BlueBerry

a blueprint for the EURACAN registry

A BLUEPRINT FOR THE EURACAN REGISTRY

August 2024

Executive Summary

Executive Summary

The BlueBerry Blueprint outlines a strategic plan for the scalability and sustainability of the EURACAN Registry, focusing on rare adult solid tumor cancers within the European Reference Network (ERN) EURACAN. The document addresses the challenges of scaling up the registry, ensuring data privacy, and harmonizing data across multiple healthcare centers in Europe, while also aligning with emerging European health data initiatives.

Key Highlights:

- ***Purpose and Goals**:*
  - *The blueprint aims to define a sustainable path for the EURACAN Registry, enabling large-scale data collection and analysis of rare cancers to improve patient outcomes.*
  - *The blueprint provides a roadmap for the transition from clinical to research data, implementation of a federated infrastructure, and the navigation of complex legal frameworks.*
- ***Challenges**:*
  - *Harmonizing data practices across healthcare centers with varying protocols and maintaining data quality*
  - *Ensuring compliance with diverse European legal frameworks*
  - *Securing sustainable funding and fostering collaboration among stakeholders.*
  - *Ensuring correct competencies are in place and that personnel have the capacity to deliver*
  - *Complexity of registry and operations means finding any one suitable coordinating center is difficult*
- ***Governance and Operations**:*
  - *The current governance, led by INT in Milan, is effective but strained by administrative and personnel constraints and funding limitations*
  - *The blueprint proposes a more sustainable operational structure, including the identification of key organizational competencies necessary for managing a federated and pan-European registry.*
- ***Data Infrastructure**:*
  - *The blueprint details the implementation of the (OMOP) to harmonize data and facilitate federated learning and analysis using vantage6.*
  - *It addresses challenges in data transformation, harmonization, and quality assurance, emphasizing the importance of maintaining data integrity and minimizing information loss.*
- ***Legal Framework**:*
  - *Navigating the complex legal landscape of European and national legislations is crucial for the registry's success. The blueprint sketches the relevant GDPR principles and provides a description of the legal documentation and framework required to allow the EURACAN registry to exist in a federated infrastructure.*
- ***Future Directions**:*
  - *The blueprint advocates for integrating the EURACAN Registry into the broader European Health Data Space (EHDS) by remaining a FAIR and federated data infrastructure.*
  - *It outlines a path towards securing long-term sustainability via an exploration of the strengths and capabilities of other European organizations.*

The BlueBerry project, running from September 2022 to August 2024, serves as a critical step in strengthening the EURACAN Registry, ensuring its scalability and positioning it as a key resource in the fight against rare cancers in Europe.

Table of Contents

1. Introduction

The primary aim of BlueBerry is to define a route towards a sustainable and impactful EURACAN Registry on rare adult solid tumour cancers. The establishment of at-scale real-world databases for rare cancers is essential to improve our understanding of the disease and identify opportunities to reduce its impact.

The strengths of the current EURACAN[1] Registry (pioneered in STARTER[2]) include its integration within the well-established EURACAN European Reference Network (ERN), recognition by the European Commission, and its collaborative approach, which is crucial for rare cancers.

Despite limited funding, the STARTER team demonstrated creativity in finding innovative solutions. A multidisciplinary team paved the way for a modern registry, and patients are actively engaged. By implementing a federated registry, without data being centralized, STARTER addressed the main concern about data control and privacy. **In the EURACAN Registry, data remains at the health care center.**

In the Blueberry project (Sept 2022 - Sept 2024), the following critical aspects have been addressed:

- The transition from clinical data (in the patient record) to research data.
- The implementation of federated technology at the EURACAN centers.
- The complexity of the legal and governance framework.
- The demonstration of impact via the EURACAN Registry, through selected case studies
- The future of the registry, including the role of the coordination

A separate challenge (as well as an opportunity) is the emerging European Health Data Space. Uncertainty around the implementation of the EHDS and emerging observational cancer data networks could challenge the priority of the registry. Opportunities for synergies with other data networks and EHDS implementation projects should be explored, and partnerships with public and private entities will be pursued.

To ensure sustainability, the EURACAN Registry aims to move from project-based to more structural funding by becoming a priority in the EURACAN activities and exploring structural funding opportunities at the national level.

1.1 The Challenges

A European rare cancer registry has been set up for two EURACAN domains: Sarcomas (Euracan domain G1, through Blueberry, IDEA4RC[3]) and head & neck cancers (Euracan domain G7, through IDEA4RC and STARTER).

Scaling up the EURACAN Registry is a complex undertaking due to several interconnected factors.

[1] https://euracan.eu/
[2] https://euracan.eu/registries/starter/
[3] https://www.idea4rc.eu/

1. Rare cancers encompass a **diverse range of malignancies** with unique biological characteristics and limited patient populations, thus it is necessary to create several registries, ultimately for all 10 Euracan domains.
2. **Establishing effective governance** requires **harmonizing practices** across multiple healthcare centers, each with varying protocols and data management procedures. Also, rules on data usage and access need to be established, including engagement with private partners.
3. Ensuring **compliance with diverse legal frameworks** across different European countries adds complexity to data sharing, patient privacy protection, and ethical considerations.
4. As the network expands, **maintaining data quality and consistency** becomes crucial, necessitating standardized data collection methods and ongoing data validation efforts.
5. **Securing sufficient funding and resources** (for the coordination and the centers involved) to support the growth and sustainability of the network represents a significant challenge.
6. **Fostering collaboration** among numerous stakeholders, including healthcare institutions, researchers, patient advocacy groups, and policymakers, demands effective communication and alignment of goals.
7. **Incentivizing expert centers** to maintain long-term data collection, either manually or via electronic capturing tools.
8. **Scaling up a federated infrastructure** adds complexity and requires expertise and understanding of the federated system across the hospitals and data users involved.

To address these challenges and to create a sustainable at-scale EURACAN Registry the following aspects should be put in place:

- The right organization for managing and coordinating the registry. (Chapter 2)
- The right data infrastructure for the registry (Chapter 3)
- The right legal framework (Chapter 4)

We conclude the blueprint with a discussion on implementing this way forward.

2. Organization and operations

Topline Summary:

The current governance of the EURACAN Registry, led by INT in Milan, is effective but faces challenges including staffing constraints, uncertain growth, administrative burden, funding limitations, engagement with public funders, and non-scientific management. Despite these challenges, INT brings strengths, such as a dedicated team, strong leadership, transparency, inclusiveness and European Commission recognition. As the EURCAN registry begins to take the long-view of its future, organizational competencies have been identified, namely expertise and knowledge in data platform management, data usage and privacy in the international context, competent organization, sustainable funding and capital management and governance. These capabilities will ultimately allow any host of the EURACAN registry to address complexity and take advantage of opportunity.

2.1. Current Governance

The EURACAN Registry is currently being led by a leading hospital in EURACAN, INT in Milan. The governance of the registry is defined on the EURACAN website[4] and defines roles and responsibilities.

The scientific coordinator is assumed by INT Milano. To support the scientific ambitions of EURACAN, the hospital is coordinating the operations of the EURACAN registry. With the growth of the registry, this is causing a significant burden on the non-scientific staff. As the coordination of a pan-European registry is not a core activity of the hospital, action is needed to further professionalize the operations and governance, and transfer responsibilities to an entity that is equipped to coordinate international cancer data infrastructures.

The following challenges in sustainability and scalability are not institutional specific, but rather outline the challenges that healthcare institutions face as the long-term hosts of a complex infrastructure such as the EURACAN registry:

1. **Staffing constraints**: The availability of essential and dedicated staff, including IT, legal, and administration, on short-term, project-based contracts, presents difficulties in establishing a dedicated and stable organization. Moreover, non-commercial salaries can hinder the attraction and retention of qualified (in particular non-scientific) personnel.
2. **Uncertain growth path:** Budgets for initiatives like the EURACAN registry often rely on incidental funds, making it challenging to commit to a sustainable growth path for the registry.
3. **Administrative burden**: The process of including centers in the network poses an administrative burden for the coordinator and may not always align seamlessly with the mission and scientific objectives of the coordinating center.
4. **Funding limitations**: Hospitals often lack dedicated strategic teams responsible for scientific research fundraising, which presents obstacles in securing funds from sources other than HORIZON and domestic grants. Attracting international funds from other member states or and their national subsidy programs, presents a challenge. Hospitals may not be best suited to engage with public funders in all EU member states, creating hurdles in seeking resources for scaling the registry at both national and EU levels. On the other hand, the nature of rare cancer research also makes attracting industry funds challenging given the perceived lack of economic value added.

[4] https://euracan.eu/registries/euracan-registry/registry-governance/

5. **Expanding Expertise**: Initiating registries for the other 8 domains requires substantial resources, scientific leadership and coordination, from various centers within EURACAN, which may easily exceed local capacities and priorities.

With these challenges in mind, five thematic areas of responsibility for an ideal EURACAN registry coordinator have been identified:

1. **Data platform Management:** the provision of a robust, scalable, interoperable data platform that includes the technological infrastructure and capabilities to handle diverse data types and to ensure seamless integration and accessibility. This included mechanisms to capture registry data from the clinical records, with or without the use of electronic tools.
2. **Competent organization:** Supply the needed expertise, experience and capacity to handle complex data-driven projects, including in-depth knowledge, skill sets and resources to support and sustain the registry
3. **Governance, oversight and steering**: Ensure that governance mechanisms are clearly defined to provide decision-making processes, accountability frameworks, and stakeholder engagement strategies in order to effectively steer the organization and provide oversight.
4. **Data usage, compliance and privacy**: Equip registry users and data providers with ethical and regulatory compliant policies, practices and measures related to data usage, protection and privacy to ensure alignment with legal requirements, industry standards, and ethical principles to safeguard data integrity and privacy rights
5. **Sustainable funding and capital management**: secure funding sources and utilize capital management practices to secure and sustain the long-term maintenance of the EURACAN registry through prudent financial planning, diversified funding streams and effective resource allocation.

A capability framework of these areas of responsibility can be found in Appendix 1. With these competencies in mind, the practical road ahead is outlined further in Chapter 5.

3. Data Infrastructure

Topline Summary

In this chapter, data inclusion and harmonization are discussed. The chosen common data model (CDM) for the EURACAN Registry, specifically the Observational Medical Outcomes Partnership (OMOP) CDM, facilitates collaborative research and large-scale analytics by standardizing data storage and relationships. The process of transitioning data from the already available database into registry data involves an "Extract, Transform and Load" (ETL) procedure that requires mapping data to standardized vocabularies, while maintaining data quality and granularity. The roadmap includes plans for federated analysis utilizing the OMOP CDM within the vantage6 federated network, focusing on secure and privacy enhancing data analysis.

A SWOT analysis of the current data strategy identifies strengths in the OMOP CDM's open-source nature and flexibility, while highlighting challenges like data loss during conversion. Opportunities include choosing suitable data elements and including (new) data sources, while threats involve scalability and tool compatibility issues. Of key importance is that validation and alignment of data mapping within the network are crucial, and that the implementation of the federated learning is carefully set up and compatible with the data harmonization efforts. Three use cases have been developed focusing on sarcomas: simple, clinically relevant, and a sustainability use case.

3.1. Introduction

Traditionally, when a researcher wants to analyze data from different sources, these datasets need to be requested, prepared and shared by each data holder to the researcher. This means that patient-level data leaves the respective organizations and is brought together on the machine of the researcher. In recent years, concerns around ensuring patient privacy have increased, making organizations more hesitant to share record-level data with third parties. On the other hand, to progress our knowledge on healthcare in general and cancer in particular, there is an increasing need to combine both horizontally as well as vertically partitioned data.

One solution to the concerns around patient privacy and data sharing is use of federated technical infrastructures. Blueberry uses Vantage6, an open-source infrastructure developed by IKNL, eScience Center, Maastro and other partners. Vantage6 enables parties to gain insights from sensitive data (individuals, patients, citizens) from different sources, without transferring the data or inspecting items from individuals within the datasets. This is done through the application of privacy enhancing technologies (PETs), including federated learning (FL), secure multi-party computation (MPC), homomorphic Encryption (HE) and differential privacy (DP). These technologies enable analysis of data, while protecting the sensitive information of individual data subjects. Each technology brings its own form of complexity to the analysis and often a mix of them is required to get the most effective result. This usually depends on the research question you would like to answer, the actors involved, the type of data, the analysis methods, computational resources, and presence of other available safeguards. Further discussion of the technical specifications for this infrastructure will be addressed in the appendix to this blueprint.

In addition to finding a means to share data insight without sharing actual data, the other technical challenge is posed by real world healthcare data, which can play a pivotal role in multidisciplinary research. However most data sources use their own unique data models and schemas, making it difficult to combine data from different sources and develop software tools for reliable and reproducible research. One solution to this problem is to create a common data model (CDM) that standardizes the storage of both the data and the relationship among data elements. Several CDMs have been developed for healthcare data, including those supported by the following organizations: Informatics for Integrating

Biology and the Bedside (i2b2), Sentinel, PCORnet (Patient Centered Outcomes Research Network), and Observation Health Data Sciences and Informatics (OHDSI, managing the Observational Medical Outcomes Partnership [OMOP] CDM). In most cases, the choice of CDM depends on the research focus/question of a study, the type and format of source data, the data element and vocabulary coverage of the CDM, the availability of tools and usability of the CDM to query and analyze the data. For the data harmonization within the EURACAN Registry, we opted for the OMOP CDM, which will be discussed further below in the first section.

For data harmonization within the EURACAN registry, we opted for the OMOP CDM. It promotes collaborative research and global communication using standardized vocabularies, including the Systemized Nomenclature of Medicine (SNOMED) and International Classification of Diseases for Oncology, 3rd Edition (ICDO-3). The OMOP CDM's oncology extension supports large-scale analytics for oncology research. Ongoing initiatives aim to improve interoperability between different CDMs (e.g. OMOP on FHIR, OMOP to PCORnet). Finally, the OMOP CDM allows researchers to answer healthcare questions with real-world evidence using data from different data sources, as they can be captured in the same schemas.

Data from medical records are integrated into electronic health records (EHRs). EHR data are not typically reusable in their raw form and must be converted into clinical registries based on predefined criteria (e.g., a core dataset) and specific characteristics (e.g., patients with sarcomas). In the Blueberry project, we include both clinical registries, which provide a comprehensive view of a patient's care, and population-based registries, which capture diagnosis and initial treatment information for all new cancer cases in a region or country, though they generally lack follow-up details.

3.2. Data Capture and Harmonization

Data Transformation

One of the biggest challenges in data harmonization is the transformation of the source data into a common data model. This process, known as "Extract, Transform and Load" (ETL), not only restructures data to the CDM, but also integrates mappings to standardized vocabularies (**Figure 1**). The ETL is developed by data owners and CDM experts, and the process is repeatable, and can be rerun whenever the source data has been updated. For the data mappings, clinical experts are consulted to ensure that source data is transformed correctly and information loss is minimized. The ETL is then implemented by technical professionals, and data quality checks are performed afterwards to ensure that the OMOP database corresponds to the source database. Once the source data of all the data partners are transformed to the OMOP CDM, one analysis script can be run on the data of each of the data partners. This increases the interpretability of the data across data sources and the reproducibility of the results.

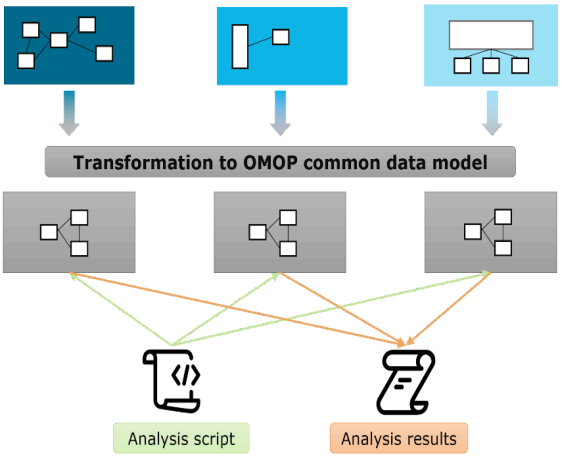
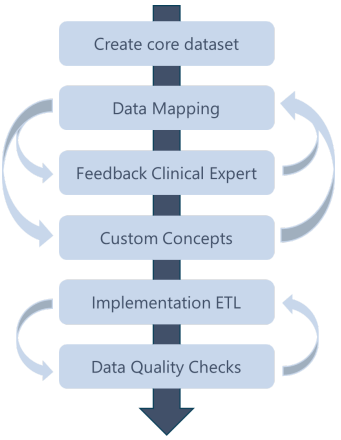


Figure 1: Data from different data sources, stored in their own data formats, are converted to one standardized common data model for observational health data (OMOP CDM).

Data Harmonization strategy



The data harmonization strategy consists of six steps (see figure), discussed in more detail below. Steps 2-6 were performed at each data partner and repeated at the network level.

Core Set of Data Elements

1) **Patient specific data/demographics:** Dates of birth/death, gender, education, marital status, ethnicity, place of residence.
2) **Risk factors and comorbidities:** Alcohol intake and smoking history, Charlson Comorbidity Index, Genetic syndromes WHO 2020, previous cancer diagnoses and/or treatment
3) **Disease specific data:** Information about the diagnosis, including pre-diagnostic lab tests and imaging. Information about the primary tumor, including tumor site, morphology, tumor depth, grading, and tumor stage.
4) **Treatment related information: I**nformation concerning treatment, such as the type of surgery, including the radicality of surgery/ resection margins, and/or medical treatment (e.g. chemotherapy, molecular target therapy, immunotherapy, radiotherapy), start and end date of treatment, treatment plan, treatment response.
5) **Follow up information:** Overall response of treatment, date of last contact, status of patient at last follow up, recurrence of disease and treatment of recurrence.

Data mapping and feedback from clinicians

Based on the core dataset, we first developed a mapping table showing which OMOP concept a data element is mapped to. This mapping table was shared within the network to ensure that all data partners used the same implementation. OHDSI tools such as “White Rabbit”, “Rabbit in a Hat” and “Usagi” were used to perform the data mapping. Data mappings were verified with clinical experts to ensure that information from the source database was correctly transformed and the information loss was minimized. Data elements not covered by existing OMOP vocabularies were considered for custom mapping.

Custom mappings

Custom mappings may be required when the desired data coverage or scientific use cases cannot be answered by the standard vocabularies. The use of such custom mappings should be minimized as they are time consuming to generate and hard to maintain. We found that diagnosis codes of less common cancers, such as sarcomas, are underrepresented in the OMOP vocabulary, potentially leading to 10-15% of sarcoma patients not being converted to OMOP. Thus, we created custom mappings for missing ICD-O3 topographies and sarcoma diagnosis concepts and shared these within the Blueberry network. These custom mappings were also communicated to the OHDSI oncology working group so that they can be added to the standardized vocabularies in the future,

ETL implementation and Quality checks

Once the data mappings were completed, we implemented extract-transform-load (ETL) pipelines tailored to each of the data sources. Since good data quality is important to obtain reliable and accurate results, we assessed the quality of the data transformation using OHDSI’s “Data Quality Dashboard” (DQD) and “Achilles” tools. Adjustments and improvements to the mapping and ETL implementation

were performed based on the results of the data quality checks for completeness, accuracy, consistency, conformance, plausibility and representativeness of the harmonized data. This process was repeated until the output of DQD- and Achilles-based assessments matched our expectations.

3.3. Federated Learning

For the technical implementation of a federated registry we used the vantage6 software. Vantage6 is a federated learning software designed to enable secure, privacy-preserving data analysis across multiple institutions without the need to share sensitive data. The principle behind vantage6 involves distributing machine learning algorithms to the data sources rather than centralizing the data itself. This allows data to remain within its original location, adhering to privacy regulations and minimizing the risk of data breaches. Among the key advantages of vantage6 are enhanced data privacy, as sensitive information never leaves its secure environment, and improved compliance with data protection laws such as GDPR. Additionally, it enables collaborative research and insights without compromising individual data security.

Implementation strategy

We used a two-step process to implement the federated network (**Figure 2**). Step 1: OHDSI Network, in which a user sends the same analysis script to each data partner individually. The data partner runs the script locally and sends back the results. Step 2: Federated Learning Network (vantage6), in which a user can choose what analysis they want to run via a user interface. This task is then sent to the server that has access to a container with analysis scripts (algorithms) and runs the script on each of the nodes connected to the network. The results of all nodes are aggregated and sent back to the user as if it was run on one data source locally. The federated learning approach allows for analyses to be run automatically and safely on a larger volume of data, making this a powerful strategy.

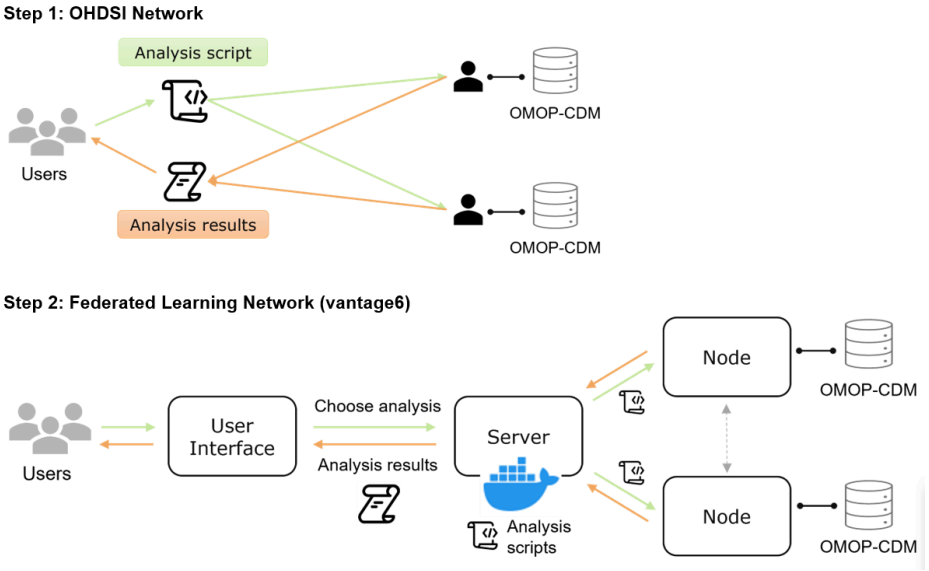


Figure 2: A two-step process is used to achieve a federated registry.

Compatibility of OHDSI tools with vantage6

The OHDSI community offers a comprehensive library of tools, with software that directly connects to an OMOP database (using SQL) for performing analyses. This approach is effective for an OHDSI-style network (**Figure 2**, step 1), where analyses are run on a device with direct access to the OMOP database. This does not work in a federated learning network like vantage6, where there is no direct access to the OMOP CDM via the vantage6 node. The initiation and execution of the analysis are separated to improve security, meaning that OHDSI tools cannot be directly transferred to the vantage6 network and require adaptation for compatibility.

OHDSI tools are also complex: they are difficult to install, modify, and debug, and are generally unstable and not production-ready. Some OHDSI packages, particularly those for using the OMOP-CDM (connecting to and reading the database), can be integrated with vantage6. We identified five core OHDSI packages—CirceR, SqlRender, Cohort Generator, DatabaseConnector, and FeatureExtraction—for which we developed wrappers to ensure compatibility with vantage6. For more information, visit python-ohdsi.readthedocs.io.

Vantage6 User Interface for the execution of analyses

The use of a complex federated learning network and the standard data model is challenging for non-technical people. It is therefore instrumental to make the registry easy-to-use. A user-friendly user-interface (UI) was developed for vantage6 that enables people to 1) select the data partners they want to include in an analysis, 2) select an algorithm from a predefined list, and 3) execute a federated learning analysis with one-click of a button. Summary statistics of some analyses are available within the user interface. Results of Kaplan-Meier survival analysis can be exported, processed and visualized in external software. In the future, simple visualizations will be added to the UI as well.

The use of the OMOP-CDM requires some knowledge and expertise to retrieve and analyze data that is stored within. The vantage6 infrastructure removes some of the difficulty by retrieving and analyzing data directly from the UI based on a cohort definition that is pasted into the UI. Cohort definitions currently need to be generated externally with OHDSI tools, such as ATLAS.

Technical considerations

Each data partner has a unique IT landscape with distinct functional requirements that provide challenges for the installation. In the blueberry project, we opted for a simplified implementation strategy. We created homogeneity in the network by ensuring that each data partner used a virtual machine with the same operating system and technical specifications (see **table 1**). The optimal implementation strategy for connecting the OMOP CDM and vantage6 node is depicted in **Figure 3**. Source data should be hosted on a local server behind a firewall. These source data are converted to a OMOP CDM at regular intervals (e.g., every 3 months) or whenever new data are available through the ETL pipeline. The OMOP CDM resides on a single virtual machine together with the vantage6 node. This allows the node to have access to the OMOP CDM, without having access to the source data (with identifiable data).

Table 1: Technical requirements

	Required	**Optional**
1 virtual machine	x	
4CPUs 64 bit	x	
8GB memory	x	expandable to 16GB if analysis requires more performance
250 GB disk space	x	
Oracle Linux 8 OS - minimal installation	x	
web access	during installation phase	

VPN access through port 22		required for Biomeris staff

Installation of vantage6 software

The central server of the federated learning network is hosted on a server maintained by one data partner or organization (in this case IKNL). This central server is connected to the node of each data partner (using the setup depicted in **Figure 3)**. The uniformity in the setup in the network provides a straightforward connection between the nodes and the central server, improves security within the network, facilitates automation, and allows for the use of standardized installation scripts.

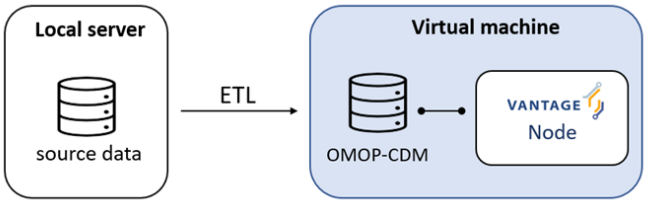


Figure 3: Connection between source data, OMOP CDM and vantage6 node

In contrast to the OHDSI tools, vantage6 decouples data from algorithms. To access the OMOP CDM, an SQL connection needed to be implemented. To establish this connection, we used two approaches: 1) Create an SSH tunnel between a locally hosted postgreSQL OMOP-CDM database and vantage6 node. An SSH tunnel is complex to set up and pose security risks, or 2) Place the OMOP CDM within a docker container and connect it to the vantage6 node using Docker services. Docker services offer the most stable connection with vantage6 and require no complicated configuration. This is therefore the most preferred implementation.

3.4. Data analysis

Use cases definitions

Several use cases were developed to test the feasibility of the legal framework, technological framework (also including the quality of the data and depth/breadth of data coverage), governance model, and business/valorization model.

1) Simple Use Case

- Provide the distribution of sarcoma subtypes according to histology
- Identify number of retroperitoneal sarcoma patients in different datasets [relevant for use case 2 and 3]

2) Clinically Relevant Use Case: ***Identification of prognostic factors for retroperitoneal sarcomas***

Prognosis of patients with primary retroperitoneal sarcoma (RPS) is variable. Tumor-related factors such as histologic type, tumor grade, tumor size and multifocality, patient-related factors such as patient age and treatment-related factors such as completeness of surgical resection determine patient survival after surgery. Outcome of patients resected for primary RPS improved significantly over a 15-year period. 5-year OS increased of about 10% from 2002-2006 (61.2%) to 2012-2017 (71.9%) due to a combination of better patient selection, improved quality of surgery, and enhanced perioperative management. Additionally, population-based and nation-wide studies have highlighted a consistent association between case volume and oncological outcome in patients with RPS.

The use case aimed to describe RPS clinical presentation and treatments and identify prognostic factors. The use case could further proceed trying to evaluate the impact of the volume of surgically treated cases on RPS outcomes.

Study-a-thon results

A study-a-thon is a catalyst event where data partners come together to make a lot of progress in a short amount of time. We organized three study-a-thons:

1) Study-a-thon 1 (May 2023)

Relying on the OHDSI style network (**Figure 2**, step 1), we resolved technical issues related to data access and software setup. With the input from clinical experts, we validated the mappings of all data partners and addressed the "fit-for-use" data quality within the network. We created cohorts for the first use case and executed the analysis with four data partners (please see the appendix for the full report).

2) Study-a-thon 2 (December 2023)

The second study-a-thon focused on defining and executing a use case 2. With the input from clinicians, we generated cohort definitions and refined the study protocol. The aim was to characterize the cohort of retroperitoneal sarcoma (RPS) patients with the following selection criteria: adult patients (≥18 years old) with a primary localized, non-recurrent, non-metastatic RPS with a surgery between 01/01/2010 and 31/12/2017. We excluded patients with other previous malignancies.

After the Study-a-thon a dedicated working group was set-up to finalise the analyses. We assessed differences and similarities of the defined RPS cohorts across contributing registries, comparing patient (age, sex), tumour (size, grade, histology, multifocality) and treatment characteristics (completeness of surgical resection, perioperative chemotherapy, perioperative radiotherapy). Further, the team contributed to develop a survival package in R and computed 5-year overall survival (OS) by sex and histology grouping to compare results with available evidence.

We identified 848 patients with RPS (444 NCR, 374 INT and 30 Graz). Sex distribution was similar across registries with a male predominance in all registers; the average age was approximately 62 years in all registries. The most common histologies were liposarcomas and leiomyosarcomas with INT having a higher percentage of dedifferentiated liposarcomas (43%) compared to Graz (37%) and NCR (27%). High-grade (G2+G3) RPS were very common (over 60%) in clinical settings (INT and Graz). Grading available in population-based registries is not comparable. Patients with RPS of the INT also had the largest tumour size (21.4±10.9). Perioperative chemotherapy was used in 26% of RPS at INT and it was rarely used in Graz. Perioperative radiotherapy was used in 18%, 19% and 47% of RPS at INT, NCR and Graz, respectively. Surgery was macroscopically complete (R0/R1) in almost all cases of INT and Graz. OS was 78% (95% CI: 60%-100%), 74% (95% CI %: 70%-79%) and 65% (95% CI: 60%-69%) in Graz, INT and NCR respectively. OS was always higher in the clinical context than in the population-based one and in detail: OS varied between 91% and 81% for well-differentiated liposarcomas, between 87% and 55% for dedifferentiated liposarcomas and between 75% and 63% for leiomyosarcomas.

This use case demonstrated that the OMOP CDM represents a reliable oncology model. The findings are consistent with expert knowledge and available evidence. The differences observed between clinical context and population are also well known, which confirm the need to refer patients affected by RPS (and sarcoma in general) to expert centers. The latter, given the greater number of cases, even the most complex ones, can guarantee better treatment and OS. It is worth highlighting that comparison between different contexts may be affected by changes in sarcoma classification over time, for example, we defined well-differentiated and dedifferentiated liposarcomas based on histology in the NCR.

3) Study-a-thon 3 (June 2024)

The focus of this study-a-thon was to test the federated learning network, user interface and run analyses based on predefined cohort definitions (same cohorts were used as in study-a-thon 2). We obtained results for cohort counts, cohort diagnostics and Kaplan-Meier survival curves from 5 centers, INT, Graz, IKNL, CLB and CRN. The results of the study-a-thon show that the federated learning infrastructure was set up correctly. However, please note that the data of the analyses have not been validated yet, thus cannot be used for clinical research (please see the appendix for the full report).

Considerations for choosing and executing use cases

When selecting a study or use case, it is crucial to consider several key factors:

1. **Understanding the data:**
   Data availability, granularity, and quality can vary across data partners, impacting cohort creation and potentially excluding data from centers with less detailed information. To address this, use cases should be tailored to balance specificity with the available data. Conducting preliminary data quality and feasibility checks, such as using the cohort diagnostics package, is essential to identify major gaps or inconsistencies before proceeding. Discussions with the data owner is essential to consider the need of data imputation and imputation rules. Finally, it is essential that data quality checks will be able to detect changes in the classification and therefore in the coding system in the DB overtime. Finally, it is important to understand the difference of the data sources and to what extent they can be comparable. In our case some variables available from the population based cancer registries were not comparable to those of the clinical registries (eg, grading).
2. **Alignment with available tools:**
   Evaluate whether the necessary tools, like survival analysis packages, are available within the OMOP-vantage6 infrastructure. If essential tools are missing or underdeveloped, executing the study may not be possible or requires significant development in the infrastructure.
3. **Expertise:**
   Multidisciplinary expertise is essential to ensure adequate mapping, analysis and interpretation of data. To properly map a database you need to involve the DB owner, the person who knows the DB and how the variables were coded. Expertise on the OHDSI tools and OMOP-CDM is also important along with clinical domain knowledge which is essential for accurate cohort definition and interpretation of results. Last but not least is the need to have adequate experience in data quality checks and analytics to ensure the execution of clinically relevant use cases.

3.5. Competencies required for data harmonization, federated learning and data analysis

	Tasks	Required knowledge	Required time
Data Harmonization	Identify Core dataset	clinical domain knowledge	
	Data Mapping, clinical input, ETL implementation and quality checks	- OHDSI tools - OHDSI vocabularies - knowledge of source data - clinical domain knowledge - SQL language	3 - 4 months (IMI EHDEN certified SME) 7 - 8 months (non-expert data partner)
Federated Learning	Installation vantage6 node	- IT knowledge - virtual machines - command line	0.5 -1 day

Data analysis	Definition of use cases	- clinical domain knowledge - knowledge of data sources - OHDSI tools	3 - 4 months
	Cohort definition	- ATLAS tool - knowledge of data sources - OMOP vocabulary	3 - 4 months
	Execution Analysis	vantage6 UI	2 - 3 days

3.6. Lessons Learned

1) OHDSI tools are complex and some lack maturity. Low level tools are robust enough to be integrated into vantage6, but alternatives should be considered where possible to ensure sustainability, interoperability of algorithms with other data standards, and improved user experience.
2) Despite using a standardized common data model, a codebook for each data partner that describes data availability and concept mapping is still important.
3) OMOP vocabularies are not exhaustive and could hamper data conversion. Close communication with the OHDSI oncology working group is important to drive change based on use cases and priorities.
4) Clear communication of expectations, responsibilities and competencies between the registry as a governance entity, the centers, data scientists, clinical experts, and other users.
5) Studyathons are a great way to achieve results and create focus and priority.
6) The ability to answer research questions and validate results requires close alignment of data experts and clinical expertise.

The following table includes specific issues encountered and the possible way forward to address them.

Transcoding to OMOP	**Process**	**Data Quality Checks (DQC)**
Missing of diagnosis codes: combinations of histology/topography in OMOP	· Repository of all missing diagnosis codes with related ICD-O vocabulary version for future updates · Periodic update of Vocabularies version and check from the repository for possible integrations	- DQC before analyses on each specific combination of site and histology - OMOP DB flags to track changes for updates - Temporary custom codes for missing diagnosis codes - Collaboration with oncology working group
Missing of event date (e.g., treatment date)	· Check for missing dates to match categorical variables (Yes/No or multi-class) in order to identify the event	- DQC on source database and discussion with clinicians to identify

	· Define event-specific imputation rules	imputation time frames
Mismatched definition of variables between clinical db and population db (e.g., grading)	· Compare the different definitions to reach agreement, if possible. If not possible, do separate analyses	- DQC before analyses highlighting all concept code differences for the same variable - Filter all concept codes in a class or leverage hierarchy
Clinical definition of missing values of a variable (e.g., missing multifocality is equal to unifocal tumor) or missing information of an event (e.g., absence of any type of metastasis is localized/nonmetastatic patient)	· Ask clinicians to define and confirm how to manage the missing values of each variable · Define event-specific imputation rules	- DQC on source database: if missing > 10% it is necessary to discuss with clinicians to understand correctness or possible corrections
Changing codification of a variable over time: a variable within the db with different codifications (e.g., excision completeness, margins)	· Ask the data manager if there are instances with different encodings of a variable · Define criteria to standardize the encodings of a variable	- DQC on source database - DQC on OMOP overtime
Defining a new definition of topography more in line with the clinicians' needs	· Identify ICD-O-3 codes of specific sites and subsites to create a new classification (according to clinicians)	- DQC on: - New Site: the most general site also seen as a group of sub-sites - New Subsite: the specific subsite (e.g., a particular organ or district)

3.7. Next Steps

<u>Vantage6</u>

1) **Infrastructure**

The current implementation requires the OMOP CDM and vantage6 node to be placed on the same virtual machine. This is not feasible for each data partner as this implementation

undermines their monitoring capabilities. The next step is to allow the vantage6 node and OMOP CDM to be stored on separate virtual machines. A network can then consist of data partners that have the vantage6 nodes and OMOP CDM on the same virtual machine and data partners that have different virtual machines for the OMOP CDM and vantage6 node.

2) **User Interface**

- Study cohorts: Future implementation will allow the user to define cohorts in two ways: 1) ATLAS generated cohorts, 2) Cohorts based on the filtering of columns within a database. The latter option is easier, comparable to how clinicians are used to working with data, and does not require knowledge of the ATLAS tool.
- Visualization: Implement simple visualization of analysis results. These guide the user on whether the chosen analysis gives the intended result. Complex visualization will remain to be executed offline.

4. Legal Framework for the EURACAN Registry

Topline Summary

Adhering to both European (GDPR) and national (Member State, henceforth MS) legislation offers both an opportunity to harmonize personal data processing and movement in the EU, as well as challenges in cross-border data processing, especially for health and scientific data and a lack of standardized interpretation of GDPR among member states. A federated infrastructure whereby aggregate analyses conducted at the local level are shared within authorized users, rather than data itself, offers a unique opportunity to comply with EU and national laws to protect privacy.

Blueberry sets forth the following legal framework for a sustainable EURACAN registry:

1) *Federated technical infrastructure*
2) *Autonomous controllership over data*
3) *Multilateral legal agreements, including a memorandum of understanding*

4.1. Context: relevant European and National Legislations

The General Data Protection Regulation (GDPR), specifically Reg. (EU) 679/2016, stands as a potent and globally impactful privacy law. The GDPR's primary focus revolves around harmonizing data processing practices and ensuring the free movement of personal data within the European Union (EU). Notably, while the GDPR aims for harmonization, it allows a certain degree of flexibility to individual member states, especially for health data and scientific research. This flexibility results in a complex landscape where different countries can adapt specific aspects of the GDPR according to their unique needs and interpretations, ultimately leading to variations in data protection regulations across the EU.

These differences among member states' legislations highlight the complexity of cross-border processing of personal data, especially in the context of health-related or genetic data, such as in the BlueBerry project. Moreover, the lack of coordination in the technological landscape presents a challenge in finding suitable tools for research data processing. Lastly, the European legal framework is enriched by guidelines, opinions and documents issued by some independent European Authorities (such as the European Data Protection Board and the European Data Protection Supervisor), as well as by the decisions of the European Court of Justice. These can aid the interpretation of the GDPR rules and obligations.

4.1.1. Reuse of Clinical Data for Research and Innovation

Under Article 9(2) of the GDPR, lawful processing of personal data for scientific research entails several conditions, the most relevant being: the consent of data subjects, the pursuing of a substantial public interest, and the pursuing of a scientific research purpose. The "Compatibility Assessment" under Article 6(4) of GDPR also aids data reuse (see Appendix 2 below for greater detail). Seeking specific consent for research purposes can be challenging due to evolving objectives and data subjects' availability and data processing for scientific research can proceed without consent if it aligns with Union or Member State law for scientific research, but there is a lack of specific laws supporting these purposes. the Compatibility Assessment mechanism determines if data processing for another purpose is compatible with the initial purpose of data collection, but it is not widely used due to a lack of guiding criteria and official interpretations.

4.1.2. Patient-level Data Sharing

European legislation lacks provisions for "donation of personal data." Currently, patients' choices pertain only to allowing or refusing the processing of their data in research projects. The "Data Altruism" concept introduced by the DGA presents a promising mechanism for the voluntary sharing of personal

data to advance public interest objectives, like healthcare and research, although it is not yet operational.

4.1.3. Anonymous Data

Anonymity prevents the need for lawful conditions, as per Recital no. 26, which states that GDPR does not apply to "anonymous information". Anonymous information is information which does not relate to an identified or identifiable person, or personal data rendered anonymous in such a manner that the data subject is not or no longer identifiable. However, distinguishing personal from anonymous data lacks standardized parameters. Unfortunately, GDPR and other European regulations, have so far not provided a binding list of items for distinguishing anonymous and personal data, leaving it for each member state to identify them individually.

4.2. Data processing in the context of the EURACAN registry

Firstly, each Participating Center (PC) shall create its own internal database, where Personal and Special categories of data will be pooled by extracting them from the original clinical sources, such as, for example, medical records, reports of medical interventions, imaging and pathological anatomy reports. Each PC shall identify and manage its own data. These data, gathered into each repository, will be elaborated through Vantage6, without any transfer outside the perimeter of each PC, which continue to maintain control of the data and any copy of them shall be made. No one except the PC will have the access to the data contained into each single internal database.

Vantage6 will elaborate data contained inside each single repository upon an express query (question) created by an authorized researcher (which be expressly appointed by his/her related PC pursuant to art. 29 of the GDPR). A template for the appointment of that person is distributed and attached to the Memorandum for the processing of personal and anonymous Data in the context of the EURACAN registry.

Upon these queries, Vantage6 elaborates a result to the various participating or authorized parties. This result will be an output, which consists only of aggregated and anonymous data. The anonymity of this output means that it will not be likely to re-identify the data subject through the use of reasonable means, by any of the PCs. Considering that the European legal framework does not provide unique criteria, in order to state and prove the anonymous nature of data, the Coordinating Center of the EURACAN Registry (INT for the purposes of Blueberry) elaborated a document compliant with criteria set forth by Italian legislation. The Coordinating Center will provide the above-described document about the anonymity of the output, to facilitate each PC to carry out their own assessments. Only this output can be the result of the query formulated by the PCs for reaching the scientific aims identified into the single Data Protection Impact Assessment. Likewise, results of these studies published in scientific magazines will be anonymous as well. The above mentioned legal and technical framework of the EURACAN Registry will be illustrated into the body of a Memorandum of Understanding, signed between all the PCs.

Role of Participating Centers: Autonomous Controller

It is crucial to highlight the “privacy roles” of PCs for personal data processing in the context of the EURACAN Registry; the role of each PC will be Autonomous Controller (AC). As data Controller, each PC shall identify, in compliance with its Member State legislation, the most suitable legal basis for processing personal data and carry out a Data Protection Impact Assessment (DPIA) of the data processing in the context of the EURACAN project, coherently with the obligations set forth into the GDPR. To comply with the European and MS legislation, each AC shall be responsible for demonstrating the anonymity of an output according to its MS criteria. The coordinating centre of the EURACAN registry will provide each PC with a DPIA template and supplementary materials on anonymity of data to facilitate local assessments in English.

Legal agreement for the EURACAN registry

Form of the Agreement: a multilateral agreement for the processing of personal and anonymous Data in the context of the EURACAN Registry

Data Sharing Agreements (DSA) or Data Transfer Agreements (DTA) are the most commonly used documentation in use to define collaborative multi-centre or cross border partnerships for the secondary use of health data. However, as the legal agreement for the EURACAN registry does not control data transfer or sharing, but rather the functioning of the technical architecture, the roles of the PCs and their commitments to respect obligations set forth in GDPR and MS legislation in the creation of their own internal repository, neither a DSA nor a DTA are adequate models of agreement. The innovative structure of the EURACAN Registry and the anonymity of the output of a query necessitates adopting a model of agreement that can guarantee flexibility to define the terms of the partnership: goals, roles, governance of the EURACAN registry, criteria for elaborating queries and risk assignment. A Data Processing Agreement (DPA), which is usually adopted in order to define roles and obligations between a Data Controller and a Data processor, is the basis for the Agreement for the EURACAN registry.

Main clauses of the Memorandum for the functioning of the EURACAN registry and its annexes

Every agreement contains several clauses which are essential to ensure the achievement of the agreed upon objectives between all the parties involved, without ambiguity. On the other hand, the “Blueberry Data Processing Agreement in the context of the EURACAN Registry” has some unique features, which will reflect the features of the Registry. The agreement contains the follow:

- A description of the EURACAN registry (Federated Architecture and Vantage6 software) and its governance;
- Description of the Data Processing in the context of EURACAN: roles, duties and rights of the Parties;
- Anonymous and aggregated Output of the single query and its guarantee by each PC;
- Intellectual Property. “Blueberry Data Processing Agreement in the context of the EURACAN Registry” shall not transfer, convey, or assign any rights in Background Intellectual Property from one party to the other party except as provided under separate written license agreements between the Parties involved;
- Confidentiality. A standard confidentiality clause shall be adopted;

Every agreement needs annexes, which constitute an essential part of them. Annexes of the “Blueberry Data Processing Agreement in the context of the EURACAN Registry” will be:

1. A technical description and feature of Vantage6;
2. EURACAN Registry Governance;
3. Results Access Form, in order to allow to a third Party not belonging to the EURACAN Registry Working Group the elaboration of a Task to Vantage6 upon the presentation of a Study Protocol to the EURACAN Registry Committee;
4. Accession form. Through this document, a third Party can access to the of the EURACAN Registry and so guarantee the scalability of the Registry;

Legal Board for the negotiation

Due to the involvement of many Parties, negotiations are an expected part of the ratification process. Therefore, a legal board may be assembled should a significant issue regarding the agreement arise. Each PC will be responsible for the appointment of a representative, duly authorized to negotiate the agreement and in charge for completing the signature process in the name of its Institution.

5. The future and ways forward

5.1. The current context

In the previous chapters we discussed several challenges related to using real-world data. However, the importance of harnessing the potential of real-world data for rare cancers is not in question.

In Europe, The European Health Data Space (EHDS) is a key pillar of the strong European Health Union and is the first common EU data space to address the challenge to unlock, share and (re)use health data.

The EHDS will:

- empower individuals to take control of their health data and facilitate the exchange of data for the delivery of healthcare across the EU (primary use of data)
- foster a genuine single market for electronic health record systems
- provide a consistent, trustworthy, and efficient system for reusing health data for research, innovation, policy-making, and regulatory activities (secondary use of data)

By doing so, the EHDS will enable the EU to fully benefit from the potential offered by a safe and secure exchange, use and reuse of health data to benefit patients, researchers, innovators, and regulators.

In spring 2024, the European Parliament and the Council reached a political agreement on the Commission proposal for the EHDS and the timetable from its adoption to its implementation for *all* intended uses provisionally reaches 2036.

5.2. Strengthen the EURACAN registry

Against the background presented, it is clear that while working towards innovative solutions for automatic data reuse, intensive manual data collection work is currently still required to process real-world data. However, manual data collection is expected to cease once most healthcare system processes are fully automated and common standards in clinical reporting have been established.

The EURACAN registry collecting and organizing real-world data in the immediate future, will help to shape the EHDS as it provides a challenging use case, where usage of pan-European data is essential.

To adequately meet the need for data now and not tomorrow, it is necessary to quickly ensure the sustainability of the registry without interrupting its activities. In this sense, as part of BlueBerry, we found it more effective to leverage existing organizations rather than develop a completely new legal entity dedicated to the registry.

- Existing organizations may be capable of supporting the EURACAN Registry. A new organization creates overhead and adds to the complexity of the health data landscape.
- Functions from an existing organization (HR, IT landscape, legal team, staff) can be leveraged by the EURACAN Registry.
- Sustainable funding for the EURACAN Registry is unsecured. The registry may better grow in an existing operation as an incubator, where functions can be gradually integrated.

A new legal entity to develop an innovative registry would require many investments, multidisciplinary skills, dedicated staff as well as a lot of time to build and expand it, gaining the full trust of EURACAN partners.

5.3. Capability assessment of existing organisation

Three organizations were considered relevant to fully support and/or manage the registry: the European Organization for Research and Treatment of Cancer (EORTC), DigiCore and the European Organization of Cancer Institutes (EOCI). These organizations were identified for their expertise in managing collaborative research studies (OECI), including clinical trials (EORTC), managing large networks and partnerships (EORTC, OECI), reusing real-world data (DigiCore), for their European mandate, for their proven financial stability (EORTC, OECI), for their focus on the cancer sector (EORTC, OECI, DigiCore) and for their innovative component (DigiCore). Finally, the role of INT, the current coordinator of the EURACAN registry, was also considered vis a vis these organizations.

After the first introductory meetings, it became clear that OECI was not aligned and interested in continuing the discussion on the registry. Therefore, only EORTC, DigiCore and INT were assessed to understand their capability to maintain and scale the registry (capability assessment). The box below summarizes, for each of the 3 organisations, advantages and disadvantages for each of the elements included in the capability assessment framework: data platform, competent organisation, data usage, governance structure and sustainable funding.

	Advantages			**Disadvantages**		
	EORTC	**DIGICORE**	**INT**	**EORTC**	**DIGICORE**	**INT**
Data platform		Alignment with EURACAN registry innovative tools and infrastructure Expertise with federated data approaches	Experience in developing and managing a federated data landscape	Lack of recent experience with federated data infrastructure Experience in centralised infrastructures		

Competent organisation	Experience in developing scalable data products Understanding of the complexity of international medical data exchange	Innovative technological partnership In-house expertise in hospital IT	Deep understanding of hospital complexity and data sources In-house legal and compliance expertise	Limited experience in legal compliance in a decentralised data environment Mainly focus on clinical trials	Limited experience in the domain and research outcomes (DigiCore is still under development) Commercial interest impacts trust from the medical community	Limited experience in scaling-up products internationally Organisational model misalignment for future needs
Data usage, compliance, privacy	Experience in data exchange Dedicated data expert employed	Experience in improving data quality and in innovative solutions for automatic data processing Experience in regulatory studies	Experience in data processing and data exchange	Limited experience data quality in a decentralised data environment		
Governance structure, oversight and steering	Utilisation of existing networks and partnership Clear Governance in place	Public-Private partnership diversity	Fully integrated in the EURACAN Network Clear governance in place		Suboptimal network coverage	
Sustainable funding	Stable fundraising and independent revenue stream				Funding stability issue	Project-based set up, lacks long-term sustainability

5.4. Future scenarios

Leveraging EORTC existing capabilities

The EORTC has strong data-driven centralised platform expertise with a well-established policy framework to ensure compliance with existing rules and regulations and a robust governance model. More importantly, the organization has demonstrated financial stability throughout its 60 years of existence with significant steps towards financial independence allowing the EORTC to independently invest in relevant projects and innovation. However, the organization's primary interests are **clinical trials** and it has no experience with data models and federated infrastructure. Considering 1) the important role of registries and real-world data for innovative clinical trial design and 2) European Medicine Agency's interest and investment in registries for regulatory purposes, the EORTC may be interested to the EURACAN registry, although investments will be needed to develop the skills and technologies needed to work with decentralized data sources.

This creates opportunities for the EURACAN registry to be part of the EORTC's future investments.

Leveraging DigiCore technological capabilities

DigiCore is a public-private partnership, bridging the gap between different sectors and bringing together diverse expertise to strengthen the use of digital advancements in the cancer research landscape. However, DigiCore is a new initiative, whose recognition and respect within the international research community is still in its formative stage. Furthermore, DigiCore has a high level of dependence on IQVIA which, in addition to representing a problem in itself in terms of funding sustainability and independence, creates resistance within the medical community in sharing data. The prioritization of activities of DigiCore may therefore not always be scientifically driven.

DigiCore's technologies and expertise may prove to be a key factor in the future of the EURACAN Registry and the EURACAN network is of potential interest for DigiCore to expand its partnership and networks. However, to turn this collaboration into a win-win situation, it will be important to adequately balance corporate interests with those of EURACAN. Expanding the partnership to other private companies and developing strong governance that keeps rare cancers and the ultimate benefit of patients and doctors as a priority could help leverage DigiCore's expertise for the registry. However, this can be a more challenging time-consuming approach. Trust should be built, where clinicians in the current DigiCore network may be able to take a key role.

Leveraging INT and other EURACAN members

A completely different scenario would be to rethink the EURACAN registry as a federation of registries where a registry corresponds to a EURACAN rare tumour domain/family. Each of these registries should be managed by its domain leader. The INT is currently coordinating the sarcoma and head and neck cancer registry also because EURACAN's leaders of tomorrow dedicated to sarcomas and head and neck cancers are INT clinical experts. The Curie Institute is coordinating the thymic tumour registry because the EURACAN Rare Thoracic Tumours domain leader is from the Curie Institute. Multiple leaders may provide better clinical leadership and incentives than a single coordinator. Furthermore, the burden related to the activities necessary for the implementation of the registry could be shared between several institutions. This decentralized vision would require the development of a strong centralized governance that leverages the current one of the EURACAN registry. The central steering/governance board will be responsible for coordination, including defining the central services needed for the registry and/or allocating dedicated EU funding for the registry. Core services could be provided by specific external service providers (leveraging those currently working with INT) or could be a topic of discussion with EORTC and DigiCore.

Continue and scale with INT

Over the past years, INT has taken on the significant responsibility of managing and developing the registry making a substantial impact in creating a federated data infrastructure, a legal and governance framework which serve as the back-bone for the registry. Furthermore, INT has adopted internationally recognised models such as OMOP to facilitate international data exchange. Most importantly, INT showed a clear commitment in advancing research on rare cancers demonstrated by health-care and clinical leadership. The hospital's research leadership is also evident in its ability to operate on an international scale. Forming collaborative alliances with other research institutions and organisations worldwide.

However, the project is still in its early stages and major challenges remain. To properly fulfill the role of data exchange facilitator on the international stage, several changed would be required including:

- a significant consolidation and reinforcement of the current data management and privacy commitment
- the integration of EURACAN activities within the hospital governance and priorities to properly balance the new role of INT at European level with the existing responsibilities of the hospital
- re-allocation of funds with a committed and long-term financial outlook to cover at least ongoing operational costs and transfer operational activities from scientific staff to a dedicated team.

These are critical challenges to face. Anyway, INT could play an important role in scenario number 3.

5.5. Founding principles of the registry

Regardless of the scenario, the experience of recent years has helped to define important principles that must always be considered:

- The governance of the registry should also include a clear role in terms of decision-making for EURACAN members.
- The registry, although currently based on intensive manual data collection, will need to continue to help define, invest in and test innovative solutions for automatic data reuse.
- The registry will continue to contribute to the data harmonization initiated under the coordination of INT. It will be important to provide input to the OMOP oncology model by finding the right balance between standardization and the necessary clinical details.
- Registry data should remain FAIR.

This will make the registry the best contributor to the implementation of the EHDS.

5.6. Way forward

Presentation of the BlueBerry project to the EURACAN Registry SC and to the EURACAN Board

Definition of a SC to initiate discussion with EORTC and DigiCore based on specific use cases

- The collaboration with the EORTC could be structured with a phased approach starting with the implementation of a new registry for a domain not managed by INT
- Discussions are ongoing with DigiCore for a possible collaboration on head and neck cancers. Engagement with DigiCore could focus on the head and neck cancer registry as a way to help shape the future of collaboration between EURACAN and DigiCore.

Appendix 1 - EURACAN registry capability framework.

Theme	No	Capability	Description of capability
Data platform Management	1.1	Robustness and stability	● Proven track record of high uptime and reliability for data platforms ● Experienced in performance optimization techniques (e.g. caching, indexing, parallel processing) ● Ensure fast query response times
	1.2	Data integration capabilities	● Manages and processes data across multiple decentralized nodes without centralized storage ● Supports a range of federated learning algorithms ● Facilitates registry growth and research on available data
	1.3	Interoperability and standards compliance	● Experienced with standard data exchange protocols, data models, ETLs/APIs ● Enables integration with external data sources (Electronic health records, population databases, laboratory systems, research databases) ● Works with diverse data types and formats (i.e. structured, semi-structured, and unstructured)
	1.4	Scalability and performance	● Experienced in building and maintaining scalable architecture ● Handles large data volumes and concurrent user requests ● Scales horizontally and vertically for growing data and user demands
	1.5	Experience with standardized data models (and in particular OMOP CDM)	● Expertise in implementing and utilizing FAIR principles ● Proficient in data standardization and interoperability ● Demonstrated experience with OMOP CDM ● Collects and maps diverse healthcare data into a common format ● Ensures data consistency, interoperability, and comparability
Competent organization	2.1	Research and innovation	● Strong commitment to continuous learning, research, and innovation in data science, machine learning, healthcare analytics ● Engages in collaborative research and scientific dissemination (publications, conferences)

	2.2	Legal, compliance and privacy expertise	● Experienced in legal complexities of healthcare data governance (consent management, DSAs, cross-border data transfer) ● In-depth understanding of data protection regulations and privacy laws (GDPR, EHDS) ● Expertise in collection, storage, processing, and sharing of healthcare data
	2.3	Domain knowledge	● Deep understanding of the healthcare landscape (clinical workflows, medical terminology, disease classifications, etc) ● Experienced staff with background in EHRs, healthcare data and patient registries
	2.4	Collaborative approach	● Proven track record in effective collaboration with diverse stakeholders (HCPs, researchers, PAGs, policymakers, technology vendors, regulatory agencies etc) ● Fosters productive partnerships ● Drives collective action toward shared goals
	2.5	Experience in data-driven projects	● Expertise in federated data platforms, data science, machine learning, AI ● Proficient in programming languages (e.g., Python, R) ● Experienced with data analytics frameworks
Data usage, compliance and privacy	3.1	Compliance with laws and regulations	● Fully compliant with data protection regulations (e.g., GDPR, AI Act, EHDS) ● Adheres to regional, national, and international laws ● Regular compliance audits and assessments of policies and controls
	3.2	Data usage	● Experienced in developing and implementing data usage policies ● Policies outline permissible and prohibited data uses ● Aligned with ethical standards and regulations ● Monitors and enforces policies to prevent misuse and unauthorized access
	3.3	Data exchange	● Experienced in international medical data exchange ● Rigorously selects and manages third-party data processors and collaborators ● Implements contractual safeguards for data protection and privacy compliance
	3.4	Data quality management	● Strong data quality management processes ● Ensures accuracy, completeness, consistency, and reliability

			● Tracks data lineage for integrity and accountability
Governance structure, oversight and steering	4.1	Transparent decision-making process	● Clear and transparent decision-making processes ● Facilitates consensus-building and effective governance ● Works within complex international stakeholder fields ● Builds on existing criteria, roles, and responsibilities ● Ensures accountability and transparency
	4.2	Representative governance bodies	● Organizes and integrates diverse stakeholder representation ● Balances various interests of stakeholders effectively
	4.3	System of governance and accountability	● Builds on strong accountability mechanisms ● Includes board of directors, internal controls, compliance oversight ● Conducts regular reviews and evaluations ● Assesses governance structures for improvement and continuous learning
	4.4	Board oversight and supervision	● Organizes ethical oversight mechanisms ● Manages risks and technological improvements ● Ensures sound data governance and management ● Complies with data protection, privacy, and confidentiality regulations
	4.5	Scientific ownership and fair acknowledgement	● Strong commitment to respecting scientific ownership and intellectual property ● Establishes fair and transparent guidelines for crediting and acknowledging contributions in publications, citations, and dissemination activities
Sustainable funding and capital management	5.1	Financial stability	● Demonstrates long-term financial stability ● Stable revenue and funding streams ● Strong financial statements ● Proven track record of effective financial management and project sustainability
	5.2	Sustainable funding model	● Develops sustainable revenue models aligned with long-term goals ● Secures and manages diversified funding sources (e.g., grants, donations, investments)

			● Experienced in applying for and managing competitive grants from various entities
	5.3	Transparency and accountability in financial reporting	● Strong commitment to transparent financial reporting ● Provides clear, accurate, and timely financial information ● Regularly publishes financial reports, including annual and audited statements
	5.4	Capital management practices	● Comprehensive long-term financial plans for maintenance and growth ● Adopts prudent capital management practices ● Manages working capital, cash flow, and investments effectively ● Ensures liquidity and financial flexibility

Appendix 2 - GDPR and Blueberry

1.1 RELEVANT EUROPEAN AND NATIONAL LEGISLATIONS

The Reg. (EU) 679/2016, also known as General Data Protection Regulation (hereinafter "GDPR"), is one of the strongest global privacy laws in effect today and the main European act which aims to regulate the processing of personal data and the free movement of such data. GDPR has replaced the previous Directive 95/46/EC, which outlined a static model of personal data processing, now outdated. Therefore, the GDPR was born with the purpose to simplify and consolidate the experiences gained in the European context, aiming to defeat the fragmentation of the legislation on personal data protection in the European Union and the widespread legal uncertainty concerning its application. Since the need was to ensure a homogeneous application of the legislation in force, the European lawmaker chose the legal instrument of the regulation that, unlike the directive, is directly applicable to EU Member States (hereinafter also "MS") according to Art. 288 of the Treaty on the Functioning of the European Union (TFEU).

But, despite its direct enforcement, the GDPR leaves, through its general provisions and principles, a certain degree of discretion to Member States for certain topics. For example, Article 9(4) of the GDPR (about the "processing of special categories of personal data") and Article 89(2) (about the "processing for archiving purposes in the public interest, scientific or historical research purposes or statistical purposes") are two clear representations of the possibility granted to the Member States to introduce restrictions, limitations and derogations to the EU provisions. In fact, these two provisions allow us to acknowledge how each Member State's legislation, with reference to processing of health data and for scientific research purposes, could be different towards each other. Even the "*Document on responses to the request from the European Commission for clarifications on the consistent application of the GDPR focusing on health research*", issued by the European Data Protection Board (hereinafter also "EDPB") on 2nd of February 2021, par. 13 and 14, highlights the difference and the impact that each Member State law could have on the harmonization level. In fact, EDPB explains: *"choices made in MS laws can have a considerable impact both on the legal basis (Article 6) and on the exemption for processing of health data (Article 9) that must be relied on when processing personal (health) data for scientific research purposes. Therefore, choices made in Member States' law can have a serious impact on the level of harmonization that can be achieved under GDPR in the domain of processing personal health data for scientific research purposes. In addition, the possibility foreseen in Article 9(4) GDPR for MS to maintain or introduce further conditions, including limitations, with regard to the processing of genetic data, biometric data or data concerning health, should be taken into account. Even though, as yet, there is no complete and detailed overview of relevant MS laws on the processing of health data, it can be observed that in Member State laws considerable differences can be found in legal bases for processing health data for scientific research purposes are either specified, prescribed or excluded and whether an exemption on Article 9(1) based on Article 9(2)(g), (i) or (j) GDPR has been foreseen (with additional requirements) in Member State law.".*

It is very important to be aware of these differences among MSs' legislations in order to understand the complexity of a cross-border processing of personal data, especially in the context of special categories of data, like data related to health or genetic data, such as in the Blueberry project.

In addition to this fragmented and uncertain legal framework, it is worth noticing that the technological scenario is not yet coordinated as well, so it is difficult so far to find adequate technological means supporting the data processing for research purposes.

Lastly, the European legal framework is enriched by guidelines, opinions and documents issued by some independent European Authorities (such as the European Data Protection Board and the European Data Protection Supervisor), as well as by the decisions of the European Court of Justice, which are useful in order to interpret the rules and obligations of the GDPR. The main references which have been compared and analyzed are:

- Article 29 Data Protection Working Party[5], *Opinion n. 4/2007 on personal data*, adopted on 20 June 2007;
- Article 29 Data Protection Working Party, *Opinion n. 5/2014 on anonymisation technique*, adopted on 10 April 2014;
- EDPS, *A preliminary Opinion on data protection and scientific research*, adopted on 6 January 2020;
- EDPB, *Guidelines 05/2020 on consent under Regulation 2016/679*, adopted on 4 may 2020;
- EDPB, *Document on response to the request from the European Commission for clarifications on the consistent application of the GDPR, focusing on health research*, adopted on 2 February 2021;
- ENISA, *Opinion on deploying pseudonymisation techniques (the case of the health sector)*, adopted on March 2022;

1.1.1 Reuse of clinical data for research and innovation

The processing of personal data for scientific research can be grounded upon several conditions of lawfulness, established in Article 9(2) of the GDPR. Among these, the most relevant ones are:

- the consent of data subjects (Art. 9(2)lit. a);
- the pursuing of a substantial public interest, on the basis of Union or Member State law (Art. 9(2)lit. g);
- the pursuing of a scientific research purpose on the basis of Union or Member State law (Art. 9(2)lit. j).

In addition, another possibility lies in the so-called "Compatibility Assessment", that will be described below.

1.1.1.1. Role of Consent

As it is widely known, the data subject's consent is the main and most used condition for the lawful processing of personal data.

Pursuant to Recital no. 11 of the GDPR, consent of a data subject must be a freely given, specific, informed and unambiguous indication of the data subject's wishes, by which he/she signifies agreement to the processing of personal data relating to him or her.

[5] The Art. 29 Working Party was the previous independent body which gathered MSs' Data Protection supervisory authorities, now replaced by the EDPB.

Even if consent is one of the possible legitimate conditions for processing of special categories of personal data (including health data), according to Art. 9(2)(a) of the GDPR, there are cases where seeking this consent reveals to be hard, such as in the context of scientific research.

Firstly, since scientific research is often grounded on the analysis of data which were previously collected, even many years ago, it can be very difficult, or even impossible, to seek data subjects' consent for further processing of data. For example, these data subjects could be in the meanwhile deceased or unable to be reached.

Secondly, it is not always easy or even possible to know the specific investigation objectives from the very beginning and this circumstance is deemed as hindering the respect of the "Specification" requirement of the consent. In fact, consent expressed in a non-specific way, which means it has been acquired with regard to general purposes, is not valid.

However, it is worth noticing that Recital no. 33 of the GDPR offers a possibility to mitigate this requirement of the "Specification", so that consent could be valid as a legal basis for the processing also in those cases where it is not possible to fully identify the purpose of personal data processing for scientific research purposes at the time of data collection. In these events, the concept of "broad consent", which derives from the interpretation of the Recital no. 33 of GDPR, allows to acquire consent "only to certain areas of research or parts of research projects" and to describe at a more general level the types of research questions and/or fields of research to be explored, as long as the Data Controller seeks other ways to ensure the essence of the consent requirements are served best.

As explained by the EDPS and the EDPB, this provision shall not take precedence over the conditions for consent set out in Articles 4(11), 6(1)(a), 7 and 9(2)(a) of the GDPR and, especially when special categories of data are processed, applying the flexible approach of Recital 33 will be subject to a stricter interpretation and requires a high degree of scrutiny (see, *EDPB's Guidelines n. 05/2020*, par. 157). Besides, even though Recital 33 offers some room for flexibility in describing the research purposes for which consent is obtained from the data subject, the requirements in Article 5 GDPR still have to be met in order for the processing to be lawful, fair and transparent.

Even if this provision seems to represent a great opportunity to process personal data, it has not being applied so far due to some strict interpretations which have also impacted on the usability of the "broad consent" approach in Member States.

A clear example of how the national legislation and its interpretation by the national Data Protection Authorities could be impactful on the European system is the case of the Italian Data Protection Authority (hereinafter also "DPA") interpretation about the Recital no. 33 of the GDPR.

In the context of a prior consultation on the processing of health data for scientific purposes, Italian DPA has adopted a "layered consent approach", according to which data subjects should be allowed to give their consent to certain areas of scientific research purposes at the time of data collection. Then, as the research advances, consent for subsequent steps in the project shall be obtained before that next stage begins.

In other words, Italian DPA expressly requires seeking specific consent for the realization of further and future research projects, which are not defined at the time of the collection. Due to this strict approach, in Italy nowadays the Italian DPA requires a primary consent for data processing at the time of collection and, then, other specific consents for every research project which may use those data.

1.2.1.2. The Purpose of Public Interest or Scientific Research

As described above, the reuse of clinical data appears to be hindered by some practical obstacles, like in the case of consent.

The current framework allows data processing for scientific research purposes without the need of seeking consent when the purpose of scientific research is provided by Union or Member State law (Art. 9(2) lit. j) or when the scientific research represents a public interest provided by Union or Member State law as well (Art. 9(2)lit. g).

However, as it is widely known, so far such laws have yet to be adopted. Besides, Member States' laws are very different from each other, so it is difficult to find a valid legal basis especially in the case of cross-border processing.

1.2.1.3. The Compatibility Assessment

Regardless of the Member States laws that, as explained above, could differently regulate this subject, the reuse of clinical data could be based on the so-called "Compatibility Assessment", according to Article 6(4) and Recital no. 50 of the GDPR.

The "Compatibility Assessment" is a new mechanism introduced by Art. 6(4) of the GDPR which entails an evaluation of compatibility in order to ascertain whether processing for another purpose is compatible with the purpose for which the personal data are initially collected. In such case, no legal basis, different from that which allowed the original collection of personal data, is required.

In order to conduct this assessment, Article 6(4) establishes criteria for determining the compatibility of further or secondary use of personal data: for example, Data Controllers shall consider "*any link between the purposes for which the personal data have been collected and the purposes of the intended further processing*; *the context in which personal data have been collected; the nature of the personal data, (…); the possible consequences of the intended further processing for data subjects; the existence of appropriate safeguards, which may include encryption or Pseudonymization*".

With a specific reference to the reuse of data for research purposes, GDPR has also envisioned a "presumption of compatibility" (article 5(1)(b)), according to which: *"further processing for archiving purposes in the public interest, scientific or historical research purposes or statistical purposes shall, in accordance with Article 89(1), not be considered to be incompatible with the initial purposes".*

However, this presumption is not a general authorization to further process data in all cases. Each case must be considered on its own merits and circumstances, as pointed out by the Article 29 Working Party in its *Opinion 03/2013*. Data collected in the commercial or healthcare context, for example, may be further used for scientific research purposes, by the original or a new controller, if appropriate safeguards are in place.

Besides, these provisions are hardly applied so far, since there are no guiding criteria or official interpretation by the authorities. For example, EDPS and EDPB talk about them ambiguously, as appears in the already mentioned "*EDPB Document on response to the request from the European Commission for clarifications on the consistent application of the GDPR, focusing on health research*", adopted on 2 February 2021, and "*A Preliminary Opinion on data protection and scientific research*", issued by EDPS on 6 January 2020. In these documents, the Authorities notice that, despite the above-mentioned Recital no. 50 states that when personal data are used for secondary compatible purposes no further legal basis is required, this Recital is not accompanied by a specific provision in the body of the GDPR.

Therefore, in the EDPS' opinion, this provision doesn't appear an exemption to the separate steps set out in the Article 8(2), but rather advisory. For this reason, in order to guarantee respect for the rights of data subject, the compatibility test under article 6(4) should still be considered prior to the reuse of data for the purposes of scientific research, especially where the data was originally collected for very different purposes or outside the area of scientific research.

Examining the reuse of personal data with a future outlook, two great opportunities at European level could be represented by the "Data Altruism", a new mechanism introduced by the Regulation n. 868/2022 ("Data Governance Act" or DGA), and by the "European Health Data Space" which aims to create a framework where health data could freely circulate and be processed for general public interest purposes. These two innovative scenarios are examined below.

1.2.2. Patient-level data sharing

Currently, European legislation does not foresee mechanisms for the "donation of personal data". Patients can only decide if they accept or not accept the processing of his/her own data in the context of specific research projects.

About the reuse of clinical data, a great opportunity for the future could be represented by the new mechanism introduced by the "Data Governance Act" (DGA), the "Data Altruism". Even if it is not operational yet, Data Altruism may be a mechanism that could help the free movement of data through the voluntary sharing of personal data donated by individuals in order to achieve the objectives of general interest, such as healthcare and scientific research purposes.

The DGA provides for a European data altruism consent form, in order to facilitate the collection of data based on data altruism. The form shall allow the collection of consent across Member States in a uniform format, by using a modular approach allowing customization for specific sectors and for different purposes. However, this form has not been issued yet. Besides, nonprofit organizations will have the opportunity to sign up into a public register of "data altruism organization".

In addition, there is an important legal draft on the table of the European lawmaker. This proposal of Regulation, so-called "European Health Data Space" (hereinafter "EHDS"), is part of the European data strategy, where the establishment of common European domain-specific data spaces was firstly advanced.

EHDS is the first of these European spaces and moves from the need to address the challenges related to access and sharing of electronic health data, the need for which has been made particularly evident during the pandemic period.

The proposal specifically regulates the secondary use of data, which were previously collected in the context of healthcare or assistance care or for purposes of public health, research, innovation and so on.

The proposal identifies the legitimate purposes for which data can be accessed, distinguishing them from the prohibited purposes.

1.1.2. Anonymous data

The above-described problems of seeking an adequate condition of lawfulness cease to exist when the processed data are anonymous data. In fact, as it is stated in Recital no. 26, GDPR does not apply to anonymous information.

"Anonymous information" is information which does not relate to an identified or identifiable natural person (namely, data subject) or personal data rendered anonymous in such manner that the data subject is not or no longer identifiable. Whereas pseudonymized data – to which GDPR applies – are personal data which have been processed in such a manner that can no longer be attributed to a specific data subject without the use of additional information, provided that such additional information is kept separately and is subject to technical and organisational measures to ensure that the personal data are not attributed to an identified or identifiable natural person.

Given this definition, the critical point is how to distinguish between personal data and anonymous data. The current framework does not provide for specific parameters or standards, which can be compared in order to assess the anonymous or personal nature of data. This assessment shall be carried out by the Data Controller under the accountability principle, which establishes that the controller shall be responsible for, and be able to demonstrate compliance with, the GDPR. Therefore, this assessment entails an evaluation of the risks to which personal data and data subjects could be exposed and for which the data controller is the exclusive responsible.

The key element which allows to distinguish between personal data and anonymous data is the possibility to attribute and link information to an identified or identifiable natural person. But, as is widely known, the concept of anonymous information is not absolute. In fact, the anonymous or personal nature of data depends on several items related to time, technical means, economic and human resources, which can be used in order to identify data subjects or to trace data subjects' identify from anonymous information back**.** Hence, information could be anonymous for some, and not for others.

Considering that absolute anonymity cannot be guaranteed, those elements shall be considered in light of their likelihood to link information, in order to verify whether and when information could be considered connectable to another and, therefore, when we may consider an information as "anonymous" or personal. This criterion seems to be the "reasonableness" as a parameter through how we could evaluate costs, resources (technical, economical and human), efforts, time and other sources available in order to re-identify anonymous data.

Unfortunately, as above mentioned, GDPR or other European act have not provided so far a binding list of items that we shall evaluate in order to distinguish between anonymous data and personal data. GDPR leaves to each Member State the opportunity to identify them on its own.

In Italy, for example, in the "Rules for Conduct on processing for scientific research purposes" (Art. 4), Italian DPA describes the means that can be reasonably used to identify a data subject referring to, in particular: economic resources; archives of names or other sources of containing identification data together with a subset of the variables to be communicated or disseminated; archives, even non-nominative ones, which provide further information beyond data which is already the subject of communication or dissemination; hardware and software resources which are necessary in order to link non nominative information to an identified subject, also taking into account the possibility of illegally obtaining his/her identification in relation to the security systems and control software adopted; the knowledge of the sampling extraction, imputation, correction and statistical protection procedures adopted for the production of data.

Once again, this shows how much each Member State's law may differ from other MSs' legislations and how complex is to find a solution that can be deemed valid for all Member States.

About this specific issue, it is worth to mention a recent judgement issued by the European Union General Court, in the lawsuit T-557/20 *Single Resolution Board (SRB) v. European Data Protection Supervisor (EDPS)*, where a new and wider interpretation has been given to the concepts of anonymous and pseudonymized data. Without going into detail, the main object of debate concerned the qualification of data as personal, and not anonymous, on the sole condition that someone held a list of additional information which, if combined with the information communicated to a third party, would have allowed the reidentification of data subjects. However, this latter party had not any reasonable means to acquire this list, so it demanded that data should have been considered anonymous. According to the European judges, "in order to determine whether the information transmitted to [a party] constituted personal data, it is *necessary to put oneself in [that party's] position* in order to determine whether the information transmitted to it relates to 'identifiable persons'". In other words, to establish whether data are personal or not and, therefore, whether it is possible to trace back to a data subject, it is not a sufficient condition that the additional information (which can be used to re-identify) is held or in the possession of a third party. To establish whether the information constitutes personal data, it is necessary to look from the point of view of the recipient of data and evaluate whether the possibility of combining the information transmitted with any additional information in the possession of the third party constitutes a reasonably feasible means of identifying data subjects.

The case here examined gives a more flexible interpretation of anonymous data and is certainly noteworthy for the potential implications it could have in the healthcare and scientific research sector, where universal and uniform criteria and standards are increasingly invoked to define when data is anonymous or not.

1.2 COUNTRY SPECIFIC CONTEXT AND EXPERIENCE: SURVEY

Blueberry is a collaboration of nine partners located in seven different countries: Austria, Italy, France, the Netherlands, Norway, Spain and Sweden. Data from the medical centers and other data sources involved will be made available with the aim of gaining insights from the collective data of the partners. The (open source) software vantage6 will be applied, through which insights can be obtained from data without actually merging the data (the functioning of the model will be explained further on).

Since the data processing takes place within the context of activities of establishments of controllers located in the European Economic Area (EEA), the General Data Protection Regulation (GDPR) applies. In addition, national laws and regulations are relevant, since the GDPR permits or requires Member States to implement national specifications or derogations on certain rules set out in the GDPR, as explained above. This means that although European data protection laws are harmonized under the GDPR, substantial national differences remain.

In order to identify these national specifications and derogations and also the challenges at the medical centers and data sources involved, a legal survey has been circulated among the consortium partners. The purposes of the survey are to:

(i) Identify risks and concerns with regard to personal data and privacy,

(ii) support data protection and privacy by design approach, and

(iii) help provide adequate guidelines and recommendations to the partners.

The findings are presented in the overview below.

	Austria	France	Italy	Nether-lands	Norway	Spain	Sweden
National laws regarding data processing for scientific research purposes?	yes, however the GDPR covers all necessary provisions	yes	yes	yes	yes	yes, however the GDPR covers all necessary provisions	yes
National laws governing centralized / federated data processing?	Yes, centralized. In the survey, the model is specified	yes, centralized: laws specific to healthcare data warehouse: strict requirements regarding centralized approach	no	no	no	no	yes, centralized. Data are stored in a "quality register". Specific rules on how to deliver data for the purpose of research
Guidelines/rules/decisions of data protection authority regarding scientific research?	yes, but not deviating from GDPR, EDPB	yes	yes	no	no	n/a	yes
Legal constraints to the possibility of communicating or transferring data among healthcare providers?	no, provided that we comply with the GDPR and no remote access to our system	yes	yes	yes	yes	yes, it is necessary to obtain a specific consent based on GDPR, as stated in its response.	yes
Lawful to require a 'unique consent', which could be deemed as valid for both the collection of data in the repository and the development of further research projects?	yes	no	no	no	no	yes	no
Which criteria are used to distinguish between anonymous data and personal data?	GDPR	GDPR, as interpreted by the Art. 29 WP	Criteria set forth by the National DPA	GDPR	Criteria set forth by the Norwegian DPS	n/a	Not specified
Joint controllers or separate controllers?	separate controllers	n/a	separate controllers	depending on governance structure (who will (jointly) determine purposes and means of data processing	joint controllers	separate controllers	separate controllers

These findings enable us to evaluate the rules of the countries involved and the interpretation of national legal protection authorities. The differences highlight that a uniform approach is not feasible and each partner will need to identify its legal basis for data processing in the context of the project. We will use these findings in order to further identify the risks and concerns with regard to personal data and privacy in the context of Blueberry and draft the relevant documents and agreements on the basis of which analyses can be performed on the data of the various partners.

1.3 CENTRALIZED VS FEDERATED APPROACH

Through the development of the Machine Learning (ML), ensuring data privacy and security have become a crucial and critical fulfillment, even due to the stringent requirements of the GDPR.

At the current state of art, a database for scientific research purposes can be created through two different technological solutions: the centralized and the federated approach. Whereas these systems

are technically examined in the previous paragraph …., in this paper, their opportunities and disadvantages will be analyzed from a general legal point of view.

The centralized approach provides for the creation of a single centrally located database, which collects and gathers all personal data from different Data Controllers. This type of approach is efficient and useful, since it allows to have a repository which contains, stores and provides secure access to great amount of data by all participating centers and may guarantee a higher data availability and a lower latency due to a reduced reliance on external systems. Also, easier management and account for data integrity and security could be a plus for adopting this system.

On the other hand, large-scale data collection, aggregation and processing at a central server not only entail the risk of severe data breaches due to single-point-of-failure, but also intensify the lack of transparency, data misuse and data abuse. Furthermore, the requirements of purpose limitations and data minimization are not always feasibly carried out.

With the federated approach, data from participating centers are temporarily, but not physically, linked in order to perform an analysis. This feature could lead to a lower risk in case of a breach. The federated system is designed in a way that does not let the service provider directly access and obtain either original training data or locally trained ML models at end-users' devices. Instead, end users, as participants in the federated system, will only send the results back to the coordination server when they are ready. With reference to the "Data Minimization" requirement (which is a challenge in the centralized system), federated system does not need to collect and process original training data; instead, a service provider only needs to gather local Machine Learning models from participants for assembling the global model.

As drawbacks, this approach may entail greater latency and lower availability of data, since its functioning depends on external systems. Besides, not every participating center has the same technical resources and assets in order to implement the federated system.

1.4 SWOT PROJECT STARTER – LEGAL ASPECTS

In order to set forth the EURACAN registry in the context of the previous project STARTER and regulate the relationship between the participants, the Coordinating Center and all the Participating Centers decided to stipulate several "Data Transfer Agreements", very similar to each other.

In this chapter, Strengths, Weaknesses, Opportunities and Threats of those DTAs will be analyzed, with the purpose to develop the next agreements in the context of the project Blueberry.

Strengths: ● Accurate description of the federated approach used to set the Registry (EURACAN) as well as the functioning of the software (VANTAGE6); ● Obligations envisioned between parties (e.g. adoption of the appropriate security measures);	**Weaknesses:** ● The Agreement for the EURACAN registry does not entail Data's flow, but the functioning of the technical architecture and how each participant can elaborate a query in order to obtain aggregated output issued by a local database owned by each Controller, thus a DTA is not adequate

● Disciplines about intellectual property and confidentiality; ● Description and functioning of the Steering Committee; ● Roles and allocation of duties and responsibilities.	because there is not any flow, transfer or communication of data between parties. All analyses will be conducted on premises and on an exclusively local level; ● For the sake of the certainty of the architecture, we cannot foresee alternatives to the federated approach ("If the federated learning will not work, the Study Data, after the validation of their quality, will be anonymized and sent to INT which will keep, manage and maintain the centralized anonymized Database"). Alternatives to the over mentioned approach shall be described in an addendum; ● Lack of accuracy regarding "studies performed outside the EU as well as by third parties" (section 1.2.); ● Lack of clearness about the notification to a relevant Regulatory Body (section 3.1); ● In the context of the BLUEBERRY project, it is not foreseen to seek the Data Subjects' consent for the disclosure of their Data, since consensus is not the only legal basis possible. Each Participating center should identify its legal basis based on the domestic legislation; ● Obligations of mutual assistance and cooperation between Parties are not coherent with the role of autonomous controller (e.g. data breach notifications, sections 3.4. and 3.5); ● Lack of a punctual description of the categories of algorithms, which will be used, and instructions to the centers, in

	order to guarantee the anonymization of outputs; ● Lack of a common Model/act for the appointment of the Processor (Software Provider, or other external entities), pursuant to Art. 28 GDPR; ● Lack of a Template for the appointment of the natural persons that will act in the name of each Center, pursuant to Art. 29 GDPR; ● Lack of Description of the aggregated and anonymized output.
Opportunities: ● Rules about Data Altruism in the Data Governance Act; ● Data sharing pursuant to European Health Data Space.	**Threats:** ● Fragmentation and uncertainty of the European and National legislation on the processing for scientific research processing; ● Stricter interpretation by the National Authorities.

1.5 ENVISIONED LEGAL FRAMEWORK

1.5.1. Legal Architecture for the EURACAN registry

1.5.1.1. Data Processing in the context of the EURACAN Registry

For the reasons above described, the EURACAN registry will be built upon a federated architecture. This technical setting reflects both the legal and data protection framework. This section provides a brief summary of the Registry and its main features.

Firstly, each Participating Center shall create its own internal database, where Personal and Special categories of Data will be pooled by extracting them from the original clinical sources, such as, for instance, medical records, reports of medical interventions, imaging (TAC or RMN diagnostic and of follow-up) and pathological anatomy reports. Obviously, this list is not mandatory and represents an exemplification, since each Participating Center shall identify and manage its own Data. These Data, gathered into each repository, will be elaborated through Vantage6, a software that allows the elaboration of Data at a local level, without any transfer outside the perimeter of each Participating Center, which continues to maintain control of the Data and any copy of them shall be made. No one

except the Participating Center will have the access to the Data contained into each single Internal Database.

Vantage6 will elaborate Data contained inside each single repository upon an express query (question) created by an authorized researcher (which be expressly appointed by his/her related Center pursuant to art. 29 of the GDPR). A template for the appointment of that person will be distributed and attached to the Memorandum for the processing of personal and anonymous Data in the context of the EURACAN registry (section 16.2.1).

Upon these queries, Vantage6 elaborates a Result to the Parties. This Result will be an output, which consists only of aggregated and anonymous data. The anonymity of this output means that it will not be likely to re-identify the Data Subject through the use of reasonable means, neither by INT nor Participating Centers with all the consequences already described (see section 1.2.3.). Considering that the European legal framework does not provide unique criteria, in order to state and proof the anonymous nature of data, the Coordinating Center of the EURACAN Registry (INT) elaborated a document compliant with criteria set forth by Italian Legislation. The Coordinating Center will provide the above described document about the anonymity of the Output, in order to facilitate each Center to carry out their own assessments.

Only this Output can be the Result upon the query formulated by the Participating Centers for reaching the scientific aims identified into the single Data Protection Impact Assessment. Likewise, results of these studies published on scientific magazines will be anonymous as well.

The above mentioned legal and technical framework of the EURACAN Registry will be illustrated into the body of a Memorandum of Understanding, signed between all the Participating Centers.

1.5.1.2 Role of Participating Centers: Autonomous Controller

Prior to moving forward, it is crucial to highlight the "privacy roles" of Participating Centers for Personal Data Processing in the context of the EURACAN Registry. Taking into account the features of the Data processing in the context of the EURACAN Registry (as well as insights obtained by Participating Centers through the Survey, as illustrated in section 1.3.), the role of each Party will be Autonomous Controller.

As Data Controller, each Participating Center shall identify, in compliance with its Member State legislation, the most suitable legal basis for processing Personal Data and carry out a Data Protection Impact Assessment (DPIA) about the Data Processing in the context of the EURACAN project, coherently with the obligations set forth into the GDPR. Considering that the Final Output of the single query elaborated through Vantage6 is deemed to be Anonymous, in order to be compliant with the European and MS legislation, each Autonomous Controller shall be responsible for demonstrating the anonymity of that Output according with its MS criteria. As Coordinating Center of the Registry and WP leader, INT will provide its Data Protection Impact Assessment and the document about the Anonymity of Data, in order to facilitate each Center to carry out their assessments. These two documents will be set up in English.

1.5.2 Agreement for the EURACAN registry

1.5.2.1. Form of the Agreement: a multilateral agreement for the processing of personal and anonymous Data in the context of the EURACAN Registry

Considering that partnership projects may take various forms, several types of legal partnership agreements have been adopted by Institutions and hospitals in order to discipline the peer to peer and

cross border partnership for the secondary use of Health Data. The most used forms are either Data Sharing Agreement (DSA) or Data Transfer Agreement (DTA). Nonetheless, as anticipated in the previous section, the Agreement for the EURACAN registry does not entail Data's flow, but the functioning of the technical architecture, roles of the Parties and their commitments to respect obligations set forth into the GDPR for the creation of their own internal repository. Therefore, neither a Data Sharing Agreement nor a Data Transfer Agreement would be the adequate models of agreement. In addition, considering the innovative structure of the EURACAN Registry and the anonymity of the output which each Participating Centers will receive as Result of the query, the main need is adopting a model of agreement that could guarantee flexibility. This feature is necessary in order to guarantee the term and details of the partnership: goals, roles, governance of the EURACAN registry, criteria for elaborating queries and risk assignment.

A Data Processing Agreement (DPA), which is usually adopted in order to define roles and obligations between a Data Controller and a Data processor seems to be the right starter model for shaping our Agreement for the EURACAN registry.

16.2.2. Main clauses of the Memorandum for the functioning of the EURACAN registry and its annexes

Every agreement contains several clauses which are essential to ensure the achievement of the agreed upon objectives between all the Parties involved, without ambiguity. On the other hand, "Blueberry Data Processing Agreement in the context of the EURACAN Registry" has some unique features, which will reflect the features of the Registry.

The aim of this section is to anticipate some of the main clauses of the final agreement. Obviously, the contract must be negotiated between the Participating Centers; therefore, the final content of the over mentioned Agreement may be slightly or greatly different. Nonetheless, the essential structure of the Agreement (as well as its annexes) can be listed as follows:

- Description of the EURACAN registry (Federated Architecture and Vantage6 software) and its governance;
- Description of the Data Processing in the context of EURACAN: roles, duties and rights of the Parties. Going into details, all the obligations that derives from GDPR as Autonomous Data Controllers of its own Personal and Special Categories of Data, among which carrying out an own Data Protection Impact Assessment (DPIA) based on a sample provided by INT;
- Anonymous and aggregated Output of the single query. Considering that the Final Output of the single query elaborated through Vantage6 is deemed to be Anonymous, each Autonomous Controller shall be responsible to demonstrate the anonymity of that Output according with its own MS criteria. Nonetheless, the Coordinating Center will provide a document about the anonymity of the Output, which could be considered a model, from which each Partner has to develop its own;
- Intellectual Property. "Blueberry Data Processing Agreement in the context of the EURACAN Registry" shall not transfer, convey, or assign any rights in Background Intellectual Property from one party to the other party except as provided under separate written license agreements between the Parties involved;
- Confidentiality. Considering the context of this collaboration, a standard confidentiality clause shall be adopted.

Every agreement needs annexes, which constitute an essential part of them. Annexes of the "Blueberry Data Processing Agreement in the context of the EURACAN Registry" will be:

1) A technical description and feature of Vantage6;

2) EURACAN Registry Governance;

3) Results Access Form, in order to allow to a third Party not belonging to the EURACAN Registry Working Group the elaboration of a Task to Vantage6 upon the presentation of a Study Protocol to the EURACAN Registry Committee;

4) Accession form. Through this document, a third Party can access the EURACAN Registry and so guarantee the scalability of the Registry.

Considering the role of Independent Controller, following discussions and considerations, none models shall be attached to the agreement, but they will be provided by INT in the context of the legal board, as described in the next chapter.

1.5.2.3 Legal Board for the negotiation

Due to the involvement of many Parties, the negotiation phase may be difficult. Therefore, a legal board might sort out this significant issue. Each Participating Center will be responsible for the appointment of a representative, duly authorized to negotiate the agreement and in charge for completing the signature process in the name of its Institution. Furthermore, in the context of this legal board, INT shall provide its DPIA and the report about the anonymity of Output, without sensitive Data.

1.6. FINAL CONSIDERATIONS

It is worth keeping an eye on the development of the European regulatory framework on health data sharing and secondary use. In the future, these pieces of law could actually provide for new and common to every Member State legal basis, which could lead to easiest (but lawful and safe) ways to process personal data for purposes of scientific research. Nevertheless, nowadays this innovative framework is either still a draft or not directly applicable without the prior creation of competent bodies or the issue of templates and models.

Likewise, anonymization techniques, criteria and requirements may be the object of an ongoing monitoring and investigation. The adoption of the latest and most up-to-date techniques and criteria of evaluation should guarantee, on one hand, data subjects the best safeguards and, on the other hand, controllers to be compliant with law.

Appendix 3 - Data Sharing in Sweden

Aim of the Blueberry project in Sweden

In Sweden, the Blueberry project aimed at integrating two separate registries, the National registry for musculoskeletal sarcomas and the National registry for intra and retroperitoneal sarcomas in the proposed federated registry. Local PI was Andreas Muth, responsible for surgical sarcoma care at Sahlgrenska University Hospital, Västra Götalandsregionen (VGR) and the VGR representative to the EURACAN G1 group.

Registry governance in Sweden

In Sweden health-care is the main responsibility of the 21 regions in Sweden, and cancer quality registries are administered by different regions. Region Skåne is the legal entity and host of both Swedish sarcoma registries through the regional cancer centre (RCC syd). The national confederation of RCCs run and develop the common INCA (Information network for cancer care) platform used by the sarcoma registries. While the registries are run by the regions through the RCCs, the record keeper and the steering committés are usually composed of health care professionals not directly affiliated to the RCCs.

Work on Blueberry in Sweden

Involvement

At the start of Blueberry the registry for abdominal and retroperitoneal sarcomas had been resting and an initiative for restarting the registry from the three centres responsible for sarcoma care nationally (Karolinska in Stockholm, Sahlgrenska in Gothenburg, and Skånes Universitetssjukhus, Lund) was underway lead by Andreas Muth. Representatives from all sarcoma centres endorsed the registry solution proposed by Blueberry as a framework for restarting the abdominal and retroperitoneal sarcoma registry.

The main responsible clinician for the musculoskeletal registry (record keeper) dr Emelie Styring was became involved early in the project, and a small working group (AM+ES) was formed. In the fall of 2022, the registry for abdominal and retroperitoneal sarcomas was closed for further registration awaiting restructuring.

In December 2022 the Blueberry project was presented at the Scandinavian Sarcoma Group (SSG) meeting in Malmö, and stakeholders from the clinical community were in favour of the project.

Access to data

Two pertinent issues were identified by the working group, the principle of a federated registry and access to the registry for mapping according to OMOP.

To ensure long term viability for the federated sarcoma registry in Sweden the first priority for the working group was to ensure that secondary health data use in a federated model was known and acceptable, in principle, by RCC Syd running the registries. We also judged that the establishment of a federated registry in association with the Swedish registries was not, as such, a research project requiring a time-limited ethical permit. The working group had several meetings with representatives from RCC Syd to present the project, and explore possibilities and potential obstacles to make the registries part of the federated model. However, after internal analysis by registry representatives from RCC Syd and Region Skåne, the conclusion of the registry holders was that secondary health data use in the proposed federated registry was not compatible with the Swedish interpretation of GDPR. A second opinion on this matter was referred to the judicial system.

In parallel the working group (mainly ES) explored the possibilities to map the registries according to OMOP and to this end had several meetings with representatives from RCC Syd including the lead statistician. Meetings were also held by the working group together with the project partners (and RCC Syd) to see how the working model proposed by the project could be employed in the INCA setting, either completely or by modification. Due to data security concerns RCC Syd were not willing to let the project partners do the mapping, and the solution for mapping proposed by RCC Syd was not acceptable for the project.

In total, failure to establish compatibility of GDPR in Sweden and the principle of a federated registry and the lack of agreement on data mapping led the working group together with the Blueberry Steering Committée to conclude that further work on the project in Sweden was not possible at that time point and the project was closed in Sweden.

Lessons learned

A few lessons can be learned from the work on Blueberry in Sweden

1) Early involvement of, not only of health-care professionals and researchers in the clinical community, but also of representatives of the platform owners is crucial.

2) The distinction between building registry infrastructure (*e.g*. the federated registry) and conducting research using registry data needs further clarification.

3) The lack of a formal structure for appeal of a decision by a RCC apart from the judicial system was highlighted. This structure risks the possibility of different interpretations by different health care regions, and, although an appeal to court is possible, this is rarely performed by individual researchers.

Finally, it should be noted that the discussion on privacy and secondary health data use is evolving in Sweden. In 2023 the Government Report SOU 2023:76 was completed with proposals to strengthen secondary health data use for clinical care and research. With the EHDS passed at the European level adjustments are underway, and we are optimistic about the possibility to employ principles of federated datause in Sweden in the future.

Appendix 4 - Results OMOP-on-Vantage6 studyathon

Status of Data Conversion to the OMOP CDM by the end of BlueBerry

Data partner	Data source	Status	Responsible party
Graz (Austria)	Clinical registry	Converted to OMOP, connected to vantage6	Biomeris
INT (Milan)	Clinical registry	Converted to OMOP, connected to vantage6	INT + Biomeris
IKNL (Netherlands)	Population-based registry	Converted to OMOP, connected to vantage6	IKNL
CRN (Oslo)	Population-based and clinical registry	Converted to OMOP, connected to vantage6	CRN
CLB (France)	Nation-wide registry with national coverage with data from clinical expert centers only (NETSARC).	Converted to OMOP, connected to vantage6	Biomeris
Sahlgrenska (Sweden)	Population-based registry	On hold due to regulatory issues	N/A
Madrid (Spain)	National database (GEIS)	OMOP mapping started	Biomeris

Technical considerations

Each data partner has a unique IT landscape with distinct functional requirements that provide challenges for the installation. In the blueberry project, we opted for a simplified implementation strategy. We created homogeneity in the network by ensuring that each data partner used a virtual machine with the same operating system and technical specifications (see **table 1**). The optimal implementation strategy for connecting the OMOP CDM and vantage6 node is depicted in **Figure 3**. Source data should be hosted on a local server behind a firewall. These source data are converted to a OMOP CDM at regular intervals (e.g., every 3 months) or whenever new data are available through the ETL pipeline. The OMOP CDM resides on a single virtual machine together with the vantage6 node. This allows the node to have access to the OMOP CDM, without having access to the source data (with identifiable data).

Table 1: Technical requirements

	Required	**Optional**
1 virtual machine	x	
4CPUs 64 bit	x	
8GB memory	x	expandable to 16GB if analysis

		requires more performance
250 GB disk space	x	
Oracle Linux 8 OS - minimal installation	x	
web access	during installation phase	
VPN access through port 22		required for Biomeris staff

Installation of vantage6 software

The central server of the federated learning network is hosted on a server maintained by one data partner or organization (in this case IKNL). This central server is connected to the node of each data partner (using the setup depicted in **Figure 3)**. The uniformity in the setup in the network provides a straightforward connection between the nodes and the central server, improves security within the network, facilitates automation, and allows for the use of standardized installation scripts.

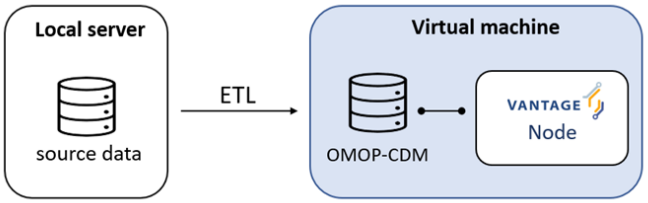


Figure 3: Connection between source data, OMOP CDM and vantage6 node

In contrast to the OHDSI tools, vantage6 decouples data from algorithms. To access the OMOP CDM, an SQL connection needed to be implemented. To establish this connection, we used two approaches: 1) Create an SSH tunnel between a locally hosted postgreSQL OMOP-CDM database and vantage6 node. An SSH tunnel is complex to set up and pose security risks, or 2) Place the OMOP CDM within a docker container and connect it to the vantage6 node using Docker services. Docker services offer the most stable connection with vantage6 and require no complicated configuration. This is therefore the most preferred implementation.

A hybrid solution was used for the blueberry project, with some instances using SSH tunnels and others using Docker services (see **Table 2**). Installation scripts for both installation strategies are publicly available on https://github.com/IKNL/v6-blueberry-installation-scripts.

Table 2: Status of vantage6 installation for each of the data partners

Data partner	Type of installation	Status	Responsible party
Graz (Austria)	SSH tunnel	Connection established	Biomeris
INT (Milan)	SSH tunnel	Connection established	Biomeris
IKNL (Netherlands)	Docker container	Connection established	IKNL
CRN (Oslo)	Docker container	Connection established	CRN
CLB (France)	Docker container	Connection established	Biomeris

Madrid (Spain)	Docker container	Not installed yet	Biomeris

Use cases

Several use cases have been developed to test the feasibility of the legal framework, technological framework (also including the quality of the data and depth/breadth of data coverage), governance model, and business/valorization model.

1) Simple Use Case

- Provide the distribution of sarcoma subtypes according to histology
- Identify number of angiosarcoma patients in different datasets [MORE COMMON]
- Identify number of retroperitoneal sarcoma patients in different datasets [NEEDED FOR NEXT USE CASES]

2) Clinically Relevant Use Case: ***Neoadjuvant Chemotherapy For Soft Tissue Sarcomas***

STRASS II[6] is a randomized phase III study of chemotherapy followed by surgery versus surgery alone to improve disease control and survival in patients with high-risk retroperitoneal sarcoma. One of the main research questions of the STRASS 2 trial is: What are the outcomes of Surgery With Or Without Neoadjuvant Chemotherapy in High Risk RetroPeritoneal Sarcoma? *Can BlueBerry data help to answer this question with real-world data?*

a. How often is neoadjuvant chemotherapy performed in high-risk retroperitoneal sarcoma patients?
b. Effectiveness question: What are the outcomes of surgery with or without neoadjuvant therapy in high-risk retroperitoneal sarcoma patients?

An additional clinically relevant question:

c. Epidemiological / natural history question: What is the incidence of distant metastases in angiosarcomas divided by 1) cutaneous (radiotherapy induced versus non-radiotherapy induced) and 2) visceral; and survival rate?

3) Sustainability Use Case*: Can We Add Value to a Currently Used Sarcoma App?*

SARCULATOR: The SARCULATOR is a validated nomogram that predicts survival of patients with resected extremity and trunk soft tissue sarcoma to aid prognostication.
It predicts 5- and 10-year overall survival and distant-metastasis-free survival. SARCULATOR can also determine the 7-year-disease-free survival[7]

Is there a volume-outcome relationship in retroperitoneal sarcomas (RPS)?

- Inclusion criteria: primary RPS treated with curative-intent surgery.
- Variables entered: age, tumor size, tumor grade, histology, multifocality, completeness of resection, case volume/year.
- Primary endpoint: overall survival
- Secondary endpoints: PFS (LR, DM if available), postoperative morbidity (if available).

[6] https://www.clinicaltrials.gov/ct2/show/NCT04031677
[7] https://ascopubs.org/doi/10.1200/JCO.2017.35.15_suppl.11016

ATLAS – Cohort definitions

- We will update the cohort entry event, considering procedure occurrence of surgery between 2010-01-01 and 2017-12-31

- We select:
  - adult (≥18-year-old) patients
  - primary localized, non-metastatic retroperitoneal and pelvic sarcoma (RPS)

    According to the clinical definition of disease extension:
    - Unifocal is localized
    - Multifocal (1+ lesion in the same organ or/anatomical compartment) is loco-regional

    Both of them are included!!

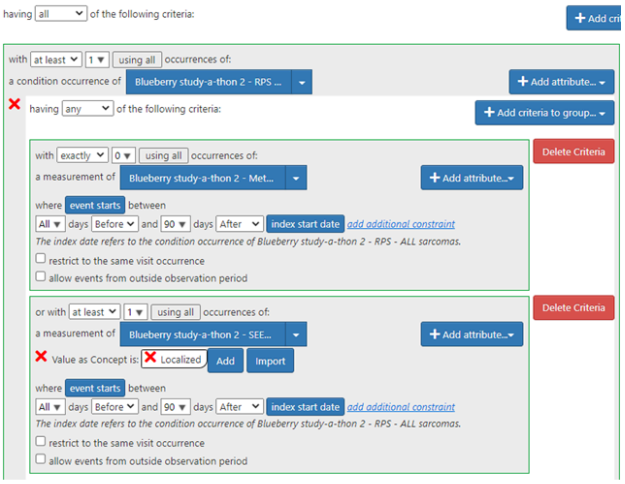
ATLAS – Cohort definitions

- Adult (≥18-year-old) patients

- Primary localized (both uni- and multifocal, according to previous definition), non-metastatic retroperitoneal sarcoma (RPS):

- with exactly 0 measurement of metastases between all days before and 90 days after diagnosis

- with at least 1 measurement of localized stage between all days before and 90 days after diagnosis



ATLAS – Cohort definitions

- with exactly 0 measurement of metastases between all days before and 90 days after diagnosis

Showing 1 to 2 of 2 entries

Concept Id	Concept Code	Concept Name	Domain	Standard Concept Caption	Exclude	Descendants
36769180	OMOP4998856	Metastasis	Measurement	Standard	☐	☑
1635142	AJCC/UICC-M1	AJCC/UICC M1 Category	Measurement	Standard	☐	☑

Property	Value
Concept Name	Metastasis
Domain Id	Measurement
Concept Class Id	Metastasis
Vocabulary Id	Cancer Modifier
Concept Id	36769180
Concept Code	OMOP4998856
Invalid Reason	Valid
Standard Concept	Standard

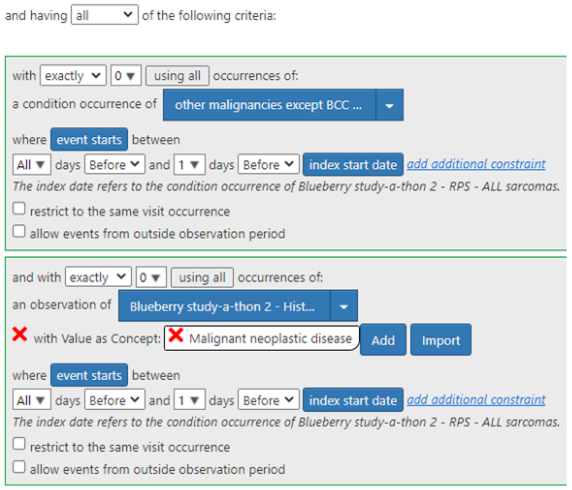
Property	Value
Concept Name	AJCC/UICC M1 Category
Domain Id	Measurement
Concept Class Id	Staging/Grading
Vocabulary Id	Cancer Modifier
Concept Id	1635142
Concept Code	AJCC/UICC-M1
Invalid Reason	Valid
Standard Concept	Standard

ATLAS – Cohort definitions

- Without any prior malignancy:

- with exactly 0 condition occurrence of other malignacies all days before and 1 days before diagnosis

- with exactly 0 observation of history of malignant neoplastic disease all days before and 1 days before diagnosis



Next steps – Cohort definitions

- INT and Graz (clinical DB) should group Liposarcoma according to the grading:
  - Grade 1 -> well differentiated
  - Grade 2/3 -> dedifferentiated
- Well-differentiated and dedifferentiated liposarcoma cohorts will include:
  - Atypical lipomatous tumour (8850)
  - Liposarcoma, well differentiated (8851)
  - Mixed liposarcoma (8855)
  - Fibroblastic liposarcoma (8857)
  - Dedifferentiated liposarcoma (8858)
- We need different cohorts based on the combination of diagnosis code and grading!

Next steps – Cohort definitions

- At present, MPNST cohort includes:
  - Malignant peripheral nerve sheath tumour NOS (9540)
  - Malignant peripheral nerve sheath tumour, epithelioid (9542)
  - Malignant schwannoma, NOS (9560)
  - MPNST with rhabdomyoblastic differentiation (9561)
  - Perineuroma, malignant (9571)
- In OMOP Malignant peripheral nerve sheath tumour, epithelioid, diagnosis codes 9542/3-C48.0 and 9542/3-C49.5 do not exist in ICDO Vocabulary and so we do not map them. Malignant peripheral nerve sheath tumour, epithelioid could be mapped as standard SNOMED Concept (4100556), but in this case it does not include our diagnosis because it seems not to be related to retroperitoneal or pelvis sites.

ATLAS – Cohort definitions

Group	WHO 5th/ICD-O-3.2	Topography	Concept sets/Cohort Name	OMOP INT DB
All retroperitoneal and pelvis sarcomas (...)	(...)		01 - Surgery - All sarcomas	616
			02 - Surgery - All sarcomas C48.0	445
			03 - Surgery - All sarcomas C49.5	171
Well-differentiated liposarcoma (8850, 8851, 8855, 8857-8858) – grade 1	Liposarcoma, NOS/atypical lipomatous tumour (8850), Liposarcoma, well differentiated (8851), Mixed liposarcoma (8855), Fibroblastic liposarcoma (8857), Dedifferentiated liposarcoma (8858)	Both	04 - Surgery - Liposarcoma well differentiated	135
		C48.0	05 - Surgery - Liposarcoma well differentiated C48.0	122
		C49.5	06 - Surgery - Liposarcoma well differentiated C49.5	13
Dedifferentiated liposarcoma (8850, 8851, 8855, 8857-8858) – grade 2/3	Liposarcoma, NOS/atypical lipomatous tumour (8850), Liposarcoma, well differentiated (8851), Mixed liposarcoma (8855), Fibroblastic liposarcoma (8857), Dedifferentiated liposarcoma (8858)	Both	07 - Surgery – Liposarcoma dedifferentiated	230
		C48.0	08 - Surgery - Liposarcoma dedifferentiated C48.0	216
		C49.5	09 - Surgery - Liposarcoma dedifferentiated C49.5	14
Leiomyosarcoma (8890,8891,8895-8896)	Leiomyosarcoma, NOS (8890), Epitheliod leiomyosarcoma (8891), Myosarcoma (8895), Myxoid leiomyosarcoma (8896)	Both	10 - Surgery - Leiomyosarcoma	103
		C48.0	11 - Surgery - Leiomyosarcoma C48.0	71
		C49.5	12 - Surgery - Leiomyosarcoma C49.5	32
Solitary fibrous tumour (8815)	Solitary fibrous tumors (8815)	Both	13 - Surgery - Solitary fibrous tumour	3
		C48.0	14 - Surgery - Solitary fibrous tumour C48.0	2
		C49.5	15 - Surgery - Solitary fibrous tumour C49.5	1

ATLAS – Cohort definitions

Group	WHO 5th/ICD-O-3.2	Topography	Concept sets/Cohort Name	OMOP INT DB
MPNST (9540, 9542, 9560-9561, 9571)	Malignant peripheral nerve sheath tumour NOS (9540), Malignant peripheral nerve sheath tumour, epithelioid (9542), Malignant schwannoma, NOS (9560), MPNST with rhabdomyoblastic differentiation (9561), Perineuroma, malignant (9571)	Both	16 - Surgery - MPNST	15
		C48.0	17 - Surgery - MPNST C48.0	10
		C49.5	18 - Surgery - MPNST C49.5	5
UPS, Undifferentiated/unclassified sarcoma (8800-8802, 8805, 8830)	Sarcoma, NOS (8800), Spindle cell sarcoma, undifferentiated (8801), Pleomorphic sarcoma, undifferentiated (8802), Undifferentiated sarcoma (8805), Malignant Fibrous histiocytoma (MFH) (8830)	Both	19 - Surgery - UPS	34
		C48.0	20 - Surgery - UPS C48.0	5
		C49.5	21 - Surgery - UPS C49.5	29
Other sarcomas (...)	(...)	Both	22 - Surgery - Other sarcomas	96
		C48.0	23 - Surgery - Other sarcomas C48.0	19
		C49.5	24 - Surgery - Other sarcomas C49.5	77

Next steps

- Definition of Hospital Volume (HV):
  - Clinical DB -> high volume (n. of surgeries >=13) vs low volume (n. of surgeries<13). Volume must be defined according to the surgeries performed in the specific center, but we need also the name (code) of the centers where other surgeries are performed
  - Population registries –> check the availability of the variable which defines the surgery performed in the specific center or in another one.
    Provide the number of surgeries done in the center and the number of surgeries performed outside the center
  - We need also the number of not surgically treated patients for both clinical DB and population registries

Next steps

- Multifocality:
  - Unifocal is localized
  - Multifocal (1+ lesion in the same organ or/anatomical compartment) is loco-regional
  - Population registries –> check availability
  - Clinical DB -> information available

Next steps - Analyses

- Descriptive analyses for the cohort of all sarcomas (separately for each database/center):
  - N. of patients stratified by sarcoma grouping and overall
  - Sex (Male/Female) by sarcoma grouping and overall
  - Age (mean, median and IQR) by sarcoma grouping and overall
  - N. of deaths by sarcoma grouping and overall
  - Tumor size (mean, median and IQR)
  - FNCLCC grade (frequencies)
  - Multifocality (Unifocal/Multifocal)
  - Completeness of resection (R0/R1=Macroscopically complete, R2=Macroscopically incomplete)
  - Tumor rupture (Yes/No)
  - Chemotherapy (Yes/No)
  - Radiotherapy (Yes/No)
  - Local recurrence (Yes/No)
  - Distant metastases (Yes/No)
  - N. of surgeries performed in the center/outside per year
  - N. not surgically treated patients per year (info derivable outside our cohorts or DB)

Furtherly differences in distributions of the above variables among the 2 groups (HV vs LV) will be assessed by χ^2 for categorical variables and T-test for continuous variables

Next steps - Analyses

- Overall/Disease free 5-year Kaplan-Meier survival analysis stratified by:
  - Hospital Volume (HV) Low vs High
  - Age classes (defined by median and quartiles)
  - Sex
  - Sarcoma grouping

  The differences in KM survival strata will be tested by Log Rank test

- 5-year Cox survival analysis model:
  - Hospital Volume (HV) Low vs High
  - Age classes (defined by median or quartiles)
  - Sex
  - Sarcoma grouping
  - Prognostic variables (to be selected by a backward stepwise regression)

  The multivariable Cox model will be defined after the analyses of the preliminary distributions of prognostic factors and Kaplan-Meier results.
  The Cox model proportional hazard assumption will be tested using Schoenfield residuals

3rd BlueBerry Study-a-thon

Re-running analyses

09-07-2024

Table of Contents

Study-a-thon participating nodes (3)
Privacy Guards (4)

Preparation (5)
- **PERSON Table Count**
- **Study-a-thon cohorts**
- **Create Cohorts**

Cohort Counts (9)
- **Cohort Counts run**
- **Counts per node**
- **Discussion for Kaplan-Meier**
- **Morphology comparison**
- **Topography comparison**

Kaplan-Meier (17)
- **Kaplan-Meier run**
- **Topography comparison**
- **Morphology comparison (expert centers)**
- **Morphology comparison (pop. registries)**
- **Survival at INT and IKNL**
- **Expert centers vs population registries**

CohortDiagnostics (25)
- **Create Cohorts**
- **CohortDiagnostics run**
- **CohortDiagnostics screenshots**

THE END (29)

iKNL BlueBerry
a blueprint for the EURACAN registry

Vantage6 for BlueBerry | Participating Nodes

Party	ID
INT (Italy)	3
GRAZ (Austria)	4
CLB (France)	17
CRN (Norway)	16
IKNL (Netherlands)	2

Origin of today's BlueBerry data partners!

Privacy Guards

PERSON COUNT

BOQ_MIN_RECORDS = 5
ALLOW_ZERO = true

COHORT COUNTS (OHDSI Cohort Generator Package)

CC_MIN_RECORDS = 5

KAPLAN MEIER

KAPLAN_MEIER_MINIMUM_NUMBER_OF_RECORDS = 10
KAPLAN_MEIER_TYPE_NOISE = "POISSON"
KAPLAN_MEIER_ALLOWED_EVENT_TIME_COLUMNS_REGEX = ".*"

COHORT DIAGNOSTICS (OHDSI Cohort Diagnostics Package)

CD_MIN_RECORDS = 5

With this study-a-thon we managed to connect 5 OMOP databases from 5 different countries. Let's not forget this is a big milestone for sarcoma research!
During this preparation step we looked at the PERSON table count: this shows the count of patients in the OMOP database per node, and for the federated network in total.

However, in this study-a-thon we are going to focus on certain cohorts of patients. In slide 7 you can find the cohorts we decided to use in this study-a-thon.

We continued with running the algorithm 'Create Cohorts' (slide 8) which creates the cohorts at the nodes. These cohort tables can then be used for algorithms like OHDSI's Cohort Diagnostics and Kaplan-Meier.

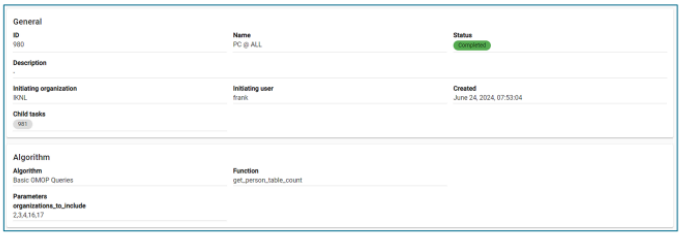
PREPARATION

PERSON Table Counts

Notebook	Table-Counts-and-Names.ipynb
Algorithm	Basic OMOP Queries
Function	get_person_table_count

Task ID	980
URL	https://portal.blueberry.vantage6.ai/#/analyze/tasks/980

Individual and global counts of the PERSON table.



Count per node

organization_id	person_count
CRN	28719
IKNL	18981
INT	10712
GRAZ	7386
CLB	1652

Global count

person_count
67450

Study-a-thon cohorts

To be able to compare different morphologies we decided to choose all cohorts not filtered on only C48.0 or C49.5. To be able to do a single comparison between topographies we added ALL sarcomas C48.0 to the list of cohorts.

The 9 chosen cohorts were:

- ALL sarcomas (1001)
- ALL sarcomas C48.0 (1002)
- Well differentiated or dedifferentiated v2 (Low grade) (1004)
- Well differentiated or dedifferentiated v2 (High grade) (1007)
- Leiomyosarcoma (1010)
- Solitary fibrous tumour (1013)
- MPNST (1016)
- UPS (1019)
- Other sarcomas (1022)

id	group
1001	ALL sarcomas
1002	ALL sarcomas C48.0
1003	ALL sarcomas C49.5
1004	Well differentiated or dedifferentiated v2 (Low grade)
1005	Well differentiated or dedifferentiated C48.0 v2 (Low grade)
1006	Well differentiated or dedifferentiated C49.5 v2 (Low grade)
1007	Well differentiated or dedifferentiated v2 (High grade)
1008	Well differentiated or dedifferentiated C48.0 v2 (High grade)
1009	Well differentiated or dedifferentiated C49.5 v2 (High grade)
1010	Leiomyosarcoma
1011	Leiomyosarcoma C48.0
1012	Leiomyosarcoma C49.5
1013	Solitary fibrous tumour
1014	Solitary fibrous tumour C48.0
1015	Solitary fibrous tumour C49.5
1016	MPNST
1017	MPNST C48.0
1018	MPNST C49.5
1019	UPS
1020	UPS C48.0
1021	UPS C49.5
1022	Other sarcomas
1023	Other sarcomas C48.0
1024	Other sarcomas C49.5

Create Cohorts

Notebook Create-Cohorts.ipynb
Algorithm OMOP Cohorts
Function create_cohort_central

Task ID 1064
URL https://portal.blueberry.vantage6.ai/#/analyze/tasks/1064

Cohorts were created for all nodes.

Metadata

task_id: 1065
shared_id: 1065000
name: 1001

task_id: 1065
shared_id: 1065001
name: 1002

task_id: 1065
shared_id: 1065002
name: 1004

task_id: 1065
shared_id: 1065003
name: 1007

task_id: 1065
shared_id: 1065004
name: 1010

task_id: 1065
shared_id: 1065005
name: 1013

task_id: 1065
shared_id: 1065006
name: 1016

task_id: 1065
shared_id: 1065007
name: 1019

task_id: 1065
shared_id: 1065008
name: 1022

Since we are going to investigate 9 cohorts we first analyzed their size. We ran the function cohorts_count_central from the algorithm OMOP Cohorts (slide 10), and after that analyzed the results.

We split the results in counts per node (slide 11), counts for morphology comparison (slide 13), and counts for topography comparison (slide 16).

For the morpohology (slide 14/15) and topography (slide 16) comparisons we made some visualizations for easy comparison of all the numbers.

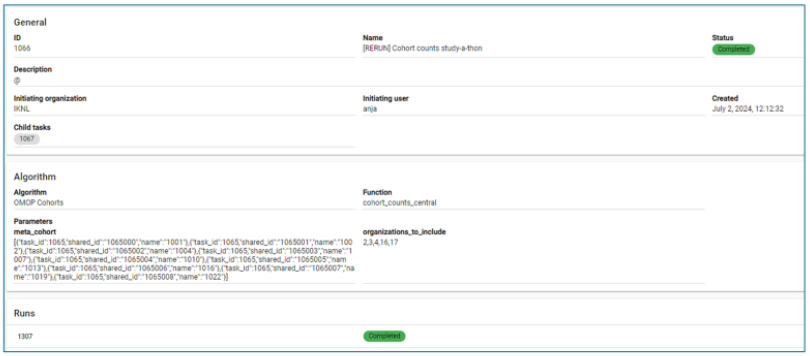
In between we also elaborated on what the results mean for the Kaplan-Meier analysis (slide 12), since, as often the case for rare cancers, we are dealing with low population numbers.

COHORT COUNTS

Cohort Counts

Notebook	Cohort-Counts.ipynb	**Task ID**	1066
Algorithm	OMOP Cohorts	**URL**	https://portal.blueberry.vantage6.ai/#/analyze/tasks/1066
Function	cohorts_count_central		

Cohort counts were retrieved. (.csv with results available)



Cohort Counts – Nodes

Expert centers

INT (IT)		GRAZ (AT)		CLB (FR)	
1001	374	1001	30	1001	57
1002	348	1002	22	1002	53
1004	89	1004	8	1004	7
1007	160	1007	11	1007	33
1010	64	1010	1-4	1010	6
1013	26	1013	1-4	1013	1-4
1016	12	1016	0	1016	1-4
1019	7	1019	5	1019	1-4
1022	16	1022	1-4	1022	1-4

Population registries

IKNL (NL)		CRN (NO)	
1001	450	1001	193
1002	256	1002	84
1004	98	1004	0
1007	54	1007	1-4
1010	88	1010	36
1013	9	1013	5
1016	0	1016	1-4
1019	38	1019	26
1022	87	1022	49

Cohort Counts – Discussion for Kaplan-Meier

Datasource type

We make a distinction between expert centers and population registries. This means that IKNL (NL) and CRN (NO) are grouped when possible, and the same applies to INT (IT), CLB (FR) and GRAZ (AT).

Privacy guards

The Kaplan-Meier algorithm currently has its privacy guard for minimum number of records set to 10. Since this is the threshold we communicated as a setting for the study-a-thon we didn't change this now. However, changing this threshold could potentially mean that more Kaplan-Meier curves can be calculated. For now:

More than 10 patients

For all cohorts that have more than 10 patients for all nodes; we ran the Kaplan-Meier algorithm.

Less than 10 patients

For cohorts that have one or multiple nodes with fewer than 10 patients, you can either 1) not run the Kaplan-Meier algorithm, 2) run the Kaplan-Meier curve excluding these nodes, or 3) lower the privacy guards. For now we went with excluding these nodes.

Cohort Counts – Morphology comparison

1001

ALL sarcomas

IKNL (NL)	450 ✓
INT (IT)	374 ✓
GRAZ (AT)	30 ✓
CRN (NO)	193 ✓
CLB (FR)	57 ✓ KM

1004

Well differentiated or dedifferentiated v2 (Low grade)

IKNL (NL)	98 ✓
INT (IT)	89 ✓
GRAZ (AT)	8
CRN (NO)	0
CLB (FR)	7 KM

1007

Well differentiated or dedifferentiated v2 (High grade)

IKNL (NL)	54 ✓
INT (IT)	160 ✓
GRAZ (AT)	11 ✓
CRN (NO)	1-4
CLB (FR)	33 ✓ KM

1010

Leiomyosarcoma

IKNL (NL)	88 ✓
INT (IT)	64 ✓
GRAZ (AT)	1-4
CRN (NO)	36 ✓
CLB (FR)	6 KM

1013

Solitary fibrous tumour

IKNL (NL)	9
INT (IT)	26 ✓
GRAZ (AT)	1-4
CRN (NO)	5
CLB (FR)	1-4 KM

1016

MPNST

IKNL (NL)	0
INT (IT)	12 ✓
GRAZ (AT)	0
CRN (NO)	1-4
CLB (FR)	1-4 KM

1019

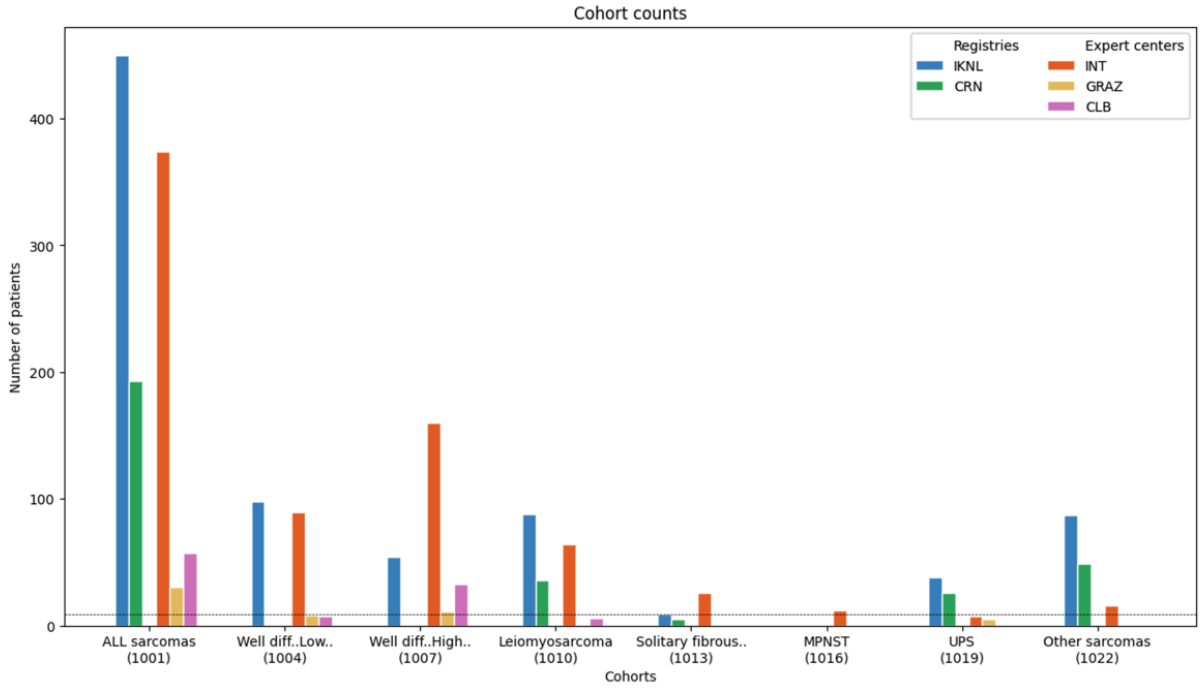
UPS

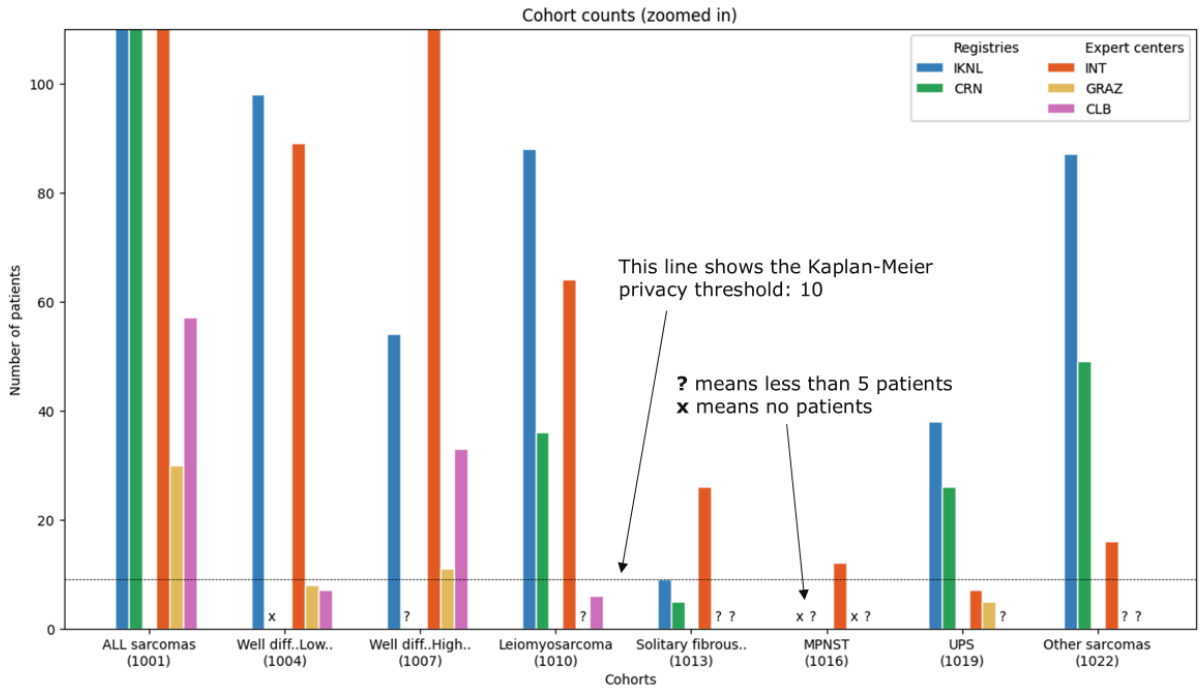
IKNL (NL)	38 ✓
INT (IT)	7
GRAZ (AT)	5
CRN (NO)	26 ✓
CLB (FR)	1-4 KM

1022

Other sarcomas

IKNL (NL)	87 ✓
INT (IT)	16 ✓
GRAZ (AT)	1-4
CRN (NO)	49 ✓
CLB (FR)	1-4 KM





Cohort Counts – Topography comparison

1001

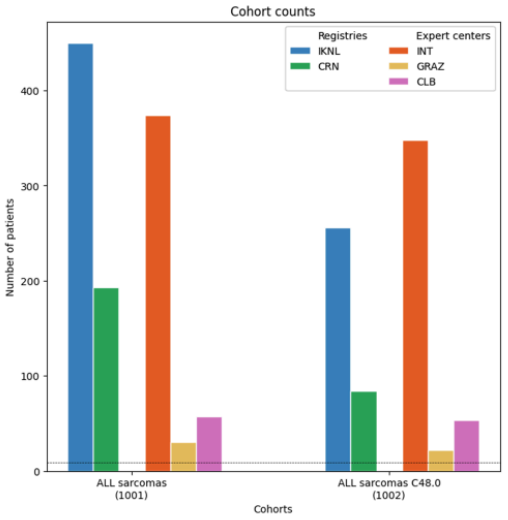
ALL sarcomas

IKNL (NL)	450 ✓
INT (IT)	374 ✓
GRAZ (AT)	30 ✓
CRN (NO)	193 ✓
CLB (FR)	57 ✓ KM

1002

ALL sarcomas C48.0

IKNL (NL)	256 ✓
INT (IT)	348 ✓
GRAZ (AT)	22 ✓
CRN (NO)	84 ✓
CLB (FR)	53 ✓ KM



Since we have 9 cohorts and also have to split up the nodes between expert centers and population registries (for clinical cohort definition reasons) there are many Kaplan-Meier curves that can be calculated. In slide 18 we show an overview of all the combinations of cohorts vs nodes we ran a Kaplan-Meier on. Some combinations of cohort vs nodes, like cohort 1013 for IKNL+CRN, cannot be calculated because of a too small cohort (in combination with privacy guards), see the discussion on slide 12.

Slides 19-24 show different plots of Kaplan-Meier curves, all these plots are accompanied by a description of what the plot is about.

If you ask us;
it's pretty awesome to see some real (clinical) results!

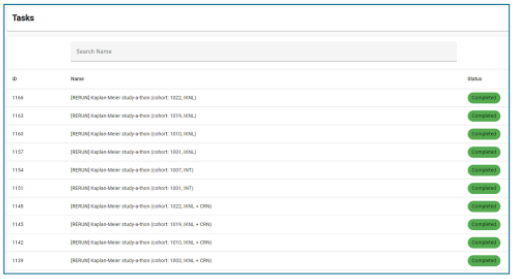
KAPLAN-MEIER

Kaplan-Meier

Notebook	Kaplan-Meier.ipynb
Algorithm	Kaplan Meier on OMOP
Function	kaplan_meier_central

Multiple Kaplan-Meier curves were created.
On the right a table with cohorts, involved nodes, and task id's.

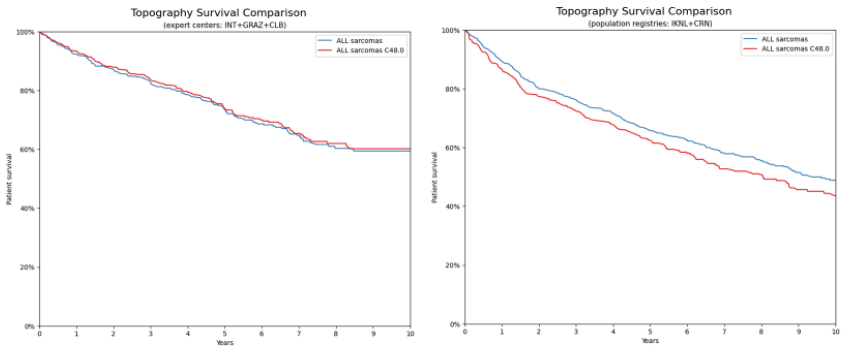
Below a screenshot of some of these tasks:



Cohort	Nodes	Task ID
1001	IKNL+CRN	1070
1001	INT+GRAZ+CLB	1080
1001	INT	1151
1001	IKNL	1157
1002	IKNL+CRN	1139
1002	INT+GRAZ+CLB	1083
1004	IKNL	1125
1004	INT	1109
1007	IKNL	1128
1007	INT+GRAZ+CLB	1089
1007	INT	1154
1010	IKNL+CRN	1142
1010	INT	1113
1010	IKNL	1160
1013	IKNL+CRN	X
1013	INT	1116
1016	IKNL+CRN	X
1016	INT	1119

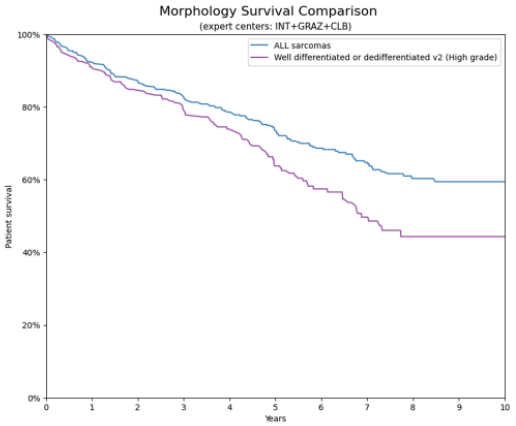
Cohort	Nodes	Task ID
1019	IKNL+CRN	1145
1019	INT+GRAZ+CLB	X
1019	IKNL	1163
1022	IKNL+CRN	1148
1022	INT	1122
1022	IKNL	1166

KM – Topography Comparison



With this topography comparison we can compare the cohort of ALL sarcomas with the cohort of ALL sarcomas C48.0. This is the only morphology we can run a topography comparison for as for all other morphologies we don't have the corresponding C48.0 cohort definition finalized. The data is split in expert centers (figure on the left) and population registries (figure on the right).
Question: Does it (clinically) make sense that the survival curves are swapped between both figures? (on the left red on top, on the right blue on top)

KM – Morphology comparison (expert centers)

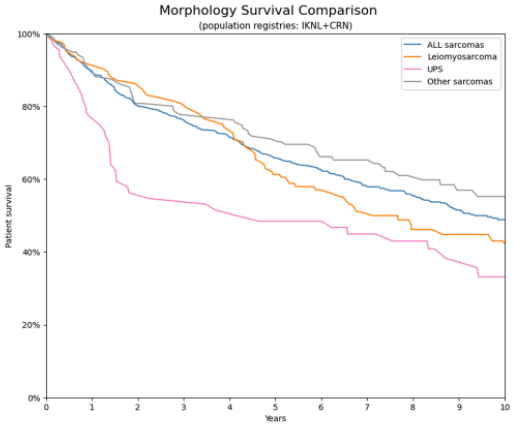


With the morphology comparison we compare different morphologies to each other, that includes 8 of the 9 chosen cohorts (1002 is excluded for this). However, we saw in slide 13 that for many cohorts we don't have a large enough population.

In this plot we plotted all cohorts for which a Kaplan-Meier could be calculated for the combination of expert centers INT+GRAZ+CLB. This means that if one of these centers did not have enough patients in the database we did not calculate a Kaplan-Meier for that cohort.

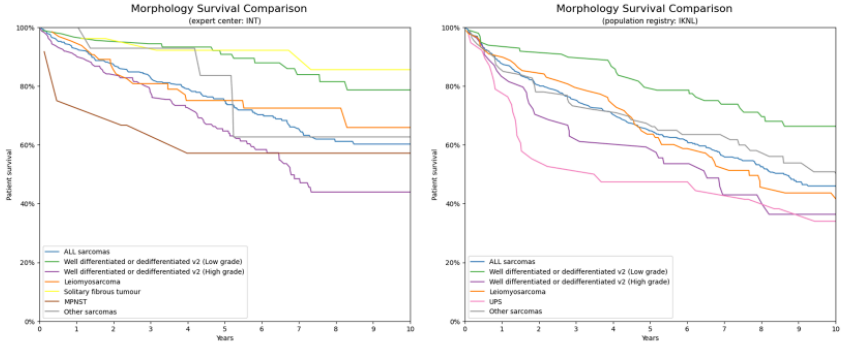
Two cohorts were left and thus can ALL sarcomas be compared to High Grade Well/de-differentiated.

KM – Morphology comparison (population registries)



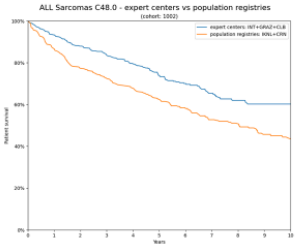
This plot is similar to the one from the previous slide, the only difference is that here we looked at the combination of population registries IKNL+CRN.
This resulted in 4 survival curves for 4 different cohorts.

KM – Survival at INT and IKNL



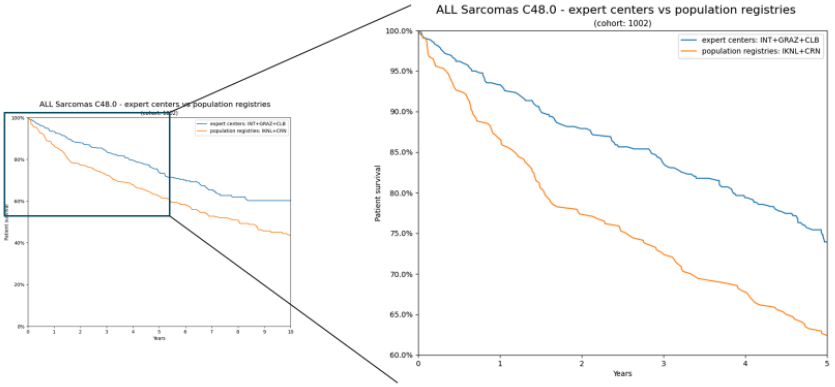
INT (figure on the left) and IKNL (figure on the right) appear to have enough patients for a Kaplan-Meier curve for most of the cohorts, so we thought it'd be interesting to see how for these single centers the survival for the different morphologies compare.

KM – Expert centers vs population registries



The comparison between expert centers and population registries can (clinically) be done for cohort 1002: ALL sarcomas C48.0. In this figure we plot the two Kaplan-Meier curves for these two groups.

KM – Expert centers vs population registries



The comparison between expert centers and population registries can (clinically) be done for cohort 1002: ALL sarcomas C48.0. In this figure we plot the two Kaplan-Meier curves for these two groups.

With a zoomed in view we can compare the two groups in a bit more detail.

The OHDSI community has built multiple tools and packages to use in combination with an OMOP database. The package CohortDiagnostics is one of them. With this package you can investigate cohorts and find difference between databases, which can result in better cohort definitions and more proper comparisons of data.

During the study-a-thon day we encountered a bug in our vantage6 CohortDiagnostics algorithm. After we fixed this bug, we created new cohorts for the nodes and could run CohortDiagnostics without problems.

The Shiny App created to analyze the results uses the .zip files that are output by the algorithm. The Shiny App has different features, in slide 28 we show two possible use cases of the Shiny App: it can be useful to compare counts, but also age or gender distribution of the data.

COHORT DIAGNOSTICS

Create Cohorts

Notebook	Create-Cohorts.ipynb
Algorithm	OMOP Cohorts
Function	create_cohort_central

Task ID	1133
URL	https://portal.blueberry.vantage6.ai/#/analyze/tasks/1133

Cohorts were created for all nodes.

Metadata

task_id: 1134
shared_id: 1134000
name: 1001

task_id: 1134
shared_id: 1134001
name: 1002

task_id: 1134
shared_id: 1134002
name: 1004

task_id: 1134
shared_id: 1134003
name: 1007

task_id: 1134
shared_id: 1134004
name: 1010

task_id: 1134
shared_id: 1134005
name: 1013

task_id: 1134
shared_id: 1134006
name: 1016

task_id: 1134
shared_id: 1134007
name: 1019

task_id: 1134
shared_id: 1134008
name: 1022

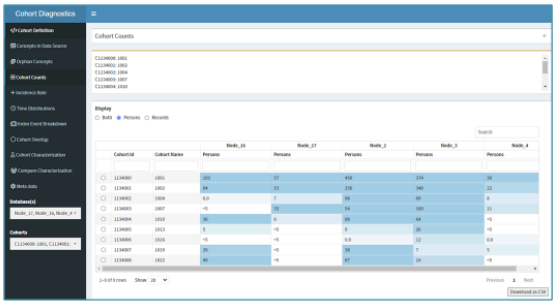
Cohort Diagnostics

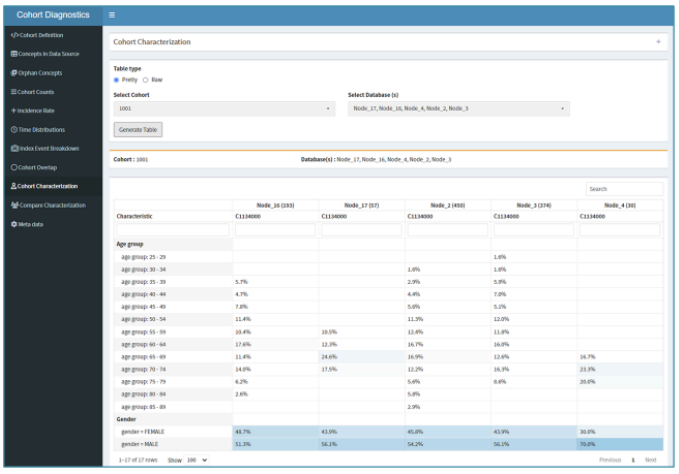
Notebook	Cohort-Diagnostics.ipynb	**Task ID**	1068
Algorithm	Cohort Diagnostics	**URL**	https://portal.blueberry.vantage6.ai/#/analyze/tasks/1068
Function	cohort_diagnostics_central		

CohortDiagnostics was successfully executed and resulted in 5 .zip result files.

CohortDiagnostics – Screenshot

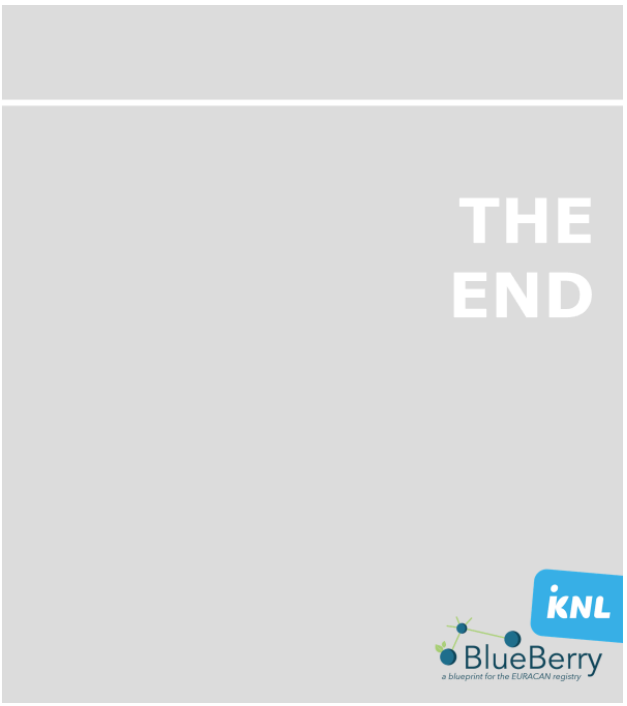
CohortDiagnostics generated per node a zip file with results. These results can be investigated with the OHDSI Shiny App for CohortDiagnostics. Here we show some screenshots of such an investigation.





First of all: the results in this powerpoint should not be used as clinical results for sarcoma patients, to achieve this the data and plots should first be checked with subject matter experts. For example: it may be that the cohort definitions do not work perfectly on CLB yet, since they were surprised that their patients numbers were so low. But that's exactly how a study like this goes: you start exploring the data and then figure out how the data should be used/compared. But with this study-a-thon we showed that the infrastructure is ready for that next step.

This study-a-thon showcased that with some effort we can harmonize our datasets, create cohort definitions, create the IT infrastructure for federated analyses, agree on legal terms, run OHDSI tooling from a distance, and finally run survival analyses to really perform research on our combined data.



Appendix 5 - Privacy Enhanced Technologies and Consent Management

D2.4. Privacy Enhanced Technologies, Anonymisation on using Rare Cancer Data

Version 1

Contents

1. Introduction ... 3
1.1 The Challenges ... 3
2. Organization and operations ... 5
2.1. Current Governance ... 5
3. Data Infrastructure ... 7
3.1. Introduction ... 7
3.2. Data Capture and Harmonization ... 8
Data Transformation ... 8
Data Harmonization strategy ... 9
3.3. Federated Learning ... 10
Implementation strategy ... 10
Compatibility of OHDSI tools with vantage6 ... 10
Vantage6 User Interface for the execution of analyses ... 11
Technical considerations ... 11
Installation of vantage6 software ... 12
3.4. Data analysis ... 12
Use cases definitions ... 12
Study-a-thon results ... 13
Considerations for choosing and executing use cases ... 14
3.5. Competencies required for data harmonization, federated learning and data analysis ... 14
3.6. Lessons Learned ... 15
3.7. Next Steps ... 17
4. Legal Framework for the EURACAN Registry ... 18
4.1. Context: relevant European and National Legislations ... 18
4.1.1. Reuse of Clinical Data for Research and Innovation ... 18
4.1.2. Patient-level Data Sharing ... 18
4.1.3. Anonymous Data ... 19
4.2. Data processing in the context of the EURACAN registry ... 19
Role of Participating Centers: Autonomous Controller ... 19
Legal agreement for the EURACAN registry ... 20
Main clauses of the Memorandum for the functioning of the EURACAN registry and its annexes ... 20
Legal Board for the negotiation ... 20
5. The future and ways forward ... 21
5.1. The current context ... 21
5.2. Strengthen the EURACAN registry ... 21
5.3. Capability assessment of existing organisation ... 22
5.4. Future scenarios ... 24
Leveraging EORTC existing capabilities ... 24

Leveraging DigiCore technological capabilities 24
Leveraging INT and other EURACAN members 24
Continue and scale with INT 25
5.5. Founding principles of the registry 25
5.6. Way forward 25
Appendix 1 - EURACAN registry capability framework 26
Appendix 2 - GDPR and Blueberry 31
Appendix 3 - Data Sharing in Sweden 44
Aim of the Blueberry project in Sweden 44
Registry governance in Sweden 44
Work on Blueberry in Sweden 45
Lessons learned 45
Appendix 4 - Results OMOP-on-Vantage6 studyathon 47
Appendix 5 - Privacy Enhanced Technologies and Consent Management 71
A5 1. Introduction 4
1.1 Purpose of the deliverable 4
1.2 Background information 4
1.3 Approach 4
1.4 Document structure 5
A5 2. Privacy Enhancing Technology, Anonymisation and Flexibility on using Rare Cancer Data 7
A5 2.1. Discussing PET and GDPR 10
A5 3. European Health Data Space 14
A5 4. Blueberry – the process flow 16
4.1. Adopting federated learning/analysis - Vantage6 18
4.2. Data storage, archiving and deletion 20
4.3. Managing Privacy Risks When Re-Running Analyses with New Data 21
A5 5. Conclusion – EURACAN complementing EHDS 24
5.1. Enhancing Data Privacy and Security through Federated Learning 24
5.2. The Role of Governance in Ensuring Compliance and Trust 24
5.3. Bridging the Gap between Local and Pan-European Data Initiatives 25
5.4. Future Directions and Potential Consent Challenges 25
5.5. Conclusion 26
A5 6. References 27
A5 7. Epilogue: Technological opportunities for consent management 28
I. Overview of key areas of innovation 28
II. Overview of key areas of innovation 28
III. Implementation Considerations 33
IV. Conclusions and recommendations 34

Executive Summary

The primary objective of the BlueBerry project is to develop a sustainable, scalable, and impactful data infrastructure for rare cancers in Europe based on data from EURACAN centres and registries. This project builds on the previous STARTER project, which established a proof of concept for a European rare cancer registry. This document addresses the (techno-)legal challenges identified in the STARTER project to ensure that sensitive patient data remains at the source while allowing for comprehensive data analysis across Europe.

The EURACAN initiative, through its innovative use of federated learning and strong governance practices, exemplifies how specialized health data can be utilized within the framework of the European Health Data Space. By balancing the need for comprehensive data analysis with stringent privacy protections, EURACAN not only contributes to the advancement of rare cancer research but also sets a standard for how health data can be shared and used across Europe. As EURACAN continues to evolve, it will undoubtedly play a key role in shaping the future of health data sharing in the European Union, demonstrating that it is possible to achieve both privacy and progress in the pursuit of better health outcomes for all.

In the pursuit of state-of-the-art analysis within the EURACAN framework, a delicate balance must be struck between ensuring robust privacy protections and maintaining the high quality of data necessary for meaningful research. Decentralizing data, as seen in the Blueberry project, aligns with stringent privacy requirements by keeping sensitive patient information at the source, thereby minimizing risks associated with data breaches or unauthorized access. However, this decentralization inherently limits the flexibility and usability of data, posing challenges for comprehensive analysis. Privacy-Enhancing Technologies (PETs) such as federated learning offer a promising solution by enabling researchers to conduct advanced analyses without direct access to raw data. Yet, these technologies must navigate the inherent trade-offs between privacy and data utility—too much anonymization or noise can degrade the data's value, while too little can compromise privacy. Thus, achieving the right balance involves not only the careful selection and implementation of PETs but also a governance framework that allows for strategic flexibility in data use, ensuring that the data remains sufficiently rich and usable to drive innovation, while still upholding the highest standards of privacy.

A5 1. Introduction

1.1 Purpose of the deliverable

This document describes the technological opportunities for consent management and data anonymization identified and evaluated for the EURACAN registry. This document is the result of task 2.7 - Consent Management and data anonymization.

The objective of this document is to contribute to the report on Patient consent and anonymization [D1.3] and the overall Blueberry blueprint [D0.6] with technological opportunities to share medical data for secondary use in a privacy enhanced way.

1.2 Background information

The BlueBerry project is a two-year initiative running from September 2022 to September 2024. Its primary objective is to develop a sustainable, scalable, and impactful data infrastructure for rare cancers in Europe based on data from EURACAN centres and registries. This project builds on the previous STARTER project, which established a proof of concept for a European rare cancer registry.

BlueBerry aims to address several challenges identified in the STARTER project, including organizational, legal, financial, and practical issues. It employs a multidisciplinary approach to create a detailed blueprint for the EURACAN registry, focusing on a federated data infrastructure. This approach ensures that sensitive patient data remains at the source while allowing for comprehensive data analysis across Europe.

The project is managed by IKNL in collaboration with the Fondazione IRCCS Istituto Nazionale dei Tumori and is funded by the Dutch Cancer Society. Longer term key goals for the EURACAN Registry include improving patient care and outcomes and supporting rare cancer research. This is done by the usage of real world data on rare cancer in Europe and established via a robust European registry for all rare adult solid cancers.

BlueBerry incorporates federated learning techniques to enhance data privacy and security while enabling extensive research capabilities. The project has already seen significant milestones, including the completion of the first version of its blueprint, which will guide the development of subsequent versions and the overall registry framework.

1.3 Approach

To investigate and analyse the technological opportunities for consent management and data anonymization for the EURACAN registry, the following approach could be implemented:

- Research and evaluate potential technological solutions regarding consent management, data anonymization technologies and secure data sharing;
- Evaluation of promising technologies focussing on compliance with GDPR and other relevant regulations;
- Collaboration with task [T1.4] and using their insights about identified strategies to facilitate/improve/simplify the patient informed consent collection and management;
- Collaboration with task [T1.5] and using their insights about developing a standard and GDPR compliant procedures for data anonymization when consent cannot be obtained.

1.4 Document structure

Chapter 1 – *Introduction* describing the objectives, context and the approach for this document;

Chapter 2 – *Privacy Enhanced Technologies, Anonymisation and Flexibility on using Rare Cancer Data*, provides a technology context for processing data in an anonymized way; Discussing also PET's and GDPR - evaluation and considerations of the technology to process data in an anonymized way and the legal aspects;

Chapter 3 – *European Health Data Space,* describes the secure data sharing context for the Healthcare domain based on a European wide initiative EHDS;

Chapter 4 – *BlueBerry – the process flow,* describes the organisational/process context for the BlueBerry project;

Chapter 5 – *Conclusions*, a summary of the key findings.

The following figure visualizes the position of this document in relation to other deliverables.

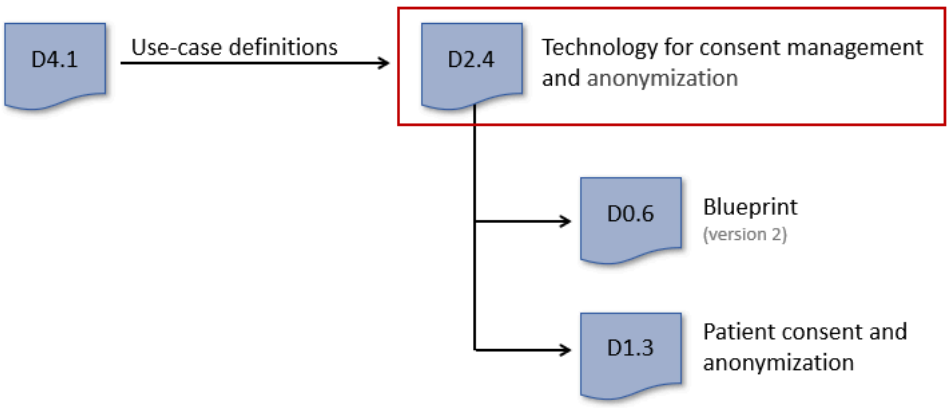


Figure - Relations of D2.4 with other documents

Deliverable D2.4 – Technology Consent Management and Data anonymization and data harmonization is based on the use cases described in deliverable [D4.1] and is input for the overall Blueprint version 2 deliverable [D0.6] and the report about Patient consent and anonymization [D1.3].

A5 2. Privacy Enhancing Technology, Anonymisation and Flexibility on using Rare Cancer Data

Secondary use of data in research on rare types of cancer requires data sharing in any form. Generally, there is simply too little data available per healthcare centre to conduct thorough research, compute statistics and ultimately collect solid evidence to impact patient care*. Is it like searching for a needle in a haystack regarding the necessary analysis, or do we need the whole haystack?*

In recent years, a lot of research has been conducted on privacy-preserving technology to ensure privacy protection. The privacy-enhancing movement is seen as a kind of "holy grail" (Stadler and Troncoso 2022). The only "way out" to open data/read access in any way, for example for research purposes, while also protecting individuals' privacy (Stadler and Troncoso 2022). Privacy Enhancing Technologies (PETs), which have advanced rapidly over the past few years, such as MPC, Federated Learning, allowing analysts to gain insights without access to sensitive, person-identifiable data or the centralisation of data.
In this way, the researcher has the opportunity to harness the full potential of multiple datasets while the data holder remains in control. However, current PETs do not address the most relevant practical issue: how to share high-quality individual data in a way that ensures privacy is maintained, but the full potential of a dataset can be utilised. This is particularly crucial to foster innovation in the field of rare types of cancer.

What is often seen in the preparation of many datasets is "cleaning" the data to ensure anonymity/pseudonymity. For example, removing certain fields so that the identity can no longer be traced. In the Netherlands, it is in general not possible to use the BSN number for secondary use. Non-government organisations may only use the BSN if this is required by law. . Besides MPC and Federated Learning, synthetic data has recently gained quite a bit of momentum. With again golden promises, but in practice faces the same challenges. The data most vulnerable to privacy attacks under anonymisation techniques, statistical outliers that often belong to minority subpopulations, can only be protected from privacy breaches if the published synthetic data does not retain the full promised value of the original dataset (Stadler and Troncoso 2022).

A lot of research has shown that sharing high-dimensional datasets in a way that preserves both privacy and high usability is nearly impossible. Recent scientific research has highlighted the significant challenges in sharing high-dimensional datasets while preserving both privacy and usability. The following key findings from recent papers substantiate that claim:

- The fundamental tension between privacy protection and data utility is particularly pronounced for high-dimensional datasets. As the number of attributes increases, it becomes increasingly difficult to maintain both strong privacy guarantees and high data utility (Shi et al. 2023), (Gadotti et al. 2024).
- Traditional de-identification techniques like k-anonymity and l-diversity have been shown to be inadequate for high-dimensional data. These methods often result in

excessive information loss or fail to provide meaningful privacy protection when applied to datasets with many attributes (Gadotti et al. 2024).
- The "curse of dimensionality" poses a major challenge for privacy-preserving data publishing. As the number of dimensions increases, data points become more sparse and unique, making it easier to re-identify individuals even from seemingly anonymized data (Shi et al. 2023), (Chu et al. 2023).

While progress is being made, the fundamental tension between privacy and utility in high-dimensional data sharing persists. Current research suggests that achieving both strong privacy guarantees and high usability for such datasets remains an open challenge in the field of privacy-preserving data publishing.

Therefore, we conclude that the ongoing pursuit of a fully flexible, highly usable, strong-privacy data release mechanism comes close to chasing rainbows. As difficult as it may be, both researchers and practitioners should finally accept the inherent trade-off between high flexibility in data usage and strong privacy guarantees, even if this means reducing the scope of data-driven applications. Depending on the data used, the goals of data sharing, and their privacy requirements, data holders will need to make explicit choices about the data-sharing methods most suitable for their use case (Stadler and Troncoso 2022).

We conclude that privacy researchers and policymakers need to reconsider their current approach to supporting data holders in their goal of sharing data in a privacy-preserving way. As a first step, both groups should give up the futile search for a panacea for all-purpose-utility high-privacy sharing of detailed data. Instead, we argue that data holders must accept that the set of use cases that can be addressed under strict privacy guarantees may be limited, and so too are the data-driven business models linked to them. Privacy researchers should therefore reorient their efforts toward developing tools that help data holders identify those use cases that can be addressed simultaneously under good privacy and good usability conditionss. Finally, we recommend that policymakers, together with technical experts, develop guidelines to help data holders navigate the complex landscape of Privacy Enhancing Technologies (PETs). These guidelines should not only focus on matching use cases with their appropriate sharing technologies but also include recommendations for empirical evaluation methods that can assure the public that any loss of privacy is balanced against the promised societal benefits.

It thus remains a kind of constant balancing act, in making trade-offs. The governance & process of how to handle (meta) data with the data holder as well as the data processor is crucial. After which technology (partly) contributes to protecting/safeguarding privacy, but it remains a balance on a thin line.

There are a few approaches in this regard, namely anonymity (pseudonymising and/or anonymising individual data) and separately perhaps obtaining prior patient consent.

At BlueBerry, the choice has been made to work with anonymous data per research centre. This is to avoid a substantial legal burden. As described above, both paths are valid, but ensuring anonymity is not straightforward.

Example: Three healthcare centres conduct a federated learning/analysis. Where anonymity is established based on several criteria (aside from what these are), but the data sets are, of course, not very large due to the rare types of cancer. If a researcher conducts this analysis at time A and then conducts another analysis at time B with a different result, it could mathematically be deduced who this new individual is. Of course, this involves complex calculations and such, but the risk cannot be completely ruled out.

Given the challenges described above, having clear and explicit agreements is crucial. Such an agreement framework or rulebook helps reduce risks or at least be aware of them. Governance is crucial.

At EURACAN, this has been thought through by making several preliminary agreements. A steering board/committee issues several conditions beforehand, such as when data is considered anonymous, algorithms and study protocols must be pre-approved, after which the researcher can use these approved building blocks. By integrating process and technology from the start, one can speak of privacy-by-design. As mentioned earlier, ruling everything out is impossible, and such risks cannot be excluded; otherwise, it becomes unworkable.

Making data available to a researcher is crucial depending on the respective goal/purpose, with the data minimisation principle[8] applied. In addition, the other requirements of the GDPR must be met.

- *adequate – sufficient to properly fulfil your stated purpose;*
- *relevant – has a rational link to that purpose;*
- *and limited to what is necessary – you do not hold more than you need for that purpose.*

(ICO definition - Principle (c): Data minimisation | ICO)

Whereby it is not always clear what is adequate, relevant and limited. This remains partly a grey area per member state within the EU. Everything obviously depends on the purpose limitation.

A lung cancer screening and risk prediction study for example illustrates the challenges in determining what data is adequate, relevant, and limited. Here's an elaboration on why these distinctions are not always clear-cut. Determining what constitutes adequate data for lung cancer research can be complex considering the Smoking history. While pack-years of smoking is crucial, it may not capture the full picture. For instance, the intensity of smoking (cigarettes per day) and duration might have different impacts on risks (Callender et al. 2023). Assessing relevance can be challenging in case of Ethnicity. Its inclusion is relevant due to varying risk factors among different groups, but it's not always clear how to categorize mixed ethnicities or how much detail is necessary (Callender et al. 2023). Limiting data collection while ensuring comprehensive risk assessment can be tricky in case of a partial

[8] Principle (c): Data minimisation | ICO

postcode for example. While limited to protect privacy, it might not provide sufficient granularity for accurate air pollution exposure assessment.

In general it is not always clear because of:

1. Evolving research: New studies constantly reveal new risk factors or refine our understanding of known ones. For example, recent research has highlighted the importance of considering younger populations for lung cancer risk (Callender et al. 2023);
2. Individual variations: What's adequate or relevant for one person might not be for another. For instance, occupational exposure might be highly relevant for some individuals but not others;
3. Technological advancements: As diagnostic technologies improve, the definition of "adequate" data may change. For example, more sophisticated CT scans might provide more detailed information, blurring the line between adequate and excessive data;
4. Balancing privacy and research needs: While limiting data collection protects privacy, it might hinder the development of more accurate prediction models. For instance, more detailed geographical data could improve risk assessment related to environmental factors but could also compromise anonymity;
5. Intersectionality of risk factors: The relevance of certain data points might depend on their interaction with others. For example, the relevance of BMI might vary depending on smoking history or ethnicity;
6. Future research potential: Data that seems irrelevant now might become crucial in future studies, making it difficult to determine what's truly limited to the current purpose.

In conclusion, the dynamic nature of medical research, individual variability in risk factors, and the need to balance comprehensive risk assessment with data protection make it challenging to definitively categorize data as adequate, relevant, or limited in lung cancer research. This complexity underscores the importance of regular reassessment of data collection practices in light of new research findings and evolving ethical standards[9], (Chandran et al. 2023)

Everything revolves around risk management, which is why a table with possible measures to reduce risks concerning personal data, while maintaining some workability for secondary use, is essential. After all, collecting all data in one place, in a bunker, with guards checking whether the researcher's identity is correct, the purpose limitation, and then on a computer without an external connection, does not help to promote innovation.

A5 2.1. Discussing PET and GDPR

There is currently legal uncertainty about how Privacy Enhancing Technologies (PETs) are related to the GDPR and whether using certain PETs, either individually or in combination,

9 https://ico.org.uk/for-organisations/uk-gdpr-guidance-and-resources/data-protection-principles/a-guide-to-the-data-protection-principles/the-principles/data-minimisation/

can ensure that data is considered anonymous in a legal sense. In this section, we provide an overview of some of the latest literature on this topic, particularly focusing on the use of federated learning.

In 2021, the Directorate-General of Health and Food Safety released a detailed report examining the rules governing the processing of health data in EU Member States in light of the GDPR (Cite, assessment). The report aimed to highlight differences, identify factors affecting cross-border exchange of health data, and propose actions to support health data use and reuse in the EU. It focused on the processing of health data for three interlinked purposes: direct patient care (primary), supporting safe and efficient healthcare systems (secondary), and driving health and research innovation (secondary).

The report found that while the GDPR was a generally appreciate, a number of legal and operational issues needed to be addressed to ensure that European healthcare systems could make the best possible use of health data while ensuring that the patients' privacy was appropriately protected. It found that variations in the interpretation and implementation of the GDPR have led to a fragmented approach, making cross-border cooperation for care provision or research difficult. While the GDPR itself is applied universally across the EU, as a regulation, it allows Member States to adopt legislation to allow for the use of data for research under Art. 9(2)(j) and 89(1). The report found that such legislation has not been implemented in a homogenous way, creating a complex and fragmented landscape for researchers to navigate.

The report found that consent is the first basis for using health data for research in the Netherlands. The release of patient data outside the treatment requires, as per the WGBO, the patient's consent - unless the requirements for making an exception are met. These are: a) if the research cannot be performed without the data, b) serves the common interest, c) the patient has not opted out, and d) sufficient safeguards have been taken to prevent re-identification of the patient and the protection of their privacy.

This brings us to the question of what counts as personal data, and how such personal data can be pseudonymised or anonymised in order to prevent re-identification under the existing regulations. There are, broadly, two approaches to identifying whether data qualifies as personal data or anonymous data under the GDPR, based on differences in interpreting Art. 4(1) and Recital 26: **the absolute approach and the relative approach**.

The absolute approach, first put forward by the Article 29 Working Party (now the European Data Protection Board), requires that data classified as anonymous carries **absolutely no risk for re-identification [Cite, A29]**. This approach has been quoted by the EDPB since, and has also been reiterated in rulings from some national authorities (for instance, Austria [cite] and France [cite]).[10]

---

[10] [More detailed:

- *In practice, the answer to the question whether data qualifies as personal data or anonymous data depends on the approach followed, arising from a difference in the interpretation of art. 4(1) GDPR and recital 26 GDPR:* ***the absolute approach and the relative approach****. The*

In contrast, the **relative approach** accepts that there is always a risk of re-identification that remains. In this approach, the personal data is only considered personal if it is in the hands of an entity that may reasonably have access to the requisite data that enables its reidentification. Essentially, this approach requires an assessment of whether the party that receives the de-identified data has means that can reasonably be used to identify (a) specific individual(s). This approach was put forward by the CJEU in the case of Breyer v. Germany [cite], and further confirmed in the case of SRB v. EDPS [cite].

The CNIL released in 2023 the results of two sandboxes that it conducted, one of which was focused on digital health (link 1, 2 and 3). In one of the four projects conducted in this project, the CNIL and Inria's Magnet research team supported Lille University Hospital in setting up a federated learning protocol for an algorithm that would use data hosted in server health data warehouses to facilitate patient care. The full results (in French) can be found here.

The results find that determining the applicable legal regime requires establishing the nature of aggregates resulting from the iterations, i.e., whether they are personal or anonymous data. They find that the data anonymity of the data should be checked as far upstream as possible. If the anonymity of the aggregates or other data used in the process cannot be verified, it becomes mandatory to perform an analysis to identify the risk of re-identification from this data. The report also notes two examples of steps that can support the risk analysis, specifically, using an explainable algorithm or using a small number of parameters.

(He 2023) traces the evolution of anonymisation and psuedonmyisation in EU regulation, from before the GDPR entered into force.

(Brauneck et al. 2023) perform a scoping review of a subset of academic literature to answer four questions: whether 'local' and 'global' models are personal data, which roles of the GDPR apply to the parties involved in a federated learning dataflow, who 'controls' at the different stages of the dataflow, and how the usage of PETs affects the answers to these questions. They find that local and global models are both likely to constitute personal data, and that while federated learning can strengthen data protection it is still likely vulnerable to attacks and data leakage. Crucially, they also find that concerns regarding attacks and leakages can be successfully addressed through the use of PETs, specifically differential privacy and secure multiparty computation.

---

different approaches determine the perspective the controller has to take into account in the assessment provided by recital 26 GDPR.

- *The absolute approach was put forward by the Article 29 Working Party (currently: the European Data Protection Board, EDPB) in its opinion on anonymization techniques from 2014, which opinion is still quoted by the EDPB. Rulings from national authorities (for instance Austria and France) follow the absolute approach as well.*

]

(Rossello, Díaz, and Muñoz-González 2021), similarly, also find that the results from the training of federated learning models may still leak personal data, and further that the raw data of federated learning models may also be vulnerable to poisoning attacks.

(Kist 2022) conducts an analysis of the Dutch legislation on the secondary use of health data for scientific research, and finds that the GDPR, its local implementation and the sectoral health legislations leave room for alternatives. She notes that the Dutch Medical Treatments Contract Act provides for the general rule of consent for the secondary use of data for research.

A5 3. European Health Data Space

The European Health Data Space (EHDS) is a regulatory framework proposed by the European Commission to facilitate the secure and standardized sharing of health data across the EU. It aims to enable better healthcare delivery, research, and innovation by making health data more accessible while ensuring robust data protection.

In the context of secondary use of data, the EHDS is designed to allow health data to be used for purposes beyond direct patient care ("primary use"), such as research, policymaking, and innovation in health technologies. This includes using data for developing AI-driven health solutions, conducting large-scale health studies, and improving public health strategies.

The EHDS aligns with other key EU legislative frameworks:

A5 4. **The General Data Protection Regulation (GDPR)**: The EHDS builds on GDPR principles, particularly around consent and data minimization, ensuring that data subjects' rights are protected when their data is used for secondary purposes.

A5 5. **The Data Governance Act (DGA)**: The DGA establishes frameworks for the re-use of sensitive data held by public sector bodies, including health data, ensuring that this data can be shared securely and ethically across different stakeholders.

A5 6. **The Data Act**: This act focuses on fair access to and use of non-personal data. While it primarily addresses industrial data, it supports the EHDS by promoting data sharing and interoperability standards that could be applied to health data in a way that respects individual privacy.

A5 7. **The AI Act**: The AI Act regulates the use of artificial intelligence in high-risk areas, including healthcare. The EHDS will interact with this by ensuring that health data used to train AI models is handled in compliance with the AI Act's requirements for transparency, safety, and non-discrimination.

Together, these legislative frameworks create a robust environment where health data can be shared and used to drive innovation while safeguarding the rights and interests of EU citizens.

If we take a closer look at a possible implementation of the EHDS, we see that TEHDAS is making an advance on process and infrastructure flows.

Figure 2 High Level Infrastructure process flow

If this approach is ultimately adopted as the implementation (Figure 2) , which seems likely given the TEHDAS project and its follow-up, TEHDAS II.

The current Blueberry project provides a valuable further elaboration of these steps within the context of EURACAN.

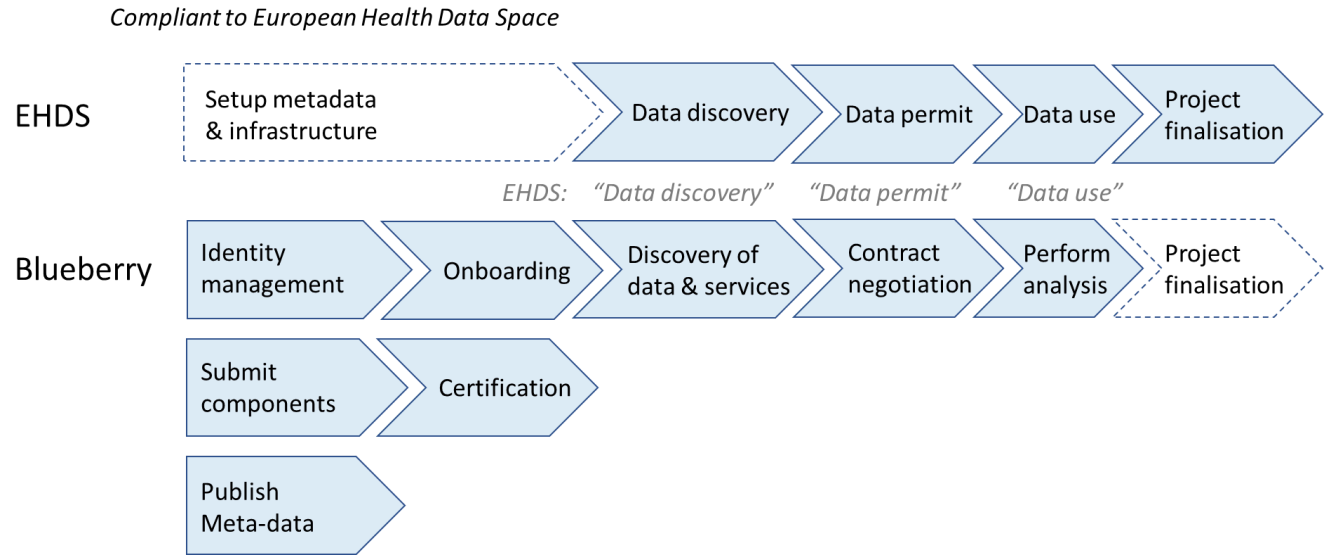


Figure 3 EURACAN as an implementation of EHDS for rare types of cancers (secondary use of data)

In Work Package 1, significant efforts have been made on the legal aspects related to conducting research for secondary use, after which Work Package 2, in particular, further elaborates and proposes how to technically implement and detail this. The following chapters will define this further.

A5 4. Blueberry – the process flow

Blueberry is a successor or an extension of the EURACAN Registry pioneered in STARTER. See the figure below. There are two key elements, namely the study protocol and the Blueberry Data Collaboration Agreement. Additionally, the DPIA is, of course, also available. The EURACAN Steering Committee is a governing body responsible for overseeing and guiding the strategic direction, policies, and operations of EURACAN, which is a European Reference Network (ERN) focused on rare adult solid cancers. The committee is typically composed of representatives from the network's member institutions, including clinical experts, researchers, and patient advocates. Its primary roles include coordinating the activities of EURACAN, ensuring compliance with the network's objectives, facilitating collaboration among its members, and making decisions on key issues such as research priorities, clinical guidelines, and patient care standards. The Steering Committee plays a crucial role in promoting the network's mission to improve the diagnosis, treatment, and management of rare adult solid cancers across Europe through a collaborative and multidisciplinary approach.

The EURACAN Registry steering committee in the figure below serves as the important gateway for approved protocols and/or algorithms. It determines the types of studies, the methods used, and the technologies involved. Some of the risks described in Chapter Two can be mitigated through this approach, which can be seen as a type of mitigating measure if data remains in various locations and at the source, but you are still attempting to find the needle in the haystack while preserving privacy. By strictly adhering to protocols and types of algorithms, a portion of the risk is covered. Additionally, for each participating healthcare centre (referred to as a party in the collaboration agreement), it remains the autonomous data controller. In other words, they decide what is and isn’t available. By keeping responsibility with each health centre, a shared responsibility is created, which offers several advantages but may also entail some risks.

Each party checks its own data quality (more than just converting it to OMOP, as described in other deliverables). However, the focus is particularly on the quality of the data per individual so that the researcher can genuinely use the data to perform analyses. A key responsibility also lies in anonymising the data at the source. It is not yet entirely clear how this will be structured, but a contribution from the approved protocols may be expected.
A potential option not specifically included in the diagram below is patient consent. For example, incorporating a simple consent from a patient for the use of scientific research in the EURACAN Registry. This could add significant value concerning the data quality of an individual, with anonymisation/pseudonymisation controls still in place. The figure below shows the process flow:

- for a researcher to execute an analysis;
- each health centre regarding anonymisation/pseudonymisation;
- Euracan registry steering committee regarding approved protocols/algorithm upfront so researcher can select them.

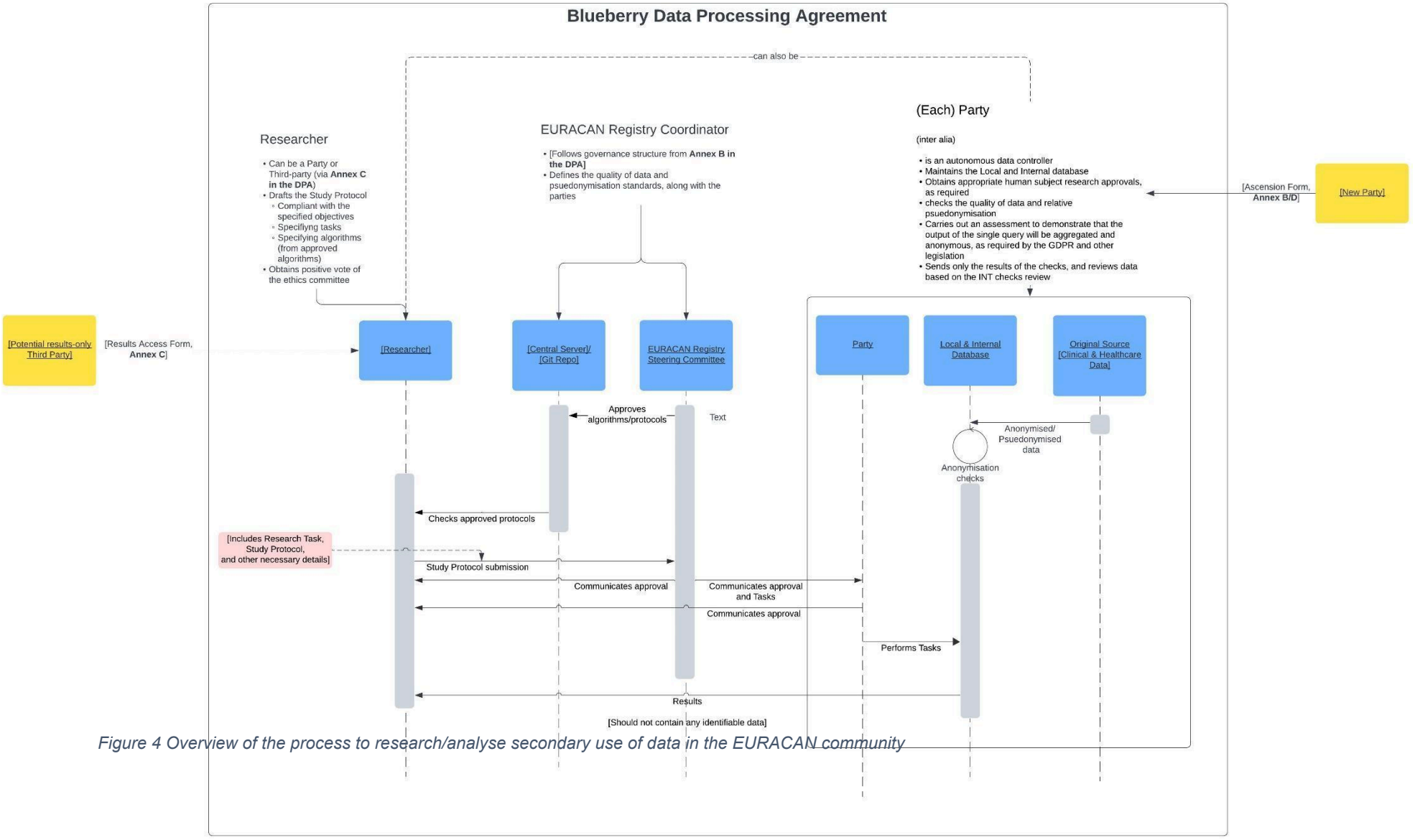


Figure 4 Overview of the process to research/analyse secondary use of data in the EURACAN community

4.1. Adopting federated learning/analysis - Vantage6

Traditional data analysis methods tend to rely on Data-to-Algorithm (D2A) approaches, where the data to be analysed has to be transferred to the entity or individual performing the analysis. This means that the data must be transferred from the data source to the researcher.

The STARTER/BLUEBERRY project for the EURACAN registry uses Privacy-Enhancing Tools such as federated learning, which implement an Algorithm-to-Data (A2D) approach. In this approach, the personal data stays at the source, and may not be accessible to the researcher. This section elaborates on the implementation of federated learning in the STARTER/BLUEBERRY project for the EURACAN registry [across the data lifecycle].

Vantage6
is a federated learning tool designed to enable secure and privacy-preserving machine learning across distributed datasets. It allows multiple organizations to collaboratively train machine learning models without the need to share or centralize sensitive data. Vantage6 facilitates this by bringing the algorithm to the data rather than bringing the data to a central location, ensuring compliance with privacy regulations while enabling powerful analytics and research across various institutions.

If we then look at further elaboration from the legal context, the EURACAN Registry Steering Committee, with the help of the available technology such as Vantage6[11], we arrive at the following process steps, see the figure below.

Where each health centre will install a Vantage6 client. The technology will continuously validate the approved study protocol from a technological perspective, ensuring that no agreements are breached and the like. This can be achieved by specifying policies in a structured way (for instance with ODRL[12]) to enable automatic policy enforcement. A possible addition could be the use of various policies, such as limiting the number of times an analysis can be performed within a certain time period to safeguard the privacy of a newly added individual. Or requiring that an analysis must always contain a minimum of Y results. Many such policies can be devised and implemented in advance, which can then be translated into machine-readable code using Open Digital Rights Language during operation, making the translation from legal to technical possible. This also helps to reduce risks while maintaining flexibility. Figure 5 is a further detailed process flow of Figure 4 where Vantage6 has been chosen a PET technology. It gives a clear overview per step what needs to be executed.

[11] https://distributedlearning.ai/vantage6/

[12] The Open Digital Rights Language (ODRL) is a policy expression language that provides a flexible and interoperable information model, vocabulary, and encoding mechanisms for representing statements about the usage of content and services. ODRL became an endorsed W3C Recommendation in 2018.

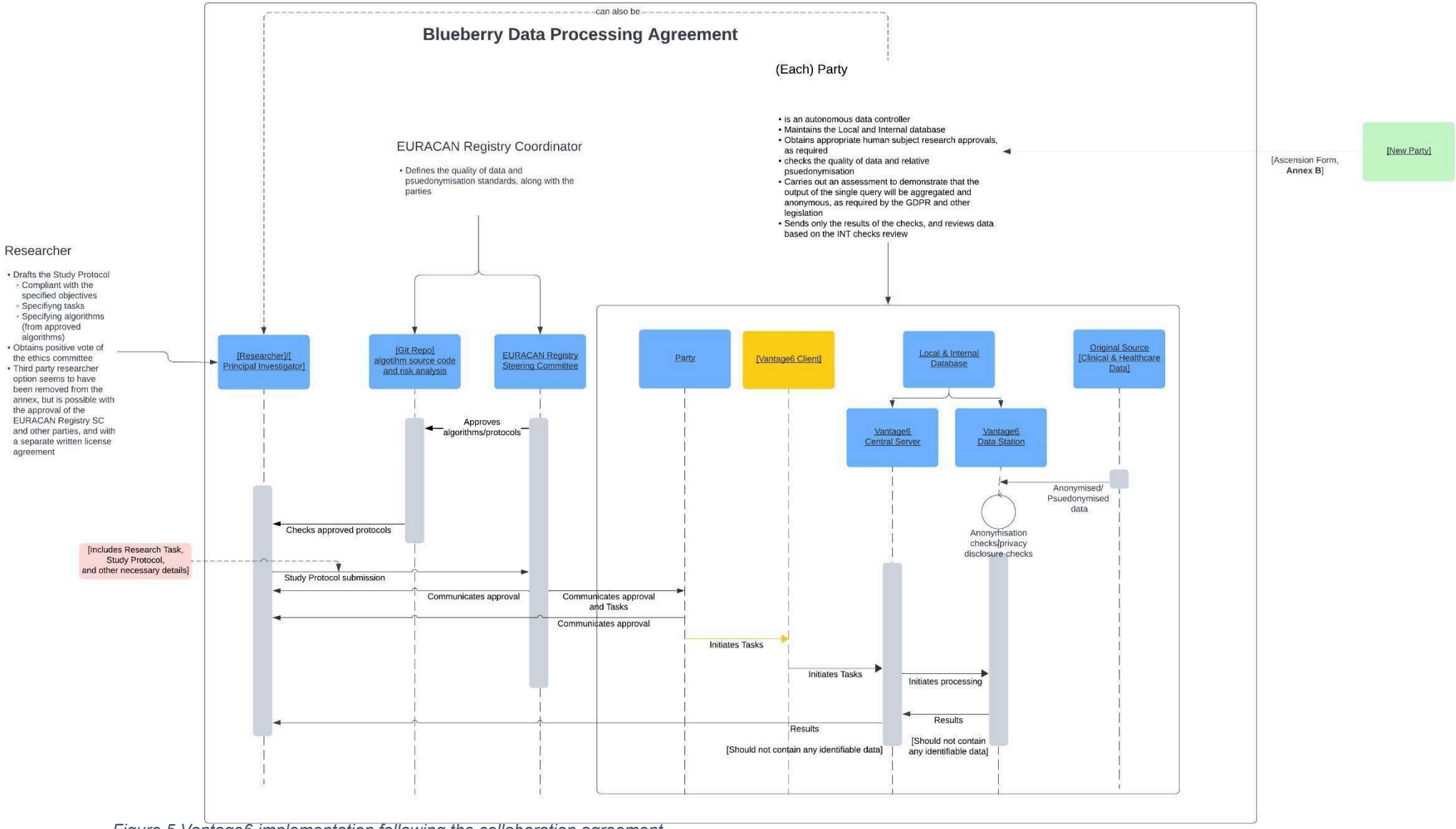


Figure 5 Vantage6 implementation following the collaboration agreement.

These technologies enable analysis of data, while protecting the sensitive information of individual data subjects. Each technology brings its own form of complexity to the analysis and often a mix of them is required to get the most effective result. This usually depends on the research question you would like to answer, the actors involved, the type of data, the analysis methods, computational resources, and presence of other available safeguards.

4.2. Data storage, archiving and deletion

In the context of the EURACAN registry and the Blueberry project, where data remains decentralized and under the control of individual health centers, the strategies for data storage, archiving, and deletion must be carefully managed to ensure compliance with privacy regulations and the integrity of the data.

Data storage in this decentralized system is maintained locally by each participating health center. This approach ensures that sensitive patient data does not leave the jurisdiction of the health center, aligning with the principles of data minimization and sovereignty. Each health center is responsible for maintaining secure storage environments that comply with national and EU regulations, including the GDPR. This includes implementing encryption both at rest and in transit, ensuring that access controls are stringent and regularly updated, and that only authorized personnel have access to the data.

Furthermore, the use of federated learning technologies like Vantage6 necessitates that the data storage infrastructure is compatible with the federated learning protocols. This includes the capability to securely run algorithms on the data without exposing the raw data itself to external entities.

Archiving data in this context involves maintaining historical data for long-term use, particularly for ongoing or future research. Given the sensitive nature of health data, archiving must be approached with strict adherence to legal and ethical guidelines. Each health center should establish clear policies for data archiving that specify the duration for which data will be retained, the security measures in place to protect archived data, and the conditions under which archived data can be reactivated for use in research.

Data archiving policies should also include provisions for auditing and tracking access to archived data, ensuring that any retrieval or use of archived data is fully documented and justifiable under the agreed-upon research protocols.

Data deletion is a critical aspect of maintaining compliance with data protection laws and ensuring that unnecessary data is not retained longer than needed. In the context of federated learning and decentralized data management, data deletion must be handled by the health center that owns the data.

Each health center must establish a clear deletion policy that outlines when and how data will be securely deleted, including the deletion of any backups or archived copies. Deletion protocols should ensure that once data is deleted, it cannot be recovered, thereby protecting patient privacy. This may involve the use of secure deletion tools that comply with recognized standards for data destruction.

In scenarios where data has been used in federated learning, the deletion policy should also consider the potential for residual data to exist in aggregated models or outputs. While federated learning is designed to prevent direct access to raw data, it is essential that health centers and researchers understand the implications of deletion in the context of derived data products.

Health Center Responsibilities

As the data controller, each health center holds primary responsibility for the data it stores, archives, and deletes. This includes ensuring that all processes align with the agreed-upon governance frameworks and that any data handling is conducted under the oversight of the EURACAN Steering Committee, particularly when it involves cross-center collaborations or the use of federated learning technologies.

Regular reviews and updates to data storage, archiving, and deletion policies are necessary to adapt to evolving legal requirements and technological advancements. The implementation of these policies should be documented in the Data Protection Impact Assessments (DPIAs) for ongoing and future research projects, ensuring transparency and accountability in the use of patient data.

This approach balances the need for maintaining the privacy and security of sensitive health data with the flexibility required for effective research in a decentralized, federated learning environment.

4.3. Managing Privacy Risks When Re-Running Analyses with New Data

In the context of secondary use of data within the EURACAN registry and the Blueberry project, a critical consideration is the potential privacy risks associated with re-running analyses after new data has been added. This section outlines strategies to mitigate the risk of re-identifying individuals, especially when datasets are updated with new patient information.

Risk of Re-Identification

When new data is added to an existing dataset and the same analysis is conducted again, there is an inherent risk that the identity of newly added individuals could be inferred, especially if the dataset is small or contains unique characteristics. This risk arises from the possibility that differences in analysis results before and after the addition of new data might inadvertently reveal information about the new individuals, even if the data is anonymized.

For example, if an initial analysis identifies certain trends or patterns in a small dataset, and these patterns shift slightly when new data is introduced, a sophisticated observer might deduce information about the new individuals based on these changes.

Mitigation Strategies

To address these risks, the following strategies should/could be implemented:

1) Aggregation and Thresholding: Implement aggregation techniques that ensure analysis results are based on groups of individuals rather than single entries. Setting a minimum threshold for the number of individuals required to produce an output (e.g., at least 5 or 10 individuals) can help obscure the contribution of any single new data point.

2) Differential Privacy: Apply differential privacy techniques that introduce controlled noise into the analysis results. This noise ensures that the inclusion or exclusion of a single individual does not significantly affect the outcome, thereby reducing the risk of re-identification.

3) Restricted Re-Analysis Frequency: Limit the frequency at which re-analyses can be conducted after new data is added. By imposing a time interval between successive analyses, the opportunity to pinpoint changes linked to the addition of specific individuals is reduced.

4) Mandatory Data Blending: Require that new data be blended with sufficient historical data before analysis. This blending ensures that the analysis reflects a broader population and reduces the likelihood that the characteristics of a few new entries will be discernible.

5) Review and Validation of Algorithms: Ensure that all algorithms used for analysis are reviewed and validated by the EURACAN Steering Committee before and after the inclusion of new data. This review should confirm that the algorithms incorporate adequate safeguards against re-identification risks, particularly in the context of small or rare datasets.

6) Role-Based Access Control: Implement strict role-based access controls to ensure that only authorized personnel have access to the raw data and analysis outputs. Researchers should only receive results that are necessary for their specific study, and these results should be subject to privacy-preserving techniques.

7) Transparency and Audit Logs: Maintain detailed logs of all analyses conducted, including the addition of new data, to ensure full transparency. Regular audits should be performed to verify that privacy-preserving measures are consistently applied and that no breaches of privacy have occurred.

8) Ethical Oversight and Informed Consent: Where possible, ensure that ethical oversight committees review the protocols for handling new data and re-running analyses. Additionally, consider obtaining broad consent from patients for the potential use of their data in ongoing research that may include re-analysis as new data becomes available.

Conclusion

Managing the privacy risks associated with re-running analyses after new data is added is essential to maintaining the integrity of the federated learning approach while safeguarding patient privacy. By implementing these mitigation strategies, the EURACAN registry can continue to leverage valuable health data for research purposes without compromising the confidentiality of the individuals involved.

A5 5. Conclusion – EURACAN complementing EHDS

The EURACAN initiative, supported by the Blueberry project, represents a significant advancement in the development and integration of a pan-European data infrastructure for rare adult solid cancers. This initiative not only enhances the quality of cancer research but also aligns closely with the broader objectives of the European Health Data Space (EHDS).

The EHDS aims to create a secure and standardized framework for the sharing of health data across the European Union, facilitating better healthcare delivery, research, and innovation. EURACAN's efforts are a practical realization of these goals, particularly within the specialized field of rare cancers. By leveraging a federated data infrastructure, EURACAN ensures that sensitive patient data remains decentralized, thereby upholding the highest standards of data privacy and security while enabling comprehensive research.

4.1. Enhancing Data Privacy and Security through Federated Learning

One of the core strengths of the EURACAN approach is its adoption of federated learning and other Privacy-Enhancing Technologies (PETs). These technologies allow for the analysis of vast datasets without the need for data centralization, a method that directly addresses privacy concerns inherent in cross-border data sharing. By keeping data at the source within individual health centers, EURACAN minimizes the risks associated with data breaches and unauthorized access, a key requirement under GDPR and other European data protection regulations.

The implementation of tools such as Vantage6 demonstrates how the EHDS's objectives can be met while maintaining compliance with strict privacy standards. This approach not only ensures that patient data is protected but also enables researchers to draw meaningful insights from large, diverse datasets, which are crucial for advancing rare cancer research.

4.2. The Role of Governance in Ensuring Compliance and Trust

Effective governance is central to the success of EURACAN and its alignment with EHDS. The EURACAN Steering Committee plays a pivotal role in overseeing the ethical and legal compliance of data usage within the network. By establishing clear protocols for data sharing, analysis, and re-use, the Steering Committee ensures that all activities within the EURACAN framework adhere to the highest standards of data protection and ethical research.

The governance framework also includes the development and enforcement of Data Protection Impact Assessments (DPIAs), which are essential for identifying and mitigating potential risks associated with the secondary use of health data. This proactive approach to governance not only fosters trust among data providers and patients but also ensures that the research outputs generated under EURACAN are robust, reliable, and ethically sound.

4.3. Bridging the Gap between Local and Pan-European Data Initiatives

EURACAN's methodology offers a blueprint for how local health centers and research institutions can contribute to and benefit from a pan-European data infrastructure. By maintaining autonomy over their data while participating in a larger federated system, these centers can contribute to a collective research effort without compromising their governance or data sovereignty. This model supports the EHDS's vision of a unified yet decentralized health data space, where data flows freely for the benefit of all, yet privacy and local control are preserved.

Moreover, the lessons learned from EURACAN's implementation provide valuable insights for the broader EHDS initiative. As the EHDS framework continues to evolve, the practices established by EURACAN can serve as a guide for other healthcare domains looking to implement similar federated data infrastructures.

4.4. Future Directions and Potential Consent Challenges

Looking ahead, the ongoing challenge for EURACAN and similar initiatives will be to continually adapt to the evolving landscape of data protection and health research. As new technologies and methods emerge, EURACAN must remain at the forefront of innovation while ensuring that all data usage remains compliant with EU regulations and ethical standards.

Additionally, the need for continuous collaboration among member states, health centers, and research institutions will be crucial. Building and maintaining trust in a federated system requires ongoing communication, transparency, and a shared commitment to the principles of data protection and research integrity.

As discussed in the Directorate-General's report, while consent is not the only basis for processing health data for research purposes, it is often the first basis used in the Netherlands. Furthermore, while the EHDS is still in the proposal stage and its final text has not been finalised, the latest text of the proposed regulation requires the setup of opt-in and opt-out mechanisms for patients regarding the secondary use of their personal data in Art. 31a, 33(5), 33(5a), and Recital 39a. The current draft of Art. 33(5) requires natural persons to have the right to opt-out of the processing of their electronic health data for secondary use, through an accessible and easily understandable opt-out mechanism. The new Art. 33 5a requires an easily accessible and understandable and user-friendly opt-in mechanism for

three categories of data.[13] The GDPR, similarly, also provides extensive requirements for consent, including that it should freely given, specific, informed and unambigious – and in the case of processing of sensitive categories of data such as health data, explicit.
Given the scale of this processing, managing such consents, including recording them from the patients and associating them with the appropriate data and databases and ensuring interoperability at the hospital, regional, national, and EU levels, is a significant task. This is now being tackled in the EHDS technical project with the Nationaal Zeggenschapsregister, and also in the Health-RI initiative as one of the steps in their Obstakel Verwijder Traject. Currently, services such as MedMij and Mitz also offer relevant solutions for consent management and the PROVES proof-of-concept service provider is also exploring the same, more details about which are available here.

4.5. Conclusion

In conclusion, the EURACAN initiative, through its innovative use of federated learning and strong governance practices, exemplifies how specialized health data can be utilized within the framework of the European Health Data Space. By balancing the need for comprehensive data analysis with stringent privacy protections, EURACAN not only contributes to the advancement of rare cancer research but also sets a standard for how health data can be shared and used across Europe. As EURACAN continues to evolve, it will undoubtedly play a key role in shaping the future of health data sharing in the European Union, demonstrating that it is possible to achieve both privacy and progress in the pursuit of better health outcomes for all.

In the pursuit of state-of-the-art analysis within the EURACAN framework, a delicate balance must be struck between ensuring robust privacy protections and maintaining the high quality of data necessary for meaningful research. Decentralizing data, as seen in the Blueberry project, aligns with stringent privacy requirements by keeping sensitive patient information at the source, thereby minimizing risks associated with data breaches or unauthorized access. However, this decentralization inherently limits the flexibility and usability of data, posing challenges for comprehensive analysis. Privacy-Enhancing Technologies (PETs) such as federated learning offer a promising solution by enabling researchers to conduct advanced analyses without direct access to raw data. Yet, these technologies must navigate the inherent trade-offs between privacy and data utility—too much anonymization or noise can degrade the data's value, while too little can compromise privacy. Thus, achieving the right balance involves not only the careful selection and implementation of PETs but also a governance framework that allows for strategic flexibility in data use, ensuring that the data remains sufficiently rich and usable to drive innovation, while still upholding the highest standards of privacy.

[13] Specifically:
1. Art. 33(1)(e) human genetic, genomic and proteomic data;
2. Art. 33(1)(fa) data from wellness applications;
3. Art. 33(1)(m) electronic health data from biobanks and dedicated databases.

A5 6. References

Brauneck, Alissa, Louisa Schmalhorst, Mohammad Mahdi Kazemi Majdabadi, Mohammad Bakhtiari, Uwe Völker, Jan Baumbach, Linda Baumbach, and Gabriele Buchholtz. 2023. ‘Federated Machine Learning, Privacy-Enhancing Technologies, and Data Protection Laws in Medical Research: Scoping Review’. *Journal of Medical Internet Research* 25:e41588.

Callender, Thomas, Fergus Imrie, Bogdan Cebere, Nora Pashayan, Neal Navani, Mihaela Van der Schaar, and Sam M Janes. 2023. ‘Assessing Eligibility for Lung Cancer Screening Using Parsimonious Ensemble Machine Learning Models: A Development and Validation Study’. *PLoS Medicine* 20 (10): e1004287.

Chandran, Urmila, Jenna Reps, Robert Yang, Anil Vachani, Fabien Maldonado, and Iftekhar Kalsekar. 2023. ‘Machine Learning and Real-World Data to Predict Lung Cancer Risk in Routine Care’. *Cancer Epidemiology, Biomarkers & Prevention* 32 (3): 337–43.

Chu, Zhiguang, Jingsha He, Xiaolei Zhang, Xing Zhang, and Nafei Zhu. 2023. ‘Differential Privacy High-Dimensional Data Publishing Based on Feature Selection and Clustering’. *Electronics* 12 (9): 1959.

Gadotti, Andrea, Luc Rocher, Florimond Houssiau, Ana-Maria Creţu, and Yves-Alexandre de Montjoye. 2024. ‘Anonymization: The Imperfect Science of Using Data While Preserving Privacy’. *Science Advances* 10 (29): eadn7053.

He, Zhicheng. 2023. ‘From Privacy-Enhancing to Health Data Utilisation: The Traces of Anonymisation and Pseudonymisation in EU Data Protection Law’. *Digital Society* 2 (2): 17.

Kist, Irith. 2022. ‘Assessment of the Dutch Rules on Health Data in the Light of the GDPR’. *European Journal of Health Law* 30 (3): 322–44.

Rossello, Stephanie, MR Díaz, and Luis Muñoz-González. 2021. ‘Data Protection by Design in AI? The Case of Federated Learning’. *SSRN*.

Shi, Wei, Xiaolei Zhang, Hao Chen, and Xing Zhang. 2023. ‘High Dimensional Data Differential Privacy Protection Publishing Method Based on Association Analysis’. *Electronics* 12 (13): 2779.

Stadler, Theresa, and Carmela Troncoso. 2022. ‘Why the Search for a Privacy-Preserving Data Sharing Mechanism Is Failing’. *Nature Computational Science* 2 (4): 208–10.

A5 7. Epilogue: Technological opportunities for consent management

I. Overview of key areas of innovation

For consent management in the context of the EURACAN registry and the secondary use of medical data for research purposes, the following key areas of innovation and application can be distinguished:

- *Dynamic consent models*, digital frameworks that allow patients to continuously manage and update their consent preferences;
- *Blockchain-based Consent management*, a decentralized system that securely records and verifies patient consent, ensuring transparency, immutability, and traceability of consent records across multiple stakeholders;
- *AI powered consent management*, using artificial intelligence to automate and optimize the process of obtaining, monitoring, and ensuring compliance with patient consent;
- *Consent as a Service*, a cloud-based solution for managing and tracking patient consent simplifying the process for both patients and researchers;
- *Federated Consent Management*, a system that allows patient consent preferences to be recognized and respected across multiple institutions or organizations, enabling seamless data sharing;
- *Interoperable Consent Standards*, standardized frameworks that enable consistent and seamless exchange of consent information across different systems and platforms;
- *Regulatory and Ethical Compliance Technologies*, systems designed to ensure that data handling and consent management processes adhere to legal standards and ethical guidelines.

II. Overview of key areas of innovation

Via desk-research a quick-scan of the technological opportunities has been performed and results are summarized in this section including a conclusion or recommendation.

Dynamic consent models

Dynamic consent models are flexible, digital frameworks that allow patients to continuously manage and update their consent preferences for the use of their personal data in research, adapting to changing circumstances and new research opportunities over time. Adaptability is an aspect to take in account for registries like EURACAN, where research needs can evolve rapidly. The following two types of applications are identified: “personalized consent management” and “ongoing patient engagement”.

Personalized Consent Management: Dynamic consent platforms allow patients to tailor their consent preferences in real-time. Patients can specify the types of research their data can be used for, set time limits on their consent, and adjust these preferences as new research projects emerge. Two recent studies initiatives from the last year that focus on personalized consent management in healthcare data are:

- **Enhancing Data Protection in Dynamic Consent Management Systems**[14], discusses the integration of advanced privacy-preserving technologies, such as differential privacy, blockchain, and zero-knowledge proofs, into dynamic consent management systems. These technologies aim to strengthen the security and privacy of personalized consent processes, making them more robust against potential adversaries;
- **Patient Perspectives and Preferences for Consent in the Digital Health Context**[15] explores how patients perceive and prefer digital health consent models, emphasizing the need for greater transparency and personalization in consent processes. The study highlights that while many patients are willing to provide consent, their preferences are highly context-dependent, which underscores the importance of personalized consent management.

Ongoing Patient Engagement: through mobile apps or web portals, patients can continuously interact with their consent choices. Notifications and updates inform them of new studies or changes in how their data might be used, enabling them to make informed decisions at any point in time. The following recent studies has been found that involves the ongoing patient engagement:

- **Dynamic Specific Consent and Ongoing Engagement**[16]: a study published in Medicine, Health Care and Philosophy discusses Dynamic Specific Consent, emphasizing its role in ongoing patient engagement. The model allows patients to engage actively in research by making decisions on a case-by-case basis about which studies they wish to participate in. This approach fosters continuous communication between researchers and participants, which can reduce mistrust and enhance research literacy among participants;
- **Blockchain-Based Dynamic Consent**[17] explores the use of blockchain technology to enhance dynamic consent systems by providing a decentralized platform for patient consent. This system supports ongoing patient engagement by enabling patients to manage and update their consent preferences in real-time, ensuring transparency and security while facilitating active participation in research over time. See also the next paragraph about Blockchain-based Consent management.

Dynamic consent models enhance patient autonomy, data protection, and research participation through personalized, flexible, and ongoing digital consent management systems but …

---

[14] Enhancing Data Protection in Dynamic Consent Management Systems: Formalizing Privacy and Security Definitions with Differential Privacy, Decentralization, and Zero-Knowledge Proofs, https://www.mdpi.com/1424-8220/23/17/7604

[15] Patient Perspectives and Preferences for Consent in the Digital Health Context: State-of-the-art Literature Review, https://www.jmir.org/2023/1/e42507

[16] Evaluating models of consent in changing health research environments, https://link.springer.com/article/10.1007/s11019-022-10074-3

[17] Blockchain-Based Dynamic Consent for Healthcare and Research, https://link.springer.com/chapter/10.1007/978-3-031-45339-7_3

the main challenge for using dynamic consent models is balancing flexibility with potential consent fatigue and addressing the digital divide among participants.[18]

Blockchain-based Consent management

Blockchain-based consent management is about decentralized systems that securely records and verifies patient consent, ensuring transparency, immutability, and traceability of consent records across multiple stakeholders. Two types of technical solutions can be distinguished / applicable here:

Immutable Consent Records: the use of Blockchain technology ensures that consent records are immutable and transparent. Once a patient provides consent, the record is stored on a blockchain, making it tamper-proof. This ensures that any changes in consent status are fully auditable and traceable, which is critical for compliance and trust in multi-institutional settings like EURACAN. In addition to the study mentioned in the previous paragraph about ongoing patient engagement, the following studies have been found concerning the use of Blockchain technology for consent management:

- **Blockchain-Based Dynamic Consent for Healthcare and Research**[19]: this study emphasizes the use of blockchain to create immutable consent records, ensuring that every consent transaction is securely recorded and cannot be altered. This provides transparency and trust, allowing participants to manage and update their consent preferences in real-time, while also maintaining a clear, unchangeable history of their consent decisions;
- **Smarter Smart Contracts: Efficient Consent Management in Health Data Sharing**[20]: This paper discusses the development of smart contracts on blockchain for managing patient consent in health data sharing. The blockchain-based system ensures that consent records are immutable, providing a secure and efficient way to handle consent that is tamper-proof and easily auditable.

Decentralized Consent Verification: using Block-chain technologies to verify consent without needing to rely on a central authority. This decentralization helps maintain patient autonomy and enhances data security. See also the previously mentioned report about “Blockchain-Based Dynamic Consent” and the following two studies which have the similar conclusions:

---

[18] Identifying facilitators of and barriers to the adoption of dynamic consent in digital health ecosystems: a scoping review, https://www.ncbi.nlm.nih.gov/pmc/articles/PMC10693132/

[19] Blockchain-Based Dynamic Consent for Healthcare and Research, https://link.springer.com/chapter/10.1007/978-3-031-45339-7_3

[20] Smarter Smart Contracts: Efficient Consent Management in Health Data Sharing, https://link.springer.com/chapter/10.1007/978-3-030-60290-1_11

- Blockchain-Based Consent Management for Decentralized Verification in Healthcare[21]
- Hybrid Blockchain Solutions for Decentralized Consent Verification[22]

The main conclusion of all these studies is that they demonstrate the feasibility of using blockchain to decentralize consent verification, enhancing security and transparency in managing healthcare data. Also that decentralized verification via blockchain significantly can improve consent management's security and trustworthiness. But …
the most critical obstacles of using Blockchain technologies are scalability, interoperability and the complexity of aligning blockchain protocols with current healthcare data standards and practices could hinder widespread adoption. Overcoming these technical and operational barriers is essential for blockchain's successful implementation in healthcare.

AI powered consent management

AI-powered consent management uses artificial intelligence to automate and optimize the process of obtaining, monitoring, and ensuring compliance with patient consent, enhancing understanding through natural language processing and ensuring adherence to consent preferences over time.

Natural Language Processing (NLP) for Consent Forms: AI-driven NLP tools can translate complex legal and medical jargon into plain language, making consent forms easier to understand for patients. This enhances patient comprehension and ensures truly informed consent.

AI-Driven Compliance Monitoring: AI systems can automatically monitor compliance with consent preferences, ensuring that data is only used within the agreed parameters. If a study's scope changes, the AI can flag it for re-consent, ensuring ongoing adherence to patient preferences.

AI-powered consent management is rapidly advancing because it enhances patient comprehension by translating complex legal jargon into plain language using NLP, ensuring truly informed consent and improving compliance monitoring. This development positively impacts the ease and accuracy of consent processes, making them more patient-friendly and aligned with individual preferences. Therefore this technology is of interest for the EURACAN registry but more in-depth research beyond a quick-scan is required to fully cover the potential.

Consent as a Service

Consent as a Service (CaaS) is a cloud-based solution that provides scalable, centralized platforms for managing and tracking patient consent across various systems and institutions, simplifying the process for both patients and researchers. This aspect is included for

[21] Dynamic Consent: a potential solution to some of the challenges of modern biomedical research, https://link.springer.com/article/10.1186/s12910-016-0162-9
[22] Blockchain-Based Dynamic Consent for Healthcare and Research, https://link.springer.com/chapter/10.1007/978-3-031-45339-7_3

completeness reasons but no further research has been done to investigate operational cloud-based services and how they can be integrated with the EURACAN registry.

Federated Consent Management

Federated consent management is about allowing patient consent preferences to be recognized and respected across multiple institutions or organizations, enabling seamless data sharing while ensuring that consent is consistently applied according to the patient's wishes. This aspect is also included for completeness reasons but not investigated as part of the quick-scan. Although relevant for a scalable solution and therefore the EURACAN registry, this topic has given a lower priority than the other aspects because this is more an organisational challenge then a technological one.

Interoperable Consent Standards

Interoperable consent standards are standardized frameworks that enable consistent and seamless exchange of consent information across different systems and platforms, ensuring that patient consent is respected and accurately applied in diverse healthcare and research environments.

FHIR-Based Consent:[23] Fast Healthcare Interoperability Resources (FHIR) standards support the interoperability of consent information across different health IT systems. FHIR-based consent modules enable seamless integration and sharing of consent data across different platforms, ensuring that consent is respected and tracked consistently in all systems involved in the research.

The most recent developments regarding FHIR-based consent modules focus on creating interoperable consent standards that enable seamless, electronic exchange of patient consent data across healthcare systems. These modules are designed to enhance the management, storage, and sharing of consent records in a standardized format, ensuring that patient preferences are respected and integrated into health data exchanges. Key features include structured consent forms that can be managed and retrieved electronically, along with enhanced security and privacy controls through integration with OAuth 2.0 for authentication and authorization. These advancements aim to streamline consent management, reduce paperwork, and improve patient data access and compliance across different healthcare settings.

Standardized Consent Taxonomies: developing standardized taxonomies for consent categories (e.g., broad consent, specific consent, withdrawal) enables more precise and interoperable management of consent preferences across various systems and jurisdictions.

One significant initiative is the publication of the CEN Workshop Agreement (CWA)[24] 17933 by CEN/TC 251, which provides a good practice guide for obtaining and documenting consent in digital health innovations. This document aims to standardize how consent is handled across different research and development stages, ensuring ethical and legal compliance while facilitating data reuse for future research. The guide emphasizes the need for clear and transparent consent processes that are adaptable to various research contexts.

[23] FHIR-Based Consent, https://build.fhir.org/consent.html

[24] CEN publishes a good practice guide for obtaining consent, https://www.ehealth-standards.eu/2023/07/21/cen-publishes-a-good-practice-guide-for-obtaining-consent/

Additionally, recent literature reviews and studies emphasize the importance of standardized terminologies and taxonomies in ensuring the interoperability of consent across different healthcare and research platforms, particularly in the context of FHIR (Fast Healthcare Interoperability Resources). These efforts are crucial for creating interoperable systems that can seamlessly manage and exchange consent information, ultimately improving patient trust and data governance in healthcare.

Interoperable consent standards, including FHIR-based modules and standardized taxonomies, are essential for ensuring consistent, secure, and seamless management and exchange of consent data across healthcare systems, improving compliance and patient trust.

Regulatory and Ethical Compliance Technologies

Regulatory and ethical compliance technologies are systems designed to ensure that data handling and consent management processes adhere to legal standards and ethical guidelines, such as GDPR, while protecting patient rights and privacy.

GDPR-Compliant Consent Management Systems is about technologies designed with GDPR compliance in mind ensure that consent management processes meet the stringent requirements for data protection and patient rights. This includes features like the right to be forgotten, data portability, and clear, affirmative consent.

Automated Consent Revocation is about automated systems can facilitate the process of consent revocation, ensuring that once a patient withdraws their consent, all associated data is flagged and excluded from future use in research. This is critical for maintaining trust and legal compliance.

Although these are useful technologies or better phrased “functionalities” are useful to implement but after implementing the core functionality of a (dynamic) consent management first. Therefore these two aspects are also mentioned for completeness reasons as well but no investigated in more detail as part of the quick-scan.

III. Implementation Considerations

For the EURACAN registry and similar initiatives, the adoption of these technologies should be guided by:

- *Scalability*: Ensuring that the chosen consent management technology can handle the large and growing volume of data and patients. For example, the complexity of aligning blockchain protocols with current healthcare data standards and practices could hinder widespread adoption;
- *Regulatory Compliance*: Adhering to local and international regulations, such as GDPR, to ensure that all consent management practices are legally sound. See also section X
- *Security and Privacy*: Prioritizing technologies that offer robust security and privacy protections to maintain patient trust. See also section Y

These technological opportunities in consent management provide the foundation for ethical, compliant, and patient-centric secondary use of medical data in research, crucial for the success of initiatives like the EURACAN registry.

IV. Conclusions and recommendations

Adopting innovative consent management technologies is essential for enhancing patient autonomy, security, and compliance in healthcare research. Dynamic consent models offer flexibility and continuous patient engagement, while blockchain ensures transparency and immutability of consent records. AI further optimizes consent processes through better understanding and automated compliance. However, challenges such as scalability, interoperability, and the digital divide must be addressed to fully realize these benefits. These technologies lay the groundwork for ethical and compliant data use, which is crucial for initiatives like the EURACAN registry. Prioritizing and addressing these challenges will ensure their successful implementation.

Given the insights from the quick-scan on consent management technologies, the following recommendations are made to guide future implementation:

- *Focus on Dynamic Consent Models*, the adoption of dynamic consent models to enhance patient engagement and flexibility, allowing consent to adapt as research needs evolve;
- *Leverage AI for Improved Compliance*: utilizing AI-powered tools can simplify consent processes, improve patient understanding through natural language processing, and ensure ongoing compliance with consent preferences;
- *Consider carefully Blockchain adoption* due to its complexity, scalability issues, and challenges in aligning with current healthcare data standards. Instead, explore simpler, more interoperable solutions.