Intelligent ecosystem to improve the governance, the sharing, and the re-use of health data for rare cancers

Deliverable 6.1

Specification of multimodality navigation

May 2024

This project has received funding from the European Union's Horizon Europe research and innovation programme under grant agreement no. 101057048

This project has received funding from the European Union's Horizon Europe research and innovation programme under grant agreement no. 101057048

This project has received funding from the European Union's Horizon Europe research and innovation programme under grant agreement no. 101057048

Distribution List

Organization	Name of recipients
1 - Coord INT	A. Trama, P. Casali, L. Buratti, P. Baili, J. Fleming, L. Licitra, E. Martinelli, G. Scoazec
2 - UDEU	A. Almeida, U. Zulaika Zurimendi, N. Kalocsay
3 - MME	F. Mercalli, S. Copelli, M. Vitali
4 - UPM	E. Gaeta, G. Fico, L. Lopez, I. Alonso, A. Estevan, V. G. Dominguez, I. Alonso, L. Hernandez, C. Vera
5 - HL7	G. Cangioli, C. Chronaki
6 - ECCP	S. Ziegler, S. Schiffner, V. Tsiompanidou
7 - ENG	P. Zampognaro, A. Sperlea, E. Mancuso, M. Melideo, F. Saccà, V. Falanga, M. Rosa
8 - CERTH	K. Votis, A. Triantafyllidis, N. Laloumis, I Drympeta
9 - UU	S. van Hees, Wouter Boon, E. Moors, M. Kahn-Parker, C. Egher
10 - DICOR	C. Lombardo, G. Pesce, G Ciliberto, A. Tonon,
10° - ACC (Affiliated)	L. Villanova, P. De Paoli, G. Piaggio, M. Pallocca. A. De Nicolo
11 - FBK	A. Lavelli, S. Poggianella, O. Mayora, A.M. Dallaserra
12 – IKNL	E. Bosma. G. Geleijnse, A. Van Gestel, F. Martin, E. Mezei
13 - CLB	A. Sans, M. Brahmi, A. Pons, J-Y Blay, H. Crochet, J. Olaz, J. Bollard, C. Chemin-Airiau, H. Crochet
14 - APHP	B. Baujat, E. Koffi
15 - FJD	J Martin-Broto, N. Hindi, M. Martin Ruiz, A. Montero Manso, C. Roldàn Mogìo, D. Da Silva, A. Herrero, B. Barrios
16 - VGR	Magnus Kjellberg, L. De Verier, A. Muth
17 - MSCI	I. Lugowska, D. Kielczewska, M.Rosinska, A KAwecki , A., P. Rutkowski
18 - MUH	R. Knopp, A. Sediva, K. Kopeckova, A. Nohejlova Medkova, M. Vorisek
19 - OUS	S. Larønningen, J. Nygård, M. Sending, O. Zaikova
20 - MMCI	J. Halamkova, I. Mladenkova, l. Tomastik, V. Novacek, T. Kazda, I. Mladenkova, O. Sapožnikov, V. Novacek
21 - CLN	R. Szmuc, J. Poleszczuk, R. Lugowski
22 - FPNS	M. Barbeito Gomez, P. Parente, L. Carrajo Garcia, P. Ramos Vieiro
23 - TNO	E. Lazovik, L. Zilverberg, S. Dalmolen
24 - INF	ML Clementi, C. Sabelli
25 - UKE	S. Bauer, S. Lang, S. Mattheis, N. Midtank, M. Kim

This project has received funding from the European Union's Horizon Europe research and innovation programme under grant agreement no. 101057048

Revision History

Revision	Date of Issue	Author(s)	Brief Description of Change
0.1	20/09/2023	UPM	ToC
0.2	27/06/2024	UPM	Sections 2, 3 and 4.1 completed. 4.2 started. Each with correspondents' annexes
0.3	28/06/2024	UPM (with UU contributions)	All sections except 4.2 completed
0.4	01/07/2024	UPM	Final and complete version
1	06/08/2024	INT	Version revised, ready for submission

This project has received funding from the European Union's Horizon Europe research and innovation programme under grant agreement no. 101057048

Addressees of this document

This document is addressed to the whole IDEA4RC Consortium. It is an official deliverable for the project and shall be delivered at the European Commission and appointed experts.

This project has received funding from the European Union's Horizon Europe research and innovation programme under grant agreement no. 101057048

TABLE OF CONTENTS

This project has received funding from the European Union's Horizon Europe research and innovation programme under grant agreement no. 101057048

LIST OF FIGURES

This project has received funding from the European Union's Horizon Europe research and innovation programme under grant agreement no. 101057048

LIST OF TABLES

This project has received funding from the European Union's Horizon Europe research and innovation programme under grant agreement no. 101057048

Abbreviations and definitions

Abbreviation	Definition
AI	Artificial Intelligence
CoEs	Centres of Excellence – Centres of Expertise
DPO	Data protection officer
EHDS	European Health Data Space
FAIR	Findability, Accessibility, Interoperability, and Reusability
H&N	Head and neck
hi-fi	High-fidelity
lo-fi	Low-fidelity
ML	Machine learning
NLP	Natural language processing
OIDC	OpenID Connect
P.I. / PI	Principal Investigator
PI	Principal Investigator
Q&A	Question-answering
RAVEN	Rare Cancer AI Virtual Exploration Navigator
SSI	Self-Sovereign Identity
STS	Soft tissue sarcomas
TEDHAS	Trusted, Efficient, Data-driven Healthcare Augmentation System
UCD	User-centred design

This project has received funding from the European Union's Horizon Europe research and innovation programme under grant agreement no. 101057048

1 EXECUTIVE SUMMARY

The IDEA4RC project aims to enhance data findability and reusability by developing an advanced augmented analytical system powered by Artificial Intelligence (AI), specifically through the Multimodal Data Navigator. This deliverable outlines the comprehensive approach taken to design and implement this system, ensuring it meets the needs of diverse stakeholders and adheres to the FAIR principles (Findability, Accessibility, Interoperability, and Reusability). The primary objective of WP6 is to support data (re)users in efficiently finding and utilizing rare cancers datasets from the federated IDEA4RC ecosystem. The Multimodal Data Navigator (implemented in T6.4) will address the complexities of the dataset search problem by leveraging AI-driven tools such as question-answering chatbots, smart visualization dashboards, and task-oriented wizards.

The methodology for developing the Multimodal Data Navigator follows a user-centred design (UCD) approach, incorporating co-creation principles and iterative validation with Centres of Expertise (CoEs). This process ensures that the system is aligned with the needs of the WP8 pilot use cases and the broader user community. The first phase involved comprehensive gathering and analysis of the requirement to link the Navigator functionalities to the identified needs and recommendations. This included understanding user perspectives, conducting market research, assessing technical functionalities, consolidating requirements, and defining user management protocols. The conceptual design phase involved defining the system's workflow by interacting with various technical partners and understanding the interconnections between different tasks. Key components analysed include metadata services, data permit and governance processes, cohort builder, and data analysis services. The architecture design followed, establishing the system's structural framework and integrating the identified tools and services. Prototyping activities involved creating low-fidelity (lo-fi) and high-fidelity (hi-fi) mock-ups. Lo-fi mock-ups helped identify and conceptualize core functionalities, while hi-fi mock-ups refined graphical elements to ensure they met user needs. Iterative validation with CoEs ensured continuous improvement and alignment with user expectations.

The results chapter details the findings from the requirement gathering and analysis, workflow and architecture definition, and presents the finalized conceptual mock-ups representing the tool's overall design. Key functionalities identified include advanced search capabilities, data permit application processes, federated data analysis, and robust data visualization tools. These functionalities address the needs for efficient data discovery, access, analysis, and reporting.

2 INTRODUCTION

In the era of big data, the challenge of finding and effectively utilizing the right datasets has become increasingly complex. This is particularly true in the context of augmented analytics and multimodal data navigation, where datasets are extracted from diverse sources and presented in various forms, encompassing different qualities, sizes, and variables. The traditional manual methods of dataset discovery and navigation are insufficient to meet the demands of modern data (re)users, necessitating advanced, Artificial intelligence (AI)-powered solutions to enhance data findability and reusability, aligned with the FAIR (Findability, Accessibility, Interoperability, and Reusability) principles.

Virtual assistants have emerged as crucial tools in the domain of data analytics due to their ability to simplify and streamline interactions between users and data systems. By leveraging natural language processing (NLP) and machine learning (ML) capabilities, virtual assistants can interpret user queries, provide relevant responses, and guide users through complex datasets efficiently. This not only saves time but also reduces the cognitive load on users, allowing them to focus on deriving insights rather than getting bogged down by the intricacies of data navigation.

The relevance of virtual assistants is underscored by their widespread adoption across various industries. They are increasingly being utilized for a range of applications, from customer service and technical support to healthcare and finance, due to their ability to deliver personalized and context-aware assistance. In the context of augmented analytics, virtual assistants are particularly valuable as they can handle diverse data formats and sources, providing users with an intuitive and interactive way to access and analyse data.

The **IDEA4RC Virtual Assistant** aims to revolutionize the way data (re)users interact with datasets by providing a multimodal data navigation experience. Its primary goal is to support users in finding the datasets they need quickly and effectively, leveraging AI-driven technologies such as question answering chatbots and smart visualization dashboards. By integrating these advanced features, the IDEA4RC Virtual Assistant addresses the complex dataset search problem, making the process faster, easier, and more efficient.

The virtual assistant is designed to align with the needs of the WP8 pilot cases (included in D8.1) and the broader user community, ensuring that it can handle the varied requirements and preferences of different user groups. This alignment is crucial for creating an advanced augmented analytical system that not only meets but exceeds user expectations, enhancing data findability and reusability.

The primary **challenge** addressed by the Multimodal Data Navigator in WP6 is the **complexity of the dataset search problem**. With virtual datasets being recomposed from multiple diverse data sources, the task of finding relevant datasets becomes highly complex. **Traditional methods are inadequate** due to the sheer volume and diversity of data, necessitating a more sophisticated approach. The Multimodal Data Navigator aims to overcome these challenges by providing an **AI-powered solution** that integrates multiple paradigms of interaction. This includes natural language interfaces for intuitive querying, smart visualizations for better data

comprehension, and task-oriented processes to guide users through specific analytical pipelines. By leveraging these technologies, the IDEA4RC Navigator ensures that users can efficiently locate and utilize the datasets they need, thereby enhancing the overall data analysis experience.

2.1 Links between the augmented analytics and multimodal navigation (WP6) and other WP/Tasks

Figure 1 illustrates the interconnectedness between the Augmented Analytics and Multimodal Navigation work package (WP6) and other work packages/tasks within the broader project framework. WP6 plays a pivotal role by interfacing with multiple work packages to ensure a seamless flow of data and functionality. In particular, WP6 aligns closely with WP2, which identifies stakeholder needs and requirements. This alignment ensures that the tools and features developed within WP6 are user-centric and meet the practical needs of different user groups. Furthermore, WP6 is designed to support and enhance the pilot cases in WP8 by providing advanced data navigation and analytical tools tailored to the specific scenarios and use cases outlined in WP8. WP6 also relies on the outputs from WP3, WP4 and WP5, which are responsible for the extraction, re-composition, and provisioning of virtual datasets from each EURACAN hospital participating to the project or from the different datasets across the 11 CoEs of IDEA4RC, and delivery of AI services and the correspondent infrastructure. These datasets and services form the foundation for the augmented analytics and multimodal navigation capabilities developed in WP6. Moreover, WP6 have strong interaction with WP7 as it serves the data governance rules and requirements and implement the tool for managing data permit to the different CoEs.

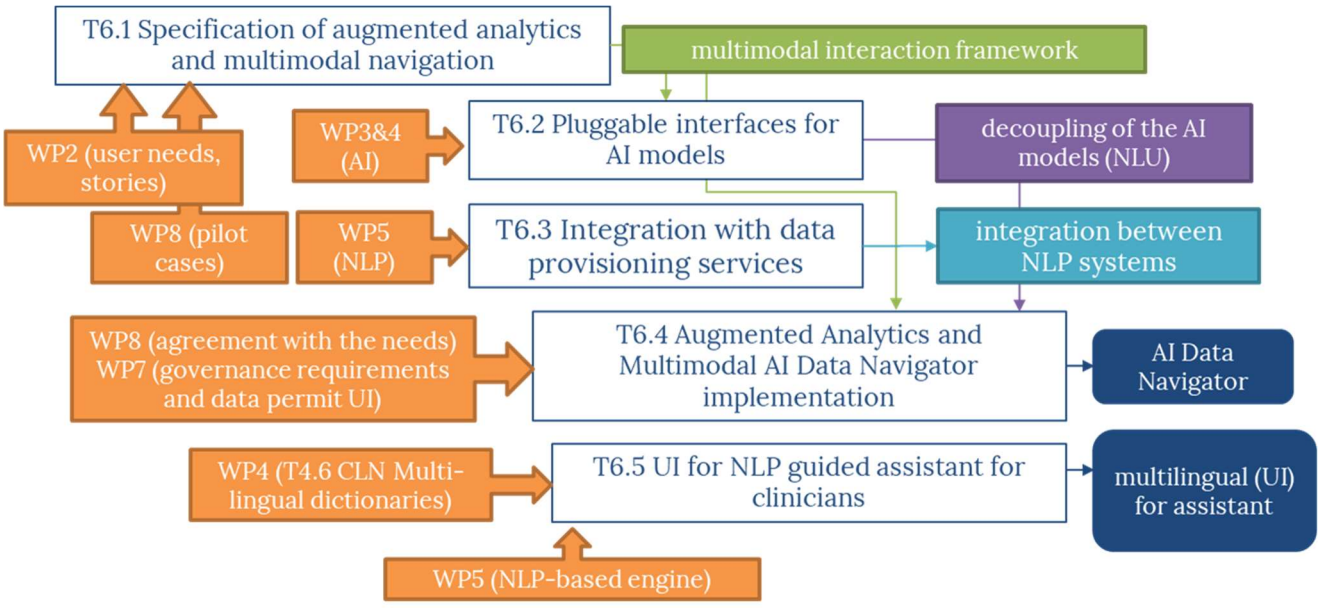

Figure 1. Relation between WP6 and other WPs and tasks

3 METHODOLOGY

The development of the Multimodal Data Navigator in WP6 is grounded in a comprehensive and iterative approach that ensures alignment with user needs and maximizes usability and effectiveness. The methodology follows a structured process based on User-Centred Design (UCD) and co-creation principles, involving continuous validation with CoEs and final users throughout all phases of the project. In the next sections, the steps followed are described.

3.1 Requirement gathering and analysis (Definition)

In this first phase the following activities have been carried out:

1) Initial specifications were gathered from the **user needs and stories identified in WP2**.
2) Input from **WP8 pilot cases**, in particular from D8.1, was incorporated to ensure the tool addresses practical scenarios and real-world applications.
3) **Two focus groups with potential users** have been launched to understand their specific needs, preferences, and challenges in data navigation and analytics. The Focus Groups were held in a remote manner using a teleconference platform. Two sessions were launched to gather input from more users. A Focus Group Script was prepared, in order to guarantee the use of the same approach in each pilot site as well as to ensure a smooth development of the Focus Group session. We asked the participants to express their consent to take part in the focus group and to be recorded. An online form was prepared to collect their consent to participate, record the meeting and also participants' profiles. The form was administered through the "Google Forms" platform to facilitate the distribution among the participants. This form is included in Annex 6.1. The session was structured (always asking users to participate and share comments) as follows:
    a. A **short introduction** of Augmented Analytics and Multimodal AI Data Navigators.
    b. Presentation of results of a small **research analysis** on tools related to what we hope to develop in this Virtual Assistant.
    c. Introduction of the **main functionalities** identified for the Virtual Assistant
    d. Specific questions about users' **professional role**, **daily activities** and **needs and challenges**.
4) **TEDHAS (Trusted, Efficient, Data-driven Healthcare Augmentation System) journey** has been analysed to be compliant with **EHDS (European Health Data Space)**, ensuring that the proposed navigator adheres to the principles and regulations aimed at facilitating the secure and efficient exchange of health data across Europe. This analysis was conducted to properly link the Navigator functionalities to the requirements and recommendations of EHDS. This ensures that TEDHAS not only improves healthcare delivery, research, and policymaking by enabling access to high-quality data but also strictly adheres to privacy and data protection standards.

All these activities served to define the ***functionalities from the user perspective (Section 4.1.1).***

5) A **market research** has been done to identify existing "Navigator-like" solutions within the eHealth market with the aim to assess what current tools provide nowadays, what is missing and what can be 'reused' or added in the IDEA4RC AI multimodal data navigator tool.

Results in **Section 4.1.2.**

6) **Technical sessions** have been launched, complemented by a survey to collect high level functionalities from technical perspectives.

These sessions and surveys served to define the ***functionalities from the technical perspective (Section 4.1.3)***.

7) **Requirements consolidation**. Once we have collected and understood the needs and initial requirements of the navigator, the ***Volere Requirements Technique***[1] was used, as a basis, to transform these needs to functional and non-functional requirements. Specifically, the identification of functional and non-functional requirements was conducted following the Volere template that has been used since 1995 and has become a commonly used methodology for requirements elicitation. The Volere template table allows to categorize and describe each requirement by ID, requirement type and sub-type (the Volere proposed requirement types are included in Annex 6.2) rationale, priority, difficulty and originator. The process followed to complete the final list of requirements can be divided into three steps:
    a. Gathering and analysis of requirements from information collected in previous activities and from the transformation of user's needs to requirements.
    b. Documentation of functional and non-functional requirements in a structured table, following the Volere template (see Annex 6.3)
    c. Validation and prioritization, by the stakeholders, of the final requirements against user needs and initial requirements.

The ***final list of requirements (Section 4.1.4)*** will be implemented by priority during the development process of the tool within T6.4.

8) To acquire additional insights related to the **user management** of the Multimodal Data Navigator, feedback from medical researchers from different CoEs was gathered through WP2 activities. The medical researchers were asked to briefly provide their opinions on the proposed user roles and permissions for the Multimodal Data Navigator, ensuring that it meets the needs of a research team. Specifically, the questions aim to determine if respondents agree with the following:

[1] VOLERE requirements type (See http://www.volere.co.uk/)

a. The Principal Investigator (PI) should have full access to all functionalities of the data navigator.
b. Other researchers should have limited access, with specific permissions tailored to their roles.
c. Some researchers should have read-only access, allowing them to view data and analyses but not perform analyses or create visualizations and reports.

Additionally, medical researchers are asked to provide:

- The number of user types (2-4) that should be considered for the IDEA4RC data navigator.
- A brief description of the main rights and functionalities that should be assigned to each user type. This includes which activities (e.g., data quality checks, certain data analyses, interpretation of results) should be accessible to each role.

3.2 Conceptual design

Based on the gathered requirements, a conceptual design of the Multimodal Data Navigator was created. This included defining the core features, interaction paradigms, and user interfaces. Throughout the design process, UCD principles ensured that user feedback was continuously incorporated. Co-creation workshops were conducted with end-users from CoEs to collaboratively refine the design and functionality. This iterative process ensured that the navigator was tailored to user needs, enhancing usability and relevance.
In particular, the following activities have been carried out:

1) **Architecture design**. The architecture design activity involved creating a detailed blueprint for the Multimodal Data Navigator, ensuring that all components are well-integrated and function seamlessly together. This process included several key steps:
   a. **Defining system components**: Identification and specification of all major components and modules required for the navigator, including data ingestion, processing, storage, and user interface components.
   b. **Integration planning**: Developing integration strategies for linking the navigator with external services and tools, such as metadata services, data permit and governance processes, the cohort builder, and Vantage6 services. This ensures seamless data flow and interoperability between systems.
   c. **Technology stack selection**: Choosing appropriate technologies and frameworks to support the development and deployment of the navigator.
   d. **Security and compliance considerations**: Incorporating security measures and ensuring compliance with relevant regulations, such as those outlined by the EHDS. This included data encryption, access control, and auditing mechanisms. *This specification will be assessed and ensured within T6.4.*
   e. **Scalability and performance planning**: Ensuring that the architecture supports scalability to handle increasing data volumes and user loads. Performance

optimization strategies were also planned to ensure efficient and fast operation of the navigator. *This specification will be assessed and ensured within T6.4.*

f. **Validation and review**: Conducting thorough reviews and validation of the architecture design with technical partners and stakeholders to ensure it meets all requirements and is aligned with project goals. *This specification will be assessed and ensured within T6.4.*

Results in **Section 4.2.1.**

2) **Workflow design.** This involves interacting with different technical partners to understand the links between various tasks and how each solution must be orchestrated within the tool. This activity involves the design of the data flow within the system, including data input, processing, storage, and retrieval. This also involved defining communication protocols and APIs for interaction between different components and external systems. Specifically, the integration and coordination with the following services/tools have been analysed and considered:
   a. ***Metadata services***: Ensuring that the tool can effectively handle and utilize metadata for improved data discoverability and management.
   b. ***Data permit and governance process***: Incorporating robust processes and *data permit tool (from WP7)* for data access permissions and governance to ensure compliance and security.
   c. ***Cohort builder***: Incorporating *cohort builder tool* (from WP3) to support the creation and management of data cohorts for analysis.
   d. ***Vantage6 services***: Ensuring compatibility and seamless operation with Vantage6, a platform for privacy-preserving federated learning.

3) **Prototyping.** Initial prototypes were developed to visualize and test the conceptual design. These prototypes ranged from low-fidelity (lo-fi) wireframes to high-fidelity (hi-fi) interactive models.
   a. **Rapid prototyping** techniques (i.e., **lo-fi mock-ups**, using power point) allowed for quick iterations and early feedback from users and stakeholders.
   b. **Hi-fi mock-ups** enable the team to determine which graphical elements are most suitable for each conceptual function approved in the Lo-fi mock-ups.
   c. Each iteration of the prototype was subjected to **refinement with final users** from the different CoEs, **and technical partners** that need to deliver the corresponding services. Feedback from iterative sessions was used for incremental improvements, ensuring the final mock-up adapt to user needs.

Results in **Section 4.2.2.**

4 RESULTS

This chapter presents the comprehensive results of our requirement gathering and analysis phase, laying the foundation for the development of the Multimodal Data Navigator. A detailed exploration of the functionalities from the user perspective is included, capturing the specific needs and preferences of users. This is followed by an analysis of the current market landscape, providing insights into existing solutions and identifying gaps that our navigator aims to fill.
Next, we delve into the functionalities from a technical perspective, outlining the capabilities and features required to support the envisioned user functionalities. The chapter then consolidates these functionalities within the defined requirements, ensuring a coherent and aligned approach to development. Additionally, results from the conceptual design phase are included, in particular, results of the workflow and architecture definition are presented, detailing the structural blueprint and integration strategies for the navigator. Finally, we showcase the high-fidelity mock-ups, representing the overall design of the tool and illustrating how the conceptual functions are visually and interactively realized.

4.1 Results of the requirement gathering and analysis (Definition)

This section is primarily structured around the TEHDAS user journey phases, as this approach effectively organizes the functionalities, requirements, and tool steps. The identified journey phases, to be considered within the navigator tool are:

- **Data discovery phase**. This phase begins with ***metadata exploration***, where users examine the available datasets published in a data catalogue. This catalogue shall manage metadata across the IDEA4RC federated environment (i.e., FHIR capsules at each CoE). This exploration should be facilitated by a Dataset Discovery Dashboard, which provides a high-level overview of the datasets available at the CoEs. This step aims to equip users with essential information, guiding them towards submitting a data permit application in the next phase of their journey.
- **Data permit application phase**: The data access application phase is efficiently coordinated by two essential components: the Data Access Application and Permit Provision. The ***Data Access Application*** process is designed to meet the needs of data users, enabling them to apply for and manage their data access requests seamlessly. This system prioritizes a user-centric approach, ensuring a smooth and efficient experience when requesting data permits. The ***Permit Provision*** process complements this by managing the review and approval workflow. It ensures that all data access requests are evaluated in accordance with regulatory requirements and institutional policies, providing a structured pathway for users to gain authorized access to data.
- **The data use phase**: The data use phase involves the utilization of health data for research, analysis, and decision-making purposes. Authorized individuals, granted access through a data permit, can analyse health data to gain insights, develop treatments, and enhance

healthcare. In the IDEA4RC user journey, federated analytics is employed during this phase. Federated analytics enables the analysis of data from multiple sources while preserving decentralization and security. It ensures privacy by allowing sensitive information to remain under the control of data owners and enables researchers to directly access data from its original sources.

- **Data finalization phase**: The data finalization phase encompasses services related to the disclosure of findings derived from data analysis. These services also help in preparing and presenting the results. The reporting document should include the requirements for accessing the original datasets, either partially or anonymized, to ensure the reproducibility of the results in research or other contexts, following the FAIR principles.

4.1.1 Functionalities from the user perspective

Table 1 includes the functionalities extracted for the **data discovery phase**. In summary, the functionalities of the Multimodal Data Navigator include: searching for variables and datasets, identifying relevant cohorts of soft tissue sarcomas and head and neck cancers, sharing searches within an institution, setting general discovery conditions, and refining exploration with keyword filters.

Table 1. Functionalities analysis – discovery phase

Nº	Theme	Title	Detailed description	Origin
1	Metadata search	**Variables exploration**	Search for variables and their associated information. This includes verifying if data exists for your target cohort.	User story from WP2
2	Metadata search	**Search and select datasets of interest**	Search for a dataset, meeting study-specific variables, to decide whether the datasets available meet the user need.	User story from WP2
3	Metadata search	**Targeted search**	Identify relevant cohorts of soft tissue sarcomas (STS) and head and neck (H&N) cancers through targeted searches. Then, assess data completeness for your research questions.	Research questions from D8.1
4	Metadata search	**Share searches**	Share existing searches in a coworking environment with members of the same institution	Focus Groups
5	Metadata search	**First general search**	General conditions at the beginning proposed to be selected, that facilitate the discovery of the data.	Focus Groups
6	Metadata search	**Granular data exploration with filtering**	Refine your exploration with keyword filters directly within relevant text areas. This allows for a more granular view of the data, helping you pinpoint specific details within free-form text fields.	Focus Groups

This project has received funding from the European Union's Horizon Europe research and innovation programme under grant agreement no. 101057048

Table 2 includes the functionalities extracted for the **data permit phase**. In summary, the functionalities of the Multimodal Data Navigator include: requesting access to research cohorts, obtaining data permits, adhering to ethical guidelines, benefiting from semi-automated governance, receiving application results, monitoring application progress, updating applications, data protection officer (DPO) management of applications, receiving notifications, sharing applications, organizing submissions, filling forms with free text, uploading documents with limitations, and receiving clear feedback on rejected applications.

Table 2. Functionalities analysis - permit phase

Nº	Theme	Title	Detailed description	Origin
7	Data access applications	**Request Access to Data**	Request access to a specific study cohort through a user-friendly data access application. Obtain a data permit for authorised use of the data. Fulfil your secondary use objective while adhering to ethical and regulatory guidelines. Benefit from semi-automated data governance processes for a faster and more efficient experience.	User story from WP2
8	Data access applications	**Receive and visualise the result of the submission**	Receive the result of the submission of the Data Access Application. The submission can be an approval, a rejection with a motivation, a request for additional information	User story from WP2
9	Data access applications	**Visualisation of a users' submissions history**	Monitor the progress of your submitted Data Access Applications. View a detailed history of all your past submissions, including their status. Stay informed about any updates or decisions regarding your applications.	User story from WP2
10	Data access applications	**Modify and resubmit a Data Access Application**	Easily update existing Data Access Applications based on feedback from CoEs. User can add the requested information to address any outstanding questions and ensure a smooth approval process for your data permit.	User story from WP2
11	Data access applications	**Manage the data applications**	The DPO and and other actors involved in application review at the CoEcan visualize the applications submitted to request data access and he/she can manage them.	Focus Groups

Nº	Theme	Title	Detailed description	Origin
12	Notifications	**Notifications about the result of the submission**	The user will receive notification with the result of the submissions of the Data Access Application (e.g., response in a submission, deadline reminder of a resubmission, etc.).	Focus Groups
13	Data access applications	**Shared data applications**	The data applications can be shared between the members of the institution and also of different institutions, to co-work.	Focus Groups
14	Data access applications	**Applications repositories**	The user can group and organize all the submissions applications in folder repositories	Focus Groups
15	Data access application response	**Data application in a form format**	The Data Application needs to be filled in a form format with some free text inputs.	Focus Groups
16	Notifications	**Notification settings**	Customizable notifications through the profile settings, where the user can set if he/she wants notifications and the channel (e.g., email, through the platform etc).	Focus Groups
17	Data access application response	**Upload documents**	The submission of the data access application should allow the upload of documents.	Focus Groups
18	Data access application response	**Limited number of documents submitted**	Number and size of the documents limited because of storage. But a reasonable number of documents and space will be provided.	Focus Groups
19	Data access application response	**Data submission response**	Clear feedback of rejected application in order to resubmit correctly the application.	Focus Groups

Table 3 includes the functionalities extracted for the **data use phase**. In summary, the functionalities of the Multimodal Data Navigator include: creating and executing queries, sharing queries, using filters for precise data cohorts, running quality checks and federated analyses, AI support for data scientists, generating and merging variables, visualizing model performance, extracting demographic descriptions, refining data analysis, offering various analysis options, unlimited analyses on data cohorts, sharing analyses and ensuring explainable visualizations and statistical results.

Table 3. Functionalities analysis – data use phase

Nº	Theme	Title	Detailed description	Origin
20	Cohort builder	**Create, edit and execute queries**	Create, edit and execute queries Creation of new queries based on existing queries	Focus Groups
21	Cohort builder	**Share queries**	Share existing queries in a coworking environment with members of the same or of different institutions collaborating in the study	Focus Groups
22	Cohort builder	**Granular Data Exploration with Filtering**	Leverage filters to select precise data cohorts. This includes options like filtering by the percentage of useful information available, ensuring you focus on the most reliable and relevant subsets of data for further analysis.	Focus Groups
23	Data quality	**Run quality checks**	Run quality checks across a selected cohort, to assess if study specific quality requirements are met.	User story from WP2
24	Federated Data Analysis	**Execute federated data analysis**	Run a federated data analysis on the selected cohort data, spanning across CoEs, to achieve the secondary use objectives.	User story from WP2
25	Federated Data Analysis	**Execute federated data analysis**	A data scientist Virtual Assistant is available, with the following functionality: - AI pipeline support for data scientists supporting the MLOps workflow - Usage of existing tools for statistical analysis already known to the Data User (examples are R, MATLAB, Python, etc.)	User story from WP2

Nº	Theme	Title	Detailed description	Origin
26	Data preprocessing	**Generate new variables**	Generate new variables from one single variable (e.g. defining specific group of one continuous variable) or merging existing variables (e.g. define multimodal treatment)	Research questions from D8.1
27	Model Evaluation	**Visualize the performance results of the models**	Visualize the performance results of the models using ROC curve, accuracy, confusion matrix, etc.	Research questions from D8.1
28	Data analysis	**Extract demographic descriptions**	Extract demographic descriptions of the cohort through graphics	Research questions from D8.1
29	Data analysis	**Filters to select which cohort to apply the data**	Refine the data analysis by applying filters like selecting specific cohorts or date ranges.	Research questions from D8.1
30	Data analysis	**Analysis options**	Endpoint: Choose the event signifying the end of the analysis, such as death or disease progression. Analysis Period: Define the timeframe for the analysis by specifying start and end dates. Follow-up End: Set a specific date to end the follow-up period for the analysis. Survival by Year: Calculate survival rates at each year following diagnosis. Median Survival: Determine the median survival time within the selected timeframe. Conditional Survival: Analyse survival probability considering other factors.	Research questions from D8.1
31	Data analysis	**Unlimited number of analysis**	There should not be a limitation on the number of analyses you can do over the same data cohort	Focus Groups

Nº	Theme	Title	Detailed description	Origin
32	Coworking	**Share the analysis results**	The analysis should be able to be shared with other users/researchers (e.g. statisticians) for further evaluation and validation if needed.	Focus Groups
33	Data analysis	**Several programming languages to perform the analysis**	The tool should be as general as possible and include as much working languages or programs as possible (e.g., Python, Matlab)	Focus Groups
34	Visualization	**Self-explainable data visualizations**	The visualization of the data should be explainable and easy to understand	Focus Groups
35	Visualization	**Self-explainable analysis' results**	The statistical results should be explainable to understand the statistical meaning behind the results	Focus Groups
36	Visualization	**Visualization of the performance results**	The visualization of the performance results should include: - Visualization of the residuals of the models and test assumption. - Visualization of the indicators of the performance of the models (confusion matrix) - Visualization of the ROC curve	Focus Groups

Table 4 includes the functionalities extracted for the **data finalization phase**. In summary, the functionalities of the Multimodal Data Navigator include: visualizing and preparing results for export, examining requirements for accessing original datasets to ensure reproducibility, and downloading and editing analysis results for publication submission.

Table 4. Functionalities analysis - data finalization phase

Nº	Theme	Title	Detailed description	Origin
37	Results preparation	**Visualize and prepare the results**	Visualize and prepare the results to export them	User story from WP2

Nº	Theme	Title	Detailed description	Origin
38	Data requirements	**Examine the requirements to reproduce the study**	Examine the requirements for accessing the original datasets, either partially or anonymized, to ensure the reproducibility of the results in research or other contexts.	User story from WP2
39	Notifications	**Incidental finding notification**	Notify concerned CoEs of an incidental finding resulting from the study conducted, so that CoEs can take relevant measures, according to current Regulations and their current practice.	User story from WP2
40	Download results	**Edit and download the results**	The results of the analysis should be able to be download and edited to be submitted for publications	Focus Groups

4.1.2 Augmented analytics and multimodal AI data navigators: current situation

The current situation of existing tools, analysed and summarised in Annex 6.4, shows a diverse array of eHealth solutions, each aimed at different aspects of clinical research and patient management. These tools are **not designed to handle specific cancer types**, except *The Rare Disease Cures Accelerator-Data and Analytics Platform* that focused on rare diseases, including some rare types of cancer, and *Cbioportal*, *xena* and *ICGC Portal* that focus on wide variety of cancer types, including solid and hematological tumors. This indicates a **need for more specialized tools tailored to particular cancer types to enhance targeted research**

The tools identified serve **various end-users, including clinicians, patients, and researchers**. They manage and analyse **different types of clinical data** such as electronic medical records, genetic information, imaging data, and physical activity metrics. These tools often feature **advanced data visualization** capabilities using dashboards, plots, graphs, tables, and diagrams, providing users with intuitive and accessible insights. Additionally, **many of these tools are web-based**, enabling remote access and collaboration. They also offer customization options for individualized treatment plans and adhere to standards like FHIR to ensure data interoperability and seamless integration with other systems. Despite these strengths, **many of these tools have not yet reached the validation phase or the market**, highlighting a significant gap in the development and implementation stages.

To address the existing gaps, the Navigator plans to incorporate several advanced functionalities, such as **on-demand guidance and support** or **natural language query building** (*under feasibility assessment within WP5, T6.2, T6.3 and T6.5*). These features are designed to create a more **user-friendly and efficient platform for clinical research application**. The goal is to ensure **better data accessibility and usability** across diverse user groups and cancer types, ultimately enhancing research capabilities and patient outcomes.

Note: *Most of the tools included here (e.g., Flan-PLM) will be assessed in the context of T6.2.*

4.1.3 Functionalities from the technical perspective

The following tables delve into the technical functionalities that drive the project, providing an idea of the internal mechanisms of the system.

Data discovery – High level functionalities (T4.2 and T6.3 related)

The ***Data discovery*** component serves as a critical tool to empower end users with a comprehensive understanding of project data and facilitate seamless navigation through searching processes. Designed to address the complex landscape of data understanding, it offers a range of features and services aimed at elucidating the underlying data model and providing users with interesting aggregate statistics. Table 5 provides technical functionalities for this phase at high level.

Table 5. Data discovery – High level tech functionalities

What is the main purpose or goal of the *Data discovery*? What problem does it solve or what value does it add?
Help end users to understand the data within the IDEA4RC project and facilitate data searching.
What are the main features or services that the *Data discovery* offers? How do they support the purpose or goal of the *Data discovery*?
It offers a set tools to understand the data model (work ongoing) and which variables can be retrieved from the navigator. E.g. provides a view of the data model for a given preliminary conditions (without getting real data) and provides aggregated statistics (metadata) for the available datasets.
How does the *Data discovery* interact with other components of the IDEA4RC system? (Wherever possible relate it with the corresponding tasks)
It should have a view on the data model, be able to query capsules to get aggregated data; and be able to query capsules to get metadata.
What are the inputs and outputs of the *Data discovery*?
Inputs would be possible queries to the navigator (or views/variables). Output would be: aggregated statistics, views on the data model and metadata output.
How do you plan to design, implement, test, and deploy the *Data discovery*? What tools, frameworks, languages, and methodologies do you use?
A distributed architecture utilizing Python-based REST APIs within each capsule will likely be employed. This approach minimizes backend processing, potentially resulting in faster response times.

Query builder – High level functionalities (T3.3 related)

The **Query Builder** has the fundamental purpose of simplifying the execution of distributed cohort queries, offering users a tool that facilitates the search and retrieval of data of a specific cohort of pts (e.g., Leiomyosarcoma in 15-39 years old, sinonasal cancers in advanced stage etc.) among multiple data capsules. With a friendly user interface, efficient backend and seamless communication with data capsules, the Query Builder seeks to optimize the user experience while providing a robust solution for the management and execution of feasibility queries in

clinical and research environments. Table 6 provides technical functionalities for this phase at high level.

Table 6. Query builder – High level tech functionalities

What is the main purpose or goal of the *Query builder*? What problem does it solve or what value does it add?
Provide a tool that makes it easier for end users to execute distributed cohort queries on a specifici cohort of patients.
What are the main features or services that the *Query builder* offers? How do they support the purpose or goal of the *Query builder*?
This includes a *cohort-builder* (i.e. web app) for managing the queries (CRUD capabilities) and related execution historical results, a *cohort-builder-api* (i.e. Rest API) that is the backend middleware for query management and distribution, and a *fhircapsule-query-executor* (i.e. Rest API) within the capsule to build and execute the query within the capsule. It is expected that the user interface simplifies the process of a search query definition. For instance, complex queries with *intercriteria* dependencies are not supported by FHIR Search. A way to overcome this limitation is to break a query into multiple smaller parts.
How does the *Query builder* interact with other components of the IDEA4RC system? (Wherever possible relate it with the corresponding tasks)
The expected interaction is between the *fhircapsule-query-executor* (which is a service deployed in each capsule) and the FHIR server of the capsule itself.
What are the inputs and outputs of the *Query builder*?
cohort-builder (i.e. web app): - Input: criteria expressed by the users - Output: the list of matching patients from each capsule *cohort-builder-api* (Rest API) - Input: the query (to save, to get, to delete, to execute) - Output: the operation result *fhircapsule-query-executor* (Rest API) - Input: the query provided in the format coming from the user interface - Output: the list of patient IDs matching criteria
How do you plan to design, implement, test, and deploy the *Query builder*? What tools, frameworks, languages, and methodologies do you use?
- User Interface (cohort-builder): This is the web application the user will interact with. It's built with Angular, a popular framework known for creating user-friendly interfaces. Think of it as the dashboard where you define your queries. - Backend Management (cohort-builder-API): This acts behind the scenes, handling communication and data flow. It is implemented as a Java microservice, meaning it's a small, independent service that efficiently manages specific tasks. - Capsule Communication (FHIRcapsule-query-executor): This component resides within each data capsule (where patient information is stored). It's built with Node.js, another efficient technology. Its role is to translate your query into a format the capsule understands and then retrieve the matching patient data.

Data permit – High level functionalities (T7.2 related)

Data permit is a component that streamlines the process for researchers to legally access and analyse sensitive data. It helps ensure that researchers comply with data protection regulations and with the data (re) use conditions established by the data holder/data controller, while allowing them to conduct valuable studies. The main function is the result of a successful data access request, authorising researchers to analyse specific data under defined terms. Among its benefits, this component aims to simplify request submission, provides custom content and templates, assists with resubmissions and modifications. Table 7 provides technical functionalities for this phase at high level.

Table 7. Data permit – High level tech functionalities

What is the main purpose or goal of the *Data permit*? What problem does it solve or what value does it add?
A data permit is the result of a successful data access application process. The researcher (data user) who is granted a data permit by one or more CoE(s) (data holder(s)) can legally proceed with the analysis of the data described in the data permit according to the terms described in the data permit.
What are the main features or services that the *Data permit* offers? How do they support the purpose or goal of the *Data permit*?
Through the WP7 solutions, researchers (data users) can seamlessly navigate the process of submitting a data access application for a selected cohort of patients, enabling them to obtain the necessary permit for legally accessing the corresponding data. This involves guiding researchers through the required steps and providing tailored content for their application, along with templates for essential documents. Assistance is also available in scenarios where a data access application requires resubmission due to additional information requested by the data holder for approval, and support can be extended for requests to amend previously approved applications (issued data permits).
How does the *Data permit* interact with other components of the IDEA4RC system? (Wherever possible relate it with the corresponding tasks)
There is interaction with tasks among T3.4, T4.2, T5.4, and T6.2, which are involved in the data discovery, and with T4.4 for the consistent and controlled access based on predefined policies and rules.
What are the inputs and outputs of the *Data permit*?
Inputs: • Metadata concerning the selected cohort. • Documents specified in T2.2 to accompany the data access application, such as the study protocol, study-specific data management plan, study-specific Data Protection Impact Assessment (DPIA), and ethics approvals. • Requested terms for data use in accordance with the EHDS proposal, including aspects like time duration, purpose, etc. Output: • Signed agreement.
How do you plan to design, implement, test, and deploy the *Data permit*? What tools, frameworks, languages, and methodologies do you use?
The upcoming activities are the design of the functional building blocks for the data access application and data permit provision processes, as well as the agreement on the detailed data flows involving them.

Single-Sovereign Identity based on the European Blockchain Services Infrastructure (EBSI) will be used for user management and Hyperledger Fabric for the auditing of data permits. The solution is planned to be tested by pilots after the provision of the first version.

Data use – High level functionalities (T4.3 related, and this T6.1)

The **Data use** component serves as a pivotal hub for enabling sophisticated data exploration and analysis while upholding stringent privacy standards and legal compliance. With a primary goal of facilitating anonymized statistical summaries and aggregated model outputs, **Data use** empowers users to delve into complex datasets, harnessing both traditional statistical techniques and cutting-edge machine learning models. Through seamless integration with other system components, such as FHIR/OMOP capsules for healthcare data and legal compliance frameworks, Data use ensures the efficient exchange and utilization of data while adhering to regulatory requirements. Additionally, the component offers a range of services, including data cleaning, augmentation, and a spectrum of analytical methods, providing users with comprehensive tools to derive actionable insights from their data. Powered by the versatile VANTAGE6 framework, the design, implementation, and deployment of Data use embody a commitment to robustness, scalability, and user-centric functionality. Table 8 provides technical functionalities for this phase at high level.

Table 8. Data use – High level tech functionalities

What is the main purpose or goal of the *Data use*? What problem does it solve or what value does it add?
Enable data exploration through anonymized statistical summaries and aggregated model outputs.
What are the main features or services that the *Data use* offers? How do they support the purpose or goal of the *Data use*?
Services offered include data cleaning and augmentation, as well as the execution of statistical analyses encompassing both univariate and multivariate techniques. Additionally, the provision extends to the implementation of machine learning models.
How does the *Data use* interact with other components of the IDEA4RC system? (Wherever possible relate it with the corresponding tasks)
The Data use component of the IDEA4RC system interfaces with other components such as the FHIR/OMOP capsules from WP3, facilitating seamless integration of standardized healthcare data. Additionally, it engages with legal aspects handled in WP7, ensuring compliance and alignment with relevant regulations throughout the data utilization process. This interaction enables efficient data exchange and utilization while adhering to legal frameworks.
What are the inputs and outputs of the *Data use*?
Input type of analysis and parameters selected by the user from the platform interface. The output will be the results of the analysis or visualization.
How do you plan to design, implement, test, and deploy the *Data use*? What tools, frameworks, languages, and methodologies do you use?
The VANTAGE6 framework will be used. Detailed information here.

Data finalization – High level functionalities (Inputs come directly from T6.1)

The **Data finalization** component within the IDEA4RC system plays a pivotal role in advancing the accessibility and interoperability of research outcomes in the scientific and medical domains. By addressing the critical need for efficient data sharing and archiving, Data finalization ensures that valuable insights generated by researchers are disseminated effectively to other users. Table 9 provides technical functionalities for this phase at high level.

Table 9. Data finalization – High level tech functionalities

What is the main purpose or goal of the *Data finalization*? What problem does it solve or what value does it add?
Answer the research question and guarantee the dissemination of results to other users. It must be guaranteed that its results are easily found, accessible, interoperable and reusable by other researchers. This includes publishing the results in standard formats, providing complete metadata, and documenting the analysis methods used. By providing a framework for effective data sharing and archiving, data completion helps ensure that the valuable insights generated by researchers are fully leveraged by the scientific and medical community.
What are the main features or services that the *Data finalization* offers? How do they support the purpose or goal of the *Data finalization*?
Archiving and results validation services: Results preparation services. Catalogue of results: Notifications to data controllers. Advisory and support services: They support the purpose of data completion by facilitating the sharing of results, improving the quality of results. Increase transparency and provide support to researchers.
How does the *Data finalization* interact with other components of the IDEA4RC system? (Wherever possible relate it with the corresponding tasks)
Requires the previous phases outputs
What are the inputs and outputs of the *Data finalization*?
Inputs: • Results Data: The results of the study analysis, which may include analysis data, visualizations, reports, and other relevant materials. • Metadata: Information about the results data, such as the source of the data and the variables included. • Documentation: Documentation that describes the research project, the methods used and the findings. • Data permissions: Permissions that give data users the right to share and archive results data. Outputs: • Archived results data: Results data stored securely in a centralised repository. • Catalogued metadata: The metadata of the results data indexed in a search catalogue. • Final Results Package: A package containing archived results data, catalogued metadata, and documentation. • Notifications: Notifications to data controllers and data subjects about project findings.
How do you plan to design, implement, test, and deploy the *Data finalization*? What tools, frameworks, languages, and methodologies do you use?
Implemented directly within the navigator.

4.1.4 Virtual assistant requirements

The requirements for a virtual assistant system are structured following the TEDHAS phases: Discovery, Permit, Use, and Finalization. The 160 requirements, listed in details in Annex 6.5, include various functional, technical, and usability aspects necessary for effective implementation and operation. Below these requirements are described grouped by the respective phases.

Note: *all the requirements are now under technical feasibility assessment, and will be re-confirmed with users. Final implementation details will be included in D6.4.*

Discovery Phase

- Implement search features to find variable information and display metadata details (Rq_01, Rq_02).
- Support general data discovery questions and include free-text areas for detailed metadata search (Rq_03, Rq_04).
- Provide search history visualization, collaborative environments, permissions management for shared searches, and ownership transfer capabilities (Rq_05, Rq_06, Rq_07, Rq_08).
- Allow users to add comments and annotations to facilitate communication and provide context for collaborators (Rq_09). *Included in all phases.*
- Implement a notification system to inform users about shared, modified, or new searches (Rq_10). *Included in all phases.*
- Design a user-friendly interface with access controls, documentation, and Single Sign-On (SSO) implementation (Rq_11, Rq_12, Rq_13, Rq_14). *Included in all phases.*

Permit Phase

- Allow users to submit and manage Data Access Applications, including status checks, additional information submission, and notification systems (Rq_15 - Rq_24).
- Support multiple notification channels and customizable preferences for notification frequency and content (Rq_25 - Rq_28).
- Enable selective sharing of Data Access Applications, access controls for shared applications, collaborative editing, and commenting (Rq_30 - Rq_33).
- Provide customizable folder structures for managing Data Access Applications, intuitive navigation tools, various folder views, and folder export features (Rq_34 - Rq_38).
- Implement secure access for folders, provide comprehensive documentation on export procedures, dataset access, and guidelines for ensuring reproducibility (Rq_39, Rq_152 - Rq_154).

Data Use Phase

- Provide interfaces for cohort selection, query execution, and exploration of data availability for research questions (Rq_60 - Rq_65).
- Facilitate query creation and execution, implement query validation mechanisms, and provide history visualization for reproducibility and auditability (Rq_66 - Rq_68).

- Incorporate quality check algorithms, detailed quality check reports, and data visualization (Rq_78 - Rq_83).
- Support federated query execution, data visualization, and collaboration among data scientists within CoEs (Rq_87 - Rq_89).
- Provide support for AI pipelines, and integration with existing tools like R, MATLAB, and Python (Rq_90 - Rq_92).
- Ensure the system handles scalability efficiently, manage computational resources for concurrent analyses, and provide detailed revision history and documentation on analytical procedures (Rq_86, Rq_115 - Rq_118, Rq_141).

Finalization Phase

- Allow users to download analysis results in various formats (Rq_149).
- Enable users to export specific subsets of the results or selected variables providing flexibility in exporting only relevant information (Rq_150).
- Allow users to specify publication restrictions, language preferences, ensure editable output formats for submission, and export data visualizations for inclusion in publications (Rq_155 - Rq_158).
- Include metadata with downloaded results, ensure compliance with publication guidelines, and provide comprehensive documentation on variable transformation methods and performance interpretation guidelines (Rq_159, Rq_160, Rq_142, Rq_143).

Overall System Features

- Design intuitive and user-friendly interfaces for various functionalities including search, metadata management, cohort selection, query execution, and analysis result presentation (Rq_11, Rq_23, Rq_75).
- Implement robust security measures to ensure secure access and data privacy throughout the system (Rq_12, Rq_49, Rq_57, Rq_76).
- Provide comprehensive documentation and training resources covering all aspects of the system for both end-users and administrators (Rq_13, Rq_50, Rq_59, Rq_105, Rq_146).

4.2 Results of Conceptual design

Based on a survey and meetings with the CoEs, the tool has been named **RAVEN (*Rare Cancer AI Virtual Exploration Navigator*)**. The subsequent sections detail the architecture, workflow, and mock-ups.

4.2.1 Architecture design

The RAVEN architecture is designed to provide a robust and secure environment for managing, accessing, and analysing data from CoEs. This architecture integrates various components, including the RAVEN web application, orchestrator, metadata APIs, and specific tools for data permits and cohort building, each playing a crucial role in the overall system. Figure 2 represents this architecture, detailing the interactions and functionalities of each component.

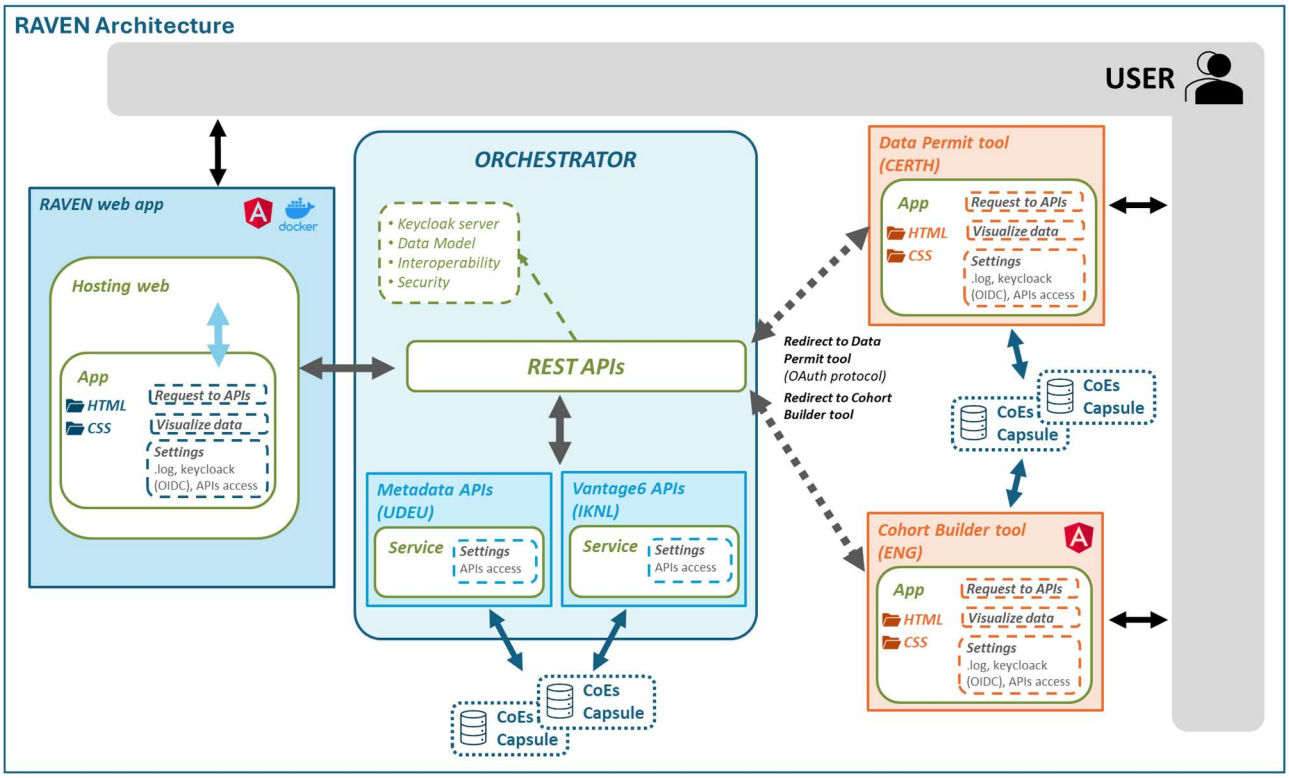

Figure 2. RAVEN architecture

Each of the components of the architecture is detailed below.

1. **RAVEN Web App:** The RAVEN web application will be hosted using Docker, ensuring a consistent and isolated environment for deployment. The application will comprise HTML, CSS, and Angular for a dynamic user interface. It handles requests to APIs, visualizes data, and manages settings such as logs, OpenID Connect (OIDC) such as Keycloak, and API access.
2. **Orchestrator:** The orchestrator will be the central hub for managing various services and ensuring interoperability and security. It includes an OIDC server (e.g., Keycloak) for user authentication, a data model for structuring information, and security protocols to protect data integrity. Moreover, the orchestrator will manage the REST APIs facilitating the communication between the web app and other services, ensuring seamless data flow and interoperability.
    a. **Metadata APIs:** This component (service) provides API access and settings for metadata management, allowing to retrieve and manage metadata information effectively.
    b. **Vantage6 APIs:** Similar to the Metadata APIs, this service provides API access and settings for the Vantage6 framework, enabling advanced data processing and analysis capabilities.
3. **Data Permit Tool:** This web tool, built with HTML, CSS, and Angular, handles requests to APIs and visualizes data. It manages settings such as logs, OIDC, and API access. This

tool interacts with the CoEs Capsule, which stores and manages the datasets accessible through the Data Permit Tool, allowing users to request access to data.

4. **Cohort Builder Tool:** Similar to the Data Permit Tool, the Cohort Builder Tool is a web app built with HTML, CSS, and Angular. It handles API requests, visualizes data, and manages settings, including logs, OIDC, and API access. This tool also interacts with the CoEs Capsule, enabling users to build and manage cohorts based on the available datasets.

Interaction and Data Flow

Users interact with the *RAVEN web app* to initiate various processes, such as searching metadata or performing data analysis. The *orchestrator* manages these interactions, ensuring secure and efficient data flow between the web app and other services. Requests are routed through the REST APIs to the relevant components, such as the Metadata APIs or Vantage6 APIs, depending on the required functionality.

For data access, users are redirected to the *Data Permit Tool*, where they can request access to datasets stored in the CoEs Capsule. The OAuth protocol is used for secure authentication and authorization. Similarly, for cohort building, users are redirected to the *Cohort Builder Tool*, where they can visualize data and manage cohorts based on the datasets in the CoEs Capsule.

This architecture ensures a streamlined and secure process for managing, accessing, and analysing data from CoEs. The interactions between the RAVEN web app, orchestrator, Metadata APIs, Vantage6 APIs, Data Permit Tool, and Cohort Builder Tool are designed to provide users with a seamless experience. By leveraging REST APIs for communication and using robust security measures like Keycloak and OAuth protocols, the Multimodal Data Navigator facilitates efficient data retrieval, management, and analysis while maintaining data integrity and security.

Finally, the defined architecture offers several key features and benefits that enhance its functionality and user experience, such as:

- **Modularity**: Each component, from the web app to the individual tools, operates as a modular unit, allowing for flexible and scalable deployment.
- **Security**: The use of Keycloak for authentication and OAuth for secure data access ensures that user data and research information are protected.
- **Interoperability**: REST APIs enable smooth communication between different components, ensuring that data can be efficiently exchanged and utilized across the system.
- **User-Friendly Interface**: The use of Angular for the web app and tools provides a responsive and interactive user interface, enhancing the overall user experience.
- **Comprehensive Data Management**: From metadata search to cohort building and analysis, the Multimodal Data Navigator offers a comprehensive suite of tools for all stages of the TEHDAS data journey.

4.2.2 Workflow management system & mock-ups

The workflow of the RAVEN tool is visually represented in Figure 3. This diagram outlines a detailed and organized process divided into four main phases (in line with the TEHDAS user journey phases): **Metadata Search** (data discovery), **Data Access** (data permit), **Data Analysis** (data use), and **Result Report** (data finalization). Each phase is designed to facilitate a comprehensive approach to managing data, from the initial search for metadata to the final reporting of analysis results.

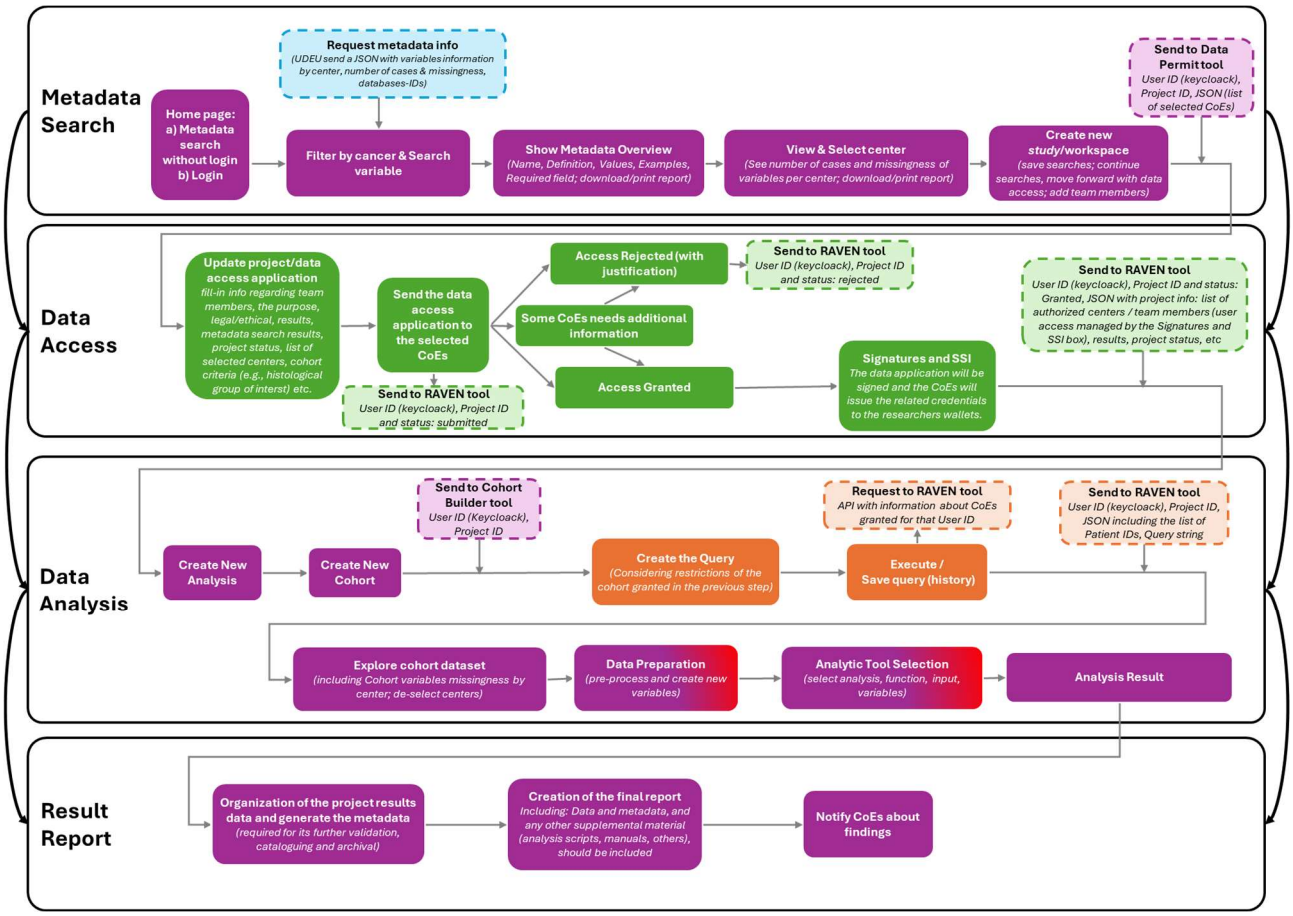

Figure 3. RAVEN workflow

The flow of each phase is detailed below, along with mock-ups illustrating the functionalities of the RAVEN tool.

Phase 1: Metadata Search

The process begins with the user accessing the welcome page (Figure 4), where they can either perform a metadata search without logging in or choose to log in. Users can filter the data by cancer type and search for specific variables. Then the Metadata Service is requested for metadata information obtained in a JSON file that includes variables by centres. This leads to the metadata overview (Figure 5), which provides detailed information such as the name, description, values and examples of the CoEs variables. Following this, users view and select

CoEs of interest based on the number of cases and the missingness of variables per centre, with the option to download or print the report (Figure 6). Notably, at various steps in the Metadata Search phase, the user has the option to download or print a report.
At this point the users can finish the metadata search or can make the request for access to the CoEs data. For this, user must proceed to create a new study or workspace (if user is logged in), allowing them to save searches, continue searches, and move forward with data access and team member management.
Finally, the study or workspace and user information, together with the list of selected CoEs will be send to the Data Permit Tool allowing the user to request the access of data.

Figure 4. Welcome page

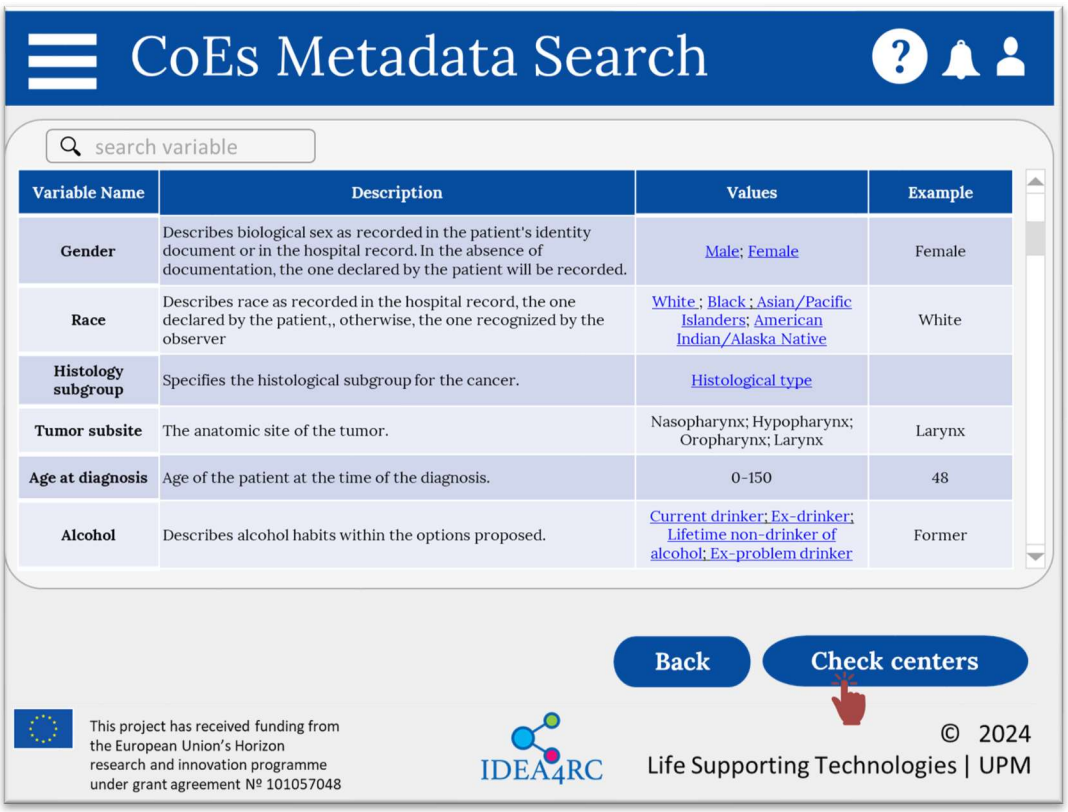

Figure 5. Metadata Overview screen

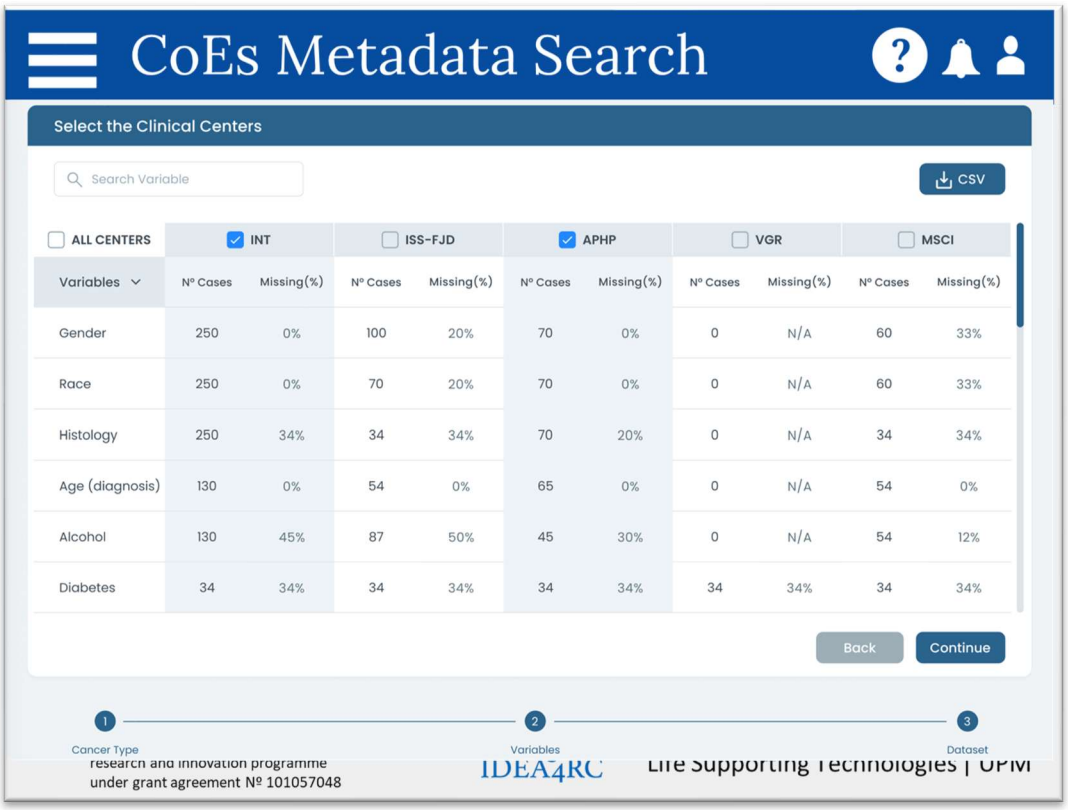

Figure 6. Visualization and selection of CoEs

Phase 2: Data Access

Once the user is in the Data Permit Tool, can update the project or data access application by filling in information regarding team members, purpose, legal/ethical aspects, metadata search results, project status, cohort criteria, etc. This updated application is sent to the selected CoEs. The data access request can result in either access being rejected where a justification for the rejection must be provided, or in the need to submit additional information for some CoEs. If access is granted, signatures and Self-Sovereign Identity (SSI) are obtained, in this point the data application will be signed and the CoEs will issue the related credentials to the researchers' wallets.

The Data Permit Tool is then sent to the RAVEN tool in a JSON file containing project information, including the list of authorized CoEs, team members, and other project details. This enables researchers to continue with the data analysis.

The user will be consistently updated about the status of the data application through both the Data Permit tool and the RAVEN tool.

Figure 7 *shows a mock-up of the Data Permit tool. More details and definition will be included in deliverable D7.2 (WP7).*

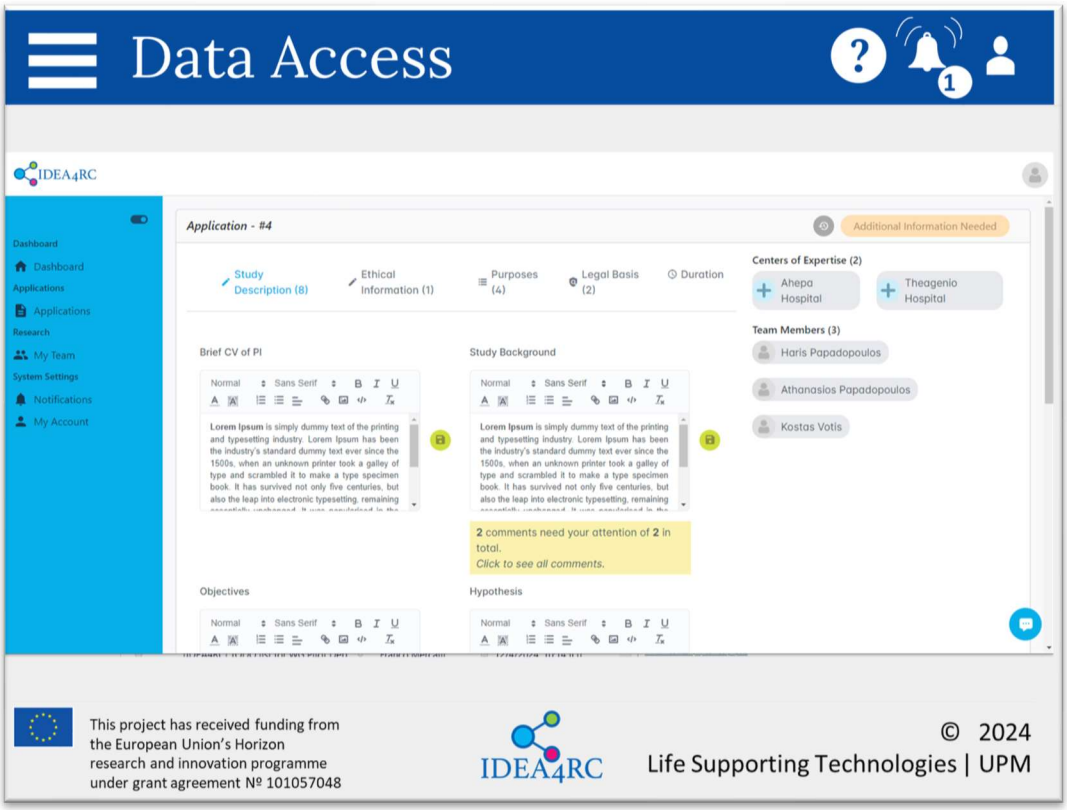

Figure 7. Data applications management (from Data Permit Tool)

Phase 3: Data Analysis

Once access is granted, users can create a new analysis (Figure 8) or a new cohort. They proceed by using the Cohort Builder tool (Figure 9), inputting their user ID and project ID. The next step involves creating the query, considering restrictions of the granted access, and executing or saving the query (history, see Figure 10). Users explore the cohort dataset (Figure

11), reviewing cohort variables and missingness by CoEs, and de-selecting centers as necessary. Data preparation follows, involving the pre-processing and creation of new variables (Figure 12). Users then select the appropriate analytic tools, functions, and input variables (Figure 13), leading to the generation of the analysis result (Figure 14).

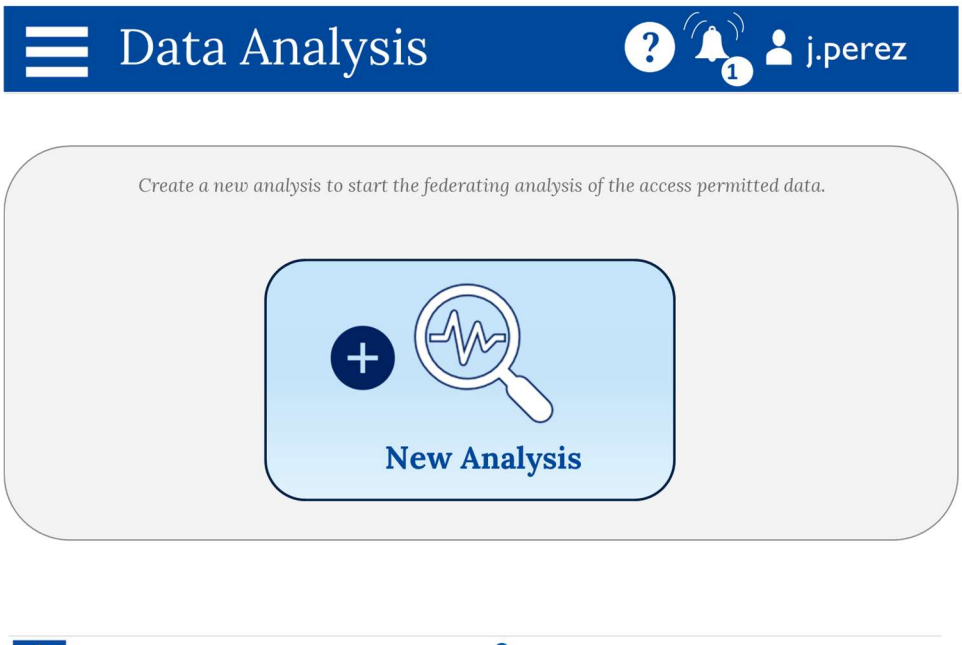

Figure 8. New analysis creation

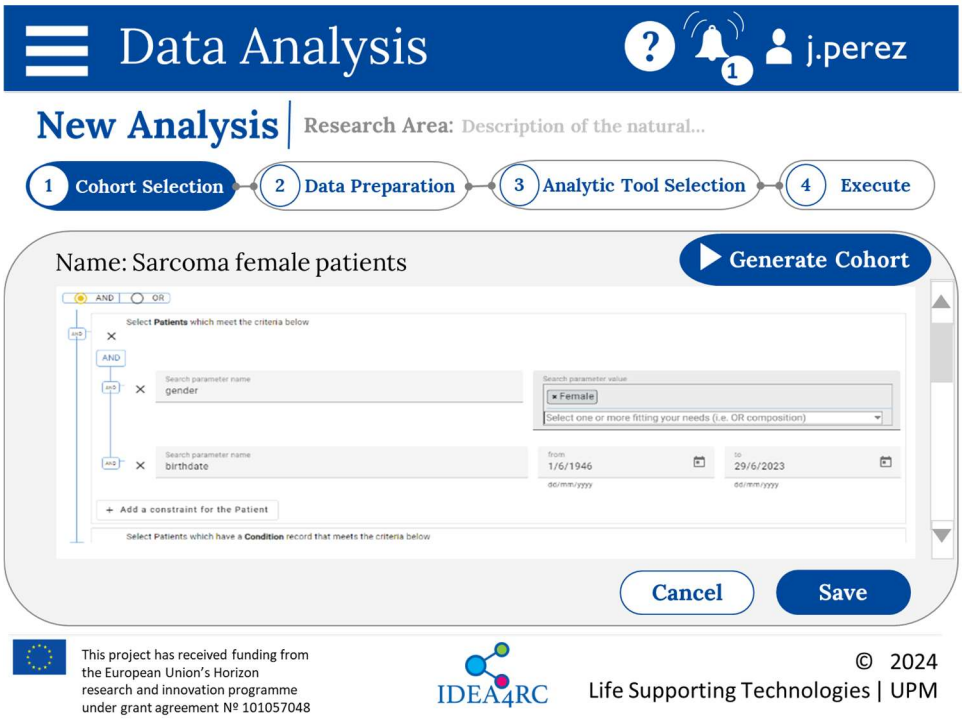

Figure 9. Cohort builder tool

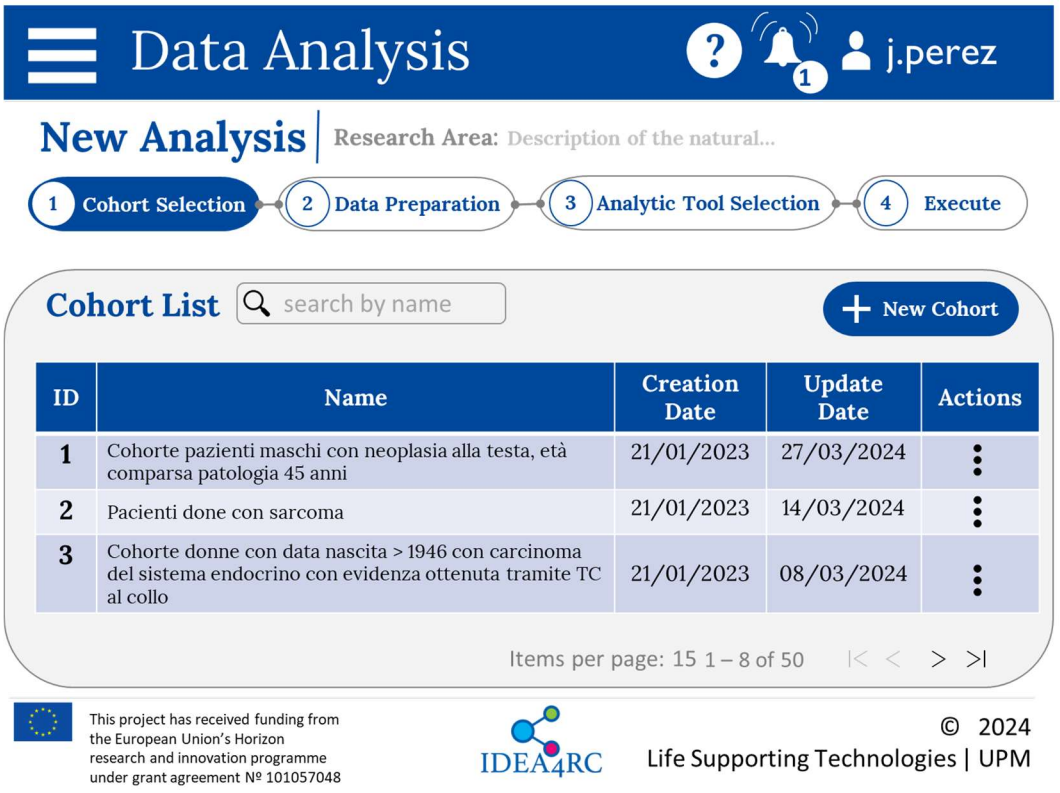

Figure 10. Cohort history - from Cohort builder tool

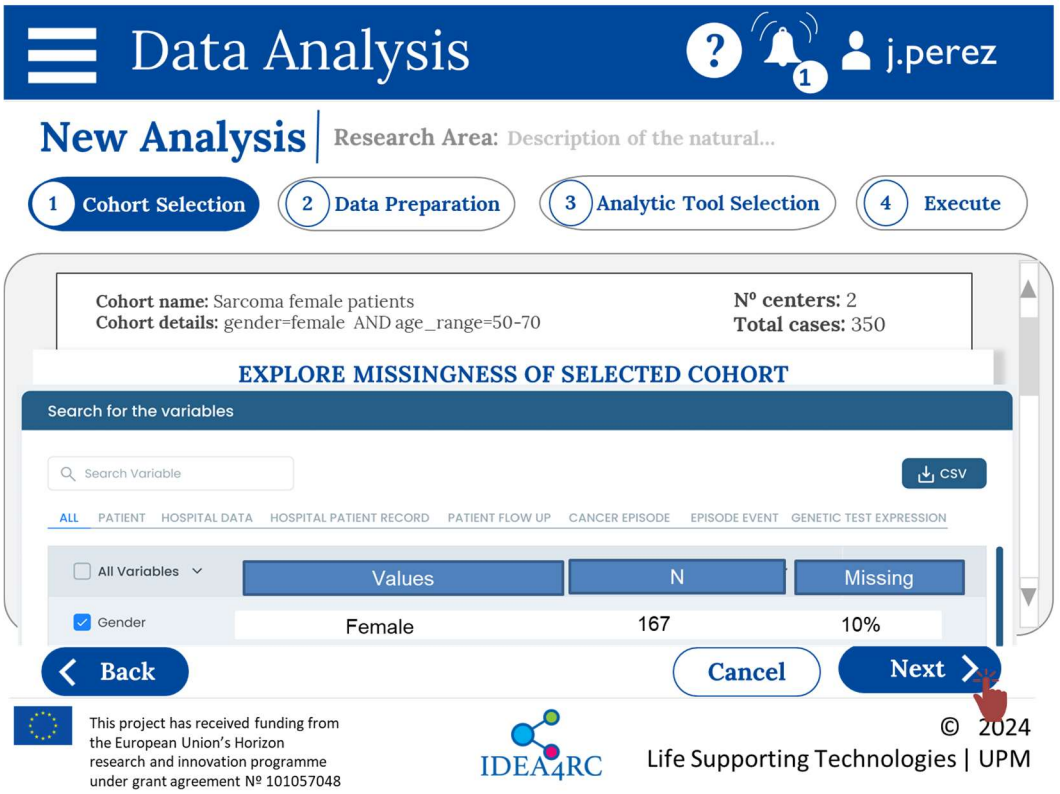

Figure 11. Explore cohort selected

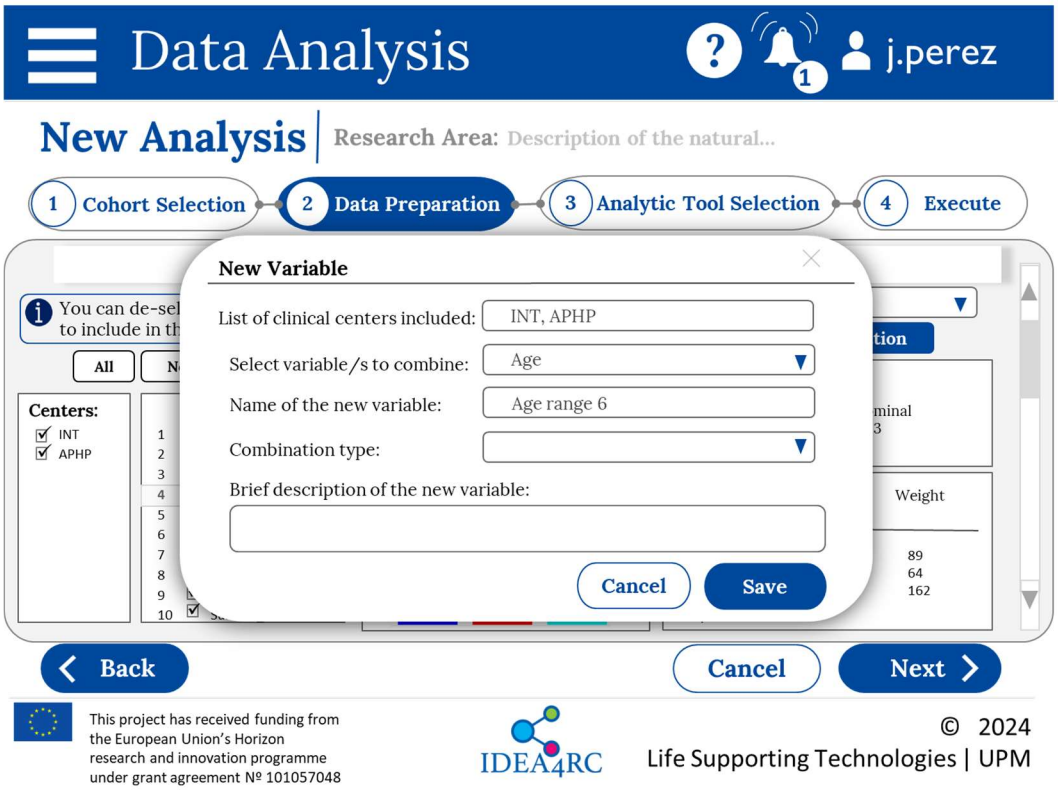

Figure 12. Data preparation - Creating new variable

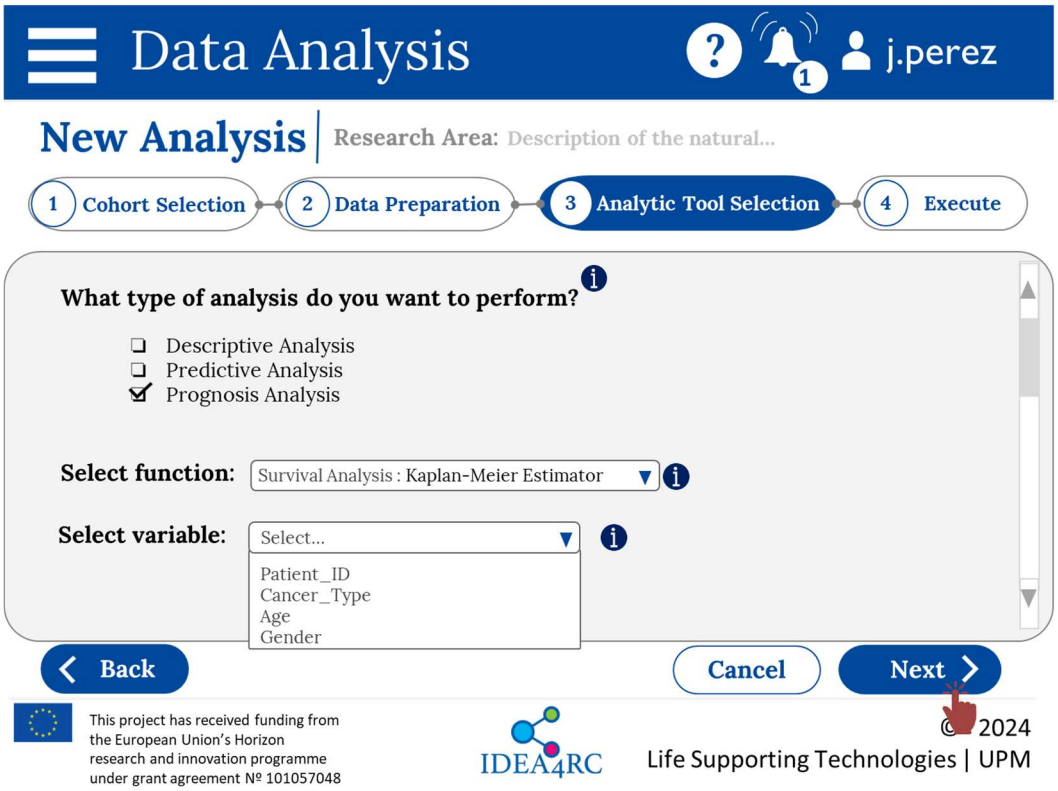

Figure 13. Selection of an analysis

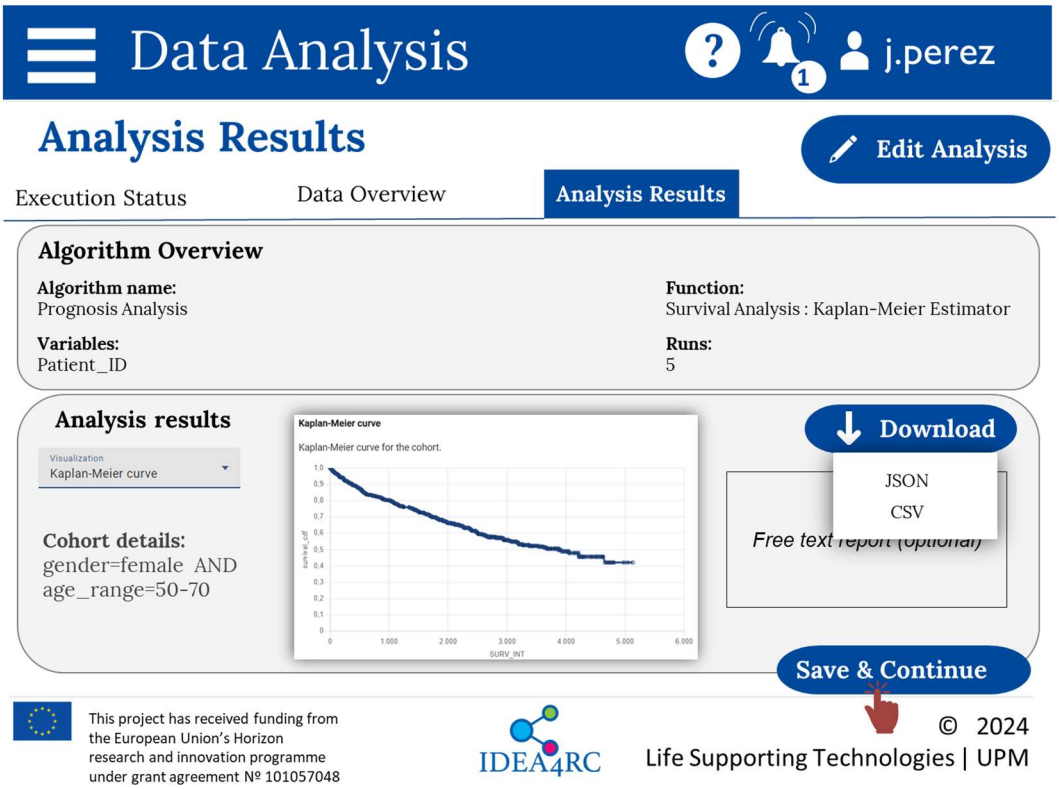

Figure 14. Analysis results

Phase 4: Result Report

In the final phase, users organize the project (based on TRIPOD guidelines[2]) or study results information and generate the data and metadata required for further validation, cataloguing, and archiving (from Figure 15 to Figure 19). The creation of the final report involves adding data and metadata or any supplementary information, such as analysis information, ethics and legal information, and quality control data, ensuring everything is included in the final output.

All of this information will be automatically included from the RAVEN tool. The user will have the ability to review it and manually add any information that cannot be extracted from the tool.

Finally, the results report can be downloaded and printed (Figure 20).

[2] https://www.tripod-statement.org/wp-content/uploads/2019/12/TRIPODAI_checklist.pdf

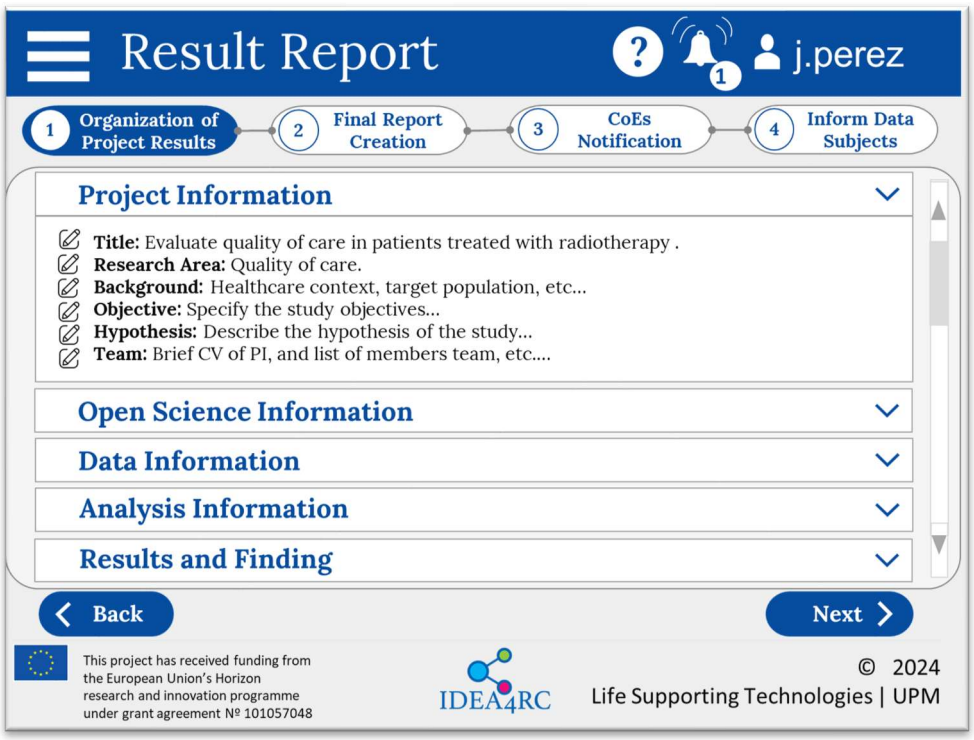

Figure 15. Organization and completion of the project results – Project information

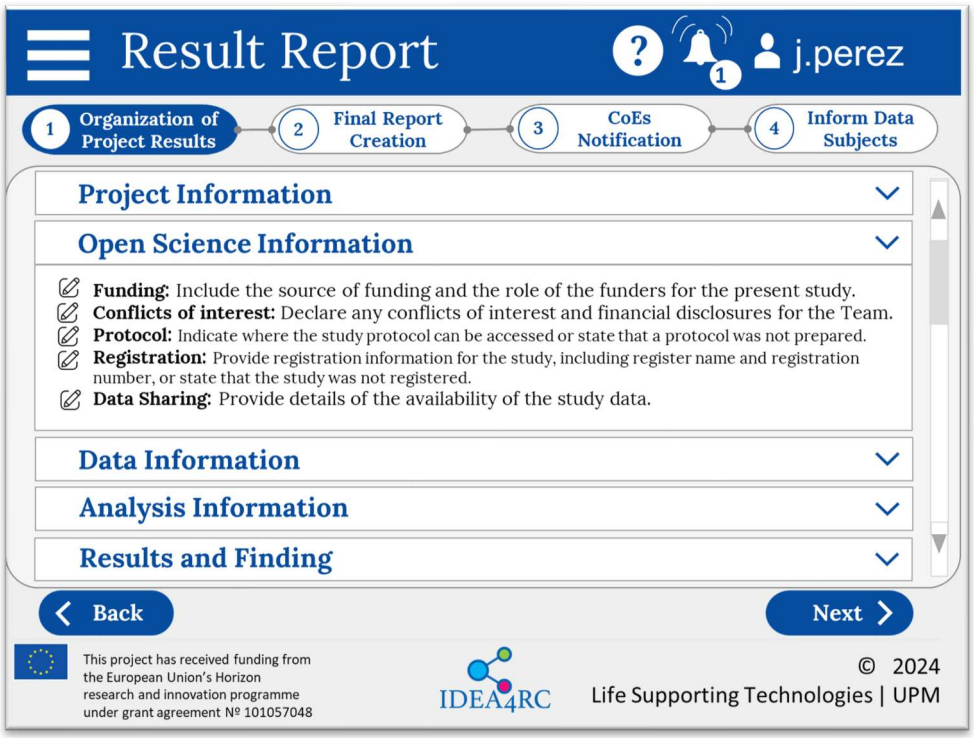

Figure 16. Organization and completion of the project results – Open science information

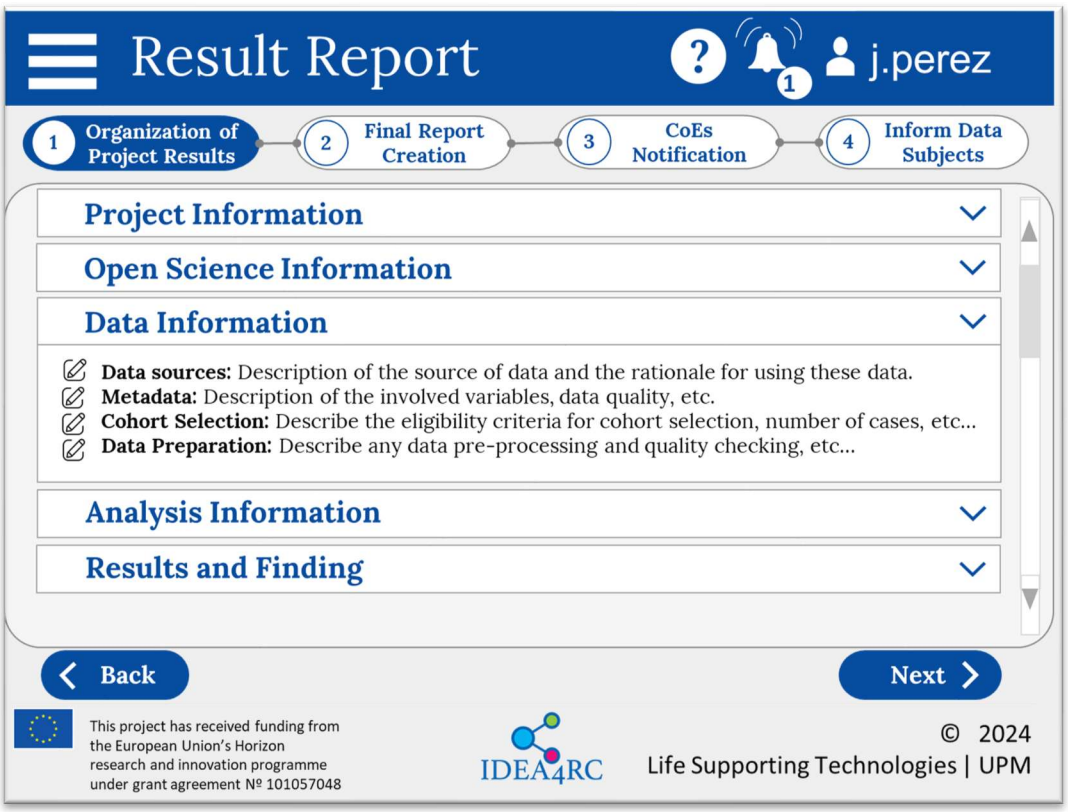

Figure 17. Organization and completion of the project results – Data information

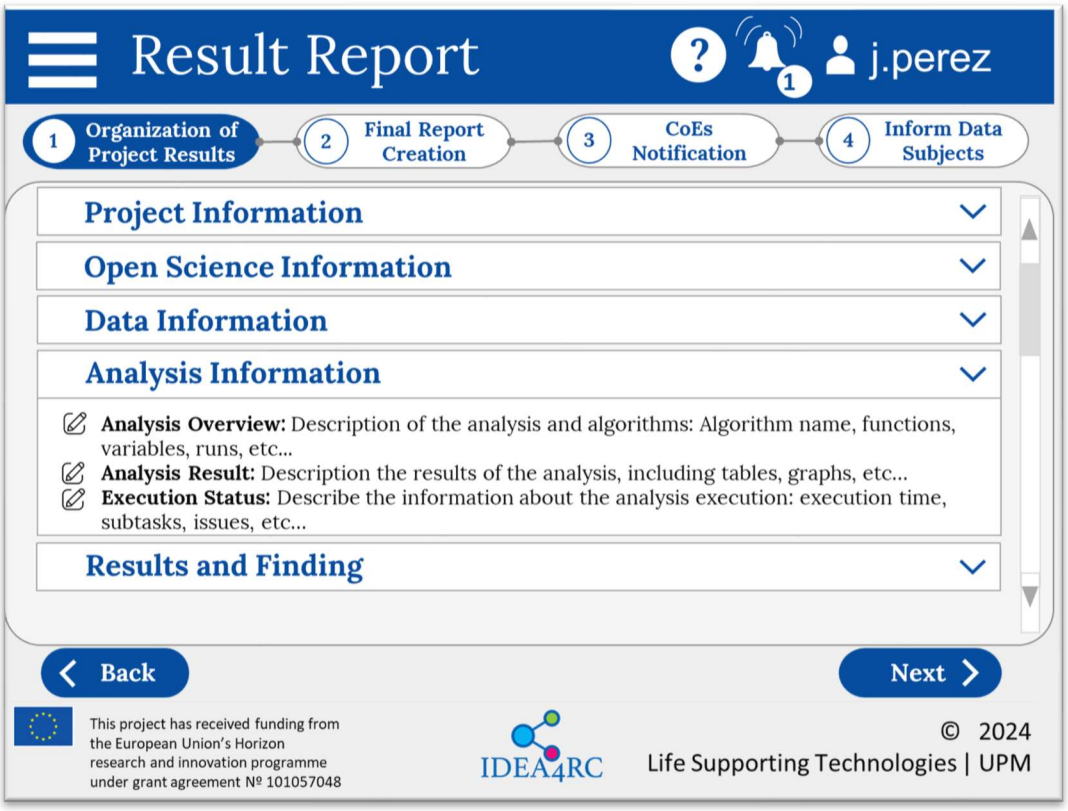

Figure 18. Organization and completion of the project results – Analysis information

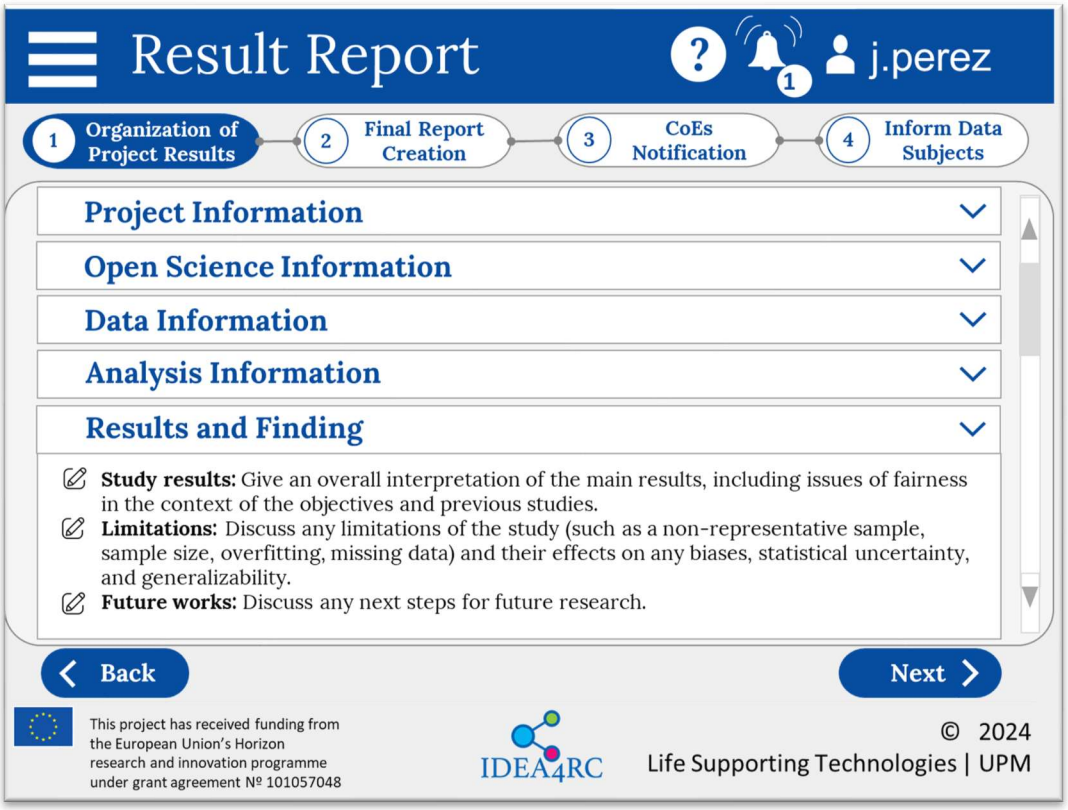

Figure 19. Organization and completion of the project results – Result and Finding

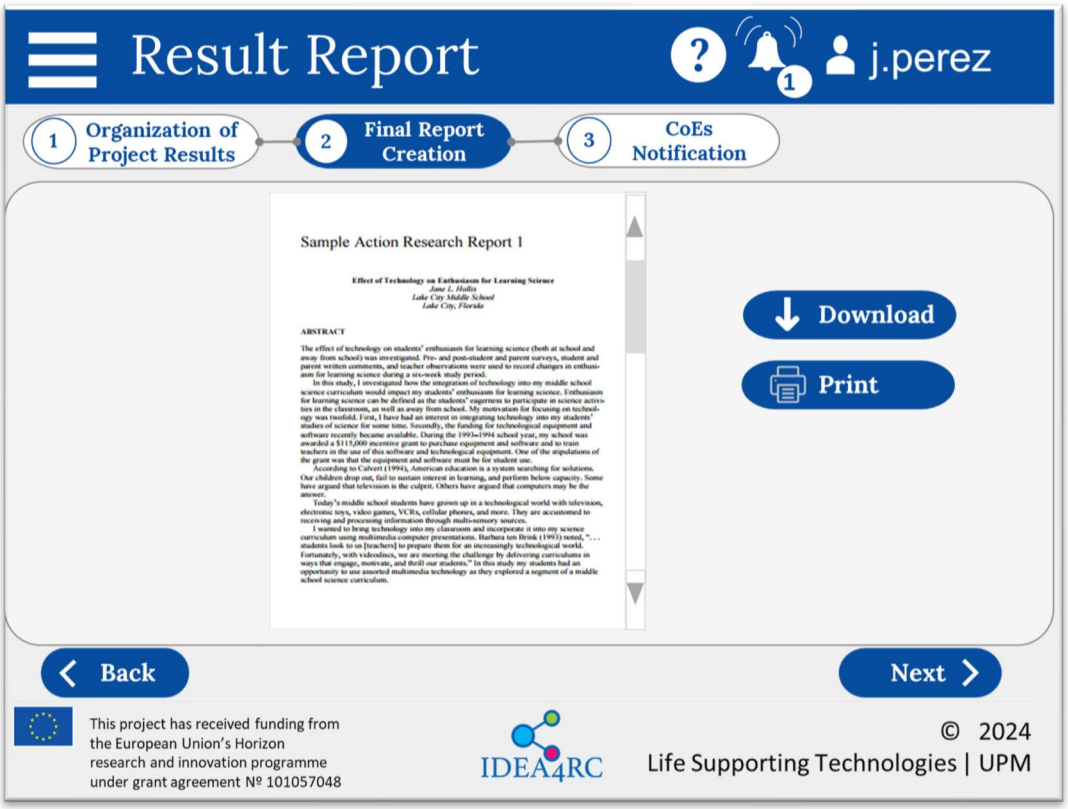

Figure 20. Final result report creation

5 CONCLUSIONS AND NEXT STEPS

The definition and design of the ***RAVEN (Rare Cancer AI Virtual Exploration Navigator) tool*** represents a significant milestone in the IDEA4RC project. By addressing the complexities inherent in dataset search and leveraging advanced AI-driven solutions, RAVEN aims to enhance the findability and reusability of data, ultimately contributing to improved research outcomes and advancements in personalized healthcare.

Throughout this deliverable, we have detailed the comprehensive approach taken to design and implement the RAVEN tool. From the initial requirement gathering and analysis phase to the meticulous conceptual design, each step was carefully planned and executed to ensure the tool meets the needs of diverse stakeholders. In particular, the low-fidelity mock-ups allowed for iterative refinement and validation with CoEs, ensuring the final design is user-centric and functionally robust.

As we move forward into Task 6.4, the focus will shift to the implementation of the RAVEN tool. This phase will encompass several key activities:

Testing and Validation:

- We will conduct usability testing, performance evaluation, and scenario-based testing to ensure that the navigator meets user expectations and project requirements.

Integration and Deployment:

- The validated design will be integrated and linked with the existing systems, ensuring seamless data flow and interoperability.
- Deployment strategies will be planned and executed in collaboration with WP8, with a focus on pilot implementations and real-world testing.

Training and Support:

- Comprehensive training sessions and support materials will be developed to facilitate user onboarding and effective utilization of the RAVEN tool.
- Ongoing support and user feedback mechanisms will be established to continually improve the system post-deployment.

Besides, AI assistant functionalities will be further explored, defined and designed, mainly In T6.2.

These activities will be crucial in ensuring that the RAVEN tool not only meets the technical specifications but also delivers a user-friendly and efficient platform for clinical and research applications. The continuous feedback loop from real-world testing and user interactions will help refine and enhance the tool, solidifying its role in advancing cancer research and improving patient outcomes.

6 ANNEXES

6.1 Focus group consent form

Virtual Assistant Focus Group (Session 1) - Consent Form

You have been asked to participate in a focus group of the Virtual Assistant (VA). The purpose of the group is to gain more insight into the perceptions and ideas of users towards this tool.

You can choose whether or not to participate in the focus group and stop at any time. Although the focus group will be video-recorded, your responses will remain anonymous and no names will be mentioned in the report. This is customary according to the GDPR privacy rules. The recordings will only be accessible to us, who are working on the project. These recordings will be used for nothing else than the mentioned purpose: Specification of augmented analytics and multimodal navigation taking in a way that it will cover your needs and expectations.

There are no right or wrong answers to the focus group questions. We want to hear many different viewpoints and would like to hear from everyone.

In respect for each other, we ask that only one participant speaks at a time in the group and that responses made by all participants be kept confidential.

* Indicates required question

1. I understand this information and agree to participate fully under the conditions stated above. *

Mark only one oval.

Mark only one oval.

- ◯ Yes *Skip to question 2*
- ◯ No

Virtual Assistant Focus Group - Participant Form

You have been asked to participate in a focus group of the Virtual Assistant (VA). The purpose of the group is to gain more insight into the perceptions and ideas of users towards this tool. With the aim of having the basic information from you, to evaluate better the results of the focus group, we ask you to fill in the following information.

2. What is your full name? *

3. Which institution are you working for? *

Mark only one oval.

- ◯ INT
- ◯ CLB
- ◯ APHA
- ◯ IIS-FJD
- ◯ VGR
- ◯ MSCI (NIO PIB)
- ◯ MUH
- ◯ OUS
- ◯ MMCI
- ◯ FPNS
- ◯ UPM
- ◯ CLN
- ◯ UDEUSTO
- ◯ ENG
- ◯ CERTH
- ◯ UU
- ◯ IKNL
- ◯ Other: _____

4. How old are you? *

Mark only one oval.

Mark only one oval.

- ◯ less than 40 years
- ◯ 40-55 years
- ◯ more than 55 years

5. What is your gender? *

Mark only one oval.

Mark only one oval.

- ◯ Female
- ◯ Male
- ◯ Prefer not to say

6. What is your specialty? *

Mark only one oval.

Mark only one oval.

- ◯ Oncologist
- ◯ Epidemiologist
- ◯ Surgeon
- ◯ Medical student
- ◯ Physician assistant (in training)
- ◯ Pathologist
- ◯ Psycologist
- ◯ Nurse
- ◯ General practicioner
- ◯ Other: _____

7. How many years of expertise in clinical practice do you have? *

Mark only one oval.

Mark only one oval.

- ◯ less than 5 years
- ◯ 5 - 10 years
- ◯ more than 10 years

8. How is your level of IT literacy? *

Mark only one oval.

Mark only one oval.

	1	2	3	4	5	
Low	◯	◯	◯	◯	◯	High

6.2 VOLERE requirement types

9. *Functional requirements*: are the fundamental or essential subject matter of the product. They describe what the product has to do or what processing actions it is to take.
Non-functional requirements: Non-functional requirements (sections 10-17) are the properties that the functions must have, such as performance and usability. Do not be deterred by the unfortunate type name (we use it because it is the most common way of referring to these types of requirements)—these requirements are as important as the functional requirements for the product's success.
10. *Look and feel requirements*:
    10a. Appearance Requirements
    10b. Style Requirements
11. *Usability and humanity requirements*:
    11a. Ease of use requirements: should cover properties such as efficiency of use, ease of remembering, error rates, overall satisfaction, feedback.
    11b. Personalization and Internationalization Requirements: should cover issues such as languages, spelling preferences and idioms; currencies including symbols and decimal conventions; and personal configuration options.
    11c. Learning Requirements: Requirements specifying how easy it should be to learn to use the product.
    11d. Understandability and Politeness Requirements: requirement for the product to be understood by its users. While "usability" refers to ease of use, efficiency, and similar characteristics, "understandability" determines whether the users instinctively know what the product will do for them and how it fits into their view of the world.
    11e. Accessibility Requirements: requirements for how easy it should be for people with common disabilities to access the product.
    11d. Convenience Requirements: requirements for things the product shall do to simplify tasks, and to expedite and make the user/customer's work easier and smoother.
12. *Performance requirements*:
    12a. Speed and Latency Requirements
    12b. Safety-Critical Requirements
    12c. Precision or Accuracy Requirements
    12d. Reliability and Availability Requirements
    12e. Robustness or Fault-Tolerance Requirements
    12f. Capacity Requirements
    12g. Scalability or Extensibility Requirements
    12h. Longevity Requirements
13. *Operational and environmental requirements*:
    13a. Expected Physical Environment
    13.b. Wider Environment Requirements
    13c. Requirements for Interfacing with Adjacent Systems
    13d. Productization Requirements
    13e. Release Requirements
    13f. Backwards Compatibility Requirements
14. Maintainability and support requirements:

- 14a. Maintenance Requirements
- 14b. Supportability Requirements
- 14c. Adaptability Requirements

15. Security requirements:
    - 15a. Access Requirements
    - 15b. Integrity Requirements
    - 15c. Privacy Requirements
    - 15d. Audit Requirements
    - 15e. Immunity Requirements
16. Cultural requirements:
    - 16a. Cultural Market Requirements
    - 16b. Cultural Diversity and Inclusion Requirements
17: Compliance requirements:
    - 17a. Legal Compliance Requirements
    - 17b. Standards Compliance Requirements

6.3 VOLERE template for requirements gathering

ID**	XXX_Rq_## XXX= Tool acronym, Rq= Requirement, ## Progressive number (e.g. PoC_Rq_01)
Description**	Short description of the current system requirement.
Requirement type**	One of the VOLERE proposed requirement types (e.g. «15. Security»),
Rationale**	The rationale behind this requirement. That is justification of the requirement
Priority**	Level of priority about the fulfillment of this requirement. The priority is a result of different contributing factors, arriving from different contexts (industrial context, business context, etc.). Possible levels: • MUST (Mandatory) • SHOULD (Of high priority) • COULD (Preferred but not necessary) • WOULD (Can be postponed and suggested for future execution)
Difficulty	Level of difficulty about the fulfillment of this requirement. Possible Levels: High, Medium, Low
Originator**	The source of this requirement.

This project has received funding from the European Union's Horizon Europe research and innovation programme under grant agreement no. 101057048

6.4 Market research – Complete results

Solution	Goal	Target Profile	Clinical Data	Technology and features	Research area
Med-PaLM	Text generation, translation, creative writing, Q&A	Patients, doctors, researchers	Medical records, test results, medical images	Transformer language model (large language model)	Any. It is not a specialized tool in this field
Flan-PaLM	Text generation, translation, creative writing, Q&A	Researchers	Scientific articles, research data	Transformer language model (large language model)	Any. It is not a specialized tool in this field
BioGPT-JSL	Text generation, translation, creative writing, Q&A	Researchers, health professionals	Scientific articles, research data, electronic medical records	Pre-trained Transformer language model with GPT-3 architecture (large language model)	Any. It is not a specialized tool in this field. Its ability to perform predictive analytics is still developing
ClinicalBERT	Text generation, translation, creative writing, Q&A	Doctors, nurses	Electronic medical records	Pre-trained Transformer language model (large language model) with BERT architecture	Any. It is not a specialized tool in this field
PubMedBERT	Text generation, translation, creative writing, Q&A	Researchers, health professionals	Scientific articles, research data, PubMed	Pre-trained Transformer language model (large language model) with BERT architecture	Descriptive analysis

This project has received funding from the European Union's Horizon Europe research and innovation programme under grant agreement no. 101057048

Solution	Goal	Target Profile	Clinical Data	Technology and features	Research area
bioconductor	Open source toolset for biological data analysis.	Researchers, bioinformaticians, data scientists.	It is not provided directly, but can be integrated with data from other sources.	R programming language Wide range of packages for cancer data analysis, including gene expression analysis, DNA methylation analysis and survival data analysis.	Descriptive analysis/Treatment effectiveness assessment
Cancer genome workbech	Visualize, analyse and explore cancer data.	Researchers, clinicians, students.	It is integrated with data from the Cancer Genome Atlas Project (TCGA).	Web platform Wide range of tools for data visualization, data analysis and data exploration.	Descriptive analysis / Prognosis and treatment response evaluation / Treatment effectiveness assessment
cbioportal	Facilitate the exploration and analysis of cancer data through an intuitive web interface.	Researchers, clinicians and patients.	Gene expression data, mutations, methylation, survival and other clinical data.	Web platform based on R Data visualization through interactive graphs. Survival analysis. Identification of relevant genes and mutations. Comparison of different types of cancer. Creating custom data sets	Descriptive analysis / Prognosis and treatment response evaluation
xena	Integrate and analyse data from multiple cancer studies.	Researchers and clinicians.	Gene expression data, mutations, methylation, survival, medical imaging and	Cloud Data visualization through heat matrices, line graphs and other tools.	Descriptive analysis / Prognosis and treatment response evaluation

Solution	Goal	Target Profile	Clinical Data	Technology and features	Research area
			other clinical data.	Analysis of correlations between different types of data. Identification of subgroups of patients with different prognoses. Exploration of biomarkers for diagnosis and treatment.	
WellBe	Comprehensive platform for cancer management	Cancer patients.	Patient health data, medical history, treatment information and other relevant data.	Mobile application and web platform. Monitoring treatment progress. Access to information about cancer and its treatment. Connection with other patients and caregivers. Emotional and psychological support.	Quality of care (WellBe focuses on improving patient care and providing emotional support, but does not perform descriptive analyzes or formal evaluations of effectiveness or quality of care)
ICGC ARGO Data Platform	Provide access to high-quality cancer genomics data globally.	Researchers, clinicians and the general public.	DNA and RNA sequencing, gene expression, methylation, proteomics and clinical data	Web platform. Search and download data. Data visualization through interactive tools. Data analysis through bioinformatics tools. Comparison of data between different types of cancer. Creating custom data sets.	Descriptive analysis

This project has received funding from the European Union's Horizon Europe research and innovation programme under grant agreement no. 101057048

Solution	Goal	Target Profile	Clinical Data	Technology and features	Research area
St. Jude Cloud19	Provide a cloud computing environment for cancer research.	Researchers	DNA and RNA sequencing, gene expression, methylation, proteomics and clinical data	Cloud platform Secure data storage. Access to high-performance computing resources. Execution of bioinformatics tools. Collaboration between researchers.	Descriptive analysis / Prognosis and treatment response evaluation
Observational Health Data Sciences and Informatics (OHDSI)	Facilitate health research using large-scale observational data.	Researchers, clinicians and epidemiologists.	Data from insurance claims, electronic medical records, biobanks and other health data.	Web platform and open source software. Development of tools and methods for the analysis of observational data. Creation of a global community of health researchers. Publication of scientific studies in high-impact journals.	Descriptive analysis / Treatment effectiveness assessment
Dataiku	Software platform for data engineering and data analysis.	Data scientists, data analysts and data engineers.	Any type of data, including clinical, genomic, imaging and other health data.	Web platform. Data preparation. Analysis of data. Data visualization. Machine learning. Artificial intelligence.	Any. It is not a specialized tool in this field but can be used for cancer research

This project has received funding from the European Union's Horizon Europe research and innovation programme under grant agreement no. 101057048

Solution	Goal	Target Profile	Clinical Data	Technology and features	Research area
Amazon web services	Cloud computing platform offering a wide range of services	Researchers, clinicians, pharmaceutical companies and start-ups	Any type of data, including clinical, genomic, imaging and other health data.	Cloud computing platform. Data storage. Cloud Computing. Analysis of data. Machine learning. Artificial intelligence.	Any. It is not a specialized tool in this field but can be used for cancer research
Google cloud platform	Cloud computing platform offering a wide range of services	Researchers, clinicians, pharmaceutical companies and start-ups	Any type of data, including clinical, genomic, imaging and other health data.	Cloud computing platform. Data storage. Cloud Computing. Analysis of data. Machine learning. Artificial intelligence.	Any. It is not a specialized tool in this field but can be used for cancer research
NLP-driven CDSS	Clinical decision support systems based on natural language processing (NLP).	Clinicians and health personnel.	Medical texts, electronic medical records, pathology reports and other health data.	Open source software and commercial platforms. Extraction of information from medical texts. Generation of alerts and clinical recommendations. Personalization of medical care.	Descriptive analysis / Prognosis and treatment response evaluation
The Rare Disease Cures Accelerator-Data and Analytics Platform	Platform for data analysis of rare diseases.	Researchers, clinicians and patients with rare diseases.	Data from different sources (e.g., research data, patient data, and clinical trial data)	Web platform Data storage. Analysis of data. Data visualization. Machine learning. Artificial intelligence.	Descriptive analysis / Prognosis and treatment response evaluation

This project has received funding from the European Union's Horizon Europe research and innovation programme under grant agreement no. 101057048

6.5 Virtual assistant requirements following Volere template

Requirements are structured following the TEDHAS phases: Discovery, Permit, Use and Finalization.

Table 10. Table of requirements following Volere template. Discovery Phase

ID	Description	Requirement type	Rationale	Priority	Difficulty	Originator
Rq_01	Search variables' information	9. Functional requirements	The system should provide a search feature to look for variables of interest.	MUST	Medium	User stories
Rq_02	Query variables' information result (meta-data)	9. Functional requirements	• Once a variable is selected or searched for, the system should display detailed meta-data information about that variable. • This information may include the definition, data type, and any specific considerations or guidelines related to the variable.	MUST	Medium	User stories
Rq_03	General data discovery questions	9. Functional requirements	Implement a set of very general questions (i.e., sarcoma or HNC) to facilitate the initial discovery of relevant data.	MUST	Low	Focus groups
Rq_04	Free text areas for detailed metadata search	9. Functional requirements	Include free-text areas in the metadata search interface to allow users to filter variables for metadata discovery.	MUST	Medium	Focus groups
Rq_05	Metadata search history	9. Functional requirements	Provide history visualization for metadata searches to track searches made over time.	SHOULD	Medium	Focus groups
Rq_06	Collaborative environment	9. Functional requirements	Facilitate knowledge sharing and collaborative search within several users. Users can share existing searches in a co-working environment, allowing others to access and leverage them.	SHOULD	Medium	Focus groups
Rq_07	Permissions management	9. Functional requirements	Allow metadata search owners to manage permissions for shared searches, specifying who can view, edit, and execute them (this requirement will be assessed for all tool functionalities).	COULD	Medium	Focus groups

This project has received funding from the European Union's Horizon Europe research and innovation programme under grant agreement no. 101057048

ID	Description	Requirement type	Rationale	Priority	Difficulty	Originator
Rq_08	Ownership transfer	9. Functional requirements	Implement the ability for search owners to transfer ownership of a search to another user (this requirement will be assessed for all tool functionalities).	SHOULD	Medium	Focus groups
Rq_09	Commenting and annotations	9. Functional requirements	Allow users to add comments and annotations, facilitating communication and providing context for collaborators (this requirement will be assessed for all tool functionalities).	COULD	Low	Focus groups
Rq_10	Notification System	9. Functional requirements	Implement a notification system to inform users when searches are shared with them, modified, or when new searches are created based on existing ones (this requirement will be assessed for all tool functionalities).	SHOULD	Medium	Focus groups
Rq_11	User-friendly interface	10. Look and feel requirements	• The interface should be designed to facilitate easy navigation and interaction for users (this requirement is transversal for all tool functionalities). • It should be intuitive, allowing users to search for variables, review information, and make decisions seamlessly.	MUST	Medium	Focus groups
Rq_12	Access Controls	15. Security requirements	Implement access controls to regulate who can view, modify, or execute shared searches (this requirement will be assessed for all tool functionalities).	MUST	Medium	User stories
Rq_13	Documentation and help	14. Maintainability and support requirements	Provide documentation to ensure effective utilization of the system, comprehensive documentation and training resources shall be provided, covering all system features (i.e., transversal requirement).	MUST	Low	User stories
Rq_14	Single Sign-On (SSO) implementation	9. Technical requirements	SSO allows users to authenticate to a single application and access multiple applications without having to authenticate again to each of them. This improves the user experience and reduces the complexity of password management (this requirement is transversal for all tool functionalities).	MUST	Low	Technical sessions

This project has received funding from the European Union's Horizon Europe research and innovation programme under grant agreement no. 101057048

Table 11. Table of requirements following Volere template. Permit Phase

ID	Description	Requirement type	Rationale	Priority	Difficulty	Originator
Rq_15	Submit data access application	9. Functional requirements	Users should be able to submit a Data Access Application for a selected cohort, providing necessary information to obtain a Data Permit for their secondary use objective.	MUST	Medium	User Stories
Rq_16	Receive result of data access application submission	9. Functional requirements	Users should receive a result after submitting a Data Access Application. The result can be an approval, a rejection with a motivation, or a request for additional information.	MUST	Medium	User Stories
Rq_17	Check status and history of submissions	9. Functional requirements	Users should have the ability to check the status and history of their submissions of Data Access Applications.	MUST	Medium	User Stories
Rq_18	Add additional information to data access application	9. Functional requirements	Users should be able to add additional information to an existing Data Access Application and resubmit it. This is done to satisfy requests from CoEs after the examination of a previously submitted application.	MUST	Medium	User Stories
Rq_19	Receive result of resubmitted data access application	9. Functional requirements	Users should receive a result after resubmitting a Data Access Application. The result can be an approval, a rejection with a motivation, or a request for additional information.	MUST	Medium	User Stories
Rq_20	Approval with Data Permit	9. Functional requirements	If the Data Access Application is approved, the system should give Data Permit to the user, granting access to the requested data for the specified secondary use objective	MUST	Medium	User Stories
Rq_21	Rejection with motivation	11. Usability and Humanity Requirements	If the request is rejected, the system provides clear feedback helping users understand the reasons for the decision and how to improve future applications.	MUST	Medium	User Stories
Rq_22	Request for additional information	9. Functional requirements	If more information is needed, the system should provide a mechanism where the user can provide the additional information to complete the evaluation process.	MUST	Medium	User Stories

This project has received funding from the European Union's Horizon Europe research and innovation programme under grant agreement no. 101057048

ID	Description	Requirement type	Rationale	Priority	Difficulty	Originator
Rq_23	User-friendly interface for application management	10. Look and Feel Requirements	Design an intuitive and user-friendly interface for managing Data Access Applications, allowing users to easily navigate, submit, and track their applications.	MUST	Medium	Focus Groups
Rq_24	Notification system for submission results	9. Functional requirements	Implement a notification system to inform users about the result of their Data Access Application submission.	MUST	Low	Focus Groups
Rq_25	Multiple notification channels	9. Non-functional requirements.	Support multiple notification channels, such as email notifications or app notifications to accommodate user preferences.	SHOULD	Medium	Focus Groups
Rq_26	Customizable notification preferences	11. Usability and Humanity Requirements	Allow users to customize their notification preferences, enabling them to choose the types of submission result notifications they wish to receive.	SHOULD	Low	Focus Groups
Rq_27	Notification frequency settings	11. Usability and Humanity Requirements	Include options for users to set the frequency of notifications, such as real-time, daily summaries, or weekly digests, based on their preferences.	SHOULD	Medium	Focus Groups
Rq_28	Notification content clarity	11. Usability and Humanity Requirements	Ensure that the content of the notifications is clear and informative, providing users with a concise summary of the result of their submission.	MUST	Low	Focus Groups
Rq_29	Notification integration with application interface	9. Functional requirements	Integrate the notification system with the application interface, allowing users to access detailed information directly from the notification.	SHOULD	Medium	Focus Groups
Rq_30	Selective application sharing	9. Functional requirements	Allow users to selectively share Data Access Applications with specific individuals or groups.	SHOULD	Medium	Focus Groups
Rq_31	Access controls for shared applications	15. Security Requirements	Implement access controls to regulate who can view, edit, or manage shared Data Access Applications, ensuring data security and privacy.	COULD	Low	Focus Groups
Rq_32	Collaborative Editing of Shared applications	9. Functional requirements	Enable collaborative editing of shared Data Access Applications, allowing multiple users to work on the same application simultaneously.	COULD	Medium	Focus Groups

This project has received funding from the European Union's Horizon Europe research and innovation programme under grant agreement no. 101057048

ID	Description	Requirement type	Rationale	Priority	Difficulty	Originator
Rq_33	Commenting and annotation on shared applications	9. Functional requirements	Allow users to add comments and annotations to shared Data Access Applications, facilitating communication and providing context for collaborators.	COULD	Low	Focus Groups
Rq_34	User-defined organization for data access applications	9. Functional requirements	Empower users to manage and organize their Data Access Applications submissions with a customizable folder structure.	SHOULD	Low	Focus Groups
Rq_35	Navigation and search	11. Usability and Humanity Requirements	Provide intuitive navigation tools and search functionality to help users quickly locate and access folders or Data Access Applications.	SHOULD	Medium	Focus Groups
Rq_36	Folder organization views	10. Look and Feel Requirements	Implement a clear and organized view to visualize the folders and the applications, such as a tree view, list view, or grid view.	COULD	Medium	Focus Groups
Rq_37	Folder Collaboration	9. Functional requirements	Allow users to share folders with other individuals or groups, facilitating collaborative organization and management of submissions.	COULD	Medium	Focus Groups
Rq_38	Folder Export	9. Functional requirements	Allow users to export or create backups of entire folders, providing a mechanism for data preservation and portability.	COULD	Medium	Focus Groups
Rq_39	Secure access for folders	15. Security requirements	Implement robust security measures to ensure that only authorized users can access, view, or manage folders, maintaining data privacy.	MUST	Medium	Focus Groups
Rq_40	Form creation interface	9. Functional requirements	Implement an intuitive interface that allows users to create Data Access Applications in a form format.	MUST	Low	Focus Groups
Rq_41	Form fields for purpose of data access	9. Functional requirements	Include free text form fields where applicants can specify the purpose of their data access, providing details on the intended use of the requested data.	MUST	Low	Focus Groups
Rq_42	Document upload feature	9. Functional requirements	Implement a feature within the data access application submission process that allows users to upload supporting documents.	MUST	Low	Focus Groups

This project has received funding from the European Union's Horizon Europe research and innovation programme under grant agreement no. 101057048

ID	Description	Requirement type	Rationale	Priority	Difficulty	Originator
Rq_43	File upload size limit	9. Non-functional requirements	Define and enforce a size limit for each uploaded document to manage storage constraints.	MUST	Medium	Focus Groups
Rq_44	Number of Documents Limit	9. Functional requirements	Set a limit on the maximum number of documents that can be uploaded for a single data access application.	MUST	Medium	Focus Groups
Rq_45	Document format support	11. Usability and Humanity Requirements	Specify and communicate the supported document formats (e.g., PDF, Word, etc.) that can be uploaded with the data access application.	MUST	Low	Focus Groups
Rq_46	Document upload validation	12. Performance Requirements	Perform validation checks on uploaded documents to ensure they meet the specified format requirements and do not exceed the size limit.	MUST	Medium	Focus Groups
Rq_47	Document preview feature	9. Functional requirements	Provide a preview option for users to review the documents they have uploaded before finalizing their data access application submission.	SHOULD	Medium	Focus Groups
Rq_48	Document removal option	9. Functional requirements	Allow users to remove or replace uploaded documents before the final submission if they wish to make changes.	SHOULD	Low	Focus Groups
Rq_49	Security and Privacy Measures	15. Security requirements	Implement robust security measures to protect sensitive information submitted in Data Access Applications and ensure compliance with privacy regulations.	MUST	High	User Stories
Rq_50	Documentation and Training	9. Functional requirements	Provide comprehensive documentation and training resources to guide users through the Data Access Application process, including submission, resubmission, and result interpretation.	MUST	Medium	User Stories
Rq_51	Application visualization for DPO	9. Functional requirements	The DPO should have the capability to visualize all Data Access Applications submitted to request data access. This includes details such as application status, applicant information, and submission history.	MUST	Medium	Focus Groups

This project has received funding from the European Union's Horizon Europe research and innovation programme under grant agreement no. 101057048

ID	Description	Requirement type	Rationale	Priority	Difficulty	Originator
Rq_52	Application management by DPO	9. Functional requirements	The DPO should be able to manage Data Access Applications, which may include reviewing, approving, rejecting, or requesting additional information from applicants.	MUST	Medium	Focus Groups
Rq_53	DPO access to application Details	13. Operational and Environmental Requirements	Ensure that the DPO has access to detailed information within each Data Access Application, allowing for thorough review and decision-making.	MUST	Medium	Focus Groups
Rq_54	Application filtering and Sorting for DPO	11. Usability and Humanity Requirements	Provide tools for the DPO to filter and sort Data Access Applications based on criteria such as status, date of submission, or applicant details, improving efficiency in application management.	SHOULD	Medium	Focus Groups
Rq_55	DPO comments and annotations	9. Functional requirements	Allow the DPO to add comments and annotations to Data Access Applications, providing a mechanism for recording insights, decisions, or additional information during the review process.	COULD	Low	Focus Groups
Rq_56	DPO dashboard for application overview	10. Look and Feel Requirements	Provide a dedicated dashboard for the DPO to have an overview of all submitted Data Access Applications, including key metrics and status summaries.	SHOULD	Medium	Focus Groups
Rq_57	Secure access for DPO	15. Security Requirements	Implement secure access controls to ensure that only authorized DPOs can visualize and manage Data Access Applications, maintaining data privacy and security.	MUST	High	Focus Groups
Rq_58	Integration with data governance processes	17. Compliance Requirements	Ensure seamless integration between the DPO's application management and the broader data governance processes, maintaining alignment with regulatory and ethical requirements.	MUST	High	Focus Groups
Rq_59	Documentation and training for DPO	14. Maintainability and Support Requirements	Provide comprehensive documentation and training resources to guide the DPO in effectively visualizing, managing, and making decisions on Data Access Applications.	MUST	Low	Focus Groups

This project has received funding from the European Union's Horizon Europe research and innovation programme under grant agreement no. 101057048

Table 12. Table of requirements following Volere template. Data use Phase

ID	Description	Requirement type	Rationale	Priority	Difficulty	Originator
Rq_60	Cohort selection interface	9. Functional requirements	Implement a user-friendly interface to select the cohort for which quality checks will be conducted.	MUST	Low	User Stories
Rq_61	Cohort search functionality	9. Functional requirements	The system provides researchers with a powerful tool to define and search for cohorts based on study-specific inclusion criteria (e.g., time period evaluation, follow-up time assessment)	MUST	Medium	User stories
Rq_62	Selection of a cohort base on query requirements	9. Functional requirements	The system should assess whether the user has all the necessary information for the cohort of interest (key variables, ensuring they meet specified criteria).	MUST	Medium	User stories
Rq_63	Query execution for cohort selection	9. Functional requirements	• The system should allow users to execute queries to select cohorts for Soft Tissue Sarcomas (STS) and Head & Neck (H&N) cancers based on specific criteria. • Users should be able to input criteria such as cancer type, treatment modalities (surgery, radiotherapy, chemotherapy), and other relevant factors.	MUST	High	D8.1
Rq_64	Cohort exploration for research questions	9. Functional requirements	• The system should facilitate the exploration of data availability for specific research questions. • Users should be able to pose questions like the number of salivary gland cancer patients treated with surgery + radiotherapy +/- chemotherapy, with information on stage, sex, age,	MUST	Medium	D8.1

ID	Description	Requirement type	Rationale	Priority	Difficulty	Originator
			comorbidities, number of recurrences, life status, and late effects.			
Rq_65	Cohort selection-decision support	14. Maintainability and Support Requirements	• The system should provide clear indications or messages to the user regarding the adequacy of information for the specified cohort, time period, and follow-up time. • Decision support may include suggestions for additional data or adjustments to criteria if needed.	COULD	Low	Focus groups
Rq_66	Query creation	9. Functional requirements	Users should be able to create, edit and execute queries to select specific healthcare data.	COULD	Medium	Focus groups
Rq_67	Query history	9. Functional requirements	Provide history visualization for queries to track creations made over time, supporting reproducibility and auditability.	SHOULD	Medium	Focus groups
Rq_68	Query validation	9. Non-functional requirements	Implement a validation mechanism to ensure that queries adhere to the FHIR and OMOP standards, preventing execution of invalid queries.	MUST	Medium	Focus groups
Rq_69	Query editor	9. Functional requirements	Facilitate knowledge sharing and collaborative search within your team. Users can share existing queries in a co-working environment, allowing others to access and leverage them.	SHOULD	Medium	Focus groups
Rq_70	Creation of new queries based on existing queries	9. Functional requirements	Users should be able to create new queries based on existing ones, leveraging shared queries as templates for further exploration or refinement.	MUST	Medium	Focus groups
Rq_71	Cohort query results display	9. Functional requirements	• The system should present query results in a clear and organized manner.	MUST	Low	Focus groups

ID	Description	Requirement type	Rationale	Priority	Difficulty	Originator
			• Information related to the specified criteria and research questions should be displayed for the selected cohorts.			
Rq_72	Cohort data availability verification	9. Functional requirements	• The system should provide a mechanism to verify the availability and quality of data for the specified criteria and research questions. • Users should be informed if certain information is missing or not available for the selected cohorts.	MUST	Medium	Focus groups
Rq_73	Data quality visualization	10. Look and feel requirements	• Incorporate visualizations to present data quality metrics in a clear and understandable manner. • Visualizations could include charts, graphs, and other visual aids to facilitate easy interpretation.	SHOULD	Medium	Focus groups
Rq_74	Advanced cohort builder	10. Look and feel requirements	• Users can refine datasets during exploration by applying filters based on various criteria, such as data quality metrics or relevance scores. This empowers them to focus on the most suitable data for their analysis. • Filters can be used to specify data cohorts based on the percentage of usable information available. This provides a clear, data-driven approach for selecting the most reliable and relevant data subsets for further exploration.	SHOULD	Low	Focus groups
Rq_75	User-friendly query interface	10. Look and feel requirements	• Design an intuitive interface that allows users to easily formulate queries, specify criteria, and interpret data quality results.	MUST	Medium	Focus groups

ID	Description	Requirement type	Rationale	Priority	Difficulty	Originator
			• The interface should support user interactions without requiring advanced technical skills.			
Rq_76	Access controls	15. Security requirements	Implement access controls to regulate who can view, modify, or execute shared queries, ensuring data security and privacy.	MUST	Medium	User stories
Rq_77	Security and privacy	15. Security requirements	Ensure that the system adheres to security and privacy standards, protecting sensitive data and maintaining confidentiality.	MUST	High	User stories
Rq_78	Quality check algorithms	9. Functional requirements	Develop algorithms to perform quality checks on the selected cohort, evaluating it against predefined study-specific quality requirements.	MUST	High	User Stories
Rq_79	Automated quality check execution	9. Functional requirements	Enable the automated execution of quality checks across the selected cohort, ensuring efficiency and consistency.	MUST	Medium	User Stories
Rq_80	Real-time quality check results	9. Functional requirements	Provide real-time feedback on the quality check results, displaying relevant information and metrics to users.	SHOULD	Low	Focus Groups
Rq_81	Detailed quality check reports	9. Functional requirements	Generate detailed reports summarizing the quality check results, including any deviations from study-specific quality requirements.	MUST	Medium	Focus Groups
Rq_82	Quality check thresholds and alerts	9. Functional requirements	Define threshold values for quality metrics and implement alerts for cases where the cohort quality falls below acceptable levels.	SHOULD	Medium	Focus Groups
Rq_83	Data visualization for quality metrics	10. Look and Feel Requirements	Implement data visualization features to represent quality metrics graphically, aiding users in interpreting and understanding the results.	SHOULD	Medium	Focus Groups

This project has received funding from the European Union's Horizon Europe research and innovation programme under grant agreement no. 101057048

ID	Description	Requirement type	Rationale	Priority	Difficulty	Originator
Rq_84	Audit trail for quality checks	9. Functional requirements	Maintain an audit trail that records details of quality checks, providing a historical record for reference and analysis.	COULD	Medium	Focus Groups
Rq_85	Documentation on quality check procedures	14. Maintainability and Support Requirements	Provide comprehensive documentation on the procedures and methodologies used for quality checks, facilitating user understanding.	MUST	Low	Focus Groups
Rq_86	Scalability planning	12. Performance Requirements	Plan for scalability to handle quality checks efficiently as the size of the selected cohorts and data sources increases.	SHOULD	Medium	Focus Groups
Rq_87	Federated query execution	9. Non-functional requirements	Develop mechanisms for executing queries across federated data sources efficiently.	MUST	High	Focus Groups
Rq_88	Data scientist collaboration	9. Functional requirements	Facilitate collaboration between data scientists within the same CoEs during the federated analysis process.	SHOULD	Medium	Focus Groups
Rq_89	Data visualization for federated analysis	9. Functional requirements	Provide data visualization tools to help users interpret and understand results from federated data analysis.	MUST	Medium	Focus Groups
Rq_90	AI pipeline support	9. Functional requirements	Provide support for AI pipelines within the MLOps workflow, assisting data scientists in building and deploying models.	MUST	High	User Stories
Rq_91	TEHDAS user journey workflow support	9. Functional requirements	Assist data scientists in navigating the TEHDAS (Technical Environment for High-dimensional and Distributed data Analysis in a Secure way) User Journey workflow for efficient analysis.	MUST	Medium	User Stories

This project has received funding from the European Union's Horizon Europe research and innovation programme under grant agreement no. 101057048

ID	Description	Requirement type	Rationale	Priority	Difficulty	Originator
Rq_92	Integration with existing tools	9. Functional requirements	Allow data scientists to seamlessly use their preferred tools for statistical analysis, such as R, MATLAB, Python, etc.	MUST	Medium	User Stories
Rq_93	Workflow automation suggestions	14. Maintainability and support requirements	Provide suggestions and automation options within the AI pipeline and TEHDAS workflow to streamline data scientist tasks.	SHOULD	High	User Stories
Rq_94	Contextual help for tool usage	14. Maintainability and support requirements	Offer contextual help and guidance within the virtual assistant for using specific analysis tools effectively.	SHOULD	Medium	User Stories
Rq_95	User authentication for personalization	11. Usability and humanity requirements	Implement user authentication for the virtual assistant to provide personalized assistance based on individual data scientist profiles or clinical profiles.	SHOULD	Medium	Focus Groups
Rq_96	Two by two table analysis	9. Functional requirements	Implement functionality to perform two by two table analysis across federated datasets.	MUST	Medium	D8.1
Rq_97	3-way contingency table analysis	9. Functional requirements	Develop functionality for analysing three-way contingency tables across federated datasets.	MUST	High	D8.1
Rq_98	Multiple functions available	9. Functional requirements	Allow, for instance: calculation of totals and percentages by row or column in contingency tables across federated datasets; Chi-Square test to evaluate the likelihood of any observed difference between sets rising by chance across federated datasets; Kaplan-Meier method to calculate observed survival and cause-specific survival rates across federated datasets; Log-Rank test to	MUST	Medium	D8.1

ID	Description	Requirement type	Rationale	Priority	Difficulty	Originator
			evaluate differences in survival curves across federated datasets...			
Rq_99	Generate new variables from single variables	9. Functional requirements	Allow users to create new variables by defining specific groups or categories based on values of a single variable.	MUST	Medium	D8.1
Rq_100	Merge existing variables	9. Functional requirements	Enable users to merge multiple existing variables to create new composite variables, such as defining multimodal treatment based on individual treatment modalities.	MUST	Medium	D8.1
Rq_101	Transformation methods	9. Functional requirements	Implement functionality to categorize continuous variables into discrete groups based on user-defined thresholds or criteria.	MUST	Medium	D8.1
Rq_102	Boolean operations	9. Functional requirements	Allow users to perform Boolean operations (e.g., AND, OR, NOT) on variables to generate new binary variables based on logical conditions.	SHOULD	Medium	D8.1
Rq_103	Preview and Validation	9. Functional requirements	Provide a preview option to visualize the results of variable transformations before applying them, along with validation checks to ensure accuracy.	SHOULD	Medium	D8.1
Rq_104	Data cleaning and pre-processing	9. Functional requirements	Perform data cleaning and pre-processing to ensure that the dataset is suitable for survival analysis, including handling missing values, outliers, and censoring.	MUST	Medium	D8.1
Rq_105	Interpretation guidelines	14. Maintainability and Support Requirements	Provide guidelines and recommendations for interpreting the results of the models, including the clinical implications of identified prognostic and predictive factors.	MUST	Low	D8.1

This project has received funding from the European Union's Horizon Europe research and innovation programme under grant agreement no. 101057048

ID	Description	Requirement type	Rationale	Priority	Difficulty	Originator
Rq_106	Outcome variable definition	9. Functional requirements	Define the outcome variable(s) of interest that represent treatment or diagnostic procedure effectiveness, ensuring they are appropriate for modelling using multilevel models.	MUST	Medium	D8.1
Rq_107	Multilevel models' parameter interpretation	9. Functional requirements	Interpret the estimated parameters of the multilevel models to assess the effectiveness of treatments or diagnostic procedures, considering their magnitude and significance.	MUST	Medium	D8.1
Rq_108	Cross-validation setup	9. Functional requirements	Configure cross-validation procedures, such as k-fold cross-validation, to evaluate the performance of analysis methods in both train and test datasets.	MUST	Medium	Focus Groups
Rq_109	Cross-validation execution	9. Functional requirements	Execute cross-validation to assess the generalizability and robustness of the analysis methods across different datasets and data partitions.	MUST	Medium	Focus Groups
Rq_110	Multilevel models' results presentation	11. Usability and humanity requirements	Present the results of the multilevel models in a clear and interpretable manner, including parameter estimates, confidence intervals, and p-values.	MUST	Medium	Focus Groups
Rq_111	Multilevel models' interpretation guidelines	14. Maintainability and support requirements	Provide guidelines and recommendations for interpreting the results of the models, including the clinical implications of treatment or procedure effectiveness.	SHOULD	Medium	Focus Groups
Rq_112	Model outcomes	9. Functional requirements	• Plot Receiver Operating Characteristic (ROC) curves to visualize the trade-off between	MUST	Medium	Focus Groups

ID	Description	Requirement type	Rationale	Priority	Difficulty	Originator
			sensitivity and specificity across different thresholds. • Calculate and visualize overall accuracy of the model predictions, providing a single metric to assess the model's performance. • Generate and display confusion matrices to visualize the performance of classification models, showing true positive, true negative, false positive, and false negative rates. • Plot Precision-Recall curves to visualize the trade-off between precision and recall across different classification thresholds. • Calculate and display the F1 score, which represents the harmonic mean of precision and recall, providing a balanced measure of model performance. • Calculate and visualize the Area Under the ROC Curve (AUC) as a single metric to quantify the overall performance of binary classification models.			
Rq_113	Interactive plots	10. Look and feel requirements	Create interactive plots to allow users to explore performance metrics dynamically, such as zooming, panning, and hovering over data points for detailed information.	SHOULD	Medium	Focus Groups
Rq_114	Heatmaps visualizations	10. Look and feel requirements	Generate heatmaps to visualize the confusion matrices, providing a clear overview of	SHOULD	Medium	Focus Groups

ID	Description	Requirement type	Rationale	Priority	Difficulty	Originator
			classification performance for different classes or categories.			
Rq_115	No limitation on analysis frequency	12. Performance requirements	Allow users to perform multiple analyses over the same data cohort without any restrictions on the frequency or number of analyses.	MUST	Low	Focus Groups
Rq_116	No limitation on concurrent analyses	12. Performance requirements	Enable users to run multiple analyses concurrently or sequentially over the same data cohort, facilitating parallel processing and workflow efficiency.	SHOULD	Medium	Focus Groups
Rq_117	Resource allocation and management	12. Performance requirements	Manage computational resources effectively to support concurrent analyses, ensuring optimal performance and preventing resource contention.	MUST	High	Focus Groups
Rq_118	Analysis queuing system	12. Performance requirements	Implement a queuing system to manage the execution order of analyses, allowing users to submit analyses sequentially and track their progress.	SHOULD	Medium	Focus Groups
Rq_119	Analysis status tracking	9. Functional requirements	Provide real-time status updates for ongoing analyses, including queued, running, and completed analyses, to keep users informed of progress.	MUST	Medium	Focus Groups
Rq_120	Persistent storage of analysis results	9. Functional requirements	Store analysis results persistently in a secure and scalable storage system, ensuring that results are accessible for future reference and analysis.	MUST	Medium	Focus Groups
Rq_121	Analysis results history	9. Functional requirements	Implement a user-friendly interface to allow users to visualize the results of previous analysis	MUST	Medium	Focus Groups
Rq_122	Shareable analysis output	9. Functional requirements	Enable users to share the results of analyses with a statistician, providing access to output files,	MUST	Medium	Focus Groups

ID	Description	Requirement type	Rationale	Priority	Difficulty	Originator
			reports, and visualizations generated during the analysis process.			
Rq_123	Analysis versioning	9. Functional requirements	Implement version control for shared analyses, allowing users to track changes, revisions, and updates made by both users and statisticians.	COULD	High	Focus Groups
Rq_124	Revision history	9. Functional requirements	Provide a detailed revision history for shared analyses, documenting all modifications, annotations, and comments made by collaborators over time.	SHOULD	Medium	Focus Groups
Rq_125	Cross-language compatibility	9. Functional requirements	Ensure compatibility and interoperability between different programming languages and software packages, allowing users to seamlessly switch between them as needed.	MUST	High	Focus Groups
Rq_126	Data exchange formats	13. Operational and environmental requirements	Support standard data exchange formats (e.g., CSV, JSON, XML) to facilitate data transfer and interoperability between different programming languages and software tools.	MUST	Medium	Focus Groups
Rq_127	Integrated development environment (IDE) support	13. Operational and environmental requirements	Provide built-in support for popular IDEs commonly used for data analysis and programming, such as Jupyter Notebooks, RStudio, Spyder, and others.	MUST	Medium	Focus Groups
Rq_128	IDE customization options	11. Usability and humanity requirements	Allow users to customize IDE settings, themes, and plugins/extensions according to their preferences and workflow requirements.	SHOULD	Low	Focus Groups
Rq_129	Version control system support	9. Functional requirements	Integrate with version control systems (e.g., Git, SVN) to enable collaborative development, code	SHOULD	High	Focus Groups

ID	Description	Requirement type	Rationale	Priority	Difficulty	Originator
			sharing, and tracking of changes across different programming languages and projects.			
Rq_130	Repository management	9. Functional requirements	Provide features for managing code repositories, including creating, cloning, branching, merging, and committing changes, within the integrated environment.	SHOULD	Medium	Focus Groups
Rq_131	User-friendly interface	10. Look and feel requirements	Implement a user-friendly interface for the overall functionality.	MUST	Low	Focus Groups
Rq_132	Visualization options	10. Look and feel requirements	Provide a variety of visualization options, such as charts, graphs, and tables, to effectively communicate the results of the analysis.	MUST	Medium	Focus Groups
Rq_133	Customization features	10. Look and feel requirements	Implement features that allow users to customize the presentation of results, including the selection of variables, colours, and formatting options.	SHOULD	Medium	Focus Groups
Rq_134	Models results presentation	10. Look and feel requirements	Present the results of analysis methods in a clear and interpretable manner, including performance metrics, variable selection, and model coefficients.	MUST	Low	Focus Groups
Rq_135	Simple graphical representation	10. Look and feel requirements	Use straightforward and familiar chart types (e.g., bar charts, line charts, scatter plots) to represent data, ensuring clarity and ease of interpretation.	MUST	Low	Focus Groups
Rq_136	Summary interpretations	11. Usability and humanity requirements	Offer concise summaries of the statistical results, highlighting key insights and trends in a manner accessible to non-experts.	MUST	Low	Focus Groups
Rq_137	Intuitive colour scheme	11. Usability and humanity requirements	Employ a colour scheme that enhances readability and avoids ambiguity, with distinct colours for different data categories or groups.	MUST	Low	Focus Groups

This project has received funding from the European Union's Horizon Europe research and innovation programme under grant agreement no. 101057048

ID	Description	Requirement type	Rationale	Priority	Difficulty	Originator
Rq_138	Data labelling	11. Usability and humanity requirements	Include clear and concise labels for data points, axes, and legends, providing context and aiding understanding.	MUST	Low	Focus Groups
Rq_139	Hover-over Information	11. Usability and humanity requirements	Provide tooltip functionality to display additional information when users hover over data points or elements, offering contextual details without cluttering the visualization.	MUST	Low	Focus Groups
Rq_140	Title and captioning	11. Usability and humanity requirements	Include descriptive titles and captions to convey the purpose and context of the visualization, guiding users in understanding its significance.	MUST	Low	Focus Groups
Rq_141	Documentation on analytical procedures	14. Maintainability and support requirements	Create comprehensive documentation on the procedures for performing analysis across federated	MUST	Low	Focus Groups
Rq_142	Documentation on variable transformation methods	14. Maintainability and support requirements	Create comprehensive documentation outlining different methods and techniques for generating new variables from existing ones.	MUST	Low	Focus Groups
Rq_143	Performance interpretation guidelines	14. Maintainability and support requirements	Provide guidelines and recommendations for interpreting performance metrics and visualizations, helping users make informed decisions based on the results.	MUST	Low	Focus Groups
Rq_145	Background information	14. Maintainability and support requirements	Include contextual information about the dataset, study design, and analytical methods used, providing necessary background for interpreting the statistical results.	MUST	Low	Focus Groups

This project has received funding from the European Union's Horizon Europe research and innovation programme under grant agreement no. 101057048

ID	Description	Requirement type	Rationale	Priority	Difficulty	Originator
Rq_146	Methodological details	14. Maintainability and support requirements	Explain the statistical methods employed, including assumptions, procedures, and limitations, to ensure transparency and comprehension.	MUST	Low	Focus Groups
Rq_147	Automated reporting	9. Functional requirements	Generate automated reports summarizing model performance metrics and visualizations for easy interpretation and sharing with stakeholders.	SHOULD	Medium	Focus Groups
Rq_148	Scalability planning	12. Performance Requirements	Plan for scalability to handle large volumes of data and increasing user demand for analytical capabilities as the user base grows.	SHOULD	Medium	Focus Groups

Table 13. Table of requirements following Volere template. Finalization Phase

ID	Description	Requirement type	Rationale	Priority	Difficulty	Originator
Rq_149	Export functionality	9. Functional requirements	Provide options for users to download the results of the analysis in various formats, including CSV, Excel, PDF, and Word documents.	MUST	Medium	User Stories
Rq_150	Data subset export	9. Functional requirements	Enable users to export specific subsets of the results or selected variables, providing flexibility in exporting only relevant information.	SHOULD	Medium	User Stories
Rq_151	Scalability planning	12. Performance Requirements	Plan for scalability to handle large volumes of exported data and increasing user demand for export functionality.	SHOULD	Medium	Focus Groups
Rq_152	Documentation on export procedures	14. Maintainability and Support Requirements	Provide comprehensive documentation on how to use the export functionality, including best practices and guidelines for exporting results.	MUST	Low	Focus Groups
Rq_153	Dataset access documentation	14. Maintainability and Support Requirements	Provide comprehensive documentation outlining the requirements and procedures for accessing the original datasets, including any necessary approvals, permissions, or agreements.	MUST	Medium	Focus Groups
Rq_154	Reproducibility guidelines	9. Functional requirements	Define guidelines and standards for ensuring the reproducibility of results obtained from accessing the original datasets, including	MUST	Medium	User Stories

This project has received funding from the European Union's Horizon Europe research and innovation programme under grant agreement no. 101057048

ID	Description	Requirement type	Rationale	Priority	Difficulty	Originator
			documentation practices, software dependencies, and analysis workflows.			
Rq_155	Publications requirements	9. Functional requirements	Allow users to specify the restrictions of the publications to be sure the results extracted meet the requirements	MUST	Medium	Focus Groups
Rq_156	Publication language preferences	9. Functional requirements	Allow users to specify language preferences for publications, ensuring that retrieved results are in languages relevant to their needs.	SHOULD	Low	Focus Groups
Rq_157	Editable output format	9. Functional requirements	Ensure that the downloaded results are in an editable format, such as Excel or Word, allowing users to make necessary adjustments and additions for publication submission.	MUST	Medium	Focus Groups
Rq_158	Data visualization export	9. Functional requirements	Allow users to export data visualizations (e.g., charts, graphs) as image files or editable formats for inclusion in publications.	SHOULD	Medium	Focus Groups
Rq_159	Export metadata	9. Functional requirements	Include metadata along with the downloaded results, providing information about the analysis parameters, dataset used, and any applied transformations or filters.	SHOULD	Medium	Focus Groups
Rq_160	Export compliance with publication guidelines	12. Performance Requirements	Ensure that the exported results comply with the formatting and style guidelines of the target publication, facilitating the submission process.	MUST	High	Focus Groups