (No. 3) in the Prospective clinical protocol. Additionally, a list of required consumables and equipment for sample collection and management was developed as an outcome of the on-site visits, which was circulated to all clinical partners.

Funded by the European Union

©2023-2026 DIOPTRA

Finally, specific limitations and potential risks were identified during this process, which were later discussed within dedicated working group and brainstorming sessions.

Requirement Workshops / Focus Groups: Dedicated discussions organised as part of project internal meeting scheduling involved different groups of the stakeholders presented above in order to define workflows, use cases, system architecture, and key requirements. Distinct clinical and technical discussions took place, with the vast majority of each one of the main 2 groups (Healthcare Staff and Technical Developers) participating in the respective sessions. More specifically, clinical experts were involved in clinical requirements' elicitation for the different studies, offering their perspectives and expertise help to better understand CRC and the related population groups, as well as to better define the research question together with technical partners in each stage of the DIOPTRA studies. Study documents, inclusion/exclusion criteria, workflows, data templates and architecture diagrams were discussed within these groups. Moreover, to adapt implementational requirements for each clinical site, partner-specific workshops will be organised during the study kick-off process to address all potential risks generated or will be generated throughout. Similarly, technical meetings defined the required system architecture and all relevant details, involving the developers of all DIOPTRA components (Section 4.2) to define operation principles and interaction mechanisms. Initial meetings were more in the form of **brainstorming sessions**, discussing general ideas on expected outcomes and required components. Group agendas became more specific later on, discussing detailed use cases, workflows, components and architecture. Moreover, clinical representatives (mainly the Clinical Manager of the project) participated in the technical discussions, and vice-versa with technical representatives attending clinical discussions. In this way, information transfer and feedback verification between the groups was ensured, establishing a **Joint Application Development** (JAD) methodology by allowing its party to express their unique viewpoint. Moreover, representatives of other groups (Biomarker Analysis Specialists, Citizen Engagement Experts) participated in both groups, actively reforming technical/clinical workflows and requirements. Finally, an additional group was established to facilitate discussion with regard to DIOPTRA exploitation process. Relevant tangible outcomes from this group are expected to occur further in the project, however representatives of this group participated in joint discussions with the other stakeholders of the project to provide comments on certain requirement aspects, mostly related with the Citizen Engagement Experts domain.

- **Structured Analysis and Design (SAD)**: Structured analysis was employed via a) diagram construction in order to depict clinical workflows, b) data flow and technical architecture diagrams, and c) the creation of the **DIOPTRA Data Template** that includes all variables to be incorporated in the clinical studies of the project. The latter essentially constitutes a **Data Dictionary**, incorporating information such as variable names, definitions, and encoding instructions. All of the above were created, circulated and reviewed using online documents which were directly accessible by all consortium specialists to ensure prompt communication of information.

    - Clinical relevance and availability of DIOPTRA data were investigated to improve the study variable list. The DIOPTRA data template then was created based on the study variable list, with data collection details for each variable in each study. Other steps in data collection such as data anonymisation, transfer etc.

©2023-2026 DIOPTRA

- **Definition of Use Cases**: Throughout the progress of the previous steps, specific workflows were developed, which were later translated into corresponding use cases, described in Section 3. These were initially developed as a **living document** from a user's perspective, considering the needs and workflow for the Healthcare Staff for the clinical interface and the study participants for the mobile application. These descriptions were formulated in collaboration with Technical Developers and were presented to members of the DIOPTRA clinical partners' teams to receive feedback and define details with regard to data management processes, information presentation and access privileges. For each use case, a full workflow description, the related system components, the actors involved, the event/circumstance(s) that may trigger the interaction, and any prerequisites that must be satisfied prior to the use case commencement have been included in Section 4.1.

- **Interface Analysis**: This method was leveraged with regard to the DIOPTRA clinical interface (Section 4.2.1) that will be developed, allowing the end-users at the collaborating clinical sites (Healthcare Staff) to upload and monitor data collected from the clinical studies' participants. Based on the outcomes of task analysis and focus groups (also considering the conjured use cases), the required interactions of the users with the clinical interface were defined within the focus groups. Moreover, requirements concerning interactions of the clinical interface with the mobile application were defined during this process.

- **Survey & Questionnaire**: In order to extract requirements with regard to the operation and functionalities of the mobile application that will be developed within DIOPTRA, an online survey was created using a free online tool[1] in order to elicit feedback from clinicians concerning the information and capabilities that should be included. This survey was distributed to the clinical partners of the DIOPTRA consortium, with a follow-up dedicated meeting organised to discuss results. A modified version of this survey was provided to external stakeholders, exploiting a communication channel established by HADEA to connect EU-funded projects that tackle cancer screening and prevention. This has been established under the 'Prevention, Including Screening' Cluster launched within the Europe's Beating Cancer Plan as part of the EU missions. Moreover, based on the 1st version of the use cases that were developed for the use of the DIOPTRA dashboards (Section 4.2.1), a short set of questions was provided to the clinical partners (end-users of the dashboard) via the living document mentioned above (#1 of this sub-list) in order to define specific details with regard to optimal data management, information presentation and access privileges.

[1] https://www.surveymonkey.com/

©2023-2026 DIOPTRA

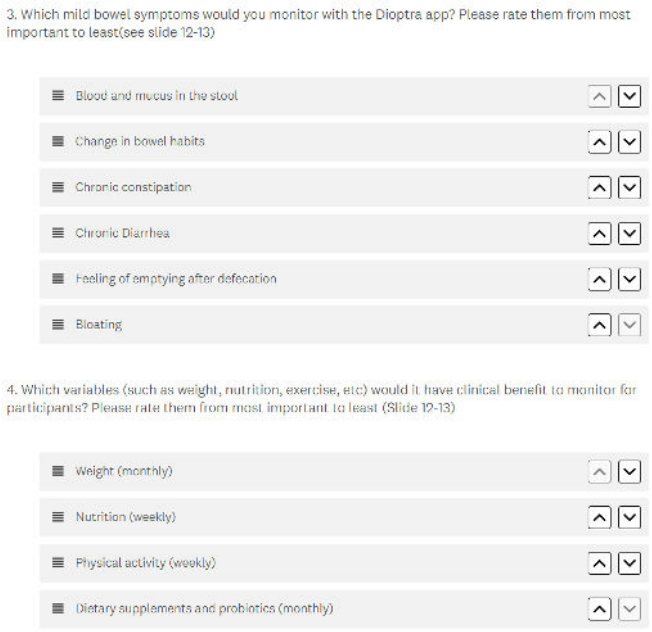

Figure 1: Online Survey for Mobile App Requirements

- **Prototyping**: This method is going to be implemented throughout the development of the DIOPTRA mobile application. Based on the initial requirements formed through the focus groups, the domain analysis and the above online survey, the initial version of the application will be formed as an interactive online service, in order to collect real-time feedback on the operation and material presentation of the application modules. This has been planned as an iterative evaluation process, progressively incorporating feedback into the prototype versioning.

4. **Documentation of Requirements**: The results of the above methodologies were utilised to encode functional and non-functional requirements for each component of the DIOPTRA system architecture. Specific table templates were used for encoding requirements, incorporating information such as **description**, **priority**, **rationale**, **verification** and **completion criteria**. Furthermore, as mentioned for the prototyping process, the elicitation of requirements and their incorporation into the final systems will be an iterative process throughout the development pipeline, with the implementation and testing of the first system versions providing feedback for refinements towards the final DIOPTRA system.

2.2 CLINICAL REQUIREMENTS

2.2.1 Clinical Workflow

In order to generate the outcomes of DIOPTRA, a general clinical workflow (Figure 2) has been conjured based on the preliminary results of the requirements' elicitation. This workflow has been built around the clinical studies of DIOPTRA, which have been organised for data collection and pilot testing of components. Specifically, the below workflow also includes the association of the collected data with the development of distinct components:

- Biomarker-based Screening Solution
- Risk Assessment Model

©2023-2026 DIOPTRA

- Full DIOPTRA Screening Model
- Mobile Application

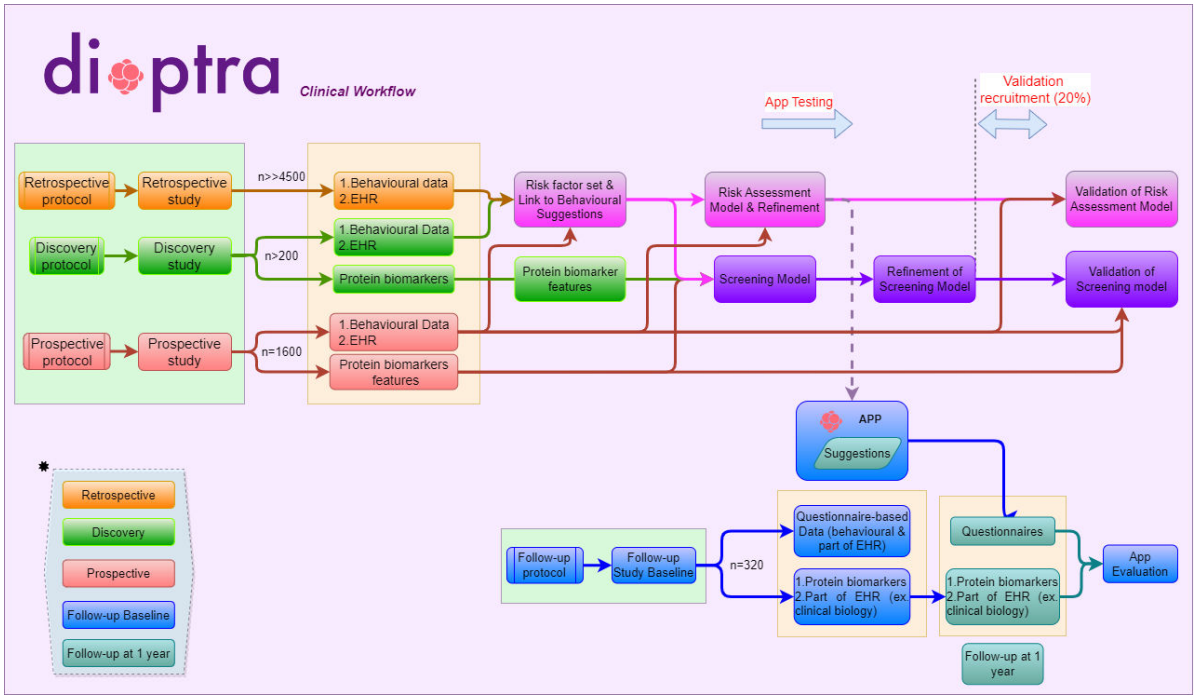

Figure 2: DIOPTRA Clinical Workflow

The first part of the DIOPTRA clinical workflow is based on the **retrospective** and **biomarker discovery** protocols of Figure 2. The former hypothesises that a set of **predictive variables** is associated with the risk of developing CRC and utilises EHR data from the clinical sites that participate in the DIOPTRA consortium to test this hypothesis and develop a **risk assessment model based on these risk factors** (Section 4.2.5). On the other hand, the biomarker discovery study utilises biological material (i.e. tissue and blood samples) to identify **biomarkers with diagnostic capacity** for early CRC. The combination of these results is used to develop a holistic CRC screening model, exploiting state-of-the-art AI methodologies (Section 4.2.6).

The second part of the DIOPTRA clinical workflow aims to refine and validate the above outcomes (i.e. risk factor model, biomarkers, screening model) via **prospective recruitment** by the clinical sites, collecting EHR, behavioural, and blood sample data. Furthermore, an additional **follow-up study** has been designed to a) test the use of the DIOPTRA **mobile application** (Section 4.2.4) by specific population groups (Table 1), and b) investigate the efficacy prospects of a behavioural change strategy (encouraged via the mobile app). The DIOPTRA app that will be used within the follow-up study includes **three primary modules** (together with behavioural questionnaires): a) Risk Assessment & Personalised Suggestion Module (RAM), b) Health Literacy Module (HLM), and c) Diary.

The population groups participating in the DIOPTRA studies include the following:

Funded by the European Union

©2023-2026 DIOPTRA

Table 1: DIOPTRA Population Groups & Study Participation

Group	Retrospective Study	Biomarker Discovery Study	Prospective Study	Follow-up Study
Healthy	✓	✓	✓	✓
Non-Advanced Adenoma (non-AA)	✓	✓	✓	✓
Advanced Adenoma (AA)	✓	✓	✓	
CRC	✓	✓	✓	

The definition of DIOPTRA clinical workflow generated a) clinical requirements that need to be addressed during the implementation, progression, and finalisation of different studies, and b) technical and system architecture requirements that aim to support the studies' implementation together with the whole workflow as well as generate the final outcomes of the DIOPTRA project. From a purely clinical perspective, additional requirements can be summarised as follows:

1. **Clinical Expert Involvement & Monitoring**: Clinical experts should be able to easily monitor the status of the clinical studies within their hospital and have access to collected data for all population groups. This will be fulfilled via the implementation of the Clinical Site Interface described in Section 4.2.1. DIOPTRA online study monitoring system will be used to monitor recruitment number during biomarker discovery study and prospective study. Amendments will be submitted during prospective study with updates in all study documents. These updates include modifications and improvements of all documents throughout the study in an organised and traceable manner. The mid-term recruitment report (D6.7) will document the study progression.

2. **Data Collection**: To ensure completeness and high quality for the obtained datasets, a general variable list was created for all DIOPTRA studies, incorporated into the **DIOPTRA Data Template** (D3.1). will be discussed in the following section 3.2 technical requirements.

3. **Clinical Study Implementation**: The operational details for conducting the clinical studies of DIOPTRA have been reflected into the corresponding study protocols (D6.2, D6.6), which also include informed consent forms and data collection requirements. These documents have been communicated to all clinical sites for local ethical approval and applicability verification (via **test runs**) considering local variations. Detailed inclusion and exclusion criteria have been incorporated as well, together with the definition of the population groups taking part in each study (healthy, non-advanced adenoma, advanced adenoma, CRC). Relevant information was also used for the registration of the DIOPTRA prospective study in ISRCTN registry[2] (reference number: 15583857). Any changes in requirements and implementation details will be incorporated in D6.3 and D6.4, as updates to D6.2 and D6.6. The ISRCTN record will be updated accordingly.

[2] https://www.isrctn.com/ISRCTN15583857

©2023-2026 DIOPTRA

4. **Data Privacy and Security**: Related requirements are depicted in D1.1 and D1.3.

2.2.2 Biomarker Analysis

2.2.2.1 Requirements for Clinical Sample Collection, Management and Storage

For **serum and plasma (blood) samples**, the following requirements have been identified:

1. Each sample should have a **unique ID** following a specific codification scheme.
2. A **Standard Operating Procedure (SOP)** with details regarding sample collection, management, storage and shipment is required. This has been included as an annex in the corresponding clinical protocol (D6.2) and should be followed by all clinical partners.
3. Any **deviations from the SOP** need to be strictly documented by clinical partners and validated by proper experts within the consortium.

Based on the on-site clinical visits and the clinical sample collection, management and storage requirements, the following table of consumable and equipment requirements was created (Table 2). The requirements presented in Table 2 are crucial in order for clinical partners to closely follow the Sample Collection and Management SOP.

Table 2: Consumable and Equipment Requirements for Sample Collection, Management and Storage

#	Description	Quantity required	Specifications	Proposed Supplier, Cat No
1	9-10mL Serum collection tubes	1 per patient	Plastic, 16x100mm, with clot activator (silica), red cap colour, transparent	BD Vacutainer, #367896 Greiner Vacuette, #455092
2	9-10mL K2EDTA Plasma collection tubes	1 per patient	Plastic, 16x100mm, with K2EDTA additive, purple cap colour, transparent	BD Vacutainer, #367525 Greiner Vacuette, #455045
3	15mL centrifuge tubes (transfer tubes)	2 per patient	nonpyrogenic and DNase-/RNase-free	Corning, #430791
4	2mL screw-cap microcentrifuge tubes (storage tubes)	10 per patient	nonpyrogenic and DNase-/RNase-free, non-sterile, freezable to -80 °C, can be centrifuged to 12,000×g, with silicone O-ring screw-caps	VWR, #525-0651 (tubes), #525-0653 (screw-caps)
5	Cryoboxes with dividers, 9x9 positions	2 boxes per 40 patients	133x133x50mm size, resistant to temperatures down to –140 °C, standard waterproof coating	VWR, #479-1417 (boxes), #479-1465 (dividers)

©2023-2026 DIOPTRA

6	Centrifuge	1	1300-1800 g (RCF) 18-25 °C For 16mm x 100mm tubes	N/A
7	Ultra-low Freezer	1	-80 °C or below	N/A
8	Pipette	1	Single channel 200-1000uL range	Rainin Pipet-Lite LTS Pipette L-1000XLS+, #17014382
9	Laminar flow hood (optional)	1	Class II A2 cabinet	N/A
10	Racks for collection/transfer/storage tubes	1 per tube size	See tube specifications	VWR, # 211-0204 (for 2mL tubes)
11	Personal Protective equipment	N/A	Lab coat, gloves etc	N/A

2.2.2.2 Sample Requirements for Olink Analysis

In the DIOPTRA biomarker discovery pipeline, blood samples will be profiled using the Olink protein analysis method. Olink analysis refers to a method used in proteomics to quantitatively analyse multiple proteins (up to 5300 proteins) simultaneously in miniscule amounts of sample. The **samples that will be used in the Olink analysis pipeline** need to fulfil the following criteria:

1. **Sample type**: The Olink Explore platform[3] is optimised and validated for plasma and serum samples.
2. **Number of samples**: In a specific run, up to 88 user samples can be analysed simultaneously.
3. **Inclusion of control samples**: A pooled plasma sample should be included in duplicate on each plate. These samples are used to assess potential variation between runs and plates, for example to calculate inter-assay and intra-assay CV, as well as for troubleshooting.
4. **Sample randomisation**: When running more than one Sample Plate (i.e. more than 88 samples across multiple 96-well plates), samples should be appropriately randomised across all plates and necessary steps for normalising and combining data are taken. Sample randomisation, i.e. ensuring that samples from different DIOPTRA groups are present in all plates, helps to ensure that technical variation does not overlap with biological variation.
5. **Volume of sample**: Minimum sample volume corresponds to ~50-100 uL/sample.
6. **Sample analysis**: An Olink analysis protocol that describes all experimental steps needs to be developed as part of WP4 (D4.2 & D4.6).

Funded by the European Union

©2023-2026 DIOPTRA

2.2.2.3 Sample and Processing Requirements for RNA Sequencing

In the DIOPTRA biomarker discovery pipeline, RNA material extracted from colon tissue samples (isolation) will be analysed using RNA sequencing. RNA sequencing is a technique used to analyse the quantity and sequences of RNA in a sample, providing insights into gene expression patterns and cellular functions. **Tissue samples that have undergone RNA isolation** should fulfil the following criteria for RNA sequencing[4]:

1. **RNA sample purity** needs to be estimated by measuring absorbance at 260 and 280nm. 260/280 ratio should be greater than 1.8. An initial estimation of sample concentration needs to be performed.

2. **Accurate quantification of RNA** should be performed using the appropriate instrument (Qubit RNA BR Assay, quantitative range 0.5 ng/µL-1200 ng/µL).

3. **RNA starting material sample integrity (RIN)** should be assessed. Samples should have a RIN≥7 (High Sensitivity RNA ScreenTape assay, quantitative range 0.5-10 ng/µL).

4. A **reference sample** needs to be processed in parallel as positive control throughout the library preparation and sequencing. Reference samples as positive controls are used to identify deviations, which allows for timely implementation of corrective actions in case of failure.

5. To create optimum **cluster densities** during sequencing, it is crucial to accurately quantify the final libraries. The final library needs to be analysed using the 4150 TapeStation system[5]. A high-quality library showing a single symmetric peak with a maximum between 250 and 350 base pairs (bp) is suitable for subsequent cluster generation and sequencing. The exact library size depends on the sample. The molarity of final libraries is calculated by the maximum peak size of the Agilent TapeStation system and the concentration measured by Qubit fluorometer. The molarity of the library is calculated using the following formula:

$$\frac{concentration\_in\_ng/\mu L}{660 g/mol \times average\_library\_size\_in\_bp} \times 10^6 = concentration\_in\_nM$$

6. **Libraries** should be normalised to 10 nM to be appropriate for RNA sequencing analysis[6].

7. A detailed **RNA sequencing analysis protocol** that describes all experimental steps needs to be developed as part of WP4 (D4.2 & D4.6).

2.2.2.4 Sample Requirements for Clinical Validation

Serum and plasma samples collected during the prospective study for clinical validation need to fulfil the following requirements:

1. Samples should be **recorded in a Biological Sample Inventory** at PAO's facilities.

2. Each sample should be accompanied by a **unique sample ID**, **DIOPTRA participant ID**, **Sample type (either serum or plasma)**, **collection date**, **sender (clinical partner)**, **lab reception date**, **date of reception by PAO**, **number of tubes collected and received**, **volume per tube**, **total volume received**, **visual inspection** (to assess if the sample is normal, haemolytic, icteric or lipemic), **storage temperature**, **storage location**, **box ID**, **comments / deviations** concerning receipt process, **comments / deviations** concerning sample collection.

©2023-2026 DIOPTRA

3. **Sample and participant documentation** related to the receipt, storage and inventory of biological samples will be archived by PAO.

4. Two aliquots of serum and two aliquots of plasma samples per participant should be **provided by the clinical partner to PAO**. Two aliquots of serum/plasma should be also retained as back-up material by each clinical partner.

5. Sample collection performed by the clinical partner should **follow the respective SOP (Annex No. 3)** described in the clinical protocol.

6. All tubes used for the collection of samples should be clearly labelled. All tubes should have certain characteristics **as described in the respective SOPs sample collection kit and Sample Collection & Management**.

7. Shipments should be arranged periodically as serum/plasma samples should be retained **at -80°C for long storage** (at the Test Facility). If samples are expected to be thawed multiple times, they should be further aliquoted in smaller volumes by PAO.

8. Shipment should be performed **in dry ice to avoid any freeze-thaw cycles** that may affect the material for analysis.

2.2.2.5 Equipment Requirements for Biomarker Analysis

The DIOPTRA biomarker pipeline ranges from early discovery and verification to clinical validation. Olink technology is based on **Proximity Extension Assay (PEA) technology** and utilises **Next Generation Sequencing (NGS)** methods to measure the concentration of thousands of human proteins[7]. The following equipment is required for biomarker analysis using **Olink, RNA sequencing and xMAP technology (Luminex Corp.)**:

1. **Illumina NextSeq 2000**: The system is required for the implementation of Olink technology and for NGS analysis of RNA samples. ILLUMINA NGS technology is used as a read-out of DNA-encoded tags conjugated in the antibodies to provide information on the type and relative amount of each biomarker in the sample of interest[8].

2. **SPT Labtech automation for Olink Explore**: The Mosquito & Dragonfly automation systems are required to perform an Olink Explore run. The system allows for accurately dispensing low volumes (nanolitre quantities) via non-contact technology, as well as plate preparation and replication.

3. **HAMILTON MOA STAR Basic Plus**: The HAMILTON automation system for NGS & OLink is a prerequisite for OLink analysis. The system is used to pool PCR1 products, prepare plates for PCR2 and then pool PCR2 products.

4. **PCR machines (high-throughput thermal cycler)** are used for the amplification of nucleic acids using the Polymerase Chain Reaction (PCR) process required for Olink analysis.

5. **Analysers & auxiliary equipment (pipettes, water baths, sonicators, rotators, magnetic separators, shakers, stirrers, etc)**: These systems are required for Olink, NGS and Luminex multiplex analysis.

6. **Tissue ruptor** is required for the simultaneous disruption and homogenisation of a sample through a combination of turbulence and mechanical shearing. It is used for protein/RNA isolation from tissue samples.

Funded by the European Union

©2023-2026 DIOPTRA

7. **PCR cabinet**: it is required for sensitive PCR amplification, manipulation of genetic material and preparation of PCR reaction mixtures for Olink and NGS analysis.

8. **Class II type A2 Biosafety Cabinet** is required for sample preparation and processing for Olink analysis.

9. **Class II Biosafety Cabinet Total Exhaust** is required for RNA isolation from human specimens (tissues, serum, etc) and use and handling of small volumes and amounts of volatile reagents during such processes.

10. **Qubit Fluorometer** for the quantitation of DNA, RNA, microRNA, and protein, as well as for the measurement of RNA integrity and quality required for Olink and NGS sample analysis.

11. **xMAP Intelliflex**: The Luminex instrumentation is required for biomarker analysis using the xMAP technology (Luminex Corp) that relies on color-coded microspheres (bead regions) to allow for the simultaneous detection of responses against multiple protein targets from the same sample.

2.2.2.6 Equipment Requirements for Assay Development

The following equipment is required for the development of multiplex assays using xMAP technology to validate the selected panel of protein biomarkers:

1. **xMAP Intelliflex** is required at all steps of assay development. xMAP technology (Luminex Corp) relies on color-coded microspheres (bead regions).

2. **Auxiliary equipment** (pipettes, water baths, sonicators, rotators, magnetic separators, shakers, stirrers, etc): These systems are required for assay development on Luminex platforms.

2.2.2.7 Infrastructure Requirements for Sample Collection, Storage and Analysis

PAO is required to undergo changes to facilitate the setup and operation of new equipment for sample collection, storage and analysis. Based on the specifications of the new equipment, various changes to the existing laboratory spaces are required to ensure the correct installation, operational and performance of the new equipment.

1. **Lab modifications**: Includes all site preparation work, additional benches, air-conditioning systems and electrical connections.

2. **Olink / NGS setup**: The handling & processing of DNA/RNA and other samples as part of the Olink activities requires special lab arrangements. At least two independent rooms, pre- and post-PCR, need to be set up to avoid contamination of samples from the post to the pre-PCR rooms. Pre- and post-rooms need to be connected only via a pass-through box for sample transfer.

3. **Special installation provisions** including dedicated bench space, dimension provisions, internet & electrical connections, storage space for consumables and temperature control are required to accommodate all the new instruments.

4. **Refrigerated centrifuges** for the manipulation of temperature-sensitive samples such as tissue lysates and cell lysates and the isolation of DNA, RNA and proteins for sample analysis: They require independent benches due to the generation of vibrations that may impact the operation of other instruments. They also require space to allow undisturbed opening & closing of the door.

Funded by the European Union

©2023-2026 DIOPTRA

5. **Flake Ice Maker** to produce flake shaped ice which maintains the conditions of biological samples during lab experiments, for a transit time outside refrigerator and deep freezers. It requires space provisions, electrical connection & water supply.

6. **Medical grade Fridge/Freezers** (-20°C freezers, -80°C freezers, +4°C refrigerators and +4°C / - 20°C refrigerator/freezers), suitable for a biotechnology laboratory to allow the storage of samples (tissue, serum, plasma) and reagents required for Olink, NGS and Luminex sample analysis: They require space provisions for undisturbed door opening/closing & stuff movement, electrical & internet connections & connection to monitoring system & temperature control.

7. **A humidity and temperature monitoring and recording system** for refrigerators, freezers and room areas is required to ensure that all samples and reagents are stored as expected, and instruments perform to specification retaining temperature levels within the specified range. It requires electrical connection for all monitored devices and rooms, internet connection, meterscope software and space on the wall for the meton display.

8. **Uninterruptible Power Supply (UPS)** to protect critical instruments for sample analysis (Illumina NextSeq 2000, Hamilton STAR, Luminex instruments, etc) from power surges and ensure their uninterrupted operation and safe shutdown: They require space provision near the instrumentation connected and electrical connection.

9. **A Network Attached Storage (NAS) device** is required to store and manage large amounts of data, such as genomic sequencing data following NGS sample analysis produced by Illumina NextSeq instruments. It requires space provision.

2.2.2.8 Regulatory Requirements for Assay Development

The development of multiplex assays involves the use of bead-based immuno-assays for the parallel quantitation of multiple proteins. Multiplex assays utilise the xMAP technology (Luminex Corp) that relies on color-coded microspheres (bead regions) to allow for the simultaneous detection of responses against multiple protein targets from the sample. Each bead region is coated with an antibody that recognises and binds to a specific part of the protein. Mixtures of bead regions are used in a sandwich-type ELISA assay to provide relative and absolute quantification of multiple proteins across the various conditions tested. These assays offer high multiplexability, sample throughput, quality of measurements and specificity for measurement of identified biomarkers in blood. The assay development pipeline follows PAO's QMS, which is required to be compliant to ISO13485, GMP and the European guidelines for IVD assays. This assay development pipeline is required to comply to the above regulations and include the following steps:

1. **Feasibility study**, where an assay is assessed for its feasibility to be developed and to detect the target protein in its recombinant and natural form in serum and plasma samples:

    - **Antibody and antigen selection and procurement**: polyclonal and monoclonal antibodies that recognise the target protein and full-length recombinant proteins will be acquired.

    - **Antibody clean-up processes** to remove free amines that interfere with the subsequent coupling and biotinylation processes, and BSA.

    - **Bead micro-coupling and QC**: an antibody is covalently coupled to a Magplex bead and a quality control for confirmation of successful micro-coupling is performed.

Funded by the European Union

©2023-2026 DIOPTRA

- **Antibody micro-biotinylation and QC**: Biotin molecules are attached to antibodies for subsequent use as detection antibodies in a sandwich ELISA format. A quality control protocol is also used for confirmation of successful micro-biotinylation.

- **Feasibility assessment and antibody pairing**: For each target, several assays are created using all possible antibody pairs. Optimal pairs are assessed using recombinant protein and samples of interest (characterised samples: positive and negative for the presence of target protein).

2. **Development and Performance Verification**, where a prototype multiplex assay is developed and verified for its analytical performance using recombinant proteins and samples of interest:

   - **Bead coupling and QC**: coupling of optimal capture antibodies to MagPlex magnetic beads (Luminex) and quality control for confirmation of successful coupling of antibodies

   - **Biotinylation and QC**: biotinylation of optimal detection antibodies using NHS-PEG4-biotin and quality control for confirmation of successful biotinylation of antibodies

   - **Optimisation** of detection antibody concentration

   - **Optimisation** of sample diluents and sample dilutions for each sample type (serum, plasma)

   - **Cross-reactivity assessment**: assays are tested for multiplexability

   - **Analytical performance**: several assay parameters are assessed including sensitivity, selectivity, precision and accuracy.

   - **Production of pilot batch**: the pilot batch is used to analyse well-characterised samples and results are analysed to assess assay performance.

3. **Clinical Performance Assessment**, where each developed assay is assessed with relation to its intended clinical use in characterised patient blood samples: Samples from different staging and negative samples, all well-characterised, are analysed with the developed multiplex assay.

2.2.2.9 Regulatory Requirements for Clinical Samples (GCLP)

The analysis of samples collected from the DIOPTRA clinical studies forms an essential part of the project and provides important data on a range of endpoints (discovery & validation). Therefore, it is essential that sample collection, analysis and reporting is performed to a standard which will ensure that data is reliable, accurate and in compliance with **Good Clinical Practice (GCP) regulations**. PAO is required to comply with **Good Clinical Laboratory Practices (GCLP)** to undertake the analyses of samples from the DIOPTRA clinical study. The GCLP guidelines include guidance on the facilities, systems and procedures that should be present to assure the reliability, quality and integrity of the work and results generated during the analysis of clinical samples from the DIOPTRA study. GCLP is intended to ensure that the requirements of GCP applicable to the analysis of clinical samples are met. PAO needs to update their QMS to adhere to GCLP standards.

2.3 TECHNICAL REQUIREMENTS

The DIOPTRA project comprises both retrospective and prospective studies, which are carried out in collaboration with eight clinical sites based on the clinical workflow presented in Section 2.2.1. The

©2023-2026 DIOPTRA

technical core of this undertaking is to provide these clinical partners with an **anonymisation tool** (Section 4.2.2), deploy a **centralised data platform** (Section 4.2.1) for data uploading and conduct **data curation** (Section 4.2.3) to validate and enhance the quality of the data. This infrastructure serves the dual purpose of **integrating heterogeneous datasets** and enabling the development of predictive models that are essential for **risk assessment and personalised suggestions** (Section 4.2.5). The project's technical framework comprises several components, including **data management**, **biological sample analysis**, **risk factor model development**, and the creation of a **mobile application** (Section 4.2.4) for the collection of behavioural data. Precise technical specifications and considerations must be applied to each component to guarantee consistent functionality, strict data privacy, and the effective achievement of the project's goals. The technical requirements necessary for the successful implementation of the DIOPTRA project are detailed in this document. It provides an extensive review of the essential tool, platform, methodologies, and models that are critical for each phase. It highlights the importance of robustness and interoperability to integrate heterogeneous datasets and obtain valuable knowledge.

2.3.1 Anonymisation and Data Storage

The **user interface** of the anonymisation tool is crucial in facilitating straightforward integration and efficient operation for clinical partners. It comprises various technical requirements that are essential for its operation. The user interface should feature a design that is simple in nature, enabling users to navigate with ease. Elements such as **explicit labelling**, **consistent guidelines**, and **easily accessible functionalities** are critical in guaranteeing a user-friendly experience when interacting with the tool. Customisation options are of the utmost importance, as they allow users to personalise anonymisation processes according to the sensitivity levels and data types. Users are granted **greater control over the process by specifying fields for anonymisation and modifying the degree of anonymisation that is applied**. The integration of **explicit feedback mechanisms** provides users with direction during the anonymisation process, whereas informative error messages facilitate the prompt resolution of issues, thereby enhancing the overall user experience. The inclusion of **instructional resources**, **tutorials**, or **tooltips** in the user interface increases user understanding and acceptance of the tool. It is imperative that **troubleshooting documentation** and support channels be easily accessible to ensure the convenience of users. The maintenance of consistency in design elements, interactions, and terminology throughout various sections of the tool plays a substantial role in fostering a unified user experience, thereby improving the usability and navigability of the user interface.

Funded by the European Union

©2023-2026 DIOPTRA

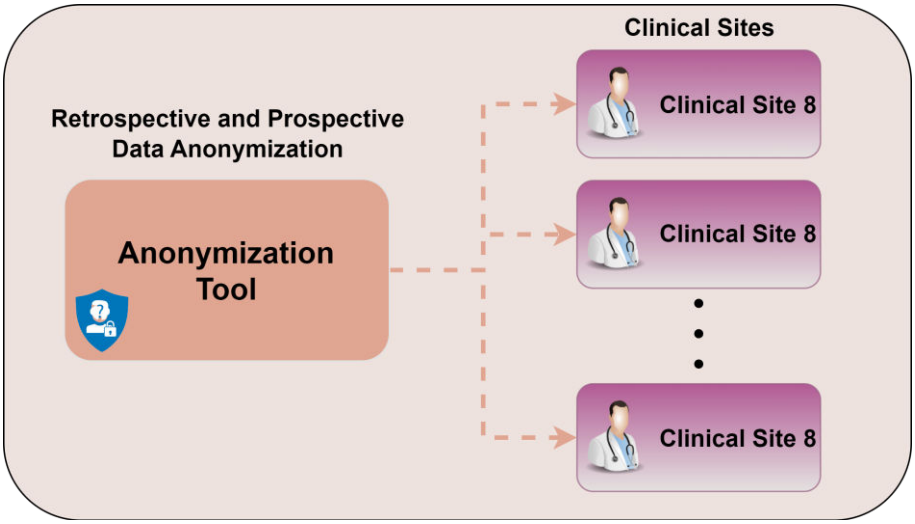

Figure 3: Implementation of Data Anonymisation Tool

Ensuring optimal utilisation and allocation of resources is vital for the sustainability of high performance. To maximise system efficiency and performance, dynamic resource allocation against processing demands is vital. It is critical to uphold satisfactory response times and throughput rates. To maintain a streamlined workflow, it is important that users encounter **minimal delays while handling the data**. It is essential to conduct extensive **scalability tests across a range of load scenarios** to evaluate the tool's capacity to manage increased data volumes and **detect any possible bottlenecks**. The implementation of redundancy and failover mechanisms ensures the dependability of the system, protecting it from disruptions and guaranteeing continuous service availability. Consistent monitoring of system performance is critical to proactively detect bottlenecks and implement subsequent optimisation strategies that will improve the overall efficacy of the system.

Integration capabilities are fundamental for the anonymisation tool, enabling **seamless interaction** with the centralised platform and other systems. It is essential to have support for **common data formats**, including CSV and JSON, to submit data to the centralised platform. By ensuring **compatibility with widely used formats**, integration processes are streamlined, and error-free data transfer is achieved, thereby improving interoperability with diverse systems. The incorporation of error handling and logging mechanisms is essential for transparent integration procedures. Clear error messages and detailed logs assist in diagnosing integration issues promptly, contributing to efficient troubleshooting and system maintenance.

The centralised data platform acts as the **primary infrastructure for the administration and curation of data collected** from clinical partners participating in the project. It requires a variety of technical requirements that are vital for its reliable operation and effectiveness. The platform should incorporate **robust data storage** mechanisms that possess the ability to effectively manage a wide range of data types and significant volumes. It is critical to **prioritise data security**, dependability, and scalability to maintain flexible data management. It is essential to have the capability to support various data formats, such as CSV, JSON, and others, to enable smooth data integration and ingestion from a wide range of sources such as the anonymisation tool. Ensuring the ability to adapt to diverse data structures is crucial in the process of data harmonisation.

Managing metadata efficiently facilitates the categorisation and organisation of datasets. The adoption of metadata standards improves the ability to locate and understand data, thereby facilitating efficient data administration. The implementation of **resilient backup and disaster recovery systems** serves to reduce the likelihood of data loss and protects the availability and integrity of the data that is stored.

Funded by the European Union

©2023-2026 DIOPTRA

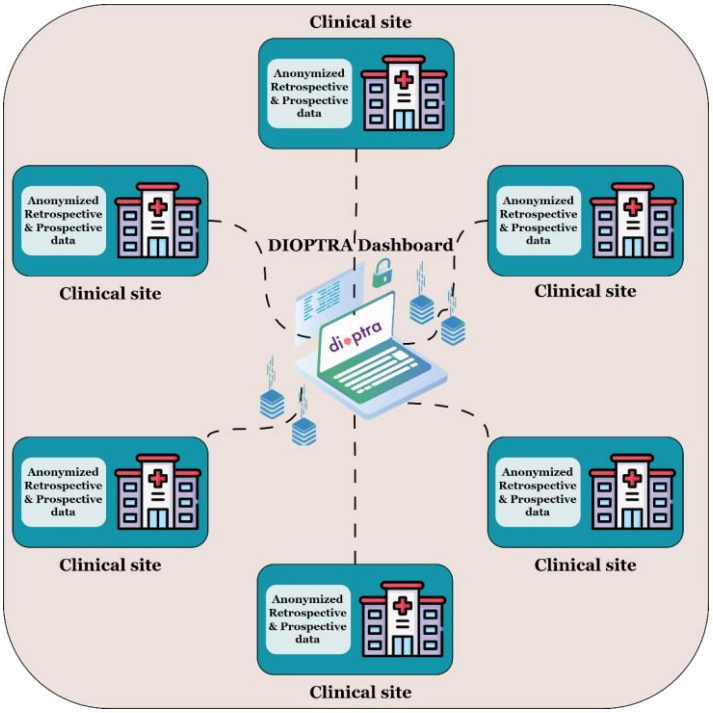

Figure 4: DIOPTRA Dashboard Service for Clinical Sites

2.3.2 Data Curation

Following the data upload to the centralised platform, the **data curation phase** within the DIOPTRA platform, encompasses several critical technical requirements. **Validation mechanisms** are critical measures that guarantee the accuracy and consistency of data submitted to the platform by ensuring that it conforms to predetermined standards. To augment the quality of the data, advanced tools and algorithms are incorporated, carrying out operations including **error identification**, **outlier detection** and **identification of any inconsistencies** on the provided data. The utilisation of automated procedures for data pre-processing and filtering is crucial in streamlining the curation process. By ensuring standardised data formats, these automated procedures not only expedite the curation process but also improve the overall consistency of the data. Strict security measures are required to protect curated data. Strict protocols are operationalised to **safeguard the curated datasets from unauthorised access or manipulation**, thereby ensuring adherence to data privacy regulations and preserving their integrity. A complete documentation is produced to provide a detailed description of the procedures used for data curation. This ensures transparency and aids in the comprehension of the curation process. Moreover, **reporting functionalities monitor the advancement of curation and quality indicators**, thereby furnishing valuable perspectives on the efficacy of the curation process.

©2023-2026 DIOPTRA

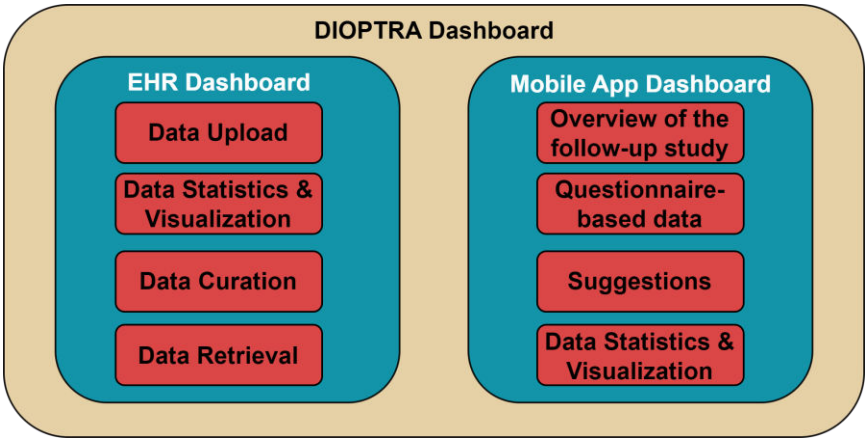

Figure 5: DIOPTRA Dashboard

2.3.3 Predictive Modelling

The **Predictive Model Development** phase within the DIOPTRA project involves a series of technical requirements. The procedure of algorithm selection is a vital initial phase in which appropriate statistical and AI-based algorithms are identified and implemented. The selection of these algorithms is predicated on their compatibility with the project's goals, which include predictive modelling and risk assessment utilising the curated dataset. Following this, **feature extraction** methods are applied to extract and enhance relevant features from the dataset. This comprises data on **medical**, **behavioural**, and **risk factors**, thereby guaranteeing the integration of pertinent information that is crucial for precise predictive modelling. Utilising the curated data, the developed models are subjected to **training and validation procedures**. The models' precision and dependability are ensured through the implementation of validation methods, including cross-validation and performance metric evaluation. Ensemble learning methods, which involve the combination of multiple models or techniques, are frequently employed to improve the accuracy of predictions. It is significant to prioritise the assurance of interpretability and explainability of the models. By integrating techniques that improve stakeholders' awareness of model outcomes, confidence in the models' predictions is increased. It is of the utmost importance to tackle imbalanced datasets, a prevalent obstacle in healthcare data. Methods such as oversampling, under-sampling, or the implementation of specialised algorithms are utilised to efficiently manage class imbalance.

2.3.4 Mobile App Services

A variety of necessary technical requirements are also associated with the **mobile application** component of the project. By utilising the risk factor model that will be developed, the application should possess the capability to compute **personalised behavioural risk scores** by analysing the gathered behavioural data. The application's features encompass the **provision of tailored suggestions** to users, which are generated using their behavioural data and risk scores. An **interface that is both straightforward and user-friendly** is critical for promoting **user engagement** and interaction. This feature guarantees effortless navigation, data entry, and comprehension of personalised recommendations. It is imperative to guarantee **compatibility among the DIOPTRA platform and diverse mobile platforms**, such as iOS and Android, to optimise functionality and performance across a range of devices. By incorporating **active user engagement mechanisms** such as reminders for data entry, notifications for personalised recommendations, and feedback systems, proactive health management is promoted, and user interaction is strengthened. Processes of continuous development, which encompass the integration of user feedback and periodic updates containing new features or enhancements, are imperative for the evolution of the application. Moreover, incorporating **comprehensive documentation and user support resources** into the application

©2023-2026 DIOPTRA

facilitates user assistance and resolution of issues, thereby guaranteeing a **smooth and uninterrupted user experience**.

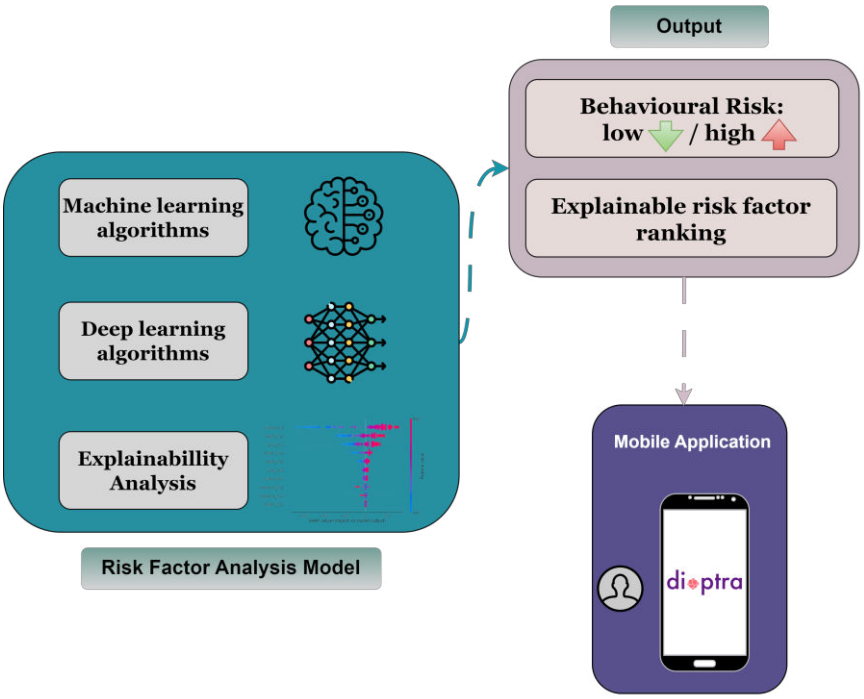

Figure 6: Risk factor model development and integration into the mobile application

©2023-2026 DIOPTRA

3 DIOPTRA USE CASES

Before defining the use cases, all actor groups relevant to the DIOPTRA ecosystem were identified:

Table 3: List of DIOPTRA Actor Groups

Group No	Group Name	Description
G1	Citizens	General population / candidates for receiving CRC screening services
G2	Patients	Individuals diagnosed with CRC or advanced adenoma (AA)
G3	Healthy / non-AA	Individuals belonging in the healthy or non-advanced adenoma groups
G4	Prospective Study Participants	Recruits for DIOPTRA prospective study (baseline/follow-up) / $\boldsymbol{G4 = G2 \cup G3}$
G5	Follow-up Study Participants	Subgroup of G4 that fulfils criteria and consents for participation in the follow-up study / $\boldsymbol{G5 \in G4, G5 \in G3}$
G6	Clinical Staff	Clinical staff of participating hospital
G7	Administrative / IT Staff	Administrative / IT staff of participating hospital
G8	Technical Partners	Partners conducting analytics on data features
G9	Biomarker Analysts	Partner analysing blood samples during prospective study & tissue/blood sample pairs during the discovery study

The above groups are involved in the use cases that are briefly summarised in the following tables:

Table 4: DIOPTRA Use Case 1 – General Info

Use Case ID	DIOPTRA_UC1
Name	Prospective / Retrospective Data Acquisition and Upload from Clinical Sites
Description	The system provides its users with the option to upload data in a predefined format. This covers both uploading new data and editing existing data (by replacing records with an updated version).
Actor(s)	G6, G7
Trigger	A clinical site has acquired a dataset that will be provided to the DIOPTRA database.
Dependencies	N/A

Funded by the European Union

©2023-2026 DIOPTRA

Table 5: DIOPTRA Use Case 2 – General Info

Use Case ID	DIOPTRA_UC2
Name	Delete Action of a Data Record from the Clinical Side
Description	The system provides the capability of deleting an existing record from the DIOPTRA database, initiated from within a clinical site.
Actor(s)	G6, G7
Trigger	A user needs to delete a record as a response to a request initiated either by the DIOPTRA partners or by a participant who wishes to exercise their right to be forgotten (RTBF)
Dependencies	DIOPTRA_UC1

Table 6: DIOPTRA Use Case 3 – General Info

Use Case ID	DIOPTRA_UC3
Name	Uploaded Data Overview
Description	The system provides the capability to overview the uploaded data, retrieving information about the data amount and quality
Actor(s)	G6, G7
Trigger	An actor within a clinical site wishes to view collective information on the data that have been uploaded to the DIOPTRA database by the clinical site (no access to data from other sites)
Dependencies	DIOPTRA_UC1

Table 7: DIOPTRA Use Case 4 – General Info

Use Case ID	DIOPTRA_UC4
Name	Data Review for Single Participant
Description	The system provides the capability to review the uploaded data for an individual
Actor(s)	G6, G7
Trigger	An actor within a clinical site wishes to view data that have been uploaded to the DIOPTRA database for a single participant of the clinical site (no access to participants from other sites)
Dependencies	DIOPTRA_UC1, DIOPTRA_UC3

Funded by the European Union

©2023-2026 DIOPTRA

Table 8: DIOPTRA Use Case 5 – General Info

Use Case ID	DIOPTRA_UC5
Name	Follow-up Study Implementation
Description	A participant that is recruited in the follow-up study uses a mobile application to provide questionnaire-based data and receive a personalised behavioural suggestion. A follow-up data acquisition is conducted after 1 year to evaluate a) behavioural change adherence and effect (via the app) and b) potential changes in the biomarker profile (via blood resampling and re-evaluation in the clinical site).
Actor(s)	G5, G6, G7
Trigger	A prospective study participant agrees to be contacted for the follow-up study and fulfils the corresponding criteria (including classification as healthy or non-AA based on the diagnosis conducted in the clinical site)
Dependencies	DIOPTRA_UC1, DIOPTRA_UC3, DIOPTRA_UC4

Table 9: DIOPTRA Use Case 6 – General Info

Use Case ID	DIOPTRA_UC6
Name	Follow-up Study Monitoring
Description	The system provides the capability to monitor current recruitment status & behavioural suggestions provided for each participant
Actor(s)	G6, G7
Trigger	A user requires to monitor the status of the study and/or data for an individual participant
Dependencies	DIOPTRA_UC1, DIOPTRA_UC3, DIOPTRA_UC4, DIOPTRA_UC5

Table 10: DIOPTRA Use Case 7 – General Info

Use Case ID	DIOPTRA_UC7
Name	DIOPTRA-based Screening
Description	Real-world scenario for CRC screening using the DIOPTRA solution
Actor(s)	G1, G6
Trigger	A citizen visits a clinical site to receive CRC screening services
Dependencies	N/A

Funded by the European Union

©2023-2026 DIOPTRA

Table 11: DIOPTRA Use Case 8 – General Info

Use Case ID	DIOPTRA_UC8
Name	DIOPTRA Mobile App Use
Description	Real-world scenario for use of the DIOPTRA mobile app
Actor(s)	G1
Trigger	A citizen downloads the DIOPTRA mobile app to gain access to CRC-related material (prevention, screening, etc.)
Dependencies	N/A

©2023-2026 DIOPTRA

4 DIOPTRA ARCHITECTURE

4.1 OVERVIEW & COMPONENTS

Considering a) the clinical workflow including all the clinical studies, b) the development and implementation of the biomarker-based screening, c) the fulfilment of the identified use cases, as well as d) the overall outcomes of the requirements' elicitation process, the list of all main DIOPTRA components was conjured:

Table 12: List of Main DIOPTRA Components

Code	Name	Description
C1	Clinical Site Interface	Interface accessible by clinical sites participating in the DIOPTRA studies for a) structured retrospective and prospective data upload, b) overall data management by clinicians, and c) review of data acquired via the questionnaires within the DIOPTRA mobile app
C2	Anonymisation Tool	Software tool for application on the EHR datasets prior to their upload to the DIOPTRA database via the Clinical Site Interface (C1)
C3	Data Curation & Storage System	Back-end Solution for dataset curation and storage
C4	Mobile Application	Application with the following components: 1) Risk Assessment & Personalised Suggestion Module, 2) Health Literacy Module, 3) Diary
C5	Risk Assessment Module	Software model utilising risk factor data for risk assessment towards behavioural recommendations
C6	Screening AI Module	Full screening model utilising risk factor data and blood biomarkers
C7	Multiplex Biomarker Assay	Assay for extracting selected biomarker features with CRC diagnostic capacity based on the results of the biomarker discovery process

C1-C6 are part of the technical infrastructure of DIOPTRA, with the specific requirements for each component being presented in Section 4.2. General requirements for development and implementation of C7 were described in Section 2.2.2.

With the definition of the components of Table 12, a more detailed description of the use cases of section 3 was conjured, listing the relevant components for each workflow:

Funded by the European Union

©2023-2026 DIOPTRA

Table 13: DIOPTRA Use Case 1 – Workflow & Related Components

Use Case ID	DIOPTRA_UC1
Name	Prospective / Retrospective Data Acquisition and Upload from Clinical Sites
Related Components	C1, C2, C3
Workflow	1. Data acquisition within the clinical site (EHR & questionnaires) 2. Assignment of participant ID 3. Organisation of data in a structured excel file based on requirements for application of C2 and upload to C1 4. Application of the anonymisation tool 5. Dashboard login, action selection & data upload 6. Review of the feedback report provided by the system, including the following: • Decision on approval or rejection of the file based on format & content criteria (required fields, allowed values, etc.) • Total number of new records uploaded and / or existing records updated (with list of participant IDs) • Availability of variable categories • Quality criteria within each upload 7. Further actions if needed (e.g. file re-formatting and re-upload based on feedback)
Comments / Prerequisites	• Excel files must go through C2 before upload to C1 • C2 uses the k-anonymity technique which requires that **upload actions should be performed in batches** • Clinical sites **should maintain local copy of the excel files** provided to the DIOPTRA system • The actor performing the upload action should **check the feedback report** provided by the system and repeat the file preparation and upload process if needed • For **data editing**, the same process is followed by uploading a new excel file to replace the record of the corresponding participant ID • A **variable category** is regarded as available if at least one of its variables is available within the dataset • If the database contains any records with unavailable variable categories (i.e. all category variables missing), a **notification** is provided to the user upon login to check the existing database records (DIOPTRA_UC3). For each clinical site, the notification concerns only the data uploaded by this specific site.

©2023-2026 DIOPTRA

Table 14: DIOPTRA Use Case 2 – Workflow & Related Components

Use Case ID	DIOPTRA_UC2
Name	Delete Action of a Data Record from the Clinical Side
Related Components	C1, C3
Workflow	1. Dashboard login & action selection 2. Selection of the participant ID for deletion 3. Action completion
Comments / Prerequisites	• Only selected users will have the required access right to delete a record from the database based on the user roles assigned

Table 15: DIOPTRA Use Case 3 – Workflow & Related Components

Use Case ID	DIOPTRA_UC3
Name	Uploaded Data Overview
Related Components	C1, C3
Workflow	1. Dashboard login & action selection 2. The system presents cumulative information on all collected data

Comments / Prerequisites	• Each clinical site will only have access to their own data
	• Specific fields for data reviewing will include:
	  ◦ Total number of participants, number of participants per diagnosis / study / sex / age range (age ranges have been defined as follows: <49, 50-59, 60-69, >70)
	  ◦ Data availability for each of the following categories defined in the DIOPTRA Data Template (availability is assumed at a category basis and not for each individual variable)
	  ▪ DIOPTRA-related
	  ▪ Sociodemographic
	  ▪ Lifestyle, Diet, Supplements and Stress
	  ▪ Family and personal medical history
	  ▪ Clinical biology
	  ▪ Colonoscopy - Symptoms and Procedural
	  ▪ Colonoscopy - Diagnosis
	  ▪ Blood sample collection

Funded by the European Union

©2023-2026 DIOPTRA

	• Data quality metrics regarding the number of missing values of the dataset based on data curation results)Other data quality metrics (to be defined based on data curation results)

Table 16: DIOPTRA Use Case 4 – Workflow & Related Components

Use Case ID	DIOPTRA_UC4
Name	Data Review for Single Participant
Related Components	C1, C3
Workflow	1. Dashboard login & action selection 2. The user is presented with a list of all individual records 3. The user can then filter among all records and/or select a specific record for review

Comments / Prerequisites	• Each clinical site will only have access to their own participants
	• Specific fields for data reviewing will include:
	  • Participant ID, diagnosis (healthy / non-AA / AA / CRC), study participation, year of birth, sex, age range (age ranges have been defined as follows: <49, 50-59, 60-69, >70)
	  • Data availability for each category as per DIOPTRA_UC3 (availability is assumed at a category basis and not for each individual variable)
	  • Upload feedback report from DIOPTRA_UC1

Table 17: DIOPTRA Use Case 5 – Workflow & Related Components

Use Case ID	DIOPTRA_UC5
Name	Follow-up Study Implementation
Related Components	C1, C2, C3, C4, C5
Workflow	1. An individual visits a DIOPTRA clinical site 2. Prospective study recruitment 3. Data acquisition (EHR, questionnaires & blood sample) 4. Diagnosis (classification into healthy / non-AA / AA / CRC) 5. Follow-up study enrolment (contacted by clinical site) 6. App download & login using participant ID 7. Risk stratification & behavioural recommendations by the app module 8. In between: App notifications for adherence & new data via a diary functionality

Funded by the European Union

©2023-2026 DIOPTRA

	9. Follow-up: Date re-entry via the app (questionnaire) & follow-up visit for blood resampling
Comments / Prerequisites	• App availability via app stores • Follow-up study participants (subset of prospective study participants) must fulfil the corresponding inclusion criteria • If the participant is eligible for the follow-up study, they are contacted by the clinical site to download the app

Table 18: DIOPTRA Use Case 6 – Workflow & Related Components

Use Case ID	DIOPTRA_UC6
Name	Follow-up Study Monitoring
Related Components	C1, C3, C4, C5
Workflow	1. Dashboard login & action selection 2. The user is presented with recruitment status (number of participants per group, data availability, etc.) 3. The user can filter among participants and view the behavioural suggestions provided
Comments / Prerequisites	• Each clinical site will only have access to their own participants • Only data from follow-up study participants will be included, information from users belonging to the general population will not be stored and provided to clinicians • The app login for DIOPTRA participants will be implemented using the participant identifier

Table 19: DIOPTRA Use Case 7 – Workflow & Related Components

Use Case ID	DIOPTRA_UC7
Name	DIOPTRA-based Screening
Related Components	C6, C7
Workflow	1. An individual (general population) visits a clinical site for CRC screening 2. Data acquisition (EHR, behavioural, blood sampling using C7) 3. Classification & decision-making (using C7 & C6) 4. Next steps in clinical workflow based on diagnosis

Funded by the European Union

©2023-2026 DIOPTRA

Comments / Prerequisites	• The following should be available at the clinical site: DIOPTRA biomarker assay (C7), DIOPTRA screening model (C6), input data model & application guidelines/templates

Table 20: DIOPTRA Use Case 8 – Workflow & Related Components

Use Case ID	**DIOPTRA_UC8**
Name	DIOPTRA Mobile App Use
Related Components	C5, C7
Workflow	1. An individual (general population) downloads the DIOPTRA mobile app 2. Anonymous access to the app modules for health literacy & personalised recommendations is granted. 3. The user answers the baseline questionnaire related to their behavioural, medical, and family history profile. 4. Following the analysis of the responses, the RAM will identify and rank the contributing risk factors and generate a personalised behavioural risk score. 5. Based on the identified risk factors, specific recommendations will be triggered and provided to the individual, to promote a healthier lifestyle.
Comments / Prerequisites	• The DIOPTRA app should be freely available via app stores • No substitution of healthcare services is offered by the app

The above information on DIOPTRA development, clinical study implementation and exploitation of components towards the final outcomes of the project was translated into the diagram of Figure 7, which initiates from base requirements and data input forms (i.e. sample collection & management, inclusion/exclusion criteria, data template, questionnaire forms) and results to the final DIOPTRA development and evaluation processes via the 4 studies (retrospective, biomarker discovery, prospective, follow-up) that are supported by specific DIOPTRA components.

Funded by the European Union

©2023-2026 DIOPTRA

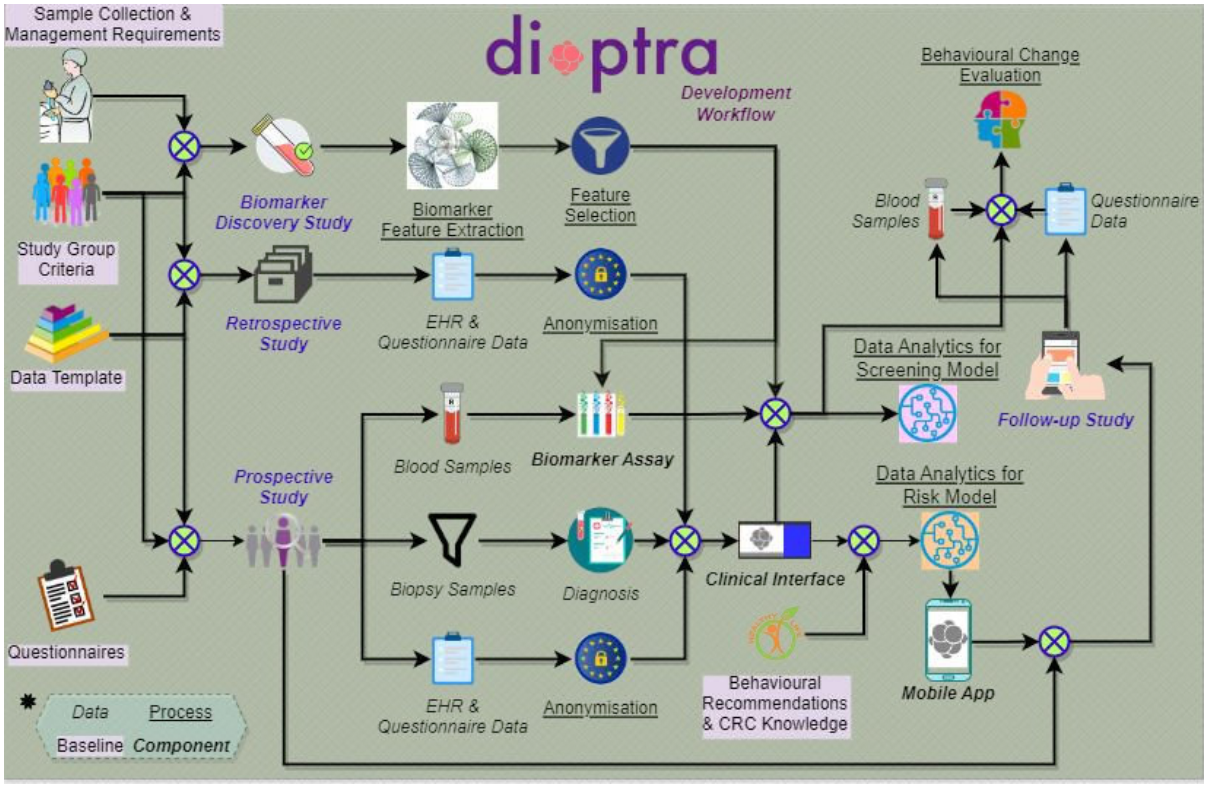

Figure 7: DIOPTRA Development within the Study Workflows

Furthermore, the additional diagram of Figure 8 was designed to depict the data flow within DIOPTRA and the corresponding components (from Table 12), highlighting the datasets that are transferred outside the acquisition site to the DIOPTRA storage infrastructure. Related groups (from Table 3) that receive and/or process data are also depicted.

©2023-2026 DIOPTRA

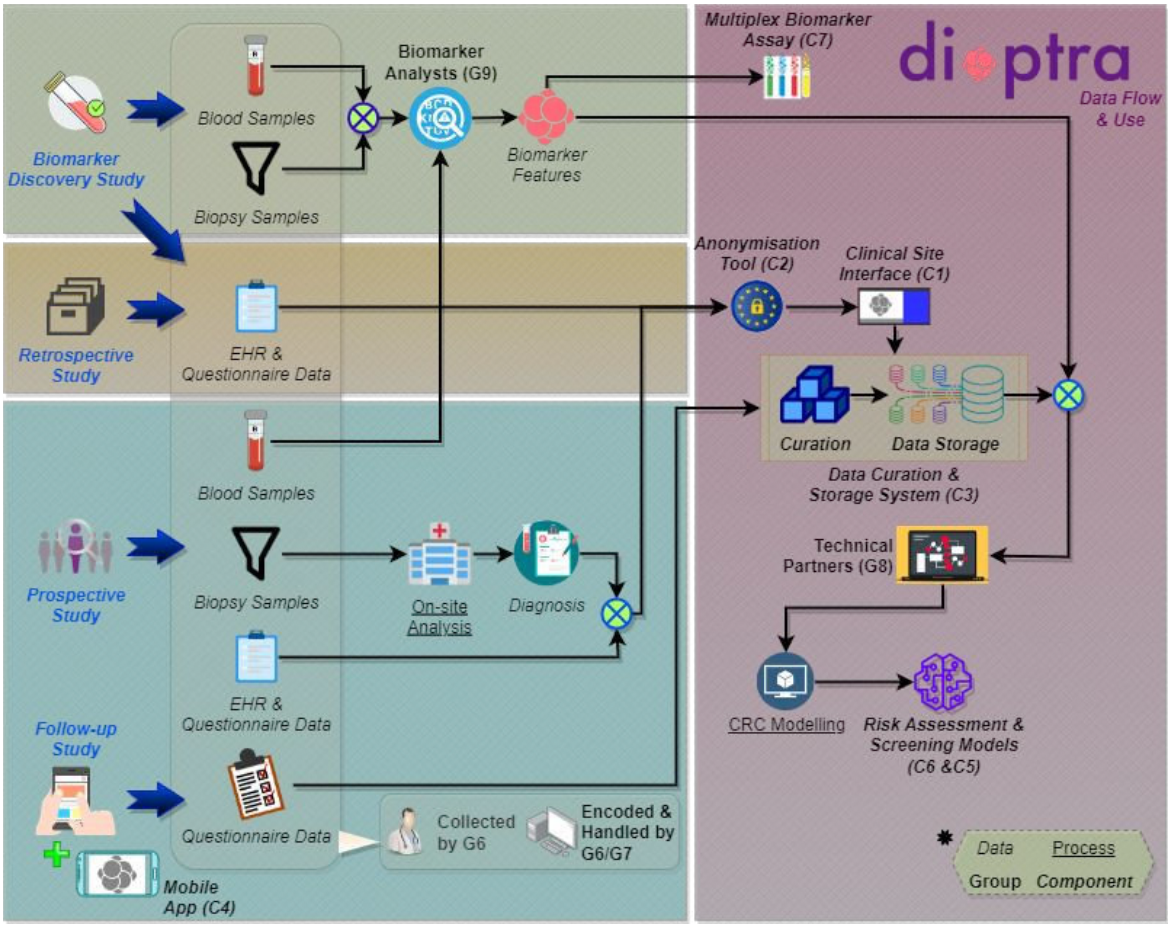

Figure 8: DIOPTRA Data Flow & Storage

4.2 DESCRIPTION OF DIOPTRA COMPONENTS

4.2.1 Clinical Site Interface

The **Clinical Site Interface** aims to support the day-to-day work for groups G6 and G7 with regard to data management and monitoring, as well as to facilitate data incorporation into the DIOPTRA database so that technical partners may access the data via the back-end. Specifically, the main functionalities of the Clinical Site Interface comprise the following:

- Uploading tabular retrospective and prospective data in predefined formats
- Access to an administrative dashboard providing information about the volume and quality of data uploaded from a clinical site
- Access to a clinical dashboard providing an overview of the follow-up study data collected via the mobile app

Funded by the European Union

©2023-2026 DIOPTRA

4.2.1.1 Functional Requirements

Table 21: Functional Requirement #01 for Clinical Site Interface

Requirement ID	DIOPTRA-CSI-FR-01
Name	Structured data upload
Description	The user should be able to add new datasets to the DIOPTRA database using the interface to upload a structured file with a predefined format.
Priority	High (it will be implemented)
Type	Functional
Rationale	Users at a clinical site should provide structured datasets to the DIOPTRA database, enabling technical partners to use the collected data to develop the risk assessment and screening AI modules.
Verification	Collection of a CRC-related dataset consistent with the format of the DIOPTRA Data Template
Completion Criteria	• Users in all clinical sites are able to successfully complete the data upload process • Data upload via the interface populates the DIOPTRA back-end storage • Content of back-end storage consistent with the data template format
Relevant Use Cases	DIOPTRA_UC1
Dependencies	DIOPTRA-CSI-FR-06, DIOPTRA-AT-FR-01

Table 22: Functional Requirement #02 for Clinical Site Interface

Requirement ID	DIOPTRA-CSI-FR-02
Name	System reporting on data characteristics
Description	When a user in a clinical site uploads a structured dataset, the system should provide a feedback report informing on the action result and the characteristics of the data uploaded.
Priority	High (it will be implemented)
Type	Functional
Rationale	The users should be informed on the successful result of their action and be able to ensure that the uploaded dataset characteristics match the intended ones (including the results of the anonymisation process), in order to identify potential errors within the data file. This is also important for the technical partners to ensure the integrity and completeness of the uploaded data.

©2023-2026 DIOPTRA

Verification	CRC-related datasets uploaded in the DIOPTRA back-end storage do not include data that do not satisfy the requirements of the data template.
Completion Criteria	• Test runs with demo data successfully reject datasets with unwanted characteristics (e.g. value types, non-anonymised variables) • Test runs with demo data accurately inform the user on the characteristics of the uploaded data and the corresponding changes to the database
Relevant Use Cases	DIOPTRA_UC1
Dependencies	DIOPTRA-CSI-FR-01, DIOPTRA-CSI-FR-06, DIOPTRA-CSI-FR-07

Table 23: Functional Requirement #03 for Clinical Site Interface

Requirement ID	**DIOPTRA-CSI-FR-03**
Name	Editing of uploaded data
Description	A user within a clinical site should be able to edit the data of a record
Priority	High (it will be implemented)
Type	Functional
Rationale	A user within a clinical site may need to add more values to the existing data of a study participant (e.g. when the diagnosis is available), or
Verification	Users are able to edit an existing data record
Completion Criteria	• Successful test runs with demo data in all clinical sites
Relevant Use Cases	DIOPTRA_UC1
Dependencies	DIOPTRA-CSI-FR-01, DIOPTRA-CSI-FR-06, DIOPTRA-CSI-FR-07, DIOPTRA-AT-FR-01

Table 24: Functional Requirement #04 for Clinical Site Interface

Requirement ID	**DIOPTRA-CSI-FR-04**
Name	Delete action of uploaded data
Description	A user within a clinical site should be able to delete all information stored in the DIOPTRA database for a participant.
Priority	High (it will be implemented)
Type	Functional

Funded by the European Union

©2023-2026 DIOPTRA

Rationale	Users should be able to delete any test data records or data from real participants upon request (e.g. right to be forgotten).
Verification	Clinical site users can successfully delete data from their clinical site
Completion Criteria	• Successful test runs with demo data in all clinical sites • Verification of data erasure from a back-end perspective
Relevant Use Cases	DIOPTRA_UC2
Dependencies	DIOPTRA-CSI-FR-01, DIOPTRA-CSI-FR-06, DIOPTRA-CSI-FR-07

Table 25: Functional Requirement #05 for Clinical Site Interface

Requirement ID	DIOPTRA-CSI-FR-05
Name	Data overview capability
Description	A user within a clinical site should be able to view characteristics and quality measures of all uploaded data both collectively and for any individual participant.
Priority	High (it will be implemented)
Type	Functional
Rationale	Clinicians need to monitor the recruitment status for all studies within their clinical site and review collected data.
Verification	User satisfaction on the data overview capabilities
Completion Criteria	• Applicability of search filters across data • Availability of data quality measures to the user • Availability of recruitment status (filtered based on the variables described in the relevant use cases) • Overview capability for data from the retrospective, prospective and follow-up studies
Relevant Use Cases	DIOPTRA_UC3, DIOPTRA_UC4, DIOPTRA_UC6
Dependencies	DIOPTRA-CSI-FR-01, DIOPTRA-CSI-FR-06, DIOPTRA-CSI-FR-07

Table 26: Functional Requirement #05 for Clinical Site Interface

Requirement ID	DIOPTRA-CSI-FR-05
Name	Provision of user notifications
Description	A user within a clinical site should be notified on pending actions required from their side.

©2023-2026 DIOPTRA

Priority	High (it will be implemented)
Type	Functional
Rationale	Different variable categories will be available by different departments within a clinical site. Data may be collected by the user responsible for upload at different timepoints for each category. Moreover, some variables (e.g. diagnosis) will be available at a later timepoint compared to the first upload of the participant data. Therefore, corresponding notifications should remind users that the database includes records with variables pending to be added.
Verification	User satisfaction on the notification functionality
Completion Criteria	• Test runs verify the successful identification of records with pending data input • Notification visualisation in the interface upon user login
Relevant Use Cases	DIOPTRA_UC1, DIOPTRA_UC3, DIOPTRA_UC4
Dependencies	N/A

Table 27: Functional Requirement #06 for Clinical Site Interface

Requirement ID	**DIOPTRA-CSI-FR-06**
Name	Differentiation of access privileges for different users within a clinical site
Description	Users within a clinical site should have different access privileges concerning the allowed actions.
Priority	High (it will be implemented)
Type	Functional
Rationale	Not all users should be able to delete data from the DIOPTRA database or have access to the detailed data of a single participant.
Verification	User satisfaction on access rights
Completion Criteria	• Test runs on availability of functionalities via the dashboard for the different access roles • Availability of an interface that allows the clinical site administrator to assign roles to users within their clinical site
Relevant Use Cases	DIOPTRA_UC1, DIOPTRA_UC2, DIOPTRA_UC3, DIOPTRA_UC4, DIOPTRA_UC6
Dependencies	N/A

Funded by the European Union

©2023-2026 DIOPTRA

Table 28: Functional Requirement #07 for Clinical Site Interface

Requirement ID	DIOPTRA-CSI-FR-07
Name	Data access across clinical sites
Description	A user within a clinical site should only have access to data from the participants of their own clinical site.
Priority	High (it will be implemented)
Type	Functional
Rationale	Each clinical site should not have access to the data collected by other clinical sites.
Verification	Access restriction for users
Completion Criteria	• Successful test runs using demo data for all clinical sites
Relevant Use Cases	DIOPTRA_UC2, DIOPTRA_UC3, DIOPTRA_UC4, DIOPTRA_UC6
Dependencies	N/A

4.2.1.2 Non-Functional Requirements

Table 29: Non-Functional Requirement #01 for Clinical Site Interface

Requirement ID	DIOPTRA-CSI-NFR-01
Title	Interface availability to all clinical sites
Description	Users within each clinical site are able to connect to the Clinical Site Interface.
Priority	High (it will be implemented)
Type	Non-functional
Rationale	All clinical sites should have access to the Clinical Interface component.
Verification	Test runs

Table 30: Non-Functional Requirement #02 for Clinical Site Interface

Requirement ID	DIOPTRA-CSI-NFR-02
Title	User access
Description	Implementation of user authentication and authorisation mechanisms for accessing functionalities
Priority	High (it will be implemented)
Type	Non-functional

©2023-2026 DIOPTRA

Rationale	To support functional requirements on user access
Verification	Test runs

Table 31: Non-Functional Requirement #03 for Clinical Site Interface

Requirement ID	DIOPTRA-CSI-NFR-03
Title	Interface communication with back-end
Description	API implementation for interaction among system components
Priority	High (it will be implemented)
Type	Non-functional
Rationale	Smooth communication and information exchange should be established.
Verification	API service testing

Table 32: Non-Functional Requirement #04 for Clinical Site Interface

Requirement ID	DIOPTRA-CSI-NFR-04
Title	Security
Description	Protection from unauthorised access
Priority	High (it will be implemented)
Type	Non-functional
Rationale	Ensure system integrity and enforcement of intended access rights for each user
Verification	Test runs

Table 33: Non-Functional Requirement #05 for Clinical Site Interface

Requirement ID	DIOPTRA-CSI-NFR-05
Title	User guidance
Description	Inclusion of clear guidance and complete information in the screens of the clinical interface
Priority	High (it will be implemented)
Type	Non-functional
Rationale	Users should be guided for easy navigation in the interface
Verification	User satisfaction

Funded by the European Union

©2023-2026 DIOPTRA

4.2.2 Anonymisation Tool

The main objective of the anonymisation app is the application of the **k-anonymity method using the Mondrian algorithm** on the input data provided by the user[9]–[11]. The anonymisation app is distributed as an executable program that can be run on a windows machine. The concept of k-anonymity for data anonymisation was developed due to the possibility of indirect identification of records from public databases. The k-anonymity method reduces the granularity of data representation, making it difficult for an intruder to determine the identity of the individuals in that data set. A k-anonymised data set has the property that each record is similar to at least another k-1 other records regarding a set of potentially identifying variables X_1, …, X_d. More formally, a table T satisfies the k-anonymity constraint with respect to its attributes X_1, …, X_d if every unique tuple (x_1, …, x_d) occurs at least k times, where X_1, …, X_d are the attribute names and x_1, …, x_d represent values of those attributes. A record in a k-anonymised data set has a maximum probability $1/k$ of being re-identified.

4.2.2.1 Functional Requirements

Table 34: Functional Requirement #01 for Anonymisation Tool

Requirement ID	DIOPTRA-AT-FR-01
Name	Personal data recording
Description	The application requires the following input from the user: 1. A delimited file containing the data, the first n rows of which contain the header lines. Currently the file types xlsx, xls are supported. 2. The definition of the fields that will be anonymised and the declaration of the data type of each field. The supported data types of the anonymisation app are the Numerical, Categorical, and Date types. Regarding the Date type, at least the following date and datetime formats are supported: • yyyy-mm-dd (ISO 8601) • yyyy/mm/dd • yyyy-mm-ddThh:mm:ss.SSS (ISO 8601) • yyyy-mm-dd hh:mm:ss (Mysql datetime) 3. The selection of a positive integer value for the k parameter 4. The selection of the number of header lines (n) included in the file. The supported options are n=1, 2, with 2 being the default value. The result of the anonymisation process is a file containing the anonymised data, as well as an error report and a loss metrics report including the loss metrics generalised information loss, discernibility metric and average equivalence class size.
Priority	High (it will be implemented)
Type	Functional

©2023-2026 DIOPTRA

Rationale	The data uploaded to the project's data backend should be adequately anonymised
Verification	All the variables chosen for anonymisation should be aggregated based on the k-value chosen.
Completion Criteria	• Output file with aggregated values in the fields chosen for anonymisation.
Relevant Use Cases	DIOPTRA_UC1, DIOPTRA_UC5
Dependencies	N/A

4.2.2.2 Non-Functional Requirements

Table 35: Non-Functional Requirement #01 for Anonymisation Tool

Requirement ID	DIOPTRA-AT-NFR-01
Title	Operating System Compatibility
Description	The anonymisation tool will be able to be used on any PC running Microsoft Windows OS.
Priority	High (it will be implemented)
Type	Non-functional
Rationale	The application should be able to be executed by workstations already present in the clinical sites.
Verification	N/A

4.2.3 Data Curation & Storage System

This component consists of the **Unified Data Asset** which is the main data infrastructure responsible for data collection, curation, and storage, providing M2M data access endpoints for the Clinical Site Interface, the Mobile App, and the ML models. Following a microservices approach, the software modules are autonomously deployed inside docker containers and the communication between them is realised through internal RESTful endpoints. All the services are deployed inside the provided GRNET infrastructure (Section 4.5).

The major functionalities of the Unified Data Asset in a nutshell:

1. **Heterogeneous Data Integration / Ingestion / Storage**
    - Retrospective, Prospective clinical data

©2023-2026 DIOPTRA

- Answers from questionnaires, outputs from other WPs
- Open-source datasets

2. **Data Curation**: Data Filtering / Harmonisation, Annotation, Cataloguing and Management
   - Field format
   - Variable value ranges
   - Correlation-based rules
   - Missing values
3. **Security**, **Data Protection**, **Secure Data Sharing**
4. **Data Visualisations**
   - Interactive Visualisation Dashboards with customised aggregated information visualisations
5. **Data Search**, **Retrieval** & **Interoperable Access** (unified data model)

The figure below depicts a high-level sequence diagram of the main workflows provided by the Data Curation & Storage System.

Funded by the European Union

©2023-2026 DIOPTRA

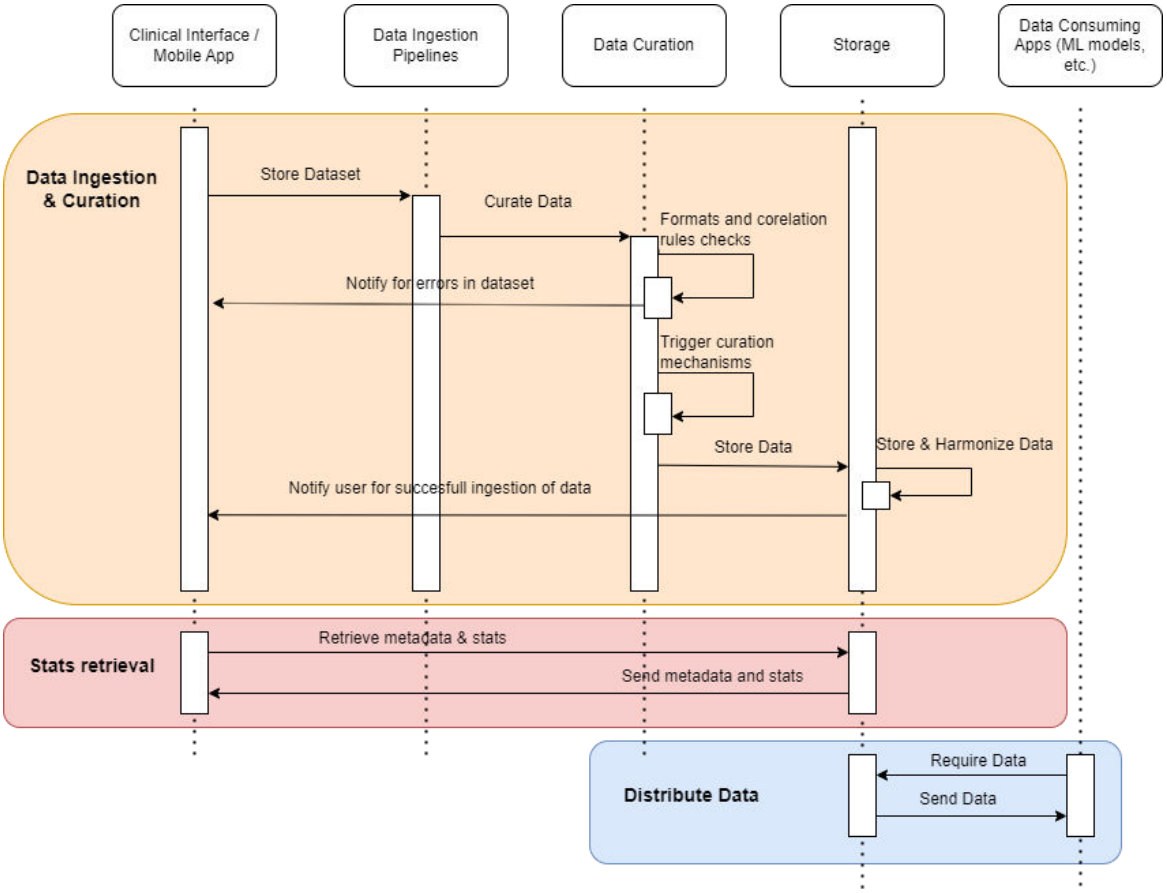

Figure 9: Data Curation & Storage System, high-level Sequence Diagram

4.2.3.1 Functional Requirements

Table 36: Functional Requirement #01 for Data Curation & Storage System

Requirement ID	DIOPTRA-DCS-FR-01
Name	Data Harmonisation & Transformation
Description	When a request for ingesting patients data arrives from the Clinical Interface, all data needs to be curated and validated against the unified data template. The pipelines provide a flexible way to collect clinical data and transform them.
Priority	High
Type	Functional
Rationale	Before storing data in the central storage, we need to make sure that every field is within the accepted values and no critical information is missing. As the stored data will be used to train ml models, we need to ensure their quality.
Verification	Defined curation rules based on data relation, critical variables, and anonymisation standards.

©2023-2026 DIOPTRA

Completion Criteria	Criteria for successful completion include: • Type & allowed values for all variables have been defined. • Technical testing for data storage has been completed. • Approval has been provided by all clinical sites
Relevant Use Cases	DIOPTRA_UC1, DIOPTRA_UC2, DIOPTRA_UC3, DIOPTRA_UC4, DIOPTRA_UC5, DIOPTRA_UC8
Dependencies	DIOPTRA-CSI-FR-01

Table 37: Functional Requirement #02 for Data Curation & Storage System

Requirement ID	DIOPTRA-DCS-FR-02
Name	Gateway Endpoints
Description	Restful API Endpoints that are accessible from the clinical interface and the data pipelines. Connects front-end with back-end applications and transfers any data required.
Priority	High
Type	Functional
Rationale	The system should be remotely accessible from other components like the clinical interface and the mobile app.
Verification	Standardised format of request and responses and error handling.
Completion Criteria	Detailed documentation of the data transferred type and requirements.
Relevant Use Cases	DIOPTRA_UC1, DIOPTRA_UC2, DIOPTRA_UC3, DIOPTRA_UC4, DIOPTRA_UC5, DIOPTRA_UC8
Dependencies	DIOPTRA_UC1, DIOPTRA_UC2, DIOPTRA_UC3, DIOPTRA_UC4, DIOPTRA_UC5, DIOPTRA_UC8

Table 38: Functional Requirement #03 for Data Curation & Storage System

Requirement ID	DIOPTRA-DCS-FR-03
Name	Data Ingestion – Integration
Description	The system should be able to provide a way to store and ingest anonymised datasets from different clinical sites and fuse them enforcing a unified format.
Priority	High
Type	Functional

©2023-2026 DIOPTRA

Rationale	The system should be remotely accessible from other components like the clinical interface and the mobile app.
Verification	Standardised format of request and responses and error handling.
Completion Criteria	Detailed documentation of the data transferred type and requirements.
Relevant Use Cases	DIOPTRA_UC1, DIOPTRA_UC2, DIOPTRA_UC3, DIOPTRA_UC4, DIOPTRA_UC5, DIOPTRA_UC8
Dependencies	DIOPTRA-DCS-FR-02

Table 39: Functional Requirement #04 for Data Curation & Storage System

Requirement ID	DIOPTRA-DCS-FR-04
Name	Data Storage and Access to stored data
Description	The main repository containing all stored data
Priority	High
Type	Functional
Rationale	Data gathered from all clinical sites should be easily accessible when required for further analysis.
Verification	
Completion Criteria	Completion Criteria include:
Relevant Use Cases	DIOPTRA_UC1, DIOPTRA_UC2, DIOPTRA_UC3, DIOPTRA_UC4, DIOPTRA_UC5, DIOPTRA_UC8
Dependencies	N/A

Table 40: Functional Requirement #05 for Data Curation & Storage System

Requirement ID	DIOPTRA-DCS-FR-05
Name	Data querying filtering and aggregations support
Description	The functionality to retrieve data filtered by certain criteria. Facilitates both prospective and retrospective data.
Priority	High
Type	Functional
Rationale	N/A
Verification	N/A

©2023-2026 DIOPTRA

Completion Criteria	N/A
Relevant Use Cases	DIOPTRA_UC1, DIOPTRA_UC2, DIOPTRA_UC3, DIOPTRA_UC4, DIOPTRA_UC5
Dependencies	DIOPTRA-DCS-FR-02, DIOPTRA-DCS-FR-03, DIOPTRA-DCS-FR-04

Table 41: Functional Requirement #06 for Data Curation & Storage System

Requirement ID	DIOPTRA-DCS-FR-06
Name	Visualisation dashboards
Description	Diverse interactive data visualisations and customised intuitive dashboards for further analysis and support
Priority	Low
Type	Functional
Rationale	N/A
Verification	N/A
Completion Criteria	N/A
Relevant Use Cases	DIOPTRA_UC1, DIOPTRA_UC2, DIOPTRA_UC3, DIOPTRA_UC4, DIOPTRA_UC5
Dependencies	DIOPTRA-DCS-FR-03, DIOPTRA-DCS-FR-04, DIOPTRA-DCS-FR-05

4.2.3.2 Non-Functional Requirements

Table 42: Non-Functional Requirement #01 for Data Curation & Storage System

Requirement ID	DIOPTRA-DCS-NFR-01
Title	Data Quality and Consistency
Description	Ingested data should meet the projects quality standards.
Priority	High
Type	Non-functional
Rationale	The data stored will be used as input for further analysis and as input for training machine-learning models
Verification	Data curation methods that include: • Data Cataloguing • Data Validation Checks (with pre-defined rules on accepted values and ranges approved by the clinical partners)

©2023-2026 DIOPTRA

	• Data Transformation (On specific fields that will not impact the ml models)

Table 43: Non-Functional Requirement #02 for Data Curation & Storage System

Requirement ID	DIOPTRA-DCS-NFR-02
Title	Resilient Storage
Description	As the platform is deployed in a distributed way it should be designed to keep working even after some components fail.
Priority	High
Type	Non-functional
Rationale	
Verification	Performa resilient deployment with: • At least 3 eligible master nodes • At least 2 nodes for each role • At least each copy of each shard

Table 44: Non-Functional Requirement #03 for Data Curation & Storage System

Requirement ID	DIOPTRA-DCS-NFR-03
Title	Security
Description	Secure and limit the access to the DIOPTRA back-end.
Priority	High
Type	Non-functional
Rationale	Data Integrity is crucial across the Project
Verification	• SSL & TLS Security Protocols with trusted certificates • Firewalls protecting the VM's where the components are deployed. • Keycloak to be used for client's verification. • Penetration tests ensuring the quality of all the protection measures.

Table 45: Non-Functional Requirement #04 for Data Curation & Storage System

Requirement ID	DIOPTRA-DCS-NFR-04
Title	Scalability and Interoperability

©2023-2026 DIOPTRA

Description	The system should be scalable, able to handle data exchanges between different systems in real time
Priority	High
Type	Non-functional
Rationale	Large data volume can be anticipated, and daily usage of the system could further stress the system. To tackle this, the system should be easily scalable and handle high volume of traffic in real time
Verification	Analyse and compare the results within different size datasets

Table 46: Non-Functional Requirement #05 for Data Curation & Storage System

Requirement ID	DIOPTRA-DCS-NFR-05
Title	User Authentication & Authorisation
Description	Restrict access to patient's data depending on the user (clinicians)
Priority	High
Type	Non-functional
Rationale	Clinicians access to patients' data should be monitored and restricted to ensure data security.
Verification	SSO Service to be used (Keycloak) for User Authentication and Authorisation

4.2.4 Mobile App

The DIOPTRA Mobile App, **primarily catered to patients enrolled in the follow-up study** but **also accessible to the broader public**, will serve as a tool for gathering risk factor data and providing personalised lifestyle recommendations to users. In addition, the application will serve as a central hub where users can find curated and up-to-date material on CRC, promoting risk awareness and justifying needed behavioural changes. To encapsulate its multifaceted functionality, the mobile app is structured around **four primary features**:

1. **Behavioural Questionnaires**: The behavioural questionnaires are a core part of the DIOPTRA project. Participants of the prospective and follow-up studies will need to answer up to two questionnaires, depending on the recruitment phase: the *baseline behavioural questionnaire* (both studies) and the *follow-up questionnaire* (follow-up study). The baseline questionnaire will be presented to all participants in a paper format at the clinical site. Since the mobile application within the follow-up study will only engage healthy and non-AA population groups and the diagnosis will not be available at the time of the baseline questionnaire, the paper format will be used for all groups at this stage of the prospective study, while the app will be provided to the follow-up study recruits. It will serve as a foundational data point to produce each user's personalised risk assessment. While participants will not be answering this questionnaire via the mobile app, their responses will be directly and seamlessly retrieved via an API endpoint of the dashboard upon the app's first usage. In contrast, the general public will be able to access and complete this questionnaire through the mobile app instead. This inclusive approach will allow broader user engagement, although responses from the public will be utilised solely for the app's

©2023-2026 DIOPTRA

basic functions and will not be included in the project's primary dataset. The follow-up questionnaire will only be presented to the follow-up study participants (Table 1), to track changes and developments post the baseline assessment. This questionnaire will be available to be completed via the mobile app only for this specific set of users.

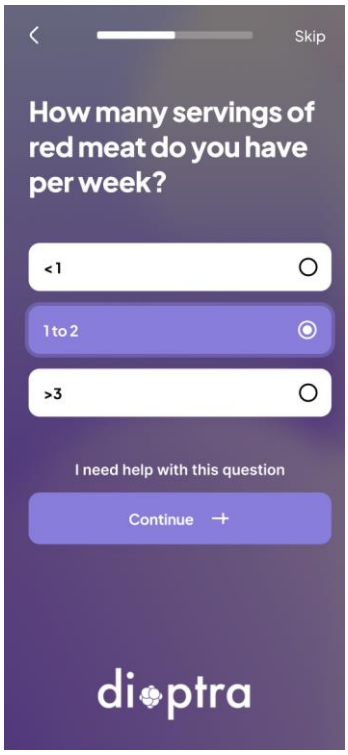

Figure 10: Mock-up showcasing a single question of the baseline questionnaire

2. **Risk Assessment and Personalised Recommendation Module**: The risk assessment and personalised recommendation module will calculate a personalised risk assessment based on the user's responses, and also identify areas that the user can improve in order to raise their score. The user will be able to explore the recommendations, provided by the Risk Assessment Module, and access additional informational content justifying these recommendations through the Health Literacy Module (Figure 11). By integrating the behavioural questionnaire within the mobile application, the RAM empowers participants to provide valuable insights about their lifestyle choices, including diet, alcohol consumption, smoking habits, physical activity, and more. The RAM analyses these responses and generate a personalised behavioural risk score, allowing for a deeper understanding of the specific areas that may require lifestyle modifications or medical attention. Based on these risk factors specific recommendations will be triggered and provided to the end users to promote a healthier lifestyle. These recommendations will be tailored to each individual's behavioural risk profile, taking into account the identified risk factors. By leveraging personalised insights, the system can suggest practical and achievable steps for end-users to make informed decisions and positive changes in their daily routines. The recommendations will also prioritise the educational media contents (in form of video, infographic, text, etc.) based on the areas of improvement of the user, which will be calculated by the risk assessment model.

Funded by the European Union

©2023-2026 DIOPTRA

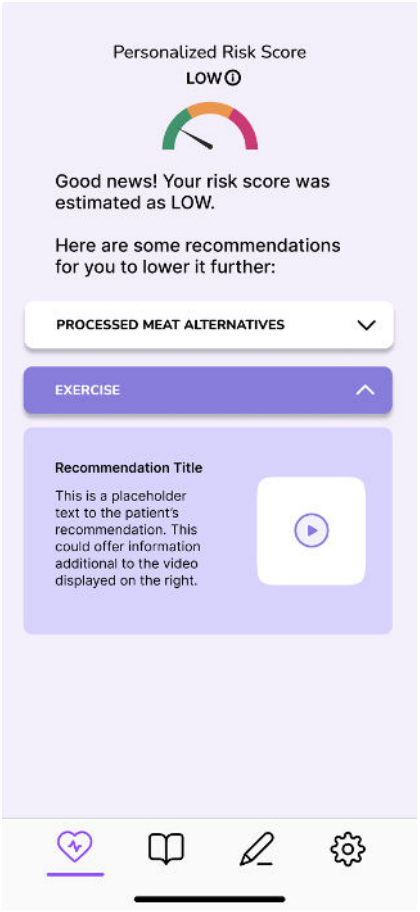

Figure 11: Example of the risk assessment results and the personalised recommendations

3. **Health Literacy Module**: The HLM within the DIOPTRA mobile app will be an educational tool aimed at enhancing public understanding of CRC and related health behaviours. It will feature a rich repository of content, including articles, infographic, and videos, curated by clinicians to provide accurate, up-to-date information about CRC such as risk factors, colonoscopy guidelines and procedures explained in a user-friendly manner (Figure 12).

©2023-2026 DIOPTRA

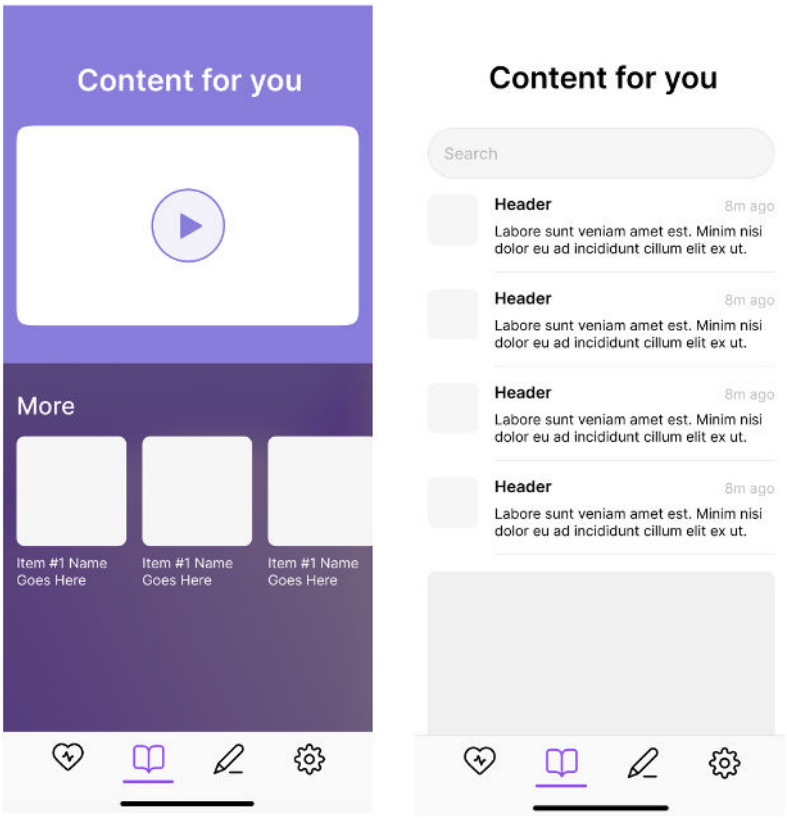

Figure 12: Example of the Health Literacy Module screen, showcasing both text and video content

4. **Diary**: The diary feature is a user-centric functionality designed to encourage and track individual lifestyle behaviours as well as mild bowel symptoms (such as chronic constipation, chronic diarrhoea, etc.) and identify any changes or recent trends in those topics. To promote user engagement and ensure that it is used by the user, it will be presented in the form of periodic set of questions that the user can answer before continuing to use the app, track behavioural improvement (such as physical activity and nutrition) and monitor mild bowel symptoms (Figure 13).

©2023-2026 DIOPTRA

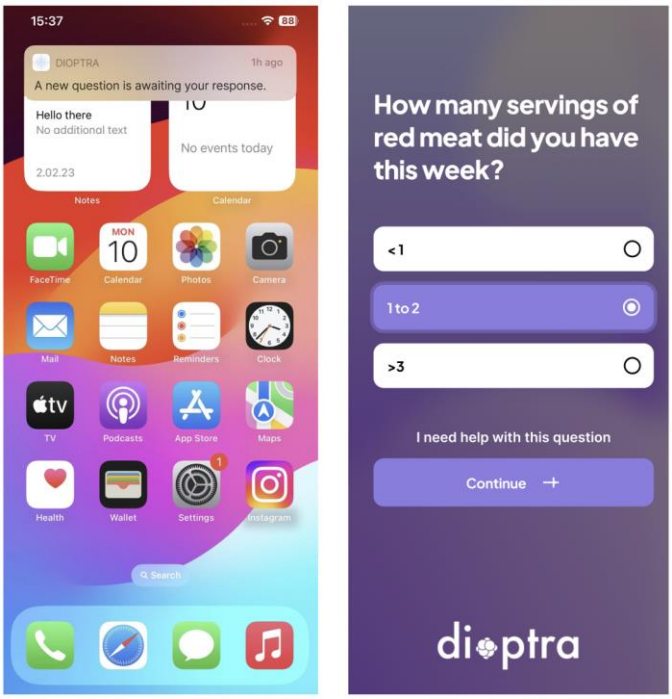

Figure 13: Mock-up screens showcasing the diary functionality. Notifications will inform the user of new questions waiting for them before continuing to use the app

4.2.4.1 Functional Requirements

Table 47: Functional Requirement #01 for Mobile App

Requirement ID	DIOPTRA-MA-FR-01
Name	User login using unique identifier
Description	Users should be prompted to enter a unique project-specific identifier, if they have received one, otherwise they can proceed as users not enrolled in the project.
Priority	High
Type	Functional
Rationale	Users enrolled in the DIOPTRA project need to be able to retrieve their past answers that have already been entered in the dashboard. In addition, these users need to be differentiated from general public users in order to collect and process their questionnaire responses.
Verification	Testing login functionality with various identifiers.
Completion Criteria	Successful login and data retrieval for different user types.
Relevant Use Cases	DIOPTRA_UC5, DIOPTRA_UC8
Dependencies	N/A

©2023-2026 DIOPTRA

Table 48: Functional Requirement #02 for Mobile App

Requirement ID	DIOPTRA-MA-FR-02
Name	Behavioural Questionnaires
Description	Present baseline and follow-up questionnaires to users, if required. Enable users to submit responses within the app. Store and manage questionnaire responses.
Priority	High
Type	Functional
Rationale	Essential for collecting user data to drive personalised recommendations and risk assessment.
Verification	Testing questionnaire presentation, response submission, and data storage.
Completion Criteria	Accurate display and storage of questionnaire responses.
Relevant Use Cases	DIOPTRA_UC5, DIOPTRA_UC8
Dependencies	N/A

Table 49: Functional Requirement #03 for Mobile App

Requirement ID	DIOPTRA-MA-FR-03
Name	Risk Assessment and Personalisation
Description	Send questionnaire responses to the Risk Assessment Module backend, retrieve the response, and present it to the user. Additionally present personalised lifestyle recommendations based on the risk assessment.
Priority	High
Type	Functional
Rationale	Core functionality for providing personalised health recommendations.
Verification	Testing integration with the Risk Assessment Module and accuracy of recommendations.
Completion Criteria	Accurate and timely presentation of risk assessments and recommendations.
Relevant Use Cases	DIOPTRA_UC5, DIOPTRA_UC8
Dependencies	DIOPTRA-RAM-FR-03

Funded by the European Union

©2023-2026 DIOPTRA

Table 50: Functional Requirement #04 for Mobile App

Requirement ID	DIOPTRA-MA-FR-04
Name	Health Literacy Module
Description	Provide educational content on CRC and related topics. Update and manage content regularly.
Priority	High
Type	Functional
Rationale	To educate users about CRC and related health topics.
Verification	Regular updates and content management system checks.
Completion Criteria	Updated, accurate, and relevant health content.
Relevant Use Cases	DIOPTRA_UC5, DIOPTRA_UC8
Dependencies	N/A

Table 51: Functional Requirement #05 for Mobile App

Requirement ID	DIOPTRA-MA-FR-05
Name	Diary
Description	Occasionally require the user to answer follow-up questions on topics they have already answered. Update the risk assessment score and personalised recommendations accordingly.
Priority	High
Type	Functional
Rationale	To track user health behaviour and adapt recommendations accordingly.
Verification	N/A
Completion Criteria	Accurate tracking and updating of user responses and risk assessments.
Relevant Use Cases	DIOPTRA_UC5, DIOPTRA_UC8
Dependencies	DIOPTRA-MA-FR-04

Table 52: Functional Requirement #06 for Mobile App

Requirement ID	DIOPTRA-MA-FR-06
Name	Data Integration and Syncing

©2023-2026 DIOPTRA

Description	Synchronise user data across the mobile app and the dashboard. Integrate with external APIs for data retrieval and sharing.
Priority	High
Type	Functional
Rationale	Ensures seamless data flow between the app and dashboard.
Verification	Testing data synchronisation and API integrations.
Completion Criteria	Reliable and accurate data syncing.
Relevant Use Cases	DIOPTRA_UC5, DIOPTRA_UC6
Dependencies	DIOPTRA-DCS-FR-04

Table 53: Functional Requirement #07 for Mobile App

Requirement ID	DIOPTRA-MA-FR-07
Name	Notification System
Description	Send notifications to users regarding questionnaire reminders and health tips. Allow users to customise their notification preferences.
Priority	High
Type	Functional
Rationale	To keep users engaged.
Verification	Testing notification delivery and customisation settings.
Completion Criteria	Timely and relevant notifications; customisable user settings.
Relevant Use Cases	DIOPTRA_UC5, DIOPTRA_UC8

4.2.4.2 Non-Functional Requirements

Table 54: Non-Functional Requirement #01 for Mobile App

Requirement ID	DIOPTRA-MA-NFR-01
Title	User Experience and Usability
Description	The application will feature an intuitive and user-friendly interface design, with easy navigation and clear prompts.
Priority	High
Type	Non-functional

©2023-2026 DIOPTRA

Rationale	An intuitive interface is critical to ensure user engagement and effective use of the app.
Verification	User testing and feedback surveys to assess ease of use and navigation.

Table 55: Non-Functional Requirement #02 for Mobile App

Requirement ID	DIOPTRA-MA-NFR-02
Title	Cross-platform availability
Description	The application will be available on both Android and iOS
Priority	High
Type	Non-functional
Rationale	Availability on both Android and iOS expands user reach and accessibility.
Verification	Testing the app on various devices and operating systems for compatibility.

Table 56: Non-Functional Requirement #03 for Mobile App

Requirement ID	DIOPTRA-MA-NFR-03
Title	Accessibility
Description	Design catering to users with various abilities and disabilities. Compatibility with OS-level accessibility features.
Priority	High
Type	Non-functional
Rationale	Ensuring the app is accessible to all users, regardless of their abilities, is crucial for inclusivity.
Verification	Accessibility testing, including compliance with standards like WCAG.

Table 57: Non-Functional Requirement #04 for Mobile App

Requirement ID	DIOPTRA-MA-NFR-04
Title	Performance
Description	Fast response times and minimal loading delays
Priority	Medium
Type	Non-functional
Rationale	Good performance ensures user satisfaction and app reliability.
Verification	Performance testing under different load conditions.

©2023-2026 DIOPTRA

Table 58: Non-Functional Requirement #05 for Mobile App

Requirement ID	DIOPTRA-MA-NFR-05
Title	Privacy and Personal Data Protection
Description	Compliance with the applicable EU data protection law, i.e. General Data Protection Regulation and e-Privacy Directive, , ensuring data protection by design and by default, transparency and, ultimately, users' control over their personal information.
Priority	High
Type	Non-functional
Rationale	Compliance of DIOPTRA partners with the applicable legal requirements is mandatory.
Verification	Performance of a Data Protection Impact Assessment (DPIA) periodically by the DIOPTRA partner responsible for processing of personal data by the Mobile App.

Table 59: Non-Functional Requirement #06 for Mobile App

Requirement ID	DIOPTRA-MA-NFR-06
Title	Internationalisation and Localisation
Description	Support for multiple languages, including English, Greek, French, Spanish, Bulgarian, Danish.
Priority	High
Type	Non-functional
Rationale	Localisation increases the app's usability and relevance across the different pilots.
Verification	Language and region-specific user testing.

Table 60: Non-Functional Requirement #07 for Mobile App

Requirement ID	DIOPTRA-MA-NFR-07
Title	User Feedback
Description	Mechanisms for user feedback collection and analysis.
Priority	Medium
Type	Non-functional
Rationale	Continuous improvement and user satisfaction are driven by user feedback.
Verification	Implementation of a feedback mechanism and its periodic analysis.

Funded by the European Union

©2023-2026 DIOPTRA

4.2.5 Risk Assessment Module

The risk assessment module developed for early screening of Colorectal Cancer represents a vital component of a comprehensive healthcare solution aimed at improving the overall health and well-being of individuals. By seamlessly integrating retrospective and prospective data this innovative module empowers participants to provide valuable insights about their lifestyle choices, including diet, alcohol consumption, smoking habits, physical activity, and more. The dataset will include demographic, family history, medical, behavioural, and clinical information derived from the Electronic Health Records of the clinical sites. The module utilises AI-based algorithms (based on the outcomes of modelling process) to analyse these data and generate a personalised risk score. This score not only quantifies an individual's potential risk of developing Colorectal Cancer but also ranks the contributing risk factors, allowing for a deeper understanding of the specific areas that may require lifestyle modifications or medical attention.

The risk assessment module, developed for the purpose of early screening of Colorectal Cancer, is a vital component of complete healthcare solutions. Its primary goal is to improve the overall health and well-being of individuals. By integrating retrospective and prospective data, this module enables users to gain significant knowledge regarding their lifestyle preferences. These factors include dietary patterns, alcohol consumption, smoking proclivities, levels of physical activity, and numerous other relevant aspects. At the core of this novel framework lies the integration of a heterogeneous dataset consisting of demographic indicators, familial medical history, extensive medical records, behavioural patterns, and complex clinical observations extracted from the Electronic Health Records (EHRs) located in the clinical sites. By applying AI-based algorithms, this module exploits the potential of these vast datasets to thoroughly examine and extract complex patterns, ultimately producing a personalised and detailed risk score for every individual.

The resulting risk score effectively rates the contributing risk factors, facilitating a deep understanding of the particular areas that may require adjustments to one's lifestyle. This all-encompassing assessment equips participants with practical knowledge, allowing them to formulate well-informed choices concerning their well-being and facilitating healthcare professionals in the provision of customised guidance and assistance. Furthermore, the incorporation of cutting-edge technologies in this module promotes a proactive approach to healthcare, which entails a shift in focus from reactive treatment to proactive prevention. Through the early identification of potential risk factors, this approach enables the implementation of focused preventive measures, which may consequently reduce the occurrence and intensity of colorectal cancer. Implementing this strategic approach not only improves the health outcomes of individuals but also offers a significant opportunity to decrease healthcare costs and burden by limiting the prevalence of this condition at a larger level in society.

Funded by the European Union

©2023-2026 DIOPTRA

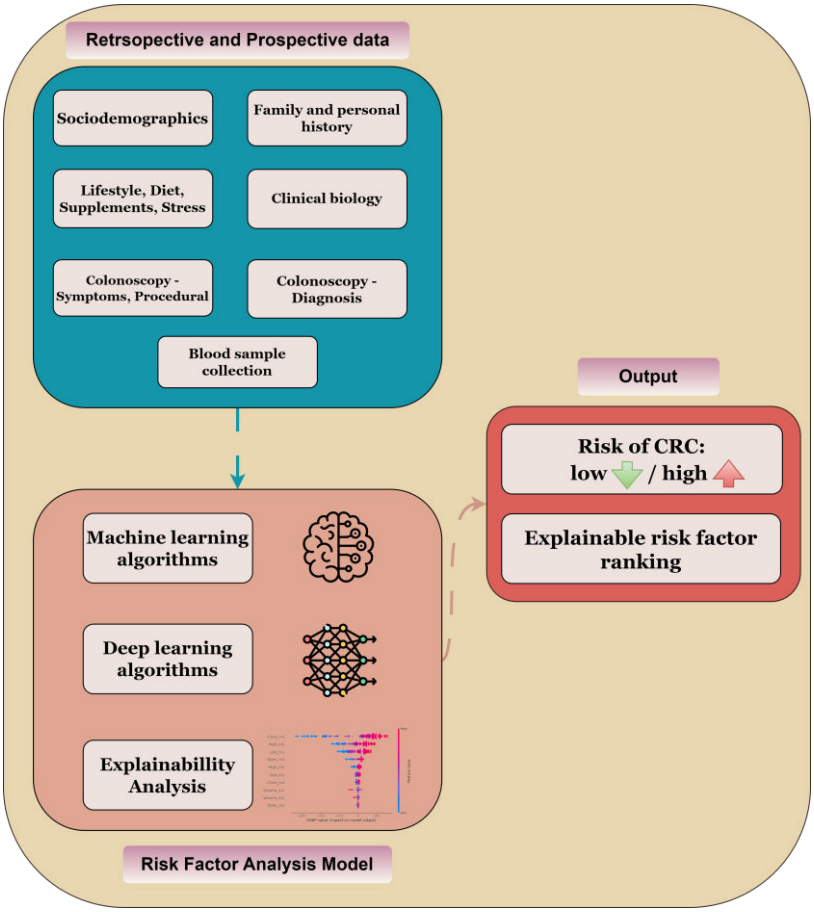

Figure 14: Risk Assessment Module Architecture

The implementation of this advanced risk assessment module represents a critical milestone in the development of healthcare solutions, clarifying a trajectory towards individualised and preventative healthcare. The organisation's diverse strategy, supported by state-of-the-art technologies and data-driven insights, serves as evidence of the profound change that can result from applying comprehensive data analytics to the healthcare sector.

4.2.5.1 Functional Requirements

Table 61: Functional Requirement #01 for Risk Assessment Module

Requirement ID	DIOPTRA-RAM-FR-01
Name	Data Integration and Collection
Description	The system must integrate retrospective and prospective data from various cohorts, including lifestyle choices like diet, alcohol consumption, smoking habits, physical activity, medical history etc.
Priority	High
Type	Functional
Rationale	Holistic integration of diverse data sources is critical to formulating comprehensive participant profiles, facilitating accurate risk assessment,

©2023-2026 DIOPTRA

	and enabling personalised recommendations for Colorectal Cancer screening.
Verification	Successful integration and storage of diverse data types (clinical, lifestyle) from multiple sources within the module's database. Verification of data consistency and completeness.
Completion Criteria	• Successful data gathering and storage of diverse data types. • Data consistency and completeness verification after the curation techniques. • Compatibility with standard data formats for seamless integration.
Relevant Use Cases	DIOPTRA_UC1, DIOPTRA_UC5
Dependencies	DIOPTRA-DCS-FR-03, DIOPTRA-DCS-FR-04, DIOPTRA-DCS-FR-05, DIOPTRA-MA-FR-02

Table 62: Functional Requirement #02 for Risk Assessment Module

Requirement ID	DIOPTRA-RAM-FR-02
Name	Advanced Analysis Techniques
Description	Utilise AI-based algorithms from scratch for robust analysis of integrated data, facilitating the generation of personalised risk scores and factors ranking.
Priority	High
Type	Functional
Rationale	Leveraging advanced analytical methods enables precise analysis of integrated data, empowering accurate risk score generation crucial for personalised recommendations in Colorectal Cancer screening.
Verification	Successful implementation of AI-based algorithms for analysis. Verification of the accuracy and reliability of generated risk scores based on test and validation sets.
Completion Criteria	• Data consistency and completeness verification after the curation techniques. • Successful implementation and validation of AI-based algorithms. • Validation of the accuracy and other related performance metrics and reliability of generated risk scores against known datasets or benchmarks.
Relevant Use Cases	DIOPTRA_UC5, DIOPTRA_UC7, DIOPTRA_UC8
Dependencies	External

©2023-2026 DIOPTRA

Table 63: Functional Requirement #03 for Risk Assessment Module

Requirement ID	DIOPTRA-RAM-FR-03
Name	Personalised Risk Score
Description	Generate an accurate risk score indicating an individual's potential risk of developing Colorectal Cancer based on the provided data.
Priority	High
Type	Functional
Rationale	The generation of a personalised risk score assists in quantifying behavioural profile with regard to Colorectal Cancer, aiding in personalised health management.
Verification	Successful computation and generation of risk scores for various lifestyle and clinical data inputs. Verification of consistency and accuracy of the generated risk scores.
Completion Criteria	• Successful computation and generation of risk scores. • Consistency and accuracy evaluation of the generated risk scores against validated benchmarks.
Relevant Use Cases	DIOPTRA_UC5, DIOPTRA_UC8
Dependencies	DIOPTRA-DCS-FR-03, DIOPTRA-DCS-FR-04, DIOPTRA-DCS-FR-05, DIOPTRA-MA-FR-02

Table 64: Functional Requirement #04 for Risk Assessment Module

Requirement ID	DIOPTRA-RAM-FR-04
Name	User Interface and Interaction
Description	Develop a user-friendly mobile application allowing participants for presenting the resulted risk scores in an understandable format.
Priority	High
Type	Functional
Rationale	A user-friendly interface enhances participant engagement, ensuring ease of data input and comprehension of risk scores, leading to better adherence and participation in Colorectal Cancer screening.
Verification	Successful implementation of an interface allowing easy data input and presentation of risk scores. User feedback and usability testing confirming satisfactory interaction.
Completion Criteria	• Implementation of an interface for data input and risk score presentation.

Funded by the European Union

©2023-2026 DIOPTRA

	• Positive user feedback and successful usability testing.
Relevant Use Cases	DIOPTRA_UC5, DIOPTRA_UC8
Dependencies	DIOPTRA-MA-FR-01, DIOPTRA-MA-FR-02, DIOPTRA-MA-FR-03, DIOPTRA-MA-FR-04, DIOPTRA-MA-FR-07, DIOPTRA-MA-NFR-01

Table 65: Functional Requirement #05 for Risk Assessment Module

Requirement ID	DIOPTRA-RAM-FR-05
Name	Scalability and Performance
Description	Ensure the system's scalability to handle increasing participant data without compromising performance, providing timely risk assessments.
Priority	High
Type	Functional
Rationale	Scalability and optimal performance are essential to accommodate a growing number of participants and provide timely risk assessments, enhancing the efficiency of Colorectal Cancer screening.
Verification	Successful system testing under varying data loads to ensure performance efficiency and scalability. Monitoring system performance metrics for responsiveness.
Completion Criteria	• System testing under varying data loads confirming performance efficiency. • Monitoring of system performance metrics to ensure responsiveness.
Relevant Use Cases	DIOPTRA_UC1, DIOPTRA_UC5, DIOPTRA_UC8
Dependencies	DIOPTRA-MA-FR-06, DIOPTRA-MA-NFR-01, DIOPTRA-MA-NFR-03, DIOPTRA-MA-NFR-04

4.2.5.2 Non-Functional Requirements

Table 66: Non-Functional Requirement #01 for Risk Assessment Module

Requirement ID	DIOPTRA-RAM-NFR-01
Title	Accuracy and Reliability
Description	Ensure high accuracy in risk score calculations and identification of contributing risk factors through validated machine learning models.
Priority	High

©2023-2026 DIOPTRA

Type	Non-functional
Rationale	High accuracy in risk assessment is crucial for providing reliable results to participants.
Verification	Cross-validation of machine learning models with known datasets, regular model performance metrics, and accuracy checks against validated benchmarks.

Table 67: Non-Functional Requirement #02 for Risk Assessment Module

Requirement ID	DIOPTRA-RAM-NFR-02
Title	Performance Efficiency
Description	Ensure optimal performance of the risk assessment module, with minimal response times during data analysis and score generation.
Priority	High
Type	Non-functional
Rationale	Efficient performance is critical for providing timely risk assessment feedback to participants, allowing for lifestyle modifications.
Verification	Performance testing under varying data loads, stress testing, unit testing and benchmarking against predefined response time thresholds. Regular monitoring of system performance metrics.

Table 68: Non-Functional Requirement #02 for Risk Assessment Module

Requirement ID	DIOPTRA-RAM-NFR-03
Title	Interoperability and Integration
Description	Ensure interoperability with the centralised platform, the mobile application, and databases for smooth integration of data.
Priority	High
Type	Non-functional
Rationale	Interoperability is important for exchanging data between systems, enabling a comprehensive view of participant health profiles. Seamless integration enhances the module's usability and effectiveness.
Verification	Successful data exchange tests with the DIOPTRA components

Table 69: Non-Functional Requirement #04 for Risk Assessment Module

Requirement ID	DIOPTRA-RAM-NFR-04
Title	User Experience

Funded by the European Union

©2023-2026 DIOPTRA

Description	Ensure a positive user experience during the engagement with the mobile application for inputting lifestyle data and presenting risk scores.
Priority	High
Type	Non-functional
Rationale	A positive user experience encourages active participation and engagement from participants, leading to accurate data input and better adherence to the screening module.
Verification	User testing, feedback collection, and usability evaluations to confirm ease of data input, comprehension of risk scores, and overall satisfaction with the module's interfaces.

4.2.6 Screening AI Module

This module intends a) to select an appropriate set of protein features from the output of the biomarker discovery pipeline (D4.2 & D4.6) that can be used as input for an AI model designed to predict the presence of non-AA, AA and CRC, b) to combine the protein features with clinical features from the analysis of T4.1 and c) to generate the aforementioned model(s). The challenge associated with this component and the mechanism-based biomarker discovery pipeline in general, lies within the fact that without having a domain-specific understanding of the biological characteristics of colorectal cancer, it is quite possible that by merely selecting the set of proteins that optimally differentiate between the health and sick subjects participating in this study, we may end up generating a model that manages to accurately detect a proxy condition or comorbidity, that only works for this particular study group but ends up generating a large number of false positives when applied to the population at large. For example, individuals with a particular disease also exhibit another condition distinguishing them from healthy subjects. However, this specific condition may not be exclusive to the disease in question. For this reason, the protein candidates will be chosen based on two characteristics: a) The construction and utilisation of the knowledge graph based on a set of approx. 30 proteins biomarkers connected to the disease mechanism (D4.2 & D4.6) to closely examine their association to CRC (this selection process will be facilitated through the proper construction and utilisation of a knowledge graph) and b) strong discriminatory power in effectively distinguishing between individuals with the disease and those who are healthy.

Knowledge graph-based (KG) AI modelling will enable the semantic creation of DIOPTRA knowledge graph of the ontologically described components of the architecture that will allow to (i) relate (including semantic characteristics) different resources and descriptions, (ii) investigate map and interlink biomarkers, proteins and extracted features, (iii) merge descriptions/ontologies by expressing semantic information, (iv) conduct queries to discover resources that match specific requirements, and (v) express user constraints to additional relationships created by the platform. The knowledge graph will be generated by integrating various publicly available PPI (Protein-Protein Interaction) databases, but depending on the availability of data, we may also incorporate other biological networks.

Table 70: Main Characteristics of Screening AI Module (to be revised within WP5)

Development Environment	Visual studio, .NET, C#, Python, Cypher, Neo4J, Docker, FastAPI
Software requirements	NET Framework, Python, NEO4J, SpaCy, Tensorflow / Scikit-learn

©2023-2026 DIOPTRA

Hardware requirements	>32 Gb RAM
Execution Time	• To build a custom graph representation in memory, a couple of minutes will be needed. • To export the graph in Neo4J or any other graph DB, a couple of minutes will be needed. (depending on the available infrastructure)
Main inputs	tabular and textual datasets, protein descriptions, biomarker descriptions, user subjects, images, texts, etc.
Nature of Expected Input	Input will be of text format. The advantage of the KG lies in the fact that the actual datasets are not stored in it. Indexes/ references to the data are being represented as nodes linked with a particular relation. The actual data can be stored elsewhere. This might also work with the hashed content developed within D4.1 and D4.2.
Main Outputs	Based on the query of a user the KG will output triplets corresponding to the query of the user. Mainly, a set of nodes and a link that describes the relationship of the nodes. The output would integrate the triplets and would be transformed into human readable format.
Nature of Expected Output	KG triplets & explanation in text

4.2.6.1 Functional Requirements

Table 71: Functional Requirement #01 for Screening AI Module

Requirement ID	**DIOPTRA-SAI-FR-01**
Title	Creation of Knowledge graph
Description	Generation of Knowledge graph
Priority	High (it will be implemented)
Type	Functional
Rationale	The use of a KG will be instrumental for the selection of the most relevant proteins to colorectal cancer
Verification	Creation of KG
Completion Criteria	• A KG has been created either in-memory or in a database • The KG can be queried

©2023-2026 DIOPTRA

Relevant Use Cases	DIOPTRA_UC71, DIOPTRA_UC7
Dependencies	DIOPTRA-DCS-FR-02, DIOPTRA-DCS-FR-04

Table 72: Functional Requirement #02 for Screening AI Module

Requirement ID	DIOPTRA-SAI-FR-02
Title	Generation of protein candidates based on the already mechanism-based biomarkers provided in WP4
Description	Selection of protein set candidate sets via the KG that are closely associated with colorectal cancer
Priority	High (it will be implemented)
Type	Functional
Rationale	The KG will dictate which proteins are mostly relevant to colorectal cancer. Since this process can be arbitrary, a number of approaches can be used, e.g. purely algorithmic through the use of network analysis, manually curated or a hybrid approach
Verification	Selection of subset of proteins
Completion Criteria	• A subset of proteins, selected in accordance to a selection criterion, that are the most relevant in terms of colorectal cancer OR • A rank ordering of the existing proteins based on a heuristic evaluation function that represents their importance to colorectal cancer
Relevant Use Cases	DIOPTRA_UC1, DIOPTRA_UC7
Dependencies	DIOPTRA-SAI-FUNC-01, DIOPTRA-DCS-FR-02, DIOPTRA-DCS-FR-04, DIOPTRA-DCS-FR-05

Table 73: Functional Requirement #03 for Screening AI Module

Requirement ID	DIOPTRA-SAI-FR-03
Title	Protein feature selection
Description	Select which of the candidate proteins will be used to generate the ML model
Priority	High (it will be implemented)
Type	Functional

Funded by the European Union

©2023-2026 DIOPTRA

Rationale	Not all of the selected proteins will be suitable candidates for the ML model. An appropriate process will have to be applied to make this distinction
Verification	Further refinement of subset of proteins
Completion Criteria	• A subset of proteins, taken out of the candidate set, that exhibit strong discriminatory power in effectively distinguishing between individuals with the disease and those who are healthy
Relevant Use Cases	DIOPTRA_UC1, DIOPTRA_UC7
Dependencies	DIOPTRA-SAI-FUNC-01, DIOPTRA-DCS-FR-02, DIOPTRA-DCS-FR-04, DIOPTRA-DCS-FR-05

Table 74: Functional Requirement #04 for Screening AI Module

Requirement ID	DIOPTRA-SAI-FR-04
Title	Protein-based ML model
Description	Generate an ML model that differentiates between individuals with the disease and those who are healthy, based on protein measurements
Priority	High (it will be implemented)
Type	Functional
Rationale	A cost-effective ML model is required that distinguishes between individuals with the disease and those who are healthy
Verification	ML model evaluation based on performance metrics on available data
Completion Criteria	• A working ML model has been established
Relevant Use Cases	DIOPTRA_UC1, DIOPTRA_UC7
Dependencies	DIOPTRA-SAI-FUNC-01, DIOPTRA-DCS-FR-02, DIOPTRA-DCS-FR-04, DIOPTRA-DCS-FR-05

Table 75: Functional Requirement #05 for Screening AI Module

Requirement ID	DIOPTRA-SAI-FR-05
Title	Medical history-based ML model
Description	Generate an ML model that differentiates between individuals with the disease and those who are healthy, based on medical data
Priority	High (it will be implemented)
Type	Functional

©2023-2026 DIOPTRA

Rationale	A cost-effective ML model is required that distinguishes between individuals with the disease and those who are healthy
Verification	ML model evaluation based on performance metrics on available data
Completion Criteria	• A working ML model has been established
Relevant Use Cases	DIOPTRA_UC1, DIOPTRA_UC7
Dependencies	DIOPTRA-SAI-FUNC-01, DIOPTRA-DCS-FR-02, DIOPTRA-DCS-FR-04, DIOPTRA-DCS-FR-05

Table 76: Functional Requirement #06 for Screening AI Module

Requirement ID	DIOPTRA-SAI-FR-06
Title	Storage endpoints
Description	APIs as a process to get the results from the ML-based models and Databases in a structured approach
Priority	High (it will be implemented)
Type	Functional
Rationale	Through the definition of clear interfaces for interaction between ML-based models and DBs, implementation of wrapper functions or classes to encapsulate native API calls, ensuring proper error handling and data validation in the bridging layer, testing functionalities to ensure compatibility and stability
Verification	Storage of results
Completion Criteria	• Deployed APIs
Relevant Use Cases	DIOPTRA_UC1
Dependencies	N/A

Table 77: Functional Requirement #07 for Screening AI Module

Requirement ID	DIOPTRA-SAI-FR-07
Title	Results explanation function
Description	An interpretation function that describes the values derived from the triplets based on the relationship between/among the nodes
Priority	Medium
Type	Functional

©2023-2026 DIOPTRA

Rationale	Though the utilisation of a LLM the results can be read from the knowledge graph (graph db) in textual description based on the query of the user
Verification	API to communicate with the module
Completion Criteria	• Deployed endpoint to query the model
Relevant Use Cases	N/A
Dependencies	DIOPTRA-SAI-FUNC-01, DIOPTRA-DCS-FR-02, DIOPTRA-DCS-FR-04, DIOPTRA-DCS-FR-05, DIOPTRA-DCS-FR-06

4.2.6.2 Non-Functional Requirements

Table 78: Non-Functional Requirement #01 for Screening AI Module

Requirement ID	DIOPTRA-SAI-NFR-01
Name	Efficiency of the module
Description	Throughput, response time, transit delay, latency
Priority	Medium
Type	Non - Functional
Rationale	Performance and efficiency in terms of latency is of significance important for the operators of the module
Verification	Evaluation of performance metrics for the models & techniques applied

Table 79: Non-Functional Requirement #02 for Screening AI Module

Requirement ID	DIOPTRA-SAI-NFR-02
Name	Scalability & flexibility
Description	The module will be able to handle an increasing number of users and volume of received data
Priority	Medium
Type	Non - Functional
Rationale	Increased maintainability
Verification	Through efficient and effective management of available resources.

©2023-2026 DIOPTRA

Table 80: Non-Functional Requirement #03 for Screening AI Module

Requirement ID	DIOPTRA-SAI-NFR-03
Name	Maintainability
Description	The code must be well documented and formatted using specific formatting rules that makes it easier to address, and the libraries must be up to date.
Priority	Medium
Type	Non - Functional
Rationale	Increased maintainability
Verification	Aligned git repo, code maintenance policy

Table 81: Non-Functional Requirement #04 for Screening AI Module

Requirement ID	DIOPTRA-SAI-NFR-04
Name	User Acceptance
Description	Enable system acceptance through fully visible adequately justified explanations. Implementation of a functionality for providing the user recommendations along with explanations
Priority	Medium
Type	Non - Functional
Rationale	High potential for adoption of the system
Verification	System usability questionnaire

Table 82: Non-Functional Requirement #05 for Screening AI Module

Requirement ID	DIOPTRA-SAI-NFR-01
Name	Reasoning capacity
Description	Expose the reasoning behind a decision that led to a result based on the knowledge graph. Presenting the path from the query to the answer
Priority	Medium
Type	Non - Functional
Rationale	Increased adoption of the system
Verification	Through the communication with the respective module / Adoption rate

Funded by the European Union

©2023-2026 DIOPTRA

4.3 COMPLETE SYSTEM ARCHITECTURE

The full system comprises the combination of all components described in Section 4.1 towards development of the project outcomes from a technical perspective, with a high-level diagram of the **overall technical architecture** being shown below:

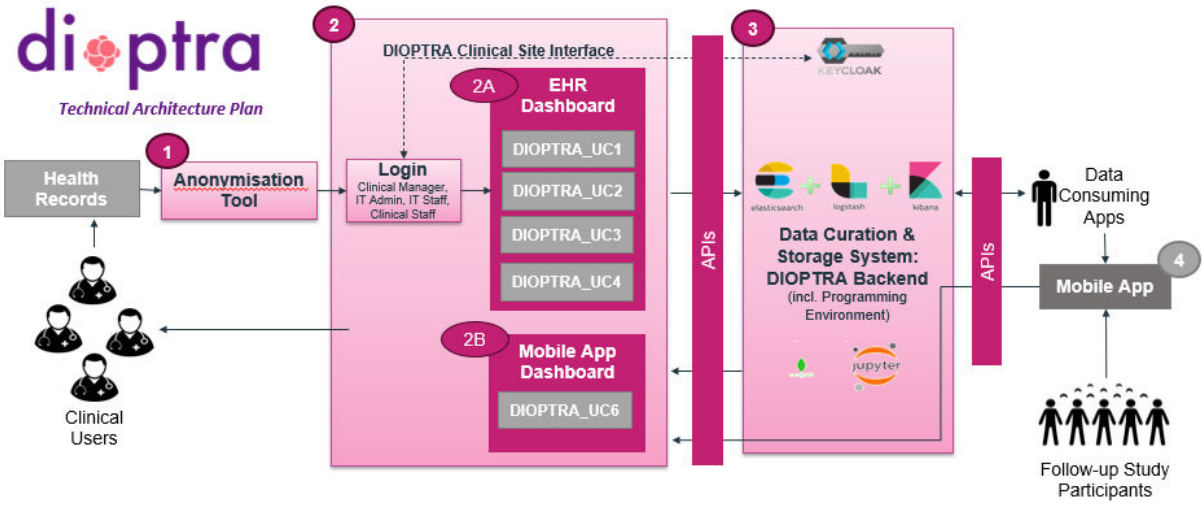

Figure 15: DIOPTRA Complete System Architecture

With regard to **user access**, technical partners will have direct access to the storage back-end for data retrieval in order to develop the Risk Assessment and Screening AI modules. **Access for users within the clinical sites** will be managed via the Clinical Site Interface, with the availability of the following actions (corresponding to DIOPTRA_UC1, DIOPTRA_UC2, DIOPTRA_UC3, DIOPTRA_UC4, and DIOPTRA_UC6):

1. Data Upload
2. Data Overview
3. Single Participant Data Review
4. Delete Participant Data

Accordingly, the following user roles have been defined:

Table 83: User Roles for Clinical Site Interface

	Clinical Manager	Clinical Staff	IT Administration Staff	IT staff
Data Upload	✓		✓	✓
Data Overview	✓	✓	✓	✓
Single Participant Review	✓	✓		
Delete	✓		✓	

©2023-2026 DIOPTRA

In addition, **an extra "Admin" role** will be provided to technical partners, granting full access for providing support & troubleshooting to clinical sites whenever needed.

4.4 GUIDELINES FOR COMPLIANCE WITH SECURITY STANDARDS

Health data holds a distinct status as a special category of personal data under the GDPR, necessitating the implementation of safeguards for their effective protection and this involves more than just applying specific techniques like pseudonymisation and encryption. The entire functionality, encompassing both data integrity and the overall solution integrity, requires protection through monitoring using state-of-the-art security and privacy assessment mechanisms.

4.4.1 Security Protocols and standards relevant to the DIOPTRA software architecture

4.4.1.1 ISO/IEC 27000 Standard Compliance

Achieving compliance with the ISO/IEC 27000 series of standards, which focuses on information security management systems (ISMS), involves a systematic approach to safeguarding sensitive information within an organisation. The general guidelines to comply with the ISO/IEC 27000 standard include:

- **Conduct a Risk Assessment**: Perform a comprehensive risk assessment to identify and assess the risks that DIOPTRA platform faces regarding information security. This forms the foundation for developing controls and implementing security measures.

- **Establish an Information Security Policy**: Develop and implement a clear and comprehensive information security policy that aligns with the organisation's objectives and the requirements of ISO/IEC 27001. The policy should define roles, responsibilities, and the overall approach to information security.

- **Define the Scope of the ISMS**: Clearly define the scope of DIOPTRA ISMS, including the boundaries, interfaces, and applicability of the system within the organisation. This ensures a focused and effective implementation of security controls.

- **Implement Security Controls**: Identify and implement the necessary security controls based on the risk assessment and the requirements outlined in ISO/IEC 27001. This may include access controls, encryption, incident response procedures, and other measures to protect information assets.

- **Train and Raise Awareness**: Provide training and awareness programs to employees at all levels. Ensuring that everyone understands the importance of information security and their role in maintaining it is crucial for the success of the ISMS.

- **Monitor and Measure Performance**: Establish monitoring and measurement processes to regularly assess the performance of the ISMS. This includes conducting internal audits and management reviews to identify areas for improvement and ensure continual compliance.

- **Document and Maintain Records**: Maintain accurate and up-to-date documentation of the ISMS, including policies, procedures, and records. This documentation provides evidence of compliance and facilitates audits.

Funded by the European Union

©2023-2026 DIOPTRA

- **Continual Improvement**: Embrace a culture of continual improvement. Regularly review and update the ISMS to address changes in the organisation, technology, and the threat landscape.

Achieving and maintaining ISO/IEC 27001 compliance is an ongoing process. For that reason, the technical partners of DIOPTRA will regularly reassess risks, update controls, and adapt to changes in the organisation and the information security landscape.

4.4.1.2 Network and Information Security Directive (NISD) 2016/1148/EU

The Network and Information Security Directive (NISD) 2016/1148/EU is a pivotal legislative framework within the European Union aimed at enhancing the overall cybersecurity resilience of critical infrastructure and digital services. Enacted to ensure a high common level of network and information security across member states, the directive mandates that essential service providers and digital service providers implement robust cybersecurity measures and report significant incidents to relevant authorities. By fostering a collaborative approach and setting standards for risk management, incident response, and cybersecurity capabilities, NISD plays a crucial role in bolstering the EU's collective defence against cyber threats.

In alignment with the Network and Information Security Directive (NISD) 2016/1148/EU, DIOPTRA platform prioritises the implementation of cybersecurity measures to fortify the resilience of critical infrastructure. Adhering to the NISD guidelines, DIOPTRA emphasises continuous risk assessments, ensuring that vulnerabilities are promptly identified and addressed. Robust incident response plans will be integral part of the DIOPTRA architecture, ensuring swift and effective responses to cybersecurity incidents.

4.4.1.3 ISO/IEC 29100 – Privacy Principles

ISO/IEC 29100 is a foundational standard that outlines Privacy Principles, providing a comprehensive framework for the protection of personally identifiable information (PII). This international standard is designed to guide organisations in establishing and maintaining privacy management programs, ensuring that the processing of PII aligns with ethical and legal considerations. ISO/IEC 29100 emphasises key privacy principles, such as consent, purpose specification, data minimisation, and accountability. By adhering to these principles, organisations can establish a privacy-centric approach to the handling of personal information, promoting transparency, user control, and responsible data governance.

Aligned with ISO/IEC 29100 – Privacy Principles, DIOPTRA platform places a paramount focus on safeguarding the privacy of individuals' health-related information. Transparent consent forms ensure that individuals have control over the use of their personal data within DIOPTRA. Purpose specification is a foundational aspect, clearly defining the purposes for which health data is collected and processed. Data minimisation principles guide the platform in collecting only the necessary information required for its intended purposes, reducing the risk associated with unnecessary data exposure. Accountability is embedded into the platform's governance structure, ensuring that all stakeholders are responsible for upholding privacy principles. By embracing the principles outlined in ISO/IEC 29100, the DIOPTRA platform endeavours to establish a robust and ethical foundation for the responsible management of personal health information.

4.4.1.4 ISO/IEC 29151:2017 – Code of Practice for Personally Identifiable Information Protection

ISO/IEC 29151:2017 provides a specific code of practice tailored for Personally Identifiable Information (PII) controllers. This international standard outlines guidelines and recommendations to assist organisations in managing and protecting PII responsibly. It focuses on the roles and responsibilities of

©2023-2026 DIOPTRA

PII controllers, emphasising transparency, fairness, and respect for individuals' privacy rights. ISO/IEC 29151:2017 serves as a valuable resource for organisations seeking to enhance their PII management practices, ensuring compliance with privacy regulations and fostering trust between PII controllers and individuals.

In alignment with ISO/IEC 29151:2017, DIOPTRA platform will adopt a meticulous approach to managing PII. DIOPTRA emphasises transparency by providing clear and accessible information about the collection and processing of PII. Fairness is ingrained in data processing practices, ensuring that individuals are treated ethically, and their privacy rights are respected. User consent is a cornerstone, with the platform implementing robust mechanisms to obtain and manage consent for the processing of PII. Accountability measures are integral, ensuring that PII controllers within the platform adhere to the established code of practice. By aligning with ISO/IEC 29151:2017, the DIOPTRA platform aims to uphold the highest standards of PII management, building a foundation of trust and compliance with international privacy best practices.

4.4.1.5 ETSI Cyber Security Technical Committee (Standards for Security and Privacy)

The European Telecommunications Standards Institute (ETSI) Cyber Security Technical Committee plays a crucial role in shaping cybersecurity standards within the telecommunications industry. ETSI, through its Cyber Security Technical Committee, develops and maintains a comprehensive set of standards and guidelines to address evolving cybersecurity challenges. This committee focuses on promoting best practices, fostering collaboration among industry stakeholders, and ensuring the interoperability and effectiveness of cybersecurity solutions. The standards produced by ETSI contribute significantly to enhancing the resilience of information and communication technologies (ICT) against cyber threats.

Aligned with the objectives of the ETSI Cyber Security Technical Committee, DIOPTRA platform is committed to implementing robust cybersecurity measures. DIOPTRA will integrate ETSI standards to enhance its resilience against cyber threats within the healthcare domain. By adopting ETSI guidelines, the platform ensures that cybersecurity best practices are embedded in its architecture, addressing vulnerabilities and safeguarding sensitive health information. Adherence to ETSI's evolving standards is central to our approach, allowing DIOPTRA software to stay abreast of the latest cybersecurity developments and maintain a proactive defence against emerging threats.

4.4.1.6 Working groups of the ISO/IEC JTC 1 / SC 27

The ISO and the IEC Joint Technical Committee 1 (JTC 1) Sub-Committee 27 (SC 27) is dedicated to information security, cybersecurity, and privacy protection standards. Within this framework, various Working Groups operate to develop and maintain standards that address the evolving landscape of information security. These working groups cover a wide range of topics, including risk management, cybersecurity controls, privacy protection, and secure coding practices. The collaborative efforts of these working groups contribute significantly to the establishment of globally recognised standards that organisations can implement to enhance their information security posture.

In alignment with the efforts of ISO/IEC JTC 1/SC 27 Working Groups, DIOPTRA will implement cutting-edge information security measures. The platform will actively integrate relevant standards developed by these working groups to ensure a robust and comprehensive approach to information security and privacy protection. By staying engaged with the outputs of specific working groups related to healthcare, the platform aligns its security protocols with globally accepted standards.

©2023-2026 DIOPTRA

4.4.2 Adhering to Privacy by Design Principles of GDPR within DIOPTRA

In accordance with the GDPR's Privacy by Design principles, DIOPTRA platform will prioritise a proactive and integrated approach to safeguarding user privacy throughout its development and deployment. From the outset, privacy considerations are embedded into the core architecture, ensuring that privacy is not an afterthought but a foundational element of the system. This approach encourages a privacy-conscious culture within DIOPTRA development team, fostering awareness of the importance of protecting individual rights and personal data.

4.4.2.1 Privacy-Enhanced Architecture

The system's architecture reflects a commitment to Privacy by Design by incorporating privacy-enhanced features at every level. This includes implementing robust access controls, data encryption mechanisms, and anonymisation techniques to minimise the risk of unauthorised access and data breaches. By adopting a privacy-centric architectural design, we aim to provide users with a secure environment that prioritises their privacy rights and protects sensitive health information throughout the entire lifecycle of the e-health platform.

4.4.2.2 Consent Mechanisms

DIOPTRA platform places a strong emphasis on user consent and control over personal data. Transparent consent forms will allow individuals to make informed decisions about the collection and processing of their data. Users have the autonomy to grant or withdraw their consent at any stage of the clinical trial. By prioritising user agency, DIOPTRA not only complies with GDPR requirements but also establishes a foundation of trust between users and the platform, emphasising ethical and user-centric data practices.

4.4.2.3 Data Integrity

Ensuring the integrity of data will be accomplished through diverse of data protection techniques. This includes the implementation of backup and replication processes, regular auditing of access logs and GDPR requests status, meticulous data and input validation measures, the removal of duplications, and the enforcement of robust access controls. These methods collectively contribute to maintaining the accuracy, consistency, and reliability of the data, safeguarding it against potential errors, unauthorised access, or compromise.

4.4.2.4 Data Minimisation

Authentication and authorisation form the fundamental framework for an end-user to log into the system and obtain user information. While acquiring user information is essential, it is equally imperative to establish an access control mechanism for its application. Role-based access control (RBAC) is a concept centred on allocating permissions to end-users based on their organisational roles. This approach minimises the risk of errors due to its simplicity and ease of management compared to individually assigning permissions to each user. Following user requirements, individuals are categorised into groups, each assigned a specific role. Subsequently, users are then assigned one or more roles, and each role is associated with one or more permissions, ensuring a structured and effective access control system. In the possibility that there are overlapping roles, the permissions that the user has is the union of the permissions of each role the user has, since RBAC is an additive model. In terms of services and how triggering them is controlled, all Restful APIs will need to first obtain a secure token. Securing RESTful APIs is paramount in today's digital landscape and utilising KrakenD provides an effective solution for robust API security. KrakenD ensures secure communication through mechanisms such as HTTPS, encrypting data in transit. Additionally, KrakenD supports authentication and authorisation protocols, allowing for the implementation of stringent access controls. By

©2023-2026 DIOPTRA

leveraging KrakenD's capabilities, developers can enforce secure communication channels, authenticate users, and authorise access to API resources, thereby fortifying the overall security posture of RESTful APIs in a streamlined and efficient manner.

4.4.2.5 Continuous Evaluation and Improvement

Adhering to Privacy by Design principles is not a one-time task but an ongoing commitment to ensuring the highest standards of privacy protection. DIOPTRA will incorporate a continuous evaluation mechanism to assess and enhance its privacy features. Regular privacy impact assessments, audits, and updates to address emerging privacy challenges are integral to our approach. By actively monitoring and improving Privacy by Design implementation, we demonstrate a dedication to staying ahead of evolving privacy concerns and maintaining the integrity of the e-health platform in compliance with GDPR requirements.

4.4.2.6 Data Transfer Outside EU

Anticipated within the DIOPTRA project is the absence of personal data transfers. Nevertheless, in instances where the transfer of personal data that is either undergoing processing or intended for processing in a third country or international organisation becomes necessary, such transfers must strictly adhere to the conditions stipulated by the GDPR. This includes compliance by the controller and processor, encompassing subsequent transfers from the third country or international organisation to yet another third country or international organisation. All GDPR provisions are to be diligently applied to ensure that the established level of protection for individuals, as guaranteed by the GDPR, remains intact. It is essential to note that any processing of personal data by the DIOPTRA solution will occur exclusively within the GRNET EU-based green data centre located in Greece. Under these circumstances, there is an unequivocal commitment to ensuring that the data remains in the EU.

4.5 PLATFORM INTEGRATION

The technical architecture presented in Section 4.3 will be set up within the infrastructure of GRNET S.A.[3], 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. More specifically, the following Virtual Machines (VMs) have been set up to host the DIOPTRA platform and uploaded data (adjustments will be considered, if necessary, throughout development and testing):

Table 84: Virtual Machines set up within GRNET Infrastructure

VM1	VM2	VM3	VM4	VM5
Master Node, Logstash, Kibana, API Gateway	Elastic Data nodes 1 & 2	Elastic Data nodes 3 & 4	Programming Environment & Interface	Staging Environment for all Services
8 or 16 cores	8 cores	8 cores	4 cores	8 cores
32 GB RAM	16 GB RAM	16 GB RAM	8 GB RAM	16 GB RAM
500 GB disk	200 GB disk	200 GB disk	100 GB disk	100 GB disk

[3] https://grnet.gr/

©2023-2026 DIOPTRA

5 CONCLUSIONS

Within the context of this document, technical and operational requirements towards the attainment of DIOPTRA outcomes were described. More specifically, relevant ecosystem actors, workflows and use cases were defined and translated into a detailed architecture featuring all the components that are needed to successfully implement the clinical studies of the project and develop the full DIOPTRA screening solution. Specific requirements were delineated on each one of the components individually, while also establishing related dependencies and rationale in relevance to the use cases. Overall, the key conclusions of Deliverable D2.1 can be summarised as follows:

- The implementation of the DIOPTRA clinical studies (biomarker discovery, retrospective, prospective, follow-up) was translated into a **clinical workflow** describing data collection and exploitation towards model development and evaluation. Study rationale and participation of different population groups were also addressed.

- Detailed requirements with regard to the **biomarker assay development for CRC screening** were documented in detail, including requirements on sample collection and management, infrastructure, equipment, as well as on the regulatory framework. Specific implementation details reflecting these requirements have been / will be included in corresponding deliverables of the project.

- **Technical requirements** aiming to fulfil the corresponding needs from the perspectives of **clinical workflow and screening model development** were identified.

- The above information has been translated into u**se cases referring to both system development and real-world use**, linking each use case with the involved actors and exploited DIOPTRA components.

- The **full technical architecture** aiming to serve the above processes was presented, together with functional and non-functional requirements linked to the implementation and testing of this architecture. Development workflows, data flows and component exploitation for development were also conjured.

- Security standards were analysed, in order to present compliance guidelines based on key protocols and privacy-by-design principles.

The above will serve as a thorough reference guide during development and evaluation, with any updates and deviations that may occur being documented (together with the related rationale) into the relevant deliverables of the project that will include the final outcomes of the corresponding work.

Funded by the European Union

©2023-2026 DIOPTRA

REFERENCES

[1] D. Zowghi and C. Coulin, "Requirements Elicitation: A Survey of Techniques, Approaches, and Tools," in *Engineering and Managing Software Requirements*, A. Aurum and C. Wohlin, Eds., Berlin, Heidelberg: Springer, 2005, pp. 19–46. doi: 10.1007/3-540-28244-0_2.

[2] Iiba, *A Guide to the Business Analysis Body of Knowledge*, 3rd ed. edition. Toronto: International Institute of Business Analysis, 2015.

[3] "Olink Explore 3072 - Olink." Available: https://olink.com/products-services/explore/

[4] "RNA-seq Sample Guidelines - The Huck Institutes (en-US)." Available: https://www.huck.psu.edu/core-facilities/genomics-core-facility/sample-recommendations/rna-seq-sample-guidelines

[5] SelectScience, "Quality control in Illumina sequencing workflows using the TapeStation system | SelectScience." Available: https://www.selectscience.net/application-articles/quality-control-in-illumina-sequencing-workflows-using-the-tapestation-system/?artid=58915

[6] "Illumina Stranded mRNA Prep Ligation Reference Guide." Available: https://support.illumina.com/downloads/illumina-stranded-mrna-reference-1000000124518.html

[7] M. Lundberg, A. Eriksson, B. Tran, E. Assarsson, and S. Fredriksson, "Homogeneous antibody-based proximity extension assays provide sensitive and specific detection of low-abundant proteins in human blood," *Nucleic Acids Res.*, vol. 39, no. 15, p. e102, Aug. 2011, doi: 10.1093/nar/gkr424.

[8] C. W. Fuller *et al.*, "The challenges of sequencing by synthesis," *Nat. Biotechnol.*, vol. 27, no. 11, pp. 1013–1023, Nov. 2009, doi: 10.1038/nbt.1585.

[9] K. LeFevre, D. J. DeWitt, and R. Ramakrishnan, "Mondrian Multidimensional K-Anonymity," presented at the Proceedings of the 22nd International Conference on Data Engineering, May 2006, pp. 25–25. doi: 10.1109/ICDE.2006.101.

[10] P. Samarati and L. Sweeney, "Protecting privacy when disclosing information: k-anonymity and its enforcement through generalization and suppression," 1998. Available: https://www.semanticscholar.org/paper/Protecting-privacy-when-disclosing-information%3A-and-Samarati-Sweeney/7df12c498fecedac4ab6034d3a8032a6d1366ca6

[11] K. El Emam and F. K. Dankar, "Protecting Privacy Using k-Anonymity," *J. Am. Med. Inform. Assoc.*, vol. 15, no. 5, pp. 627–637, Sep. 2008, doi: 10.1197/jamia.M2716.

Funded by the European Union

©2023-2026 DIOPTRA