EUCAIM
CANCER IMAGE EUROPE

Project title: European Federation for Cancer Images

Project acronym: EUCAIM

Grant Agreement: 101100633

Call identifier: DIGITAL-2022-CLOUD-AI-02

D2.3: Requirement analysis of Real World Data Holders

Partner(s): HULAFE

Author(s): Patricia Serrano Candelas (HULAFE), Carina Soler (HULAFE), Silvia Flor (HULAFE), Ana de Marco (HULAFE), Luis Martí Bonmatí (HULAFE), Javier Soto (SERMAS), Víctor Sóñora (BAHIA), Celia Martin (QUIBIM), Marta Martínez (QUIBIM), Hanna Leisz (DKFZ), Laure Saint-Aubert (MEDEX), Diepriye Charles-Davies (FPG)

Contributors: Sophia Schulze-Weddige (Charité), Ignacio Gómez-Rico (HULAFE)

Date of delivery: 23/12/2024

Version: 1

Reviewers: Ignacio Blanquer (UPV), Ana Miguel (MAT)

Document revision history

Version	Revision date	Change description	Section	Author

Table of Contents

List of abbreviations and acronyms

AI4HI: AI for Health Imaging

CDM: Common Data Model

DB: Database

DFF: Data Federation Framework

DH: Data Holder

DPMT: Data Population Monitoring Team

DPO: Data Protection Officer

DSA: Data Sharing Agreement

DTA: Data Transfer Agreement

DUs: Data Users

DW: Data Warehouse

ET: Engagement Team

FIST: Technical Implementation Support Team

FAIR: Findable, accessible, interoperable, re-usable

GDPR: General Data Protection Regulation

KPI: Key Performance Indicator

LMS: Learning Management System

PACS: Picture Archiving and Communication System

RIS: Radiology Information System

RWD: Real World Data

RWDH: Real World Data Holders

TST: Technical Support Team

TB: Terabyte

TT: Training Team

VPN: Virtual Private Network

WP: Work Package

1. Introduction

Aim and scope of the deliverable

The European Federation for Cancer Images (EUCAIM) represents a groundbreaking effort to enable the secondary use of cancer-related medical images and associated clinical data, fostering research and innovation in precision medicine. A cornerstone of this initiative is the seamless connection of Real World Data Holders (RWDHs), such as hospitals, to the EUCAIM infrastructure that ensures compliance with high standards of privacy, security, and interoperability. By leveraging their health data systems, these institutions play a crucial role in generating high-quality datasets to support the development of AI-driven tools and enhance healthcare outcomes.

In deliverable D4.2, *Final EUCAIM Operational Platform*, two main types of DH were identified based on the origin of their data: the research and innovation environment and the Real World Data (RWD) environment. This deliverable focuses on RWDHs (left side of *Figure 1*), specifically on hospitals with access to primary health data, because on the right hand side, where secondary-use data repositories already exist, DH can apply the work carried out in the technical packages to align with EUCAIM requirements and standards, recently collected in the *Final Rules for Participation report* (D4.4) and the *Minimum Data Federation and Interoperability Framework* (D5.6) among others. The open challenge that EUCAIM is trying to address is just this integration of data collected from RWD studies into a single Atlas of Cancer Images.

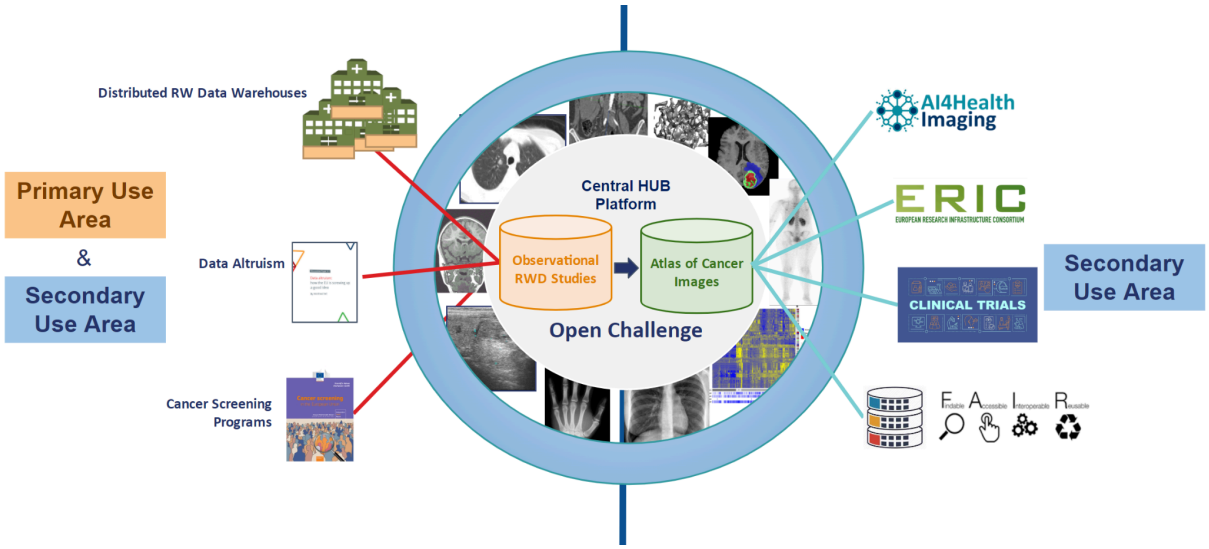

Figure 1: EUCAIM hybrid platform

Therefore, the scope of this deliverable specifically is the analysis of the clinical partners currently involved in the EUCAIM project to understand their present status and to assess their readiness to generate datasets dynamically in alignment with the requirements of the Data federation and interoperability framework defined in D4.2, D4.4 and D5.6. The insights gained will be also applicable to other RWDHs, including those participating through the project Open Calls focused on the onboarding of new cancer images to increase data incorporation with different geographic dimensions, data modalities and targets. This broader

applicability ensures that the framework developed can guide Data Holders aiming to join EUCAIM, although it will be continuously improved to meet their needs through the different WP2 support teams participating during the engagement and liaison of DH.

While significant progress has been made in defining the compliance levels of datasets once they are available for secondary use in the Atlas of Cancer Images, this deliverable addresses the critical prior step of dynamically generate datasets, which in this context applies to the capacity of hospitals to create or locate datasets tailored to the specific needs of research and innovation projects. This process requires a robust infrastructure, standardised data models, vocabularies and protocols, as well as workflows capable of aggregating, processing and de-identifying data to ensure it is compliant with EUCAIM's framework.

At the heart of this integration effort are the Data Warehouses (DWs) as the foundational systems for structuring and preparing data. These DWs serve as centralised hubs where clinical and imaging data from multiple primary health information systems, such as Picture Archiving and Communication System (PACS) and Radiology Information System (RIS), is integrated and consolidated over time, ensuring that it is organised and prepared for secondary use and data analytics.

For clinical partners to participate effectively in EUCAIM, their DWs must meet specific requirements for interoperability, data quality and documentation, among others. These prerequisites ensure that datasets are structured in a way that enables their reuse in observational studies and other research or innovation initiatives. To bring together all aspects that can help to understand how prepared these clinical partners are to facilitate the secondary data transfer or sharing within EUCAIM, in this deliverable the maturity of their DWs is assessed and based on the different aspects included in the questions described in ANNEX I. Additionally, a scoring system was developed to quantitatively assess the answers of this questionnaire and to classify the participants in different categories according to the maturity status of their DWs defined in this context.

Thus, this deliverable establishes a framework to analyse and guide the capabilities of RWDHs to dynamically generate or prepare high-quality datasets. It identifies the maturity levels of current clinical partners, highlighting areas where further support might be needed, as well as the main challenges they may find in the creation or improvement of their DW. This iterative exercise is designed to adapt to the evolving needs of DH, ensuring that all participants in EUCAIM are equipped to meet the demands of research and innovation projects. In fact, the Data User (herein referred to as End-users) requirements and their relationship with the DH are also analyzed in order to understand the user needs and be able to adapt to ongoing developments in the EUCAIM project and related upgrades of the platform.

Finally, it is worth noting that the data provision model for making new Real World Data available, in which this deliverable is focused, aligns closely with the principles outlined in the new European Health Data Space (EHDS) proposal. Under the EHDS framework, the dynamic creation of new datasets of health data by DHs is emphasized, enabling the secondary use of data. This deliverable, therefore, ensures alignment with the EHDS regulation on facilitating secure and efficient health data exchange. Therefore, Data Warehouses will be key components to deal with the obligations of the EHDS.

Relation with other deliverables and Work Packages

This deliverable is built on the analysis of the questionnaire designed to evaluate the status of hospital information systems for the secondary use of data, which initial version was presented in D2.1 (*Onboarding Invitation Package*). Furthermore, it aligns with the work carried out by various WP2 groups, as outlined in the workflow presented in Figure 2, which also references D2.4 (*Training Evaluation: Guidelines, Best Practices, Lessons Learned*). Additionally, a complementary survey to D1.4 (*Stakeholder Survey*) is proposed to identify the needs of the end users and effectively communicate these needs to the DH.

This deliverable also considers the roles defined in D4.2 (*Final EUCAIM Operational Platform*) and their interactions, highlighting actions related to RWDHs and the functional requirements involved. The technical requirements for the *Minimum Data Federation and Interoperability Framework* (min-FIF) are detailed in D5.6, which provides guidelines on the tools, services, and workflows needed for compliance across all tiers of the framework. Additionally, D4.4 (*Final Rules of Participation report*) outlines the high-level and accessible rules for DH, focusing not only on technical aspects, but also on the legal and organizational requirements they must meet.

However, the analysis presented in this deliverable prioritises the preparatory phase before creating EUCAIM-compliant datasets for secondary use. As described, its aim is to understand the current state of each hospital before becoming a node within the EUCAIM infrastructure, which was initially outlined in D5.11 (*Interim Set-Up of Local Nodes for Data Federation*).

2. Workflow in the engagement of DH

Drawing on the experience gained from large-scale research initiatives such as those conducted within the AI for Health Imaging (AI4HI) projects, a dedicated work package in EUCAIM was established to focus on the Engagement and Liaison of DH (WP2). This strategic decision reflects the need to leverage the know-how developed in prior projects to facilitate effective interaction with these critical stakeholders. WP2 is designed to coordinate and support DHs through specialised teams tasked with ensuring seamless collaboration, monitoring, and assistance throughout the entire data provision process. In the context of this Deliverable, these teams play distinct roles that collectively enable the integration of the entire workflow, as outlined in the subsequent sections. All these teams will work collaboratively to help DH overcome all the obstacles and ensure compliance with EUCAIM's Data Federation Framework (DFF). This approach not only streamlines communication but also ensures that all operational aspects are aligned with the needs and expectations of the DH, thereby enhancing their participation and contribution to EUCAIM.

Figure 2 shows the workflow of the interaction and support of the different teams to the DH. These steps, along, reflect the roadmap required for the integration of a given DH in EUCAIM.

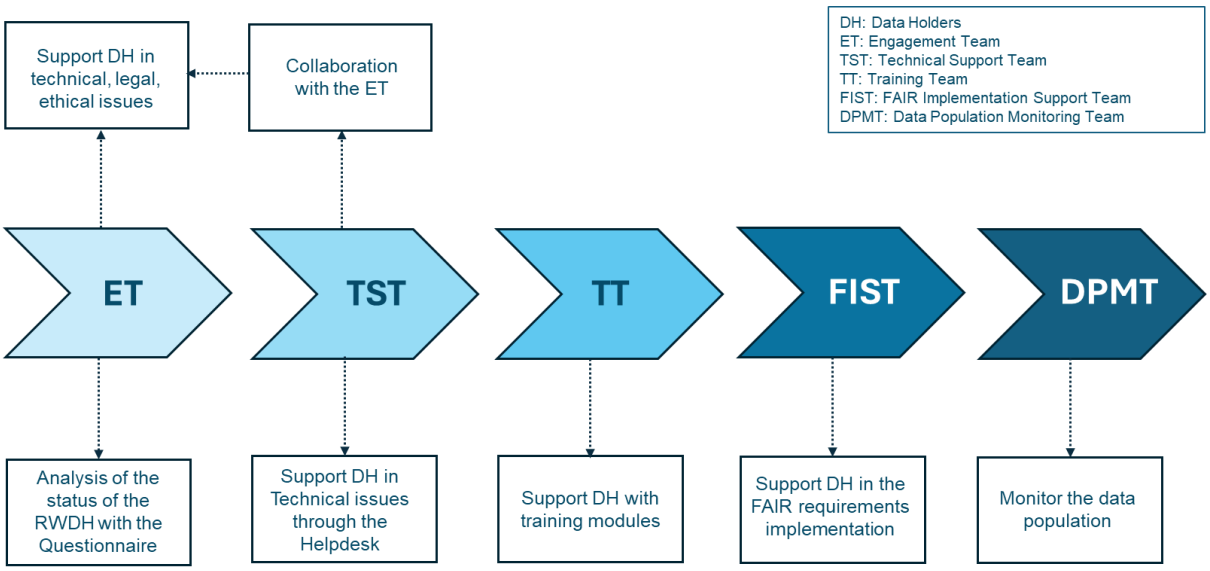

Figure 2: Workflow of the support and interaction process of the teams with the DH

2.1. Engagement Team (ET)

The objectives of the Engagement Team are, firstly, to ensure that partners that are DH populate the infrastructure and, secondly, to increase the number of DH as stakeholders.

The general responsibilities of this team are:

- Establish first contact with DH (partners and stakeholders).
- Analysis of the status of the existing health information systems for secondary use of data.
- Control the digital transformation for creating the DW and for adapting their data structure to EUCAIM's Common Data Model (CDM).
- Audit the actual data integration along the process.
- Active support in the design of the solution and intermediation in the resolution of integration problems with the organisation's platform (including legal, technical and clinical aspects) together with the Technical Support team.

The team is made up of 9 consortium members, a coordinator, 3 technicians and 4 clinicians with these specific tasks:

Coordinator:

- Coordinate the relationships with DH
- Establish contact with DH
- Manage meetings and invite participants (DH, technicians and if necessary, clinicians)
- Monitor the entire process

Technicians:

- Analyse the status of the existing health information systems and data structure with the questionnaire developed for this task

- Analyse if the hospital prefers to be a federated node or transfer data
- Define what changes the DH needs for creating their DW
- Control the creation of DW and audit the transfer
- Collaboration with Technical support team to solve specific doubts

Clinicians:

- Support the definition of what clinical data is needed. Clinicians are contacted as a last resort if there are specific doubts that the technicians cannot solve
- Solve specific doubts related to hospitals ecosystems

2.2. Technical Support Team (TST)

A Technical Support Team, composed of a group of 5-10 consortium members with technical expertise in various domains, has been created to provide technical assistance to DH. Any question or issue they may have with data provision, such as questions on the technical requirements, or an issue in setting up a federated node, may be addressed to the Technical Support Team via the EUCAIM Helpdesk ticketing system.

The Helpdesk is accessible to the public via a webform on the EUCAIM web page[1]. A demo video and written guidelines on how to submit a request using the Helpdesk ticketing system are available in the training material shown in D2.4, *Evaluation of Training*. The rules for ticket management by the Technical Support Team, as well as for any other support unit, are accessible in the EGI space[2].

2.3. Training Team (TT)

The training team is responsible for creating and delivering the training for all EUCAIM stakeholders. This team is currently implementing a Moodle Learning Management System[3] (LMS) that allows for providing user profile-specific training. Moodle is a widely used open-source LMS designed to facilitate the creation, management, and delivery of online courses. It supports a diverse range of educational and training needs across various sectors, including education, corporate training, and non-profit organizations.

For DH there are two Moodle courses offered by the training team. Both courses require registration through the Life Science Authentication and Authorization Infrastructure (LS AAI):

(1) A mandatory training course on legal and ethical aspects
(2) A general training course on platform use

Deliverable2.4, *Evaluation of Training* includes further information on these courses.

[1] Helpdesk URL: https://help.cancerimage.eu/#login
[2] Rules for ticket management: https://confluence.egi.eu/display/EUCAIM/EUCAIM+-+Helpdesk
[3] EUCAIM Moodle LMS: https://training.eucaim.cancerimage.eu/

2.4. FAIR Implementation Support Team (FIST)

The mission of this team is to educate and assist the DH in the adoption of the FAIR principles (Findability, Accessibility, Interoperability and Reusability), which are described in more detail in ANNEX 5 of deliverable D4.4. It is formed by 8 members of the technical WPs (WP4,5 and 6), including data scientists, engineers and clinical researchers.

This team has the following responsibilities:

- Design a checklist for DH to evaluate the FAIR compliance of their (meta)data and infrastructure.
- Perform a quantitative analysis of the checklist of compliance.
- Provide feedback to the DH and a list of corrective measures to be applied to continuously improve their (meta)data quality
- Hold periodic meetings to check the FAIR adoption rate (based on the Data Object Assessment metrics).
- Discuss strategies to enhance the assimilation of the recommendations by the DH based on the experience in the project EOSC-Synergy and use its developments on automated FAIR data evaluators implementing the Research Data Alliance FAIR compliance principles[4].

In the context of EUCAIM the FAIR principles refer to:

- Findability: The data that a DH exposes is findable, which is achieved through two means: The registration of the datasets in the central catalogue (tier 1), publicly exposing the metadata and the integration of the federated search (tier 2). If the DH manages a registry on its Data Warehouse, this registry should expose a DCAT-AP FAIR Data Point.
- Accessibility: The data exposed by a DH is accessible by an authorised user. This involves two main actions: datasets are registered in the negotiator service so a user can request access to them, a dataset is exposed in a Virtual Research Environment, federated processing or any other processing environment.
- Interoperability: The DH will expose the data and metadata according to the EUCAIM hyperontology schema, either directly (e.g. directly coded following this schema) or indirectly through mediator or materialisation components.
- Reusability: The DH will provide access to the data under reasonable conditions that will enable the Data Users (DU) to conduct their research properly. The conditions are expressed “a priori” on the EUCAIM Catalogue.

2.5. Data Population Monitoring Team (DPMT)

The data population monitoring team is responsible for coordinating EUCAIM’s monitoring activities related to the relevant key performance indicators (KPI) to evaluate the achievement of the objectives of the project. With respect to the integration of DH and data in the platform, the EUCAIM project has committed to reach the following KPIs as defined in the Grant Agreement:

[4] FAIR eva tool to check the FAIR requirements: https://github.com/EOSC-synergy/FAIR_eva

KPI1: Number of hospitals and imaging data repositories linked to the central hub. The project starts with 21 clinical sites from 12 countries and aims to have at least 30 distributed data providers from 15 countries by the end of the project.

KPI2: Both common (such as breast, lung, prostate, colorectal, lymphoma, multiple myeloma) and rare (e.g., ovarian, paediatric) cancers will be included with anonymized images and annotations through this pan-European Cancer Images infrastructure. More than 100,000 cases are expected to be included.

Specific KPIs to monitor the engagement of DH are listed in table 1.

Table 1: Specific KPIs related to DH

KPIs related to DH	**Description**
Number of engaged DH	Number of DH who submitted the Expression of Interest[5] via EUCAIM's webpage
Overall number of federated nodes	Number of federated nodes in total (repositories, hospitals, projects, research institutions)
Number of DH providing data to the reference nodes	Number of DH who have made data available in one of the EUCAIM reference nodes
Number of federated nodes per tier level	Number of federated nodes that comply with either tier 1, 2 or 3
Number of DH' countries of origin	Number of countries of origin of the DH engaged in EUCAIM (i.e. who have submitted the Expression of Interest)

3. Predefined requirements for Data Holders

3.1. Functional requirements

Depending on whether they already have datasets prepared or plan to make them available on demand for new observational studies proposed by Data Users, DHs will follow one of the

[5] Expression of interest for stakeholders: https://dashboard.eucaim.cancerimage.eu/expression-of-interest

data provision models defined in D4.2: the data push model or the data harvest model. These two different scenarios are envisioned to happen as follows:

- **"Data push"**: In the case of the RWDHs, if they already have datasets prepared for specific use cases, such as those proposed in the Internal and External Calls, they can register them directly in the EUCAIM Public Metadata Catalogue, which would position them as repositories within the research and innovation environment following a data push model.
- **"Data harvest"**: However, the case to be analysed especially in this deliverable is when RWDHs are contacted individually when a new proposal arises. In this case, a data harvest model approach is followed, where RWDHs do not register their entire Data Warehouse (DW) in the EUCAIM Public Metadata Catalogue. Instead, if they are chosen to participate in a new observational study, they must prepare the necessary datasets within their environment and share them with the federation, either through a federated node or by uploading them to the Reference Nodes.

Figure 3 illustrates these two scenarios for RWDHs. On the left hand side, DHs have an active role, as they will need to extract, transform and load the necessary information from primary-use hospital systems (represented in orange) and generate secondary-use datasets in their DWs (depicted in blue) on demand, which can be, for example, when a DU wishes to build an observational study with new data not collected in previously existing repositories. In this context, it is important to evaluate how prepared their Data Warehouses (DWs) or infrastructure are to generate these datasets dynamically. Once a dataset is prepared by fulfilling one of the Tiers of Compliance within the EUCAIM Data Federation Framework (indicated in the figure as EUCAIM-compliant data), they will transition to the right hand side of the figure (shown in green). This also applies to datasets that RWDHs already have available. This stage is equivalent to acting as a repository within the research and innovation environment and applies to both Federated and Reference Nodes.

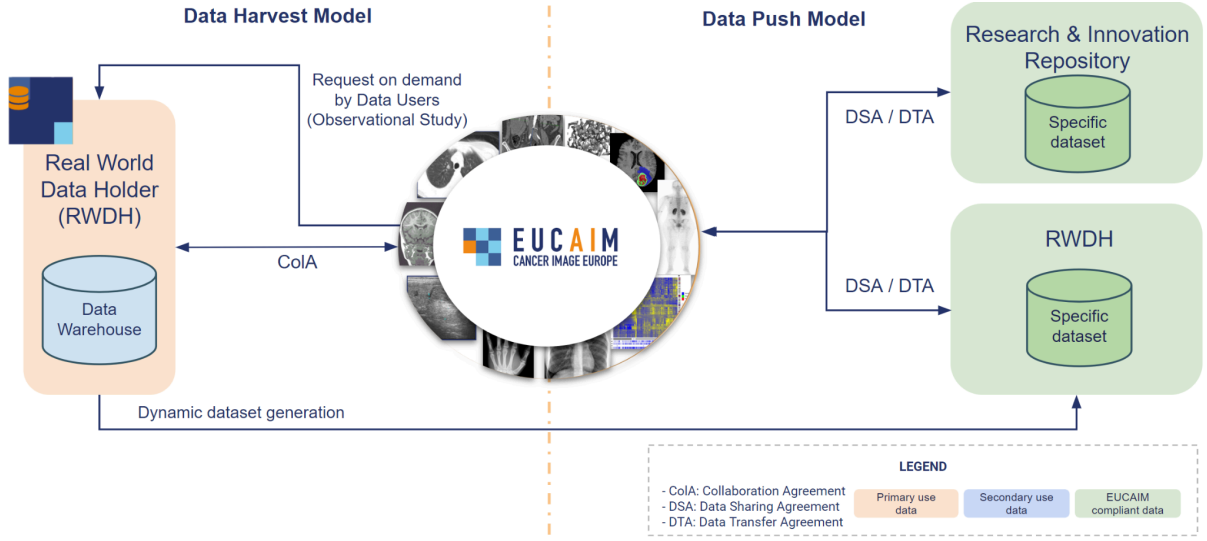

Figure 3: Data provision models and scenarios for RWDHs

Thus, a Data User can request access to an existing dataset or request the creation of a new observational study involving RWD via the Negotiator, providing the necessary information requested during the Data Access Request process (as explained in the next section of end-user requirements). When the EUCAIM Access Committee receives such a request, it can either approve or reject it depending on the alignment of the research or innovation project and the agreements already in place with the DH, including the data access conditions. If the datasets are located in federated nodes, the EUCAIM Access Committee, via the Negotiator, may forward the request to the Data Holder Access Committee if needed.

In the context of this deliverable, it is essential to analyse how the EUCAIM Access Committee interacts with RWDHs. When a new request to create an observational study is submitted, this Access Committee will check the catalogue of DHs connected to EUCAIM and will contact them. This process is currently done manually via email, but is planned to be integrated within the workflow of the Negotiator component over the next few months. The functional requirements for RWDHs in this context are as follows:

- Ability to run queries on their data and filter for specific inclusion and/or exclusion criteria specified by the DUs. The DH must provide an aggregated result with the number of available cases or expected volume and any other required information necessary to design the project.
- Providing a response via the Negotiator. Although a two-week response time for DHs would be preferable, the actual duration may vary depending on the complexity of the request and the filters to be applied. The Access Committee will evaluate the request, and within approximately one month of the initial submission, the DU will receive a report listing centers that meet the required data volume criteria and are willing to participate.
- Creating datasets if selected. If chosen to participate, the DH must generate datasets containing the committed number of cases and ensure the provision of corresponding metadata. This must align with the timelines agreed during the negotiation with the DU to ensure smooth project execution.

From then on, DHs can perform all the actions described in the User Stories defined in the Section 3 of the deliverable D4.2 (numbered as #usDH), which correspond to their functional requirements. These operational steps and workflows are expected to be performed by the Local Data Manager, an authorized technical expert or team of experts at the DH's site. The Local Data Manager is responsible for installing, configuring, operating, and maintaining local services that support the Federation, as well as managing data ingestion into Reference Nodes when applicable, as defined in the following technical requirements.

3.2. Technical requirements

The EUCAIM Data Federation Framework comprises three technical levels (tiers) that define the integration of the DH's data and services with the federation services of EUCAIM. This model of tiers is extensively described in the *Rules for Participation* (D4.4), but in a nutshell define:

- Tier 1: Compliance at the level of the metadata of the datasets and the central catalogue. DH can decide how they organize the data (e.g. by disease, purpose, etc.), being a dataset a coherent collection of data (coherent regarding format and access conditions). By being compliant to tier 1, the datasets that the DH would expose in the EUCAIM catalogue will follow the EUCAIM specification and the datasets will be registered in the central catalogue. This will be achieved by connecting the registry of the Data Warehouse, if it exists, or by requesting its registration in the central catalogue.
- Tier 2: Compliance at the level of the discoverability of the data, by integrating the data searching endpoints of the Data Warehouses with the federated search of EUCAIM. This will enable retrieving the number of subjects that fulfill the inclusion criteria defined in the searching fields. This is especially important for users to identify the best sources for a specific project, and it will facilitate the transition towards the EHDS, in which national or regional registries will be defined.
- Tier 3: Compliance at the level of the processing. By being compliant to tier 3, DH committed to expose the data fully following the EUCAIM Data Schema. DH may provide processing resources connected to the federated processing system of EUCIM or transfer the data for the processing to secure environments.

As previously mentioned, it is the responsibility of each DH to prepare both their Data Warehouse (DW) and the datasets to be shared or transferred to EUCAIM. However, they may request assistance from the WP2 teams described in the previous section if they have any questions via the Helpdesk. The technical requirements can be grouped, based on the scenarios discussed, as follows:

- **Technical requirements for adapting and integrating local Data Warehouses**

  These requirements pertain to the preparation of relevant datasets coming from the DWs of the hospitals. Since all these processes will be fully dependent on how the source data is stored and the characteristics of the DW, it will be followed by a case by case approach. In fact, Section 5 of this deliverable will analyze the current status of EUCAIM's clinical partners and recommend specific actions to improve the maturity of their DWs based on the results of the distributed questionnaire. This analysis will serve as a starting point for the Engagement Team and Technical Support Team to develop guidelines as a living document, incorporating the answers to questions raised by RWDHs in the various scenarios and settings and providing tailored assistance for each case.

  To illustrate that these requirements will depend on each case, a specific example of the HULAFE infrastructure is shown in Figure 4. This hospital has a Data Warehouse which integrates a daily copy of the data collected from various healthcare services and systems. Within the Data Warehouse, subsets of data are organized into specific Data Marts, which are smaller, subject-oriented structures designed to meet the needs of particular areas or departments, using a star schema model. Additionally, an OMOP database is used to structure clinical data and to link it to imaging data through a unique identifier (Accession Number), since imaging studies from the PACS are not directly part of the DW itself, but rather the metadata from the DICOM headers. In this way, they download the DICOM studies under the approval of

research or innovation projects and generate the datasets for each one, enriching them with clinical data and annotations if appropriate and meeting the minimum requirements in terms of FAIR compliance, de-identification and quality control.

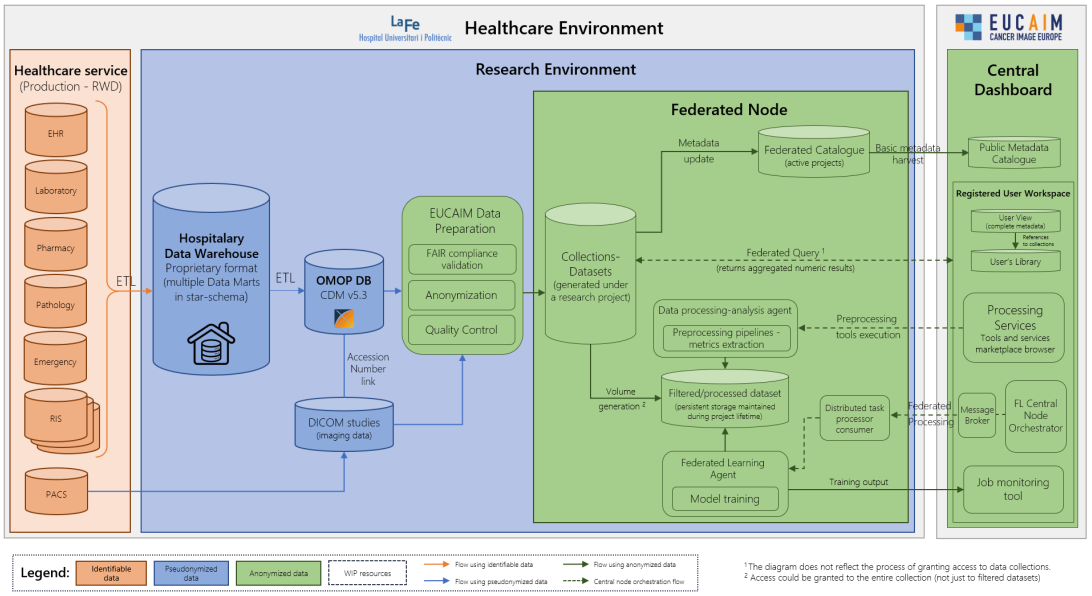

Figure 4: HULAFE DW infrastructure in the context of the use case of data sharing with EUCAIM

- **Technical requirements for sharing or transferring data from a local node**

Once the DHs have the data they want to share or transfer located and the datasets prepared, they may choose to deploy a Federated Node at their facility for federated querying and/or processing or upload their data to the Reference Nodes. In both cases, modalities of data sharing will depend on which Tier (1, 2, or 3) the datasets comply with, according to the level of compliance with the EUCAIM Data Federation Framework. Different technical requirements apply to the different tiers, both in terms of hardware and software. It should be noted that DHs choosing to upload their data to the Reference Nodes will have very little requirements to comply with, while the deployment of Federated Nodes is expected to be more resource-intensive. All requirements defined so far in the different technical work packages have been summarised by the Technical Support Team in the *ad-hoc* documentation dedicated to DHs[6], which will be accessible to them through the Dashboard once the deliverables due M24 are considered final.

3.3. Legal and ethical requirements

DHs contributing with data to the EUCAIM project are required to comply with a set of legal and ethical standards to ensure alignment with the GDPR and with EUCAIM's data governance framework. These requirements apply to all DHs, as defined by the forthcoming

[6] Technical requirements document: Technical_requirements_Data_Holders_internal_v3

European Health Data Space (EHDS) regulation, and vary slightly based on the selected data contribution model: Reference Nodes (for which it will be needed a Data Transfer Agreement - DTA) or Federated Node (for which it will needed a Data Sharing Agreement - DSA). However, before these documents are signed there is a whole process before that needs to be completed.

The approval process for DHs involves two phases: an initial legal and ethical assessment, followed by pre-transfer compliance checks. The first phase requires DHs to submit documentation to demonstrate GDPR compliance and ethical adherence. This includes a self-declaration from the Data Protection Officer (DPO), certifying that the data meets GDPR standards and outlining any legal limitations on data use, such as restrictions related to intellectual property rights or conditions established by the data subject's consent. DHs achieving Tier 3 compliance, specifically in federated processing scenarios, should also provide a security certification, such as ISO 27001, or at least they must provide an equivalent report from their Chief Security Officer. This report should cover key areas of data security, interoperability, and cataloguing. Additionally, DHs operating at Tier 3 compliance are required to submit a summary of their Data Protection Impact Assessment (DPIA), certified by the DPO, unless such an assessment was not legally required, in which case the DPO should confirm this in the compliance report. Ethical review documentation must also be included where required by national law. In cases where a country's legislation does not require formal ethical approval, the DPO report should state this explicitly, while those required to complete ethical self-assessment must submit relevant documentation. Certification of legal representation must also be provided, verifying the identity and authority of individuals empowered to bind the institution legally.

Once the initial review is completed and documentation approved, DHs must meet further conditions before transferring data. For data intended for centralised storage in one of the reference nodes, proof of data anonymization is mandatory. If EUCAIM provides anonymization services, a data processor contract will be signed to formalise the processing relationship.

EUCAIM's governance framework ensures data integrity and security through formal collaboration agreements. Each DH is required to sign a DSA or DTA, which formalises data handling responsibilities, usage rights, and intellectual property obligations. Additionally, EUCAIM's Access Committee oversees all data access requests to ensure they comply with the project's legal, ethical, and scientific guidelines and collaborates with the Steering Committee to grant data access only to users who meet GDPR and ethical standards.

To maintain ongoing compliance, EUCAIM conducts regular audits of DHs, verifying adherence to GDPR and other relevant regulations. DHs must participate in these audits, supplying necessary documentation on request. In the event of a data security incident, DHs are required to report the incident immediately to EUCAIM, enabling the consortium to manage risks and notify affected parties if needed.

For all data contributions, the relevant Data Sharing Agreement (DSA) or Data Transfer Agreement (DTA) must be signed to define the rights, responsibilities, and usage rights of both EUCAIM and the DH, ensuring secure data management within the consortium. For research use cases where anonymization is not feasible, pseudonymized data may be

provided, with EU and national regulatory requirements governing its use. This option requires DH to justify the need for pseudonymized data and adhere strictly to the forthcoming EHDS requirements once in effect.

DHs choosing the Reference Nodes (signing a DTA) model will upload anonymized datasets to EUCAIM. This model is advantageous for institutions that prefer simplified data management within a centrally controlled environment. Alternatively, DHs can retain control over their datasets by hosting them locally as a Federated Node signing a DSA. In this model, datasets remain on-site, but data can be accessed through EUCAIM's federated querying and processing capabilities. Federated Nodes are required to meet at least Tier 2 compliance standards within EUCAIM's Data Federation Framework, ensuring secure user authentication, data anonymization, and interoperability compliance.

For both Reference Nodes and Federated Node models, data security and interoperability standards are essential. DHs contributing as Federated Nodes must implement secure authentication, encryption, and regular auditing procedures, with Tier 2 security compliance as the minimum standard, which include medium-term duration digital certificates to connect to the federated service. Tier 3 will require user authentication compliant to the LS-AAI OpenID specification. EUCAIM's interoperability requirements further specify that contributed datasets be catalogued and structured according to EUCAIM's metadata standards to facilitate seamless integration within the federated infrastructure.

In summary, DHs contributing to EUCAIM must adhere to comprehensive legal, ethical, and technical standards that ensure data privacy, security, and ethical management, as represented in Figure 5. Through these requirements, EUCAIM promotes a responsible and compliant infrastructure, positioning as a pan-European hub for cancer research data.

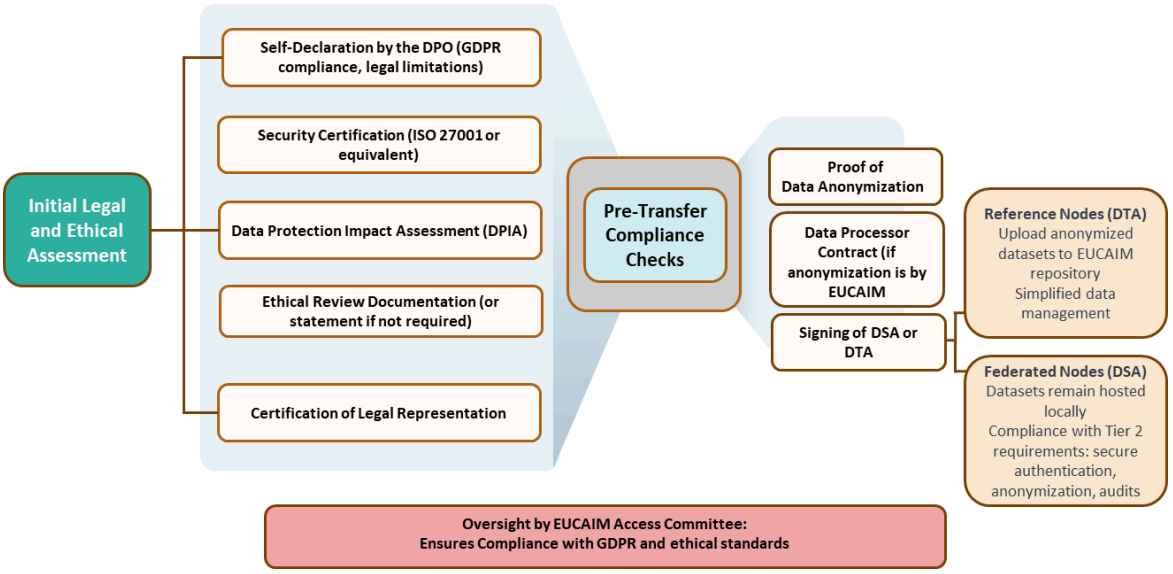

Figure 5: Legal and ethical Requirements for EUCAIM Data Contribution

4. End-user requirements

This section focuses on the end-users of the EUCAIM platform, the DUs, who have been defined as any natural or legal person who wants to make use of the data that is made accessible through EUCAIM infrastructure for research, development and innovation purposes.

The aim is to understand how they request access to data and the information and documentation they are asked to provide, so that DHs can be aware of their needs and realize whether they are able to provide data accordingly. In particular, this section has focused not only on the role of researchers, but also on the role of innovators, by proposing a new survey that could be passed to understand their requirements as well as specific documentation that may be requested during the data access process.

4.1. Data access process

As mentioned in the functional requirements of the previous section, DUs can request access to datasets either from existing repositories in the Public Catalogue or by proposing new observational studies within the RWD environment. The EUCAIM negotiator processes the requests of DUs which need to undergo the evaluation and approval of the Access Committee. Approval grants usage rights from the dataset. The main steps in the data access process involve:

1. Exploring the public catalogue: DUs start as anonymous users, reviewing the dataset's metadata to identify datasets that are aligned with their objectives.
2. Registration and authorization: DUs complete the registration process, agreeing to the Terms and Conditions and gain authenticated access to the infrastructure services.
3. Data searching: for existing data, DUs select the datasets from the public catalogue and submit detailed project documentation through the Negotiator. For new data, DUs must select the option of "Build an Observational Study" and provide the required documentation which may include agreements with DHs among others.
4. Accessing Datasets and tools: DUs can analyze the data using the processing softwares provided by EUCAIM or use the data to train or validate their own AI models.

Some fees may apply for accessing the data or services. All the process is more extensively described in Deliverable D4.2, section 5. Within the different purposes for requesting data access and software two different profiles may play a role; namely the researchers who typically conduct studies, research or analysis with the intention of generating new knowledge in the field and the innovators that typically request data to develop or validate AI algorithms with special focus on creating new products and services.

4.2. Survey to analyse the Data User requirements

Initial insights were gathered from the Stakeholders Survey conducted by EIBIR on November 17, 2023. The findings, documented in Deliverable D1.4, are based on 233 valid

responses. Among these 100 respondents, 43% were identified as potential DUs, categorized as follows: 20 innovators or health technology developers (9%), 61 medical/academic researchers (26%), and 19 AI researchers (8%). Innovators and health technology developers, representing the industry sector, demonstrated a significant interest in accessing images for research and exploitation, consistent with their commercially driven objectives. In contrast, researchers—both medical/academic and AI-focused—exhibited strong interest in academic research and collaborative opportunities.

The survey also highlighted differing priorities for the EUCAIM initiative based on stakeholder roles. Medical and academic researchers identified facilitating cross-border data sharing as the primary goal, while AI researchers emphasized the importance of advancing AI research in healthcare. Innovators and health technology developers shared this focus on advancing AI research, aligning it with their industry-oriented objectives.

In order to enhance understanding of the specific requirements of innovators, a new survey will be proposed. This survey will cover key areas, including data accessibility, usability for product development and research, and compliance with regulatory frameworks, to better align EUCAIM's offerings with DUs needs.

1. Data User General Information

This set of questions is aimed at better understanding the profile of the DUs that will be interested in using EUCAIM in the future.

- Name, contact details.
- Type of Data User: Data User-Researcher or Data User-Innovator.
- Size of the organization/group.
- Field of expertise: Collect information about the user's area of expertise (e.g., oncology, radiology, artificial intelligence, drug development, treatment planning, research on pure AI, precision medicine).
- Data modality: which types of data are most relevant to their work (e.g., imaging data, clinical data, pathology data, other).
- Intended use of the data:
    - Training/developing new software using EUCAIM infrastructure.
    - Validation/benchmarking software using EUCAIM infrastructure.
    - Regulatory software approval using EUCAIM.
    - Using built-in state of the art algorithms provided by EUCAIM for research purposes .
    - Other.

2. Data accessibility and usability

The goal of this section is to identify how users prefer to access and interact with the data.

- Preferred access mechanism to data: Federated processing and/or, centralized and/or download data.
- Annotation of data: which types of annotation would DUs like to have, and their interest on using EUCAIM to annotate.

- Interest on using pre-processing tools provided by EUCAIM: which tools would you be interested in using (automatic segmentation algorithms, harmonization algorithms, data curation, quality assessment…).
- Own resources: understand whether DU have their own resources (tools/methods..) to curate the data and would like it to be applicable within EUCAIM. Would they be interested in contributing as software providers?
- Data quality: understand which requirements do DUs have in terms of data quality
- Standards within EUCAIM: analyze the level of confidence of DUs working with data standards defined in EUCAIM (DICOM, DICOM Seg)
- Traceability: importance of having detailed information about provenance and curation process of data.

3. Technical infrastructure requirements

Based on the following questions, it can gather information on the technical capabilities of DUs and their infrastructure preferences.

- Use of EUCAIM trusted environments for training/validation of AI models: understand DU needs to run their own algorithms in EUCAIM (programming language, possibility of using their own tools, challenges of training using federated processing…).
- Ethical Compliance: Should EUCAIM provide tools or guidelines to ensure ethical use of AI models trained or validated on the platform?
- Model passport: Would a feature to trace the history of algorithmic development (e.g., versions, contributors) on EUCAIM be beneficial?
- Results: understand how DUs would like to access the result of a given training/validation.
    - Accessibility: Explore whether outputs are expected in specific formats.
    - IPR on developments done within EUCAIM.
    - Storage of results: understand whether DU need that results can be exported outside of the institution and if they have the infrastructure for storing results on premise.
- Training needs
    - Guidance on AI model training and validation.
    - Benchmarking of models/databases.
    - Regulatory compliance (e.g., MDR).

4. Privacy, security, and regulatory needs

- Understand critical concerns of DUs about privacy (e.g., GDPR compliance) and security measures.
- Explore preferences for data security: anonymization, data encryption.
- Regulatory considerations: Assess needs related to compliance with current and future regulatory frameworks, such as the EHDS or MDR.
    - Do they require specific certifications for the deployment of software on premise (e.g. CE certificates).
    - Do they require specific documentation (e.g. a Cyber Security Assurance Plan or a Data Protection Impact Assessment).

5. Platform value proposition expectations

The following questions can help to understand the DUs preferences related with the type of use cases and datasets and the interest in participation in a Research Community.

- Intended use cases: Identify which use cases they would be more interested in implementing using data from EUCAIM.
    - Anatomical region/cancer type.
    - Use case area (e.g: segmentation, diagnosis, treatment planning…).
    - Expected implementation timeline of the use case.
- Understand in which area they think EUCAIM can contribute more to their business/research and to the whole community.
- Specific interests: Interest on specific cancer types, study types (e.g., longitudinal studies), etc.
- Interest in RWD studies.
- Interest in building a community for sharing insights or best practices.

6. Project sustainability

These potential questions in the survey could provide valuable information on the importance for DUs of the sustainability of the platform and the business models that will govern the relationship between them and EUCAIM.

- Explore the value for that DU of enhanced collaboration opportunities using EUCAIM (e.g., partnerships with researchers, healthcare providers and policy-makers or regulators).
- Understand the different services that the DU would like to find available through the use of the infrastructure.
- Explore the possibility of paying for services in exchange for platform features (e.g., access to multicentric datasets to train and validate AI tools, access to state-of-the art tools available within EUCAIM and support for navigating regulatory approvals).
- Future upgrades they would like to see implemented in the future.
- How they would measure the success of the use of EUCAIM (faster model validation, regulatory approval, published research, improved AI models performance..).
- Measuring Ethical Success: Should ethical adherence (e.g., inclusivity, transparency, privacy compliance) be included as a metric for measuring EUCAIM's success?

4.3 Documentation requested to Data Users

The application documents that DUs must provide through the Negotiator when requesting access to data to create a new observational study involving Real-World Data (RWD) can be found in this link[7] accessible through the Dashboard. However, it has been detected that this documentation was very focused on **Researchers as DUs** and not on the profile of innovators in this role. Therefore, in this section the list of application documents has been adjusted to also address the needs and objectives of projects led by Innovators, as well as the specific requirements that the EUCAIM platform must fulfil for the Validation of AI-Based

[7] Documentation required by the Negotiator: Negotiator_documentation_v2-build-datasets.pdf - Google Drive

Medical Devices. It is foreseen that these requirements will be taken into account when creating a new version of the Negotiator's documents, unifying them with the information currently available.

Application Documents: Innovators as DUs

1. Project title.
2. Company details. Provide the company's name, legal entity type, company registration number and company address.
3. Authorized representative details. Include the name, position, and contact information of the person authorized to represent the company for this application.
4. Institutional Letter of Commitment (signed by the authorized representative).
5. Team composition: team members and roles in the project.
6. Brief CV of the team members.
7. Cover letter, explaining the purpose of the request, the expected benefits and contributions from the project, relevant prior experience in similar projects (maximum 500 words).
8. Project aim and objectives: Define the overall scope, the specific objectives of the project and the work plan. Include, if possible, the medical device classification in Europe and the Single Registration Number (SRN).
9. Proposed Use of Data (maximum 600 words). Provide details about:
    a. The type of data requested (imaging modalities, case report forms).
    b. Planned methodologies for data analysis or software development.
    c. Data filtering criteria, recruitment period, and anticipated number of cases. This requirement only applies to observational projects with RWDH.
    d. Use of annotations, tools, computational resources and temporary storage.
10. Expected outcomes and business applicability. Summarize the expected outcomes of the project and their relevance to the company’s objectives and business strategy. (Maximum 200 words).
11. Timeline and Milestones: Provide a project timeline, specifying key milestones and deliverables.
12. Budget.
13. Data security measures. Outline the company’s policies and procedures to ensure data security, including measures for data storage, encryption, and access control, as well as temporary storage systems and protocols for data withdrawal. *
14. Proof of compliance with relevant data protection regulations. *
15. Execution of the relevant license agreement is required to establish the terms for accessing the data or tools, and it will limit the authorized uses allowed.

* Items 13 and 14 apply only when innovators transfer the data from EUCAIM to their own systems.

EUCAIM Requirements for the Validation of AI-Based Medical Devices for Regulatory Approval

EUCAIM offers access to diverse datasets and tools, enabling the development and training of AI models within the infrastructure. Therefore, EUCAIM provides a robust infrastructure to support the validation of AI-based medical devices, ensuring compliance with European regulation for medical devices (MDR). To achieve this aim, the platform must meet the following specific requirements to ensure the validation of AI-based medical devices:

1. Imaging datasets must adhere to the DICOM format, and data transfer will operate using the DICOM protocol.
2. Inclusion of imaging acquisition parameters in the metadata of the datasets.
3. Information regarding the certifications of scanners used for image acquisition (i.e., EANM Research Ltd. [EARL] accreditation for PET/CT) must be also included, if applicable, to ensure the reliability of the imaging data used for validation.
4. Access to annotated medical imaging datasets. Software companies must use datasets with annotations that should be appropriately documented, including the metadata on image annotations, information about number of radiologists involved in the annotation process, radiologist's expertise field and years of experience, as well as any qualification in the field.
5. Support for performing subgroup analysis based on demographics, clinical variables, imaging parameters, equipment vendors.
6. EUCAIM must support the provision of bias-free, diverse datasets.
7. Enable users of the platform to swiftly enter into the necessary license agreements to access the platform's datasets and/or tools. Establish clear guidelines on the authorized uses of these resources and the intellectual property rights for any resulting innovations, which shall always remain with the user unless otherwise specified in the related legal agreements.

To comply with the EU Medical Device Regulation (MDR), EUCAIM must ensure data traceability for medical device validation. This includes enabling dataset downloads, supporting auditability throughout clinical investigation. The necessity of downloading data is supported by:

- ISO 14155:2020 "Clinical investigation of medical devices for human subjects. Good clinical practice" Clause 7.7 highlights the need for direct access to source data during and after the clinical investigation for monitoring, audits, EC review, and regulatory authority inspections.

- MDR demands strong evidence of clinical performance from the legal manufacturer. Without direct access to raw data, assessing population-specific impacts, image quality, or artifact-related performance issues is challenging. These evaluations are necessary to determine the target population and establish applicable warnings or contraindications.

- Access to clinical data facilitates post-validation improvements of the device through comparative analysis with the pre-existing cases.

In addition, EUCAIM must allow downloads of datasets in accordance with the European Data Strategy established by the European Commission, which aims to make the EU a leader in a data-driven society by creating a single market for data to allow it to flow freely within the EU and across sectors for the benefit of businesses, researchers and public administrations. EUCAIM understands that there are clashing regulations (such as the conditions of output preservations of the EHDS which requires the destruction of the data used in a period of six months and the IA Act restrictions that require data used in the training of a model to be available for auditing), and will work towards identifying the legal ground to make innovation possible. The EUCAIM platform will implement the technical means to ensure traceability and privacy preservation.

5. Status of the existing health information systems

5.1. Questionnaire used to analyse the current status of the hospitals Data Warehouses

To effectively run federated learning algorithms over a decentralised network of hospitals, substantial computing power and specialised hardware are required. DHs who agree to establish their own data node will need to procure or own the necessary infrastructure, following local procurement procedures to obtain management and technical resources. Since these requirements were not made mandatory during the course of application, the Engagement team decided to evaluate the status of the health information systems of hospitals already part of the EUCAIM consortium. To do this, the questionnaire that was included in the D2.1 *Onboarding invitation package* has been improved by the Engagement team and has been distributed among the hospitals detailed in the following section. The new version of the questionnaire has been included in ANNEX I.

The main objective of the questionnaire is to conduct a self-assessment of the current state of the hospital's Data Warehouse, to determine its preparedness and maturity to be part of a federated European data infrastructure for research. This questionnaire was to be filled by competent persons from each centre, specifically hospital staff in charge of IT (Data Scientists, Systems scientists) and Legal (Data Protection Officers (DPO) or Chief data officers).
This questionnaire sought to evaluate aspects related to data modelling, governance, processing, interoperability and accessibility of the DW. This will help determine strengths and ways to ameliorate the hospital's system to produce a more mature and reliable data infrastructure, thereby enabling its participation in the EUCAIM federation. The questionnaire covered the key areas represented in Figure 6:

Figure 6: Sections of the questionnaire used by the Engagement team to analyse the current status of the hospitals

The questionnaire consisted of both closed and open questions in relation to the different key areas in focus.

1. **Technical characteristics:** The questions within this category were asked to understand if the hospitals had a functioning DW and the type of data incorporated into the DW, if available; to know if the DW was centralised or distributed; the type of database used by the hospitals and how often data is normally incorporated into the system; if the hospitals had documented means for data extraction, maintenance, updating and backing-up the system; and if changes could be identified and tracked into source systems.
2. **Data storage and analytics:** This section covered questions related to the type of data domain (i.e., pathology, laboratory, imaging, genetics) used within the hospital's DW. In addition, the hospitals were required to provide the total data volume and expected data volume growth over the next 12 months for medical imaging and all available data marts.
3. **Standards, Common Data Models, and vocabularies:** Questions in this section were asked to know if the hospitals have a documented process for data cleaning and validation prior to its upload into the DW. Questions also addressed the use of standard models, terminologies and the format in which data was stored within the hospitals.
4. **Data accessibility:** Questions involved the management of documented data security and access control within the DW of the hospitals. These were asked to understand the mechanisms implemented to manage the authentication and authorization of data services and resources available within the hospital's DW.

5. **Data Governance:** This section was aimed at verifying if each hospital has a specific documented data governance layer and de-identified process. Each hospital was also requested to provide the name and e-mail of the person responsible for governing access permissions and enabling secure remote connections for the purposes of configuring the federated data node.
6. **IT policies:** The questions under this section addressed policies regarding the use of virtual private networks (VPNs) and personal devices to establish remote access to the organisational network and federated data node; the implementation of specialised firewall policies to safeguard the network; the restriction of network ports from external access; and if regular audits of the DW are performed in each hospital.
7. **Privacy, Security and Legal requirements:** Hospitals were required to provide details of their data protection officer (DPO). The questions were aimed at knowing if risk assessments had been performed on the DWs, and if an ethics committee approval had been obtained for the DWs and the related use of data in each centre.
8. **Hardware requirements for federated nodes:** In this section, a table outlining the indicative hardware requirements for a single node as defined in the context of the project was provided. These requirements represent the minimum hardware requirements to guarantee that the majority of EUCAIM use cases will be supported in terms of node performance, and that no critical performance bottlenecks will occur during platform operations. Each hospital was asked to provide details regarding the model and specifications of the servers that will host federated data nodes, including CPU, RAM, GPU, storage, motherboard, server provider and model.

5.2. Real World Data Holders in EUCAIM

There are 21 hospitals in the EUCAIM consortium, 4 of which are not expected to assume DHs roles, although one of them has answered the evaluation questionnaire and plans to do so in the future (Fondazione Policlinico Universitario Agostino Gemelli IRCCS (FPG)). Table 2 lists the 3 hospital partners collaborating in the consortium with roles related to the development of AI solutions that are not DH.

Table 2: Hospitals in EUCAIM that are not DH in the Grant Agreement

NUMBER OF BENEFICIARY	SHORT NAME	LEGAL NAME	COUNTRY
44	TUM-MED	KLINIKUM RECHTS DER ISAR DER TECHNISCHEN UNIVERSITAT MUNCHEN	GERMANY
6	ERASMUS MC	ERASMUS UNIVERSITAIR MEDISCH CENTRUM ROTTERDAM (University Medical Center)	NETHERLANDS
17	CF	FUNDACAO D. ANNA DE SOMMER CHAMPALIMAUD E DR. CARLOS MONTEZ CHAMPALIMAUD	PORTUGAL

The table 3 shows the **17 hospitals** that have answered the questionnaire, on whose answers the analysis performed in the following sections of this deliverable is based. Unfortunately, there was one partner, PSD (POLICLINICO SAN DONATO SPA) that did not send the questionnaire on time to conduct the analysis of its status.

Table 3: List of the partners that have answered the questionnaire

NUMBER OF BENEFICIARY	SHORT NAME	LEGAL NAME	COUNTRY
2	HULAFE	FUNDACION PARA LA INVESTIGACION DEL HOSPITAL UNIVERSITARIO LA FE DE LA COMUNIDAD VALENCIANA	SPAIN
26,1	HCB	HOSPITAL CLÍNIC DE BARCELONA (AE-FUNDACIO CLINIC PER A LA RECERCA BIOMEDICA (FCRB))	SPAIN
28	SERMAS	SERVICIO MADRILEÑO DE SALUD (Hospital Ramón y Cajal/Hospital Clínico San Carlos)	SPAIN
29	SAS	SERVICIO ANDALUZ DE SALUD (Hospital Virgen del Rocío)	SPAIN
32	SYNLAB	SYNLAB SDN SPA	ITALY
59,3	IFO	ISTITUTI FISIOTERAPICI OSPITALIERI	ITALY
22	FPG	FONDAZIONE POLICLINICO UNIVERSITARIO AGOSTINO GEMELLI IRCCS	ITALY
12,1	Neuromed	INSTITUTO NEUROLOGICO MEDITERRANEO NEUROMED SOCIETA PER AZIONI	ITALY
47	UKA	UNIVERSITAETSKLINIKUM AACHEN (University Hospital)	GERMANY
61	Charité	CHARITE - UNIVERSITAETSMEDIZIN BERLIN	GERMANY
39	NKI	STICHTING HET NEDERLANDS KANKER INSTITUUT	NETHERLANDS
42	RADBOUDUMC	STICHTING RADBOUD UNIVERSITAIR MEDISCH CENTRUM	NETHERLANDS
58	CHUP	CENTRO HOSPITALAR UNIVERSITARIO DO PORTO EPE	PORTUGAL
33	KI/KS	KAROLINSKA INSTITUTET	SWEDEN
14	APHP	ASSISTANCE PUBLIQUE HOPITAU DE PARIS	FRANCE
20	MUW	MEDIZINISCHE UNIVERSITAET WIEN	AUSTRIA
23	GOC	LINAC-PET SCAN OPCO LIMITED	CYPRUS

Figure 7 shows the number of hospitals analyzed in each country.

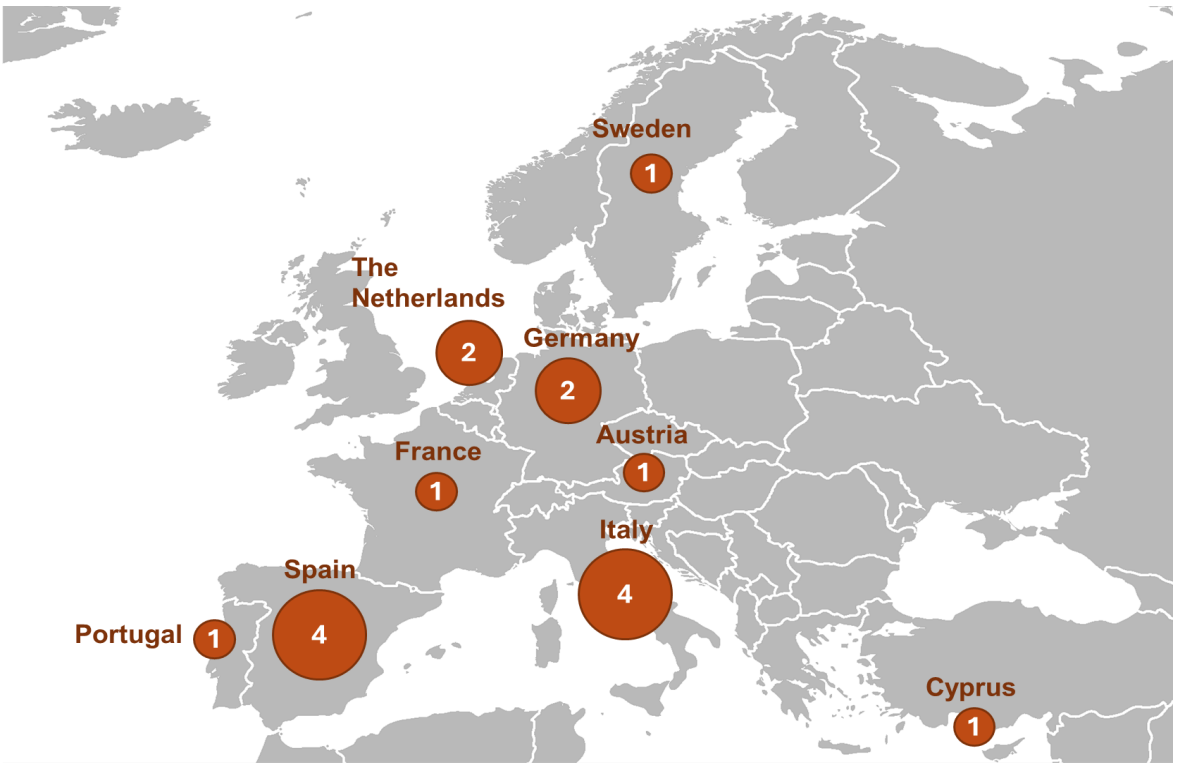

Figure 7: Distribution of analysed hospitals by country

5.3. Overview of the results

To have a better understanding regarding the current state of the hospital's Data Warehouse (DW) this section present the descriptive analysis carried out for each of the question blocks, indicating in percentages the answers of the 17 hospitals for each one of the questions (example: percentage of positive, negative or unanswered answers: NA).

5.3.1. Technical characteristics

- Use of an existing Data Warehouse (DW): **71%** of the clinical partners answered positively compared to the 29% which answered "No".
- Data types incorporated into their DW: Figure 8 shows that 76% of the clinical partners manage structured data, where each variable or measurement has its own independent field (e.g., laboratory reports); 82% handle semi-structured data, which may include some text with different variables in the same field but they are clearly identifiable (e.g., pathology reports); and 71% work with unstructured data, such as free-text reports, which lack a predefined format. All these data types are not mutually exclusive, meaning that a single DW can contain different types of data on a non-exclusive basis. Additionally, 58.8% incorporate medical images into their DW, either the studies themselves, associated reports, or the corresponding DICOM tags.

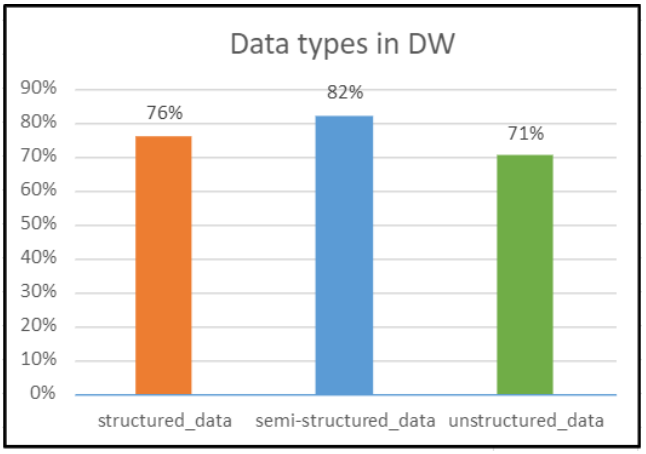

Figure 8: Representation of responses on data structuring

The terms "structured," "semi-structured," and "unstructured" have been used here to describe the data types at the attributes level based on their format and not to indicate whether the data is structured within a data model (asked later in the CDM questions). For example, data in OMOP are structured, even if one attribute of a specific entity has free-text/unstructured information (e.g. Note entity, of the note_text field).

- Data integration process of ETL documented: **70,6%** answered positively compared to the 17,6% who did not and the 11,8% did not answer.
- Database management system to support the DW: the most commonly used is **"SQL Server"**, used by 41% of the clinical partners, considering that one single

center could have more than one database management system. Also, other database management systems mentioned by the clinical partners (being the 18% represented) were: HDFS, Hbase, Impala, Hive, Blaze, Amazon S3, and Dedalus (Figure 9).

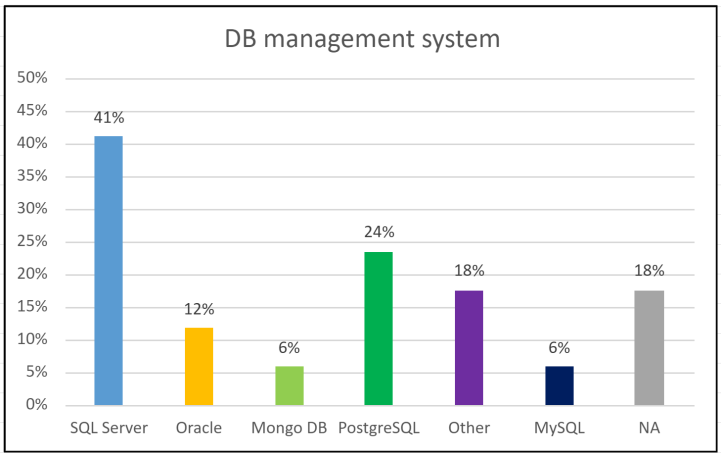

Figure 9: Representation of responses on the database management systems used to support the DW

- Type of DW: most of the clinical partners have a **centralized DW (70,6%)** rather than distributed (5,9%), the 23,5% did not answer.
- Deployment of data marts: **70,6%** deployed **on premises** rather than in the cloud (11,8%), the 17,6% did not answer.
- Frequency in which new data is incorporated into the DW: the Figure 10 represents that is mainly **daily (41%)** and the 17% representing "Other" refers to "upon request".

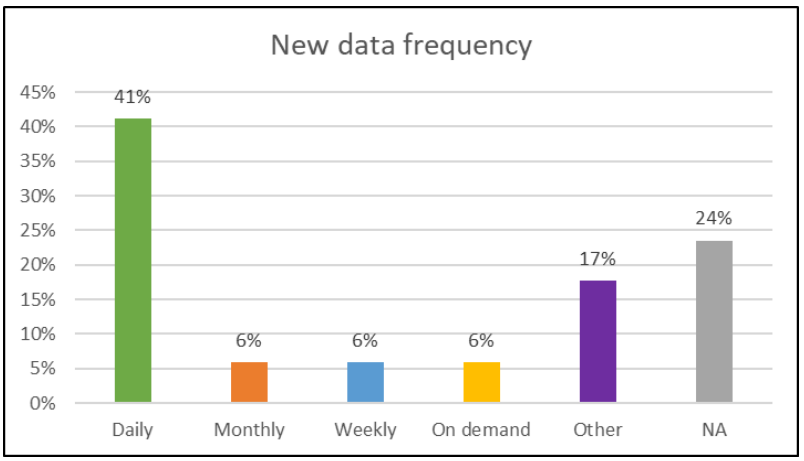

Figure 10: Representation of responses on the new data incorporation frequency

- Documenting backup policies: **47,1%** answered positively compared to the 35,3% that did not and the 17,6% left the question unanswered.

- Identifying and tracking changes into source systems: this was the most unanswered question of the section with 29,4%. **35,3%** answered positively and the other 35,3% answered "No".

5.3.2. Data storage and analytics

- Data domain covered in their DW: the most common is "**Oncology**" being the **76%** of the clinical partners covering it (Figure 11). The 29% representing "Others" implied answers such as: "Radiology workflow, Microbiology, Demography, Diagnosis, Transfers, Vital Signs, Devices, and Social History and Lifestyle Factors.

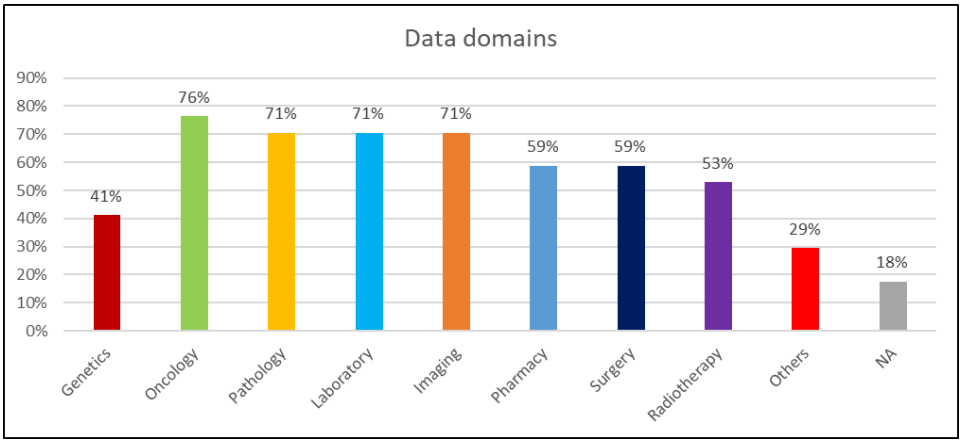

Figure 11: Representation of responses on the different domains covered in the DW

- Data mart and imaging current volume and its growth: it has received a miscellaneous response which also led to understanding the necessity of reformulating those questions to conduct their answers. Also, the percentage of data mart volume and growth questions unanswered reached 53% and 59% respectively, while the percentage of medical imaging volume and growth reached 65% of unanswered questions.

5.3.3. Standards, CDM and vocabularies

- Documenting data cleaning and validation prior loading data into DW: **64,7%** answered "Yes" compared to the 17,6% which answered "No" the 17,7% left did not answer the question.
- The most commonly used Standard Data Model was **OMOP** by 47% of the clinical partners but almost half of them left the question unanswered (Figure 12). The 24% representing "Other" implied answers such as: "OpenEHR", "i2b2", "MIMIC", and "EN/ISO13606").

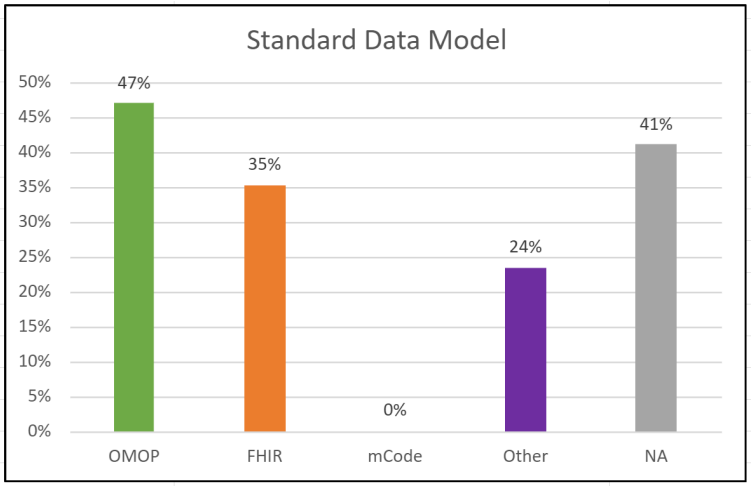

Figure 12: Representation of responses on the Standard Data Model used

- Using any standard data model: **29,4%** have a complete **documentation of the Common Data Elements and Common Data Model used**, the other 23,5% don't have it and 47,1% left the question unanswered.
- Using Standard Vocabulary: Figure 13 shows that **59%** of the clinical partners answered positively and the most common one used was the **ICD** with 47% (implying version such as ICD-9, ICD-10, ICDDIAG, ICDPROC), but almost half of them left the question unanswered.

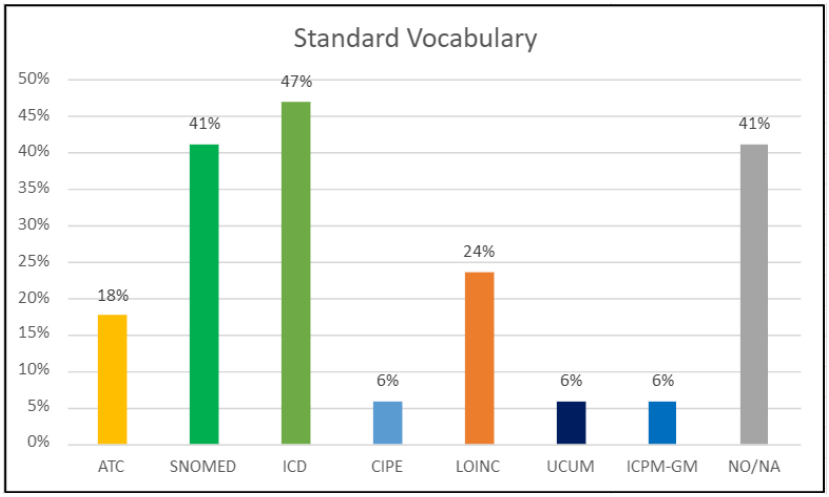

Figure 13: Representation of responses on the Standard Vocabularies used

- Storing data annotations: **64,7%** confirmed and the most commonly used was the **DICOM-SEG** with a **41%** (Figure 14). The 24% representing "Other" implied answers such as: DICOM-RTS, NRRD, ROI data, segmentations and heatmaps as either MHA (SimpleITK format) or TIFF (pathology images) and annotations such as bounding boxes, lines, polygons, etc. in JSON format.

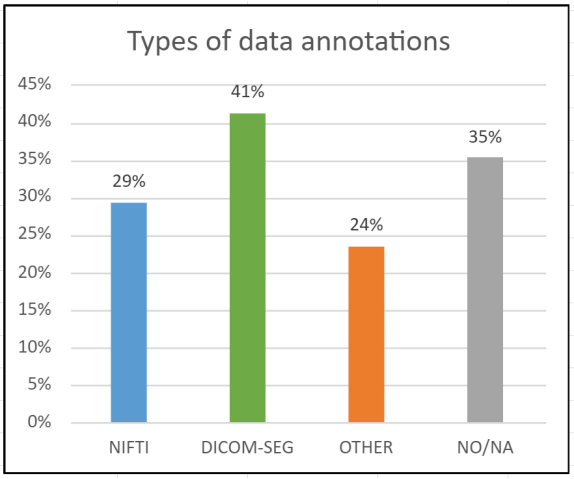

Figure 14: Representation of the types of data annotations stored

5.3.4. Data accessibility

- Documenting how data security and access control are managed within the DW: **64,7%** answered positively, 17,6% answered "No", and 17,7% did not answer.
- Data availability through APIs and metadata catalog: just **29%** answered their services are available to provide access to data within the DW compared to the **47% which did not** and the 24% left did not answer the question.
- Availability level of data accessibility: Figure 15 shows 35% during working hours, and the other 35% did not answer.

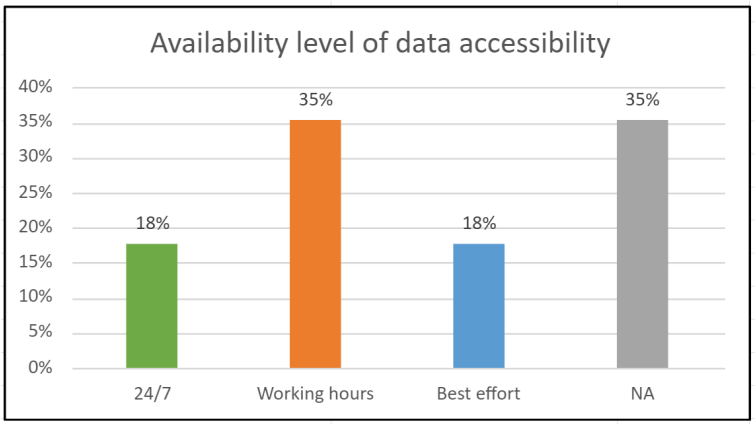

Figure 15: Representation of responses on the availability levels of data accessibility

5.3.5. Data governance

- Having a specific data governance layer: **35,3%** confirmed, 23,5% confirmed but it's not documented, 23,5% answered "No" and 17,7% did not answer.
- Document de-identification: **53%** confirmed compared to 35% which answered "No" and 12% did not answer.

5.3.6. IT policies

- Having formal policies regarding VPNs and personal devices to establish remote access, have also implemented specialized firewall policies, and have network ports restricted from external access: 94%, the 6% left, did not answer.
- Their organization ensures regular audits of the Data Warehouse: 59% answered "Yes", 18% answered "No" and 24% did not answer.

5.3.7. Privacy, security and legal

- Risk assessment: **71%** confirmed to have it compared to 24% which did not and 6% did not answer.
- Ethics committee approval: **64,7%** confirmed to have it compared to 11,7% which did not and 23,6% did not answer.

5.3.8. Hardware requirements

- Having the minimum hardware requirements: **47%** confirmed compared to 35% which did not, and 18% did not answer.

5.4. Analysis of Data Holders and Data Warehouses maturity

5.4.1. Objective

The objective of this section is to evaluate the maturity of DHs' Data Warehouses using a structured scoring system based on responses to the questionnaire presented. This evaluation identifies areas of strength and opportunities for improvement, enabling better integration into the federated data infrastructure. It is worth mentioning that this analysis does not aim to evaluate the situation of each individual DH in order to propose specific plans for each one. Rather, it seeks to draw conclusions that can be generally applicable to all DHs, or at least to several subsets of them. Additionally, it is important to note that the focus of this analysis is not on the quality of infrastructure for a specific set of datasets, but rather on supporting a DW and its integration into the EUCAIM federation.

5.4.2. Methodology

When studying the responses both individually and collectively, a scoring system has been developed to quantitatively assess the different categories of questions, which were already organized in the questionnaire. In general, the structure of the questionnaire was followed, with a few variations: the IT Policies category was omitted, and 2 questions (such as the de-identification process and cleaning and validation processes) were grouped differently for greater logical cohesion.

It was decided that, for each category, it would be best to simplify the score by minimizing the number of possible levels for each category. Additionally, comments provided by the DHs have been taken into account to better interpret some of the results and to clarify the rating that corresponds to each item.

5.4.3. Scoring system

The scoring criteria includes these categories, in a very similar structure to the questionnaire:

1. Technical Characteristics
    - 0: No DW.
    - 1: DW exists.
    - 2: DW includes medical imaging related data (images and/or DICOM metadata), relational databases, and frequent data updates (≥ weekly).
    - 3: Includes additional data types, reports, non-relational databases (e.g., HDFS), and documented ETL processes.
        - Plus: Documented maintenance, updates, and backup processes.
        - Minus: Data update frequency < daily.
2. Data Storage and Analytics
    - 0: Data mart size <20TB.
    - 1: Includes domains like Oncology, Pathology, and Imaging.
    - 2: Expanded domains (e.g., Laboratory, Surgery).
3. Standards, Common Data Models, and Vocabularies for Clinical Data and Image Metadata
    - 0: No DW or Common Data Model (CDM).
    - 1: Non-standard CDM but documented.
    - 2: Standard CDM used.
        - Plus: Use of vocabularies and annotations.
4. Data Accessibility
    - 0: No access control.
    - 1: Access control implemented.
    - 2: Access control and API availability.
        - Plus: Ensured availability.
5. Data Governance
    - 0: None.
    - 1: Governance layer.
    - 2: Governance layer and cleaning and validation processes.
        - Minus: Lack of documentation.
6. Privacy, Security, and Legal Requirements
    - 0: No DW.
    - 1: De-identification performed.
    - 2: De-identification and risk assessment performed.
7. Hardware Requirements for Federated Nodes
    - 0: Implementation not currently feasible.
    - 1: Feasible but lacking resources.
    - 2: Resources available and planned accordingly, but transition to DW not implemented yet.
    - 3: Requirements are met.

It should be noted that, for each category, a higher rating assumes the requirements of the previous levels are met.

5.4.4. Results

DW capabilities vary widely among hospitals, with some excelling in technical characteristics and data accessibility, while others lack basic infrastructure. However, some patterns were observed, suggesting varying degrees of progress for different subsets of DHs.

Some limitations of the current analysis will be addressed in a subsequent iteration, progressing in at least two complementary directions. On the one hand, it was identified that certain questions caused some initial confusion among the DHs, leading to different interpretations. The Engagement Team (ET) will rework some of the questions and update the self-assessment questionnaire accordingly. On the other hand, as part of an ongoing effort, a more individualized complementary analysis is being conducted to better understand the motivations and specificities behind some of the responses.

Heatmap and analysis

In order to help understand the big picture of the DH situation in relation to their ability to support or create a DW, a heatmap has been created for visual representation of the overall results (Figure 16).

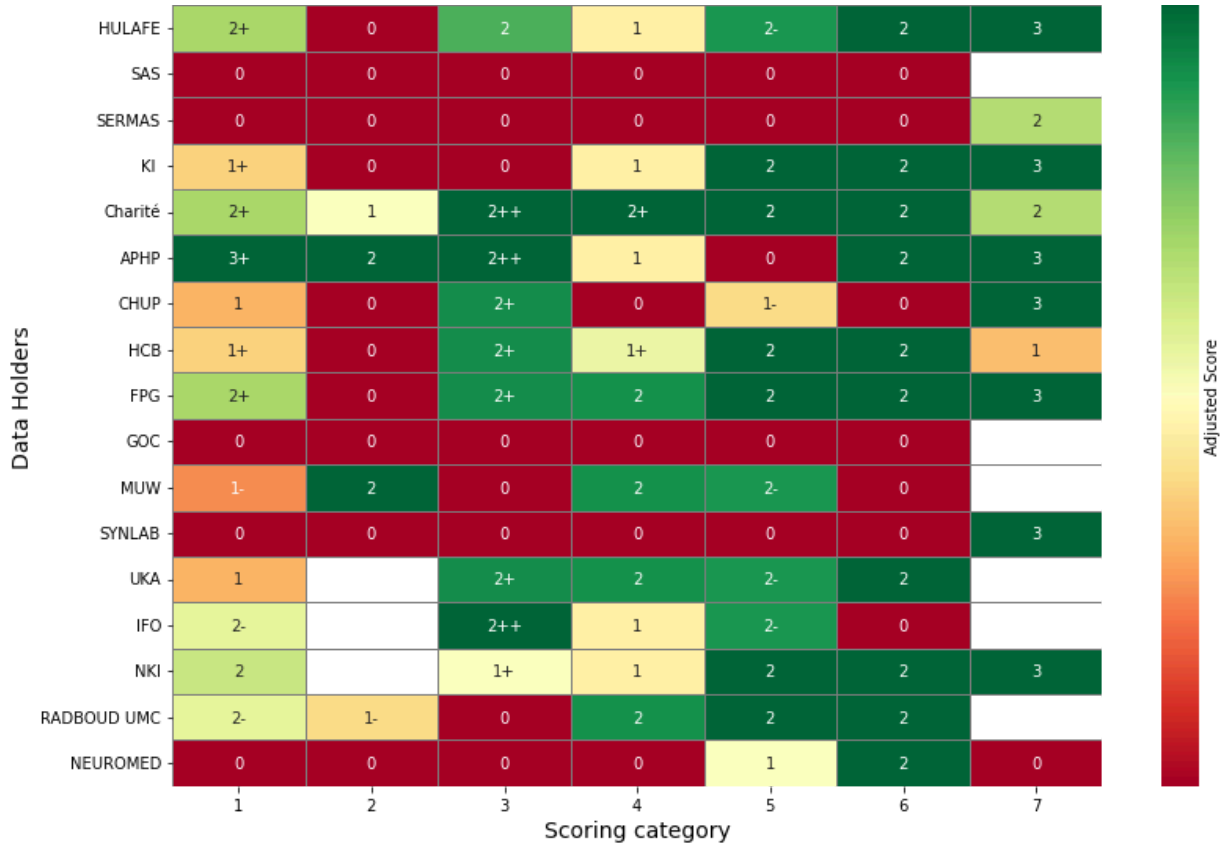

Figure 16: Heatmap representing the DW maturity assessment score by DH

The results are based on the scoring system indicated in section 5.4.3. For visualization purposes, '+' and '-' have been assigned a quantitative value of ±0.2. A shade of green/red is assigned to each cell of the heatmap based on its quantitative value. Since the scoring categories have different ranges, the color scale was normalized within each category. When not enough information was available to compute the score, the cell was left blank.

The heatmap analysis reveals significant variability in DW maturity among DHs. Scores in all categories range from minimum to maximum, highlighting diverse levels of progress and priorities.

Categories such as Standards, Common Data Models, and Vocabularies (3), Data Governance (5), and Privacy, Security, and Legal Requirements (6) generally achieve high scores. However, these categories also display notable disparities, with scores clustering at the extremes. This pattern suggests that these areas are often prioritized in early development stages, focusing on integrating data into a CDM and establishing governance, privacy, and security frameworks. It is worth noting that these disparities could also stem from the design of the questionnaire or scoring methodology.

In contrast, Technical Characteristics (1) and Data Accessibility (4) show more gradual distributions, with scores spread across the spectrum. This suggests incremental progress in these areas, reflecting varied starting points and approaches among DHs.

Data Storage and Analytics (2) stands out for consistently low scores across most DHs, indicating a widespread limitation in managing and analyzing large-scale, multi-domain datasets.

Finally, a subset of DHs scores 0 across all categories (1 to 6), indicating an absence of DW infrastructure and an inability to meet even basic requirements for integration into the federation.

Overall, the heatmap highlights the uneven distribution of DW maturity among DHs. This variability serves as a basis for subsequent clustering analysis to identify meaningful groupings, uncover common patterns, and provide tailored recommendations for each group.

Clustering and analysis

To further illustrate the maturity level of the different DH a clustering analysis was conducted on the set of questionnaire scores analyzed. For this analysis, the categories were re-grouped into two super-groups., Figure 17 shows the Cluster Map of DH. On the X-axis is represented the weighted sums of the values related with the technical capabilities obtained in the scoring categories 1) Technical Characteristics, 2) Data Storage and Analytics, and 3) Standards, Common Data Models, and Vocabularies, giving greater weight to the categories of Technical Characteristics and Data Storage. The Y-axis represents the weighted sum of the features related to 4) Data Accessibility, 5) Data Governance and 6) Privacy, Security, and Legal Requirements.

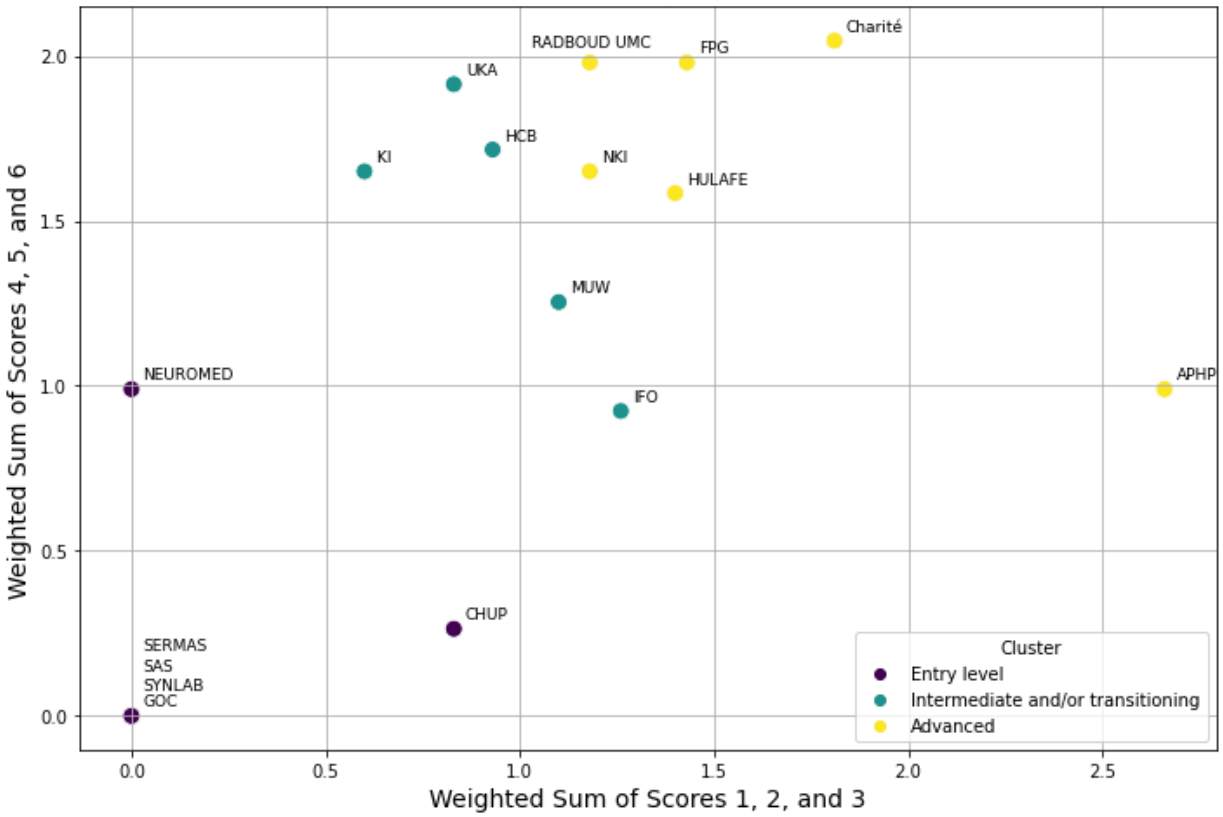

Figure 17: Cluster Map of maturity level of DHs

Weighted sums of the values related with the technical capabilities obtained in the scoring categories are represented on the X-axis, accounting for 1) Technical Characteristics, 2) Data Storage and Analytics, and 3) Standards, Common Data Models, and Vocabularies, giving greater weight to the categories of Technical Characteristics and Data Storage. The Y-axis represents the weighted sum of the features related to 4) Data Accessibility, 5) Data Governance and 6) Privacy, Security, and Legal Requirements.

Weighted sums were calculated as follows:

- x-axis: category 1 score * 0.5 + category 2 score * 0.35 + category 3 score * 0.15.
- y-axis: category 4 score * 0.33 + category 5 score * 0.33 + category 6 score * 0.33.

The weights were selected to best represent the relative relevance of the different score categories. Missing scores were assigned a value of zero.

Clusters were initially formed using the k-means algorithm. However, in some cases, DHs have been manually reassigned to a different cluster, which was considered a better fit, taking into account the overall maturity level of their DW.

The following clusters were observed:

Cluster 1: Entry level

- Partners: SERMAS, SAS, SYNLAB, GOC, NEUROMED, CHUP.
- Characteristics:
  - ➔ Low sums on both axes (X and Y), indicating limited infrastructure in terms of technical capabilities (X-axis) and governance, accessibility, and security (Y-axis).
  - ➔ DHs in this cluster likely have rudimentary DWs or lack a DW altogether.
- Explanation:
  - ➔ These DHs may be in the early stages of implementation, with no documented processes or advanced functionalities like API-based accessibility or robust governance controls.
  - ➔ They will require significant support to reach the maturity level needed for full participation in the EUCAIM federation.

Cluster 2: Intermediate and/or transitioning

- Partners: UKA, KI, HCB, MUW, IFO.
- Characteristics:
  - ➔ Moderate sums on both axes, indicating some implementation of technical capabilities (X-axis) and governance, accessibility, and security (Y-axis).
  - ➔ DHs in this cluster likely have functional but incomplete infrastructure, with specific gaps in standards, accessibility, or security processes.
- Explanation:
  - ➔ These DHs represent a transitional stage, with partially functional systems and basic governance. They could integrate into the EUCAIM federation more easily with targeted adjustments.

Cluster 3: Advanced

- Partners: HULAFE, NKI, RADBOUD UMC, FPG, CHARITÉ, APHP.
- Characteristics:
  - ➔ High sums on both axes, reflecting robust technical capabilities (X-axis) and a mature approach to governance, accessibility, and security (Y-axis).
  - ➔ DHs in this cluster likely have well-documented DWs, API-based accessibility, implemented governance processes, and storage that meets standards.
- Explanation:
  - ➔ These DHs are probably ready to participate in EUCAIM and can serve as examples or leaders to support other DHs in their transition.

Recommended actions by cluster

Based on the provided cluster analysis, some general recommendations tailored to each cluster are as follows:

Cluster 1 (entry level):

- Plan first and build a foundation for a Data Warehouse: defining core elements and data model, selecting the appropriate data storage architecture and technologies, designing ETL processes, etc. Consider using standardized data exchange formats in the first iterations already.
- Seek technical assistance, building the necessary internal capacity in the initial stages.

Cluster 2 (intermediate and/or transitioning):

- Enhance data interoperability: implementing standardized data formats and terminologies and improving data quality and consistency through automated checking and validation processes.
- Strengthen data security and privacy: implementing advanced security measures such as intrusion detection systems or anomaly detection systems and conducting regular security assessments and vulnerability scans.
- Keep regular updates and revisions for a set of data governance policies.

Cluster 3 (advance candidates):

- Maintain standards and adapt according to federation requirements.
- Lead by example, acting as mentors in the federation supporting other DHs in their development.

6. Constraints for the DH in the creation of a DW and potential solutions

The implementation of a DW in the context of a hospital presents numerous challenges that span across ethical, legal, technical, operational, and organizational domains. This report summarizes the findings from diverse healthcare settings to provide a comprehensive overview of the main limitations encountered throughout the lifecycle of a DW and the strategies employed to address them.

6.1. Ethical and legal constraints

One of the primary challenges relates to data protection laws, which vary significantly across nations. Stricter regulations in some areas demand a clear distinction between the use of data for research and commercial purposes. Data sharing is often permissible only under specific conditions, such as explicit patient consent or active participation of the data-holding

institution in the research project. This becomes more evident in the case of cross-border data transfer, with varying legal requirements for data storage, use and sharing.

Additionally, the debate surrounding anonymization and pseudonymization complicates compliance efforts. In many cases, there is no consensus on the applicability or adequacy of these methods for ensuring data privacy. For instance, some data protection officers argue that true anonymization of medical data is not feasible, leading to complexities in establishing data-sharing agreements. Moreover, it is difficult to solve this type of problem, as full anonymization would greatly diminish the value of the data for analysis and training models.

6.2. Technical and operational constraints

Hospitals face technical challenges related to the integration and management of disparate data sources. Access to legacy systems, such as Picture Archiving and Communication Systems (PACS), is often limited, necessitating the creation of separate, curated servers for data extraction and sharing. While technical implementation may not always pose direct barriers, significant resources—both financial and human—are required to establish, supervise, and maintain the necessary infrastructure.

Operational barriers include labor-intensive tasks such as obtaining patient consent and curating datasets, which require specialized staff and adequate compensation. Furthermore, the historical lack of standardized and consistent data entry practices has resulted in heterogeneous data quality, adding complexity to the processes of data cleaning and preparation.

6.3. Organizational constraints

Hospitals frequently struggle with adapting their organizational structures to accommodate data-driven projects. This is particularly evident in the integration of new professional profiles specializing in data science, engineering, and analytics. Traditional hiring regulations, especially in public healthcare institutions, often lag behind technological advancements, making it difficult to recruit and retain qualified personnel.

Moreover, the culture within many hospitals has not traditionally supported data-driven approaches. As a result, there is often resistance to change and a lack of alignment among administrative staff, clinicians, and technical teams.

6.4. Economic constraints

Limited financial resources represent a persistent challenge, particularly in public healthcare systems where budgets are constrained. Indeed, demonstrating the value of a Data Warehouse project to stakeholders can be difficult, particularly in the early stages. While investments in hardware and general IT infrastructure are often prioritized, funding for stable, dedicated data teams remains insufficient. This imbalance hinders progress in achieving higher levels of data maturity.

6.5. Strategies for overcoming constraints

Organizations have employed several strategies to address these challenges. Educational initiatives have been launched to raise awareness and promote a data-driven culture within institutions. Existing staff are often trained in new technologies to bridge skill gaps, reducing reliance on external recruitment.

To address organizational barriers, some institutions have established centralized data management boards to govern data strategy and ensure alignment among stakeholders. These boards play a critical role in cataloging and understanding disparate data sources, as well as planning and securing resources for future developments.

In cases where funding is limited, pilot projects have been used to demonstrate the value of DWs and secure additional resources. These projects often focus on specific use cases, such as predictive analytics for patient outcomes, to establish a proof of concept and build momentum for larger-scale implementation.

7. Conclusions

This deliverable analyses the current status of the RWDH that are beneficiaries of EUCAIM as of project M24 (December 2024). The same analysis will be conducted with new partners that will join the consortium due to the Open Call and with the stakeholders that will be interested in the creation of a DW. This analysis will be recurrent, and the number of DH connected will be reflected in the Dashboard.

In the process of analyzing the DH with the questionnaire it has been noted that it needs some improvements in order to obtain useful information about the status of the hospitals. The percentage of unanswered questions goes up to almost 50% in some of the sections, in some cases even over this percentage, which suggests that the questionnaire should be rethought not just by rephrasing the questions but also restructuring it. An open field text or “Unknown” in those questions which only lead to answers “Yes” or “No” can be included, and on the contrary, limitate the open field text in those questions in which it would be better that the centers will be more specific. Another improvement can be the reformulation of some questions to better classify and support the centers regarding their level of maturity, and also deleting some of the questions and adding new ones. These improvements will be implemented by the Engagement Team (ET). In addition, the final version of the “Self-Assessment Questionnaire” will be included on the EUCAIM Dashboard so that DHs can fill in the form online and it will be automatically received by the ET members for further analysis.

The analysis of DWs maturity shows significant variability in technical and governance capabilities. DHs can be classified into three groups: entry-level, with rudimentary or no DW; intermediate, with functional but incomplete infrastructure; and advanced, with well-documented DWs, good accessibility, and governance processes. These distinctions highlight the varying degrees of readiness for integration into the EUCAIM federation. Each group requires targeted actions to reach the maturity needed for federation participation.

In addition to the technical maturity of DWs, it is crucial to consider each DH perspective and level of commitment to EUCAIM. Regardless of their maturity level, some DHs are initially not inclined to become a federated node (i.e., remain as Tier 1) or to join the federation without connecting their DW to the infrastructure (i.e., uploading specific datasets to the federated node). Conversely, some DHs lacking DW infrastructure have shown a stronger commitment to EUCAIM, aiming to align their DW design with the consortium's guidelines.

As it has been analysed in Section 6, hospitals encounter various difficulties in setting up a DW. This is a process that requires considerable financial investment, as well as qualified personnel to carry out the transformation in the hospital, but insufficient funding is a critical constraint in many regions and the resistance to change and outdated hiring practices are prevalent issues, particularly in public healthcare systems. The hospitals also find Technical and Operational obstacles, given that the integration of disparate data sources and the labor-intensive nature of data curation are common barriers, necessitating significant investment in infrastructure and personnel. Another limitation is related to Ethical and Legal issues, since data protection laws and the lack of consensus on anonymization and pseudonymization remain universal challenges that require careful navigation. All these obstacles will be taken into account by the Engagement team to support the DHs that initiate or continue this process.

Despite these challenges, institutions that prioritize education, centralized governance, and strategic pilot projects have demonstrated progress in overcoming barriers. These efforts underscore the importance of a coordinated approach to designing, developing, deploying, and maintaining a Data Warehouse in hospitals.

ANNEX I - Questionnaire for the evaluation of the status of the existing health information systems for secondary use of data (Data Warehouse)

1. Technical Characteristics

- Does your healthcare organisation use a Data Warehouse to store and analyse data for secondary use? (if no, you can skip the rest of the questions)
    - o Yes
    - o No

- Which of the following data types are incorporated into your Data Warehouse? Select all that apply.
    - o Structured data (e.g. laboratory results)
        - ■ Yes
        - ■ No
    - o Semi-structured data (e.g. clinical-pathological reports, treatment plans)
        - ■ Yes
        - ■ No
    - o Unstructured data (e.g. clinical notes)
        - ■ Yes
        - ■ No
    - o Medical Images
        - ■ Yes
        - ■ No
    - o Medical Images associated radiological/nuclear medicine reports
        - ■ Yes
        - ■ No
    - o Medical Images DICOM tags:
        - ■ Yes
        - ■ No

- Do you have documented the process by which data is extracted from source systems, transformed into the appropriate format, and loaded into the environment (commonly referred to as ETL)?
    - o Yes
    - o No

- Which database management system is used to support the Data Warehouse?
    - o MySQL
    - o PostgreSQL
    - o SQL Server
    - o Oracle
    - o Mongo DB
    - o Other (please specify)

- Do you have a centralised or distributed Data Warehouse?
    - o Distributed
    - o Centralised

- Are the data marts deployed on-premise or in the cloud?
    - o On-premises
    - o Cloud

- How frequently is new data incorporated into the Data Warehouse?
    - o Daily
    - o Weekly
    - o Monthly
    - o Other

- Have you documented the process for maintaining, updating and backing-up of the Data Warehouse over time as source systems evolved?
    - o Yes
    - o No

- Did you identify and track changes into source systems?
    - o Yes
    - o No

2. Data Storage and Analytics

- **What data domains are currently covered in your organisation's Data** Warehouse?
    - ○ Genetics
    - ○ Oncology
    - ○ Pathology
    - ○ Laboratory
    - ○ Imaging
    - ○ Pharmacy
    - ○ Surgery
    - ○ Radiotherapy
    - ○ Others (please specify)

- For all data mart available, please provide:
    - ○ Total data volume:
    - ○ The expected data volume growth over the next 12 months:

- For Medical Imaging, please provide:
    - ○ Total data volume:
    - ○ The expected data volume growth over the next 12 months:

3. Standards, Common Data Models, and vocabularies

- Have you documented the data cleaning and validation processes performed prior to loading data into the Data Warehouse?
    - Yes
    - No

- Are you using any standard Data Model? (select all that apply)
    - OMOP (Observational Medical Outcomes Partnership)
    - FHIR (Fast Health Interoperability Resources)
    - mCode (Minimal Common Oncology Data Elements)
    - Other (please specify)

- If you are not using any standard data model, do you have a complete documentation of the Common Data Elements and Common Data Model used?
    - Yes
    - No

- Are you using any standard vocabularies or terminologies?
    - Yes (please specify)
    - No

- Do you store any type of data annotations (e.g. image segmentations)?
    - Yes (select all that apply)
        - NIFTI
        - DICOM-SEG
        - Other (please specify)
    - No

4. Data Accessibility

- Do you have documented how data security and access control are managed within the Data Warehouse?
    - Yes
    - No

- Data Availability through APIs and Metadata Catalog: Are services available to provide access to data within the Data Warehouse?
    - Yes
    - No

- What specific mechanisms and tools does your organization use to manage authentication and authorization to use the Data Access Services and resources:

- Which level of availability would you be able to guarantee?
    - 24/7
    - Working hours
    - Best effort

5. Data Governance

- Who within your organisation is responsible for governing access permissions and enabling secure remote connections for the purposes of configuring the federated data node? Please provide name and email

- Do you have a specific data governance layer implemented and documented?
    - Yes
    - Yes, but not documented
    - No

- Do you have a documented de-identification process?
    - Yes
    - No

6. IT policies

- Does your organisation have formal policies regarding the use of virtual private networks (VPNs) and personal devices to establish remote access to the organisational network and federated data node?
    - Yes
    - No

- Is there any Does your organisation have formal policies regarding the use of virtual private networks (VPNs) and personal devices to establish remote access to the organisational network and federated data node?
    - Yes
    - No

- Has your organisation implemented specialised firewall policies to safeguard the network and federated data node?
    - Yes
    - No

- Are there any network ports restricted from external access?
    - Yes
    - No

- Does your organisation ensure regular audits of the Data Warehouse?
- Is there any additional information that you consider relevant?

7. Privacy, Security and Legal requirements

- Has a risk assessment been performed on your Data Warehouse and computing infrastructure (e.g. Data Protection Impact Assessment)?
    - Yes
    - No

- Provide the contact information (name and email) of your organisation's Data

Protection Officer (DPO).

- Do you have an ethics committee approval for your Data Warehouse and related uses of data?
  - Yes
  - No

8. Hardware requirements for federated nodes

Below you will find a table outlining the indicative Hardware requirements for a single node as they have been defined so far in the context of the project. The requirements below have been suggested to guarantee that the majority of EUCAIM use cases will be supported in terms of node performance, and that no critical performance bottlenecks will occur during platform operations. These requirements are subject to updates during the EUCAIM lifetime.

- **Do you have the minimum hardware requirements described in the table?** (In next steps, you will be asked to provide details regarding the model and specifications of the servers that will host federated data nodes, including CPU, RAM, GPU, storage, motherboard, server provider and model).
  - Yes
  - No

Hardware	Option 1	Option 2	Notes
CPU	Minimum Cores: 16 >=1.8GHZ	Minimum Cores: 12 >=3.0Ghz	• If a GPU is not present, a server-grade, high core-count CPU is necessary for the Second Prototype. • If not comparable by cores, ideal thread count is 24+.
RAM	64GB	64GB	• DDR5 is ideal. • ECC memory is highly recommended for stability.
Motherboard	4+ Ram Slots	4+ Ram slots	• Make sure to double check the compatibility of selected CPUs with the Chipset of the motherboard. • In the case of DDR5, double check motherboard compatibility with DDR5.

Storage	521 GB SSD Drive for Operating System (Either NVMe M.2 PCI Gen4 or SATA III) 1TB++ SATA III Drive (SSD or HDD) for local storage of medical data		• M.2, NVMe, Gen4 Drives are suggested for the OS • For data storage size, DHs (DH) are expected to plan their purchase depending on the size of the Data they will provide. 1TB is a minimum, with some DPs already planning for 2 TB + datasets. • For data storage, SSD are preferred for speed but are not mandatory.
Graphics card	NVIDIA Quadro	NVIDIA RTX 3XXX	• 12GB RAM+ is preferred. • Maximizing the amount of Tensor Cores is a priority, most recent GPUs will generally have higher Tensor Core counts. • Ampere and Volta architectures are preferred.
Operating System	Linux		• The latest version of any mainstream Linux distribution is acceptable: Ubuntu, Alpine or other. • Windows is NOT acceptable, unless absolutely impossible for a DP to setup a Linux environment
Power Supply	-		Each DP must make calculations depending on the hardware setup that will be selected to make sure that needed Wattage is covered and ideally exceeded to prepare for any future upgrades to the machine.
Internet	100mbps (baseline)		Each DP must make best efforts to provide the best possible connection to their Node. Network performance will directly affect node stability and can invalidate AI training or prevent successful demonstrations of the platform.