In-depth Analysis of the Homogenization of Disparate Retrospective Data & the CRC Knowledge Model in ONCOSCREEN

Authors: ServTech

Working Paper

1. Introduction

The purpose of Retrospective CRC screening Data Collection and Analysis in ONCOSCREEN is to collect retrospective hospital data on CRC screening and early detection from the clinical sites in the project. It will, in addition, design and shape the data models and repositories that will be used during testing, diagnosis, and clinical validation.

Health data in ONCOSCREEN consists of many data elements from multiple data sources and these data must be unified using standards and terminology mapping. The key element of the healthcare industry's data-sharing challenges is data integration. Data integration is the process of combining data from various sources into one unified view, safeguarding data security and integrity, and protecting patient confidentiality for efficient data management to derive meaningful insights and gain actionable intelligence. Aim to aggregate data regardless of its type, structure, or volume, unify this data to make it compliant with the standards of codification, and present this data in a meaningful and user-friendly way.

ETL, which stands for Extract, Transform, and Load [1], is a popular data integration process commonly used in data warehousing and business intelligence to consolidate and transform data from multiple sources into a unified and consistent format for efficient querying and analysis, ultimately facilitating informed decision-making and reporting across a federated data environment.

Extract: In the first step, data is extracted from various source systems, which can include databases and other data repositories. These sources may be distributed across different hospital departments, organizations, or geographical locations. Extracting data involves reading and copying the data from the source data systems, which may be in various formats and structures, into a staging area. The staging area acts as an interim storage location for the extracted data, e.g., a warehouse repository..

Transform: The transformation step is where data from the staging area is cleaned, standardized, and transformed to ensure consistency and compatibility across a distributed database. Data transformation may include data cleansing (removing duplicates and correcting errors), data enrichment (adding missing information), data integration (combining data from different sources),

and data validation (ensuring data quality and integrity). Data transformation processes often involve using ETL tools and scripting languages to perform operations like filtering, aggregation, data type conversion, and calculations.

Load: Once the data has been extracted and transformed, it is loaded into a centralized database. Loading can be achived in various ways, such as appending new data, updating existing data, or creating new tables or views in the federated database to store the integrated data. Loading should be done efficiently to ensure that the data is readily available for querying and reporting by end-users.

Unfortunately, the ETL approach can raise significant concerns related to autonomy and privacy when used for the integration of healthcare data due to the following reasons [2]:

- Violation of data autonomy: ETL typically involves aggregating data from various sources into a centralized data warehouse or database. In the context of healthcare data, this aggregation can lead to concerns about individual data autonomy. Patients may be uncomfortable with their health data being collected, stored, and analyzed in a centralized location without their explicit consent.
- Privacy and Security Risks: Healthcare data is highly sensitive and contains personal information about patients' medical history, diagnoses, treatments, and other confidential details. ETL processes involve data extraction, transfer, and loading, which can expose data to potential security risks and breaches if not properly secured.
- Data Linkability: ETL can link healthcare data from different sources to create a comprehensive patient profile. While this may seem beneficial for healthcare professionals, it also poses a risk to patient privacy as it becomes easier to identify individuals and their medical histories, potentially without their knowledge or consent.

To avoid these serious drawbacks, T2.4, whose aim is to collect and homogenize retrospective data from hospitals & open databases on CRC screening and early detection, relies on an improved federated database approach described later in this document. Section 5 describes the federated approach that will be used to homogenize heterogeneous data. Section 6 describes healthcare standards used for data integration. Section 7 the use of advanced meta-data for CRC data source. Finally, Section 8 describes the CRC Knowledge Model.

This document is structured as follows. Section 2 describes typical disease profiles for CRC patients, Section 3 describes how digital CRC data can be used to assist with CRC diagnosis, while Section 4 focuses on patient risk assessment.

2. Typical Disease Profile Data for Colorectal Cancer

A typical disease profile for colorectal cancer includes a comprehensive set of data that provides a detailed description of the disease, its characteristics, and its impact on the patient. This profile is essential for diagnosis, treatment planning, and monitoring. The typical disease profile data for colorectal cancer (CRC) include [3], [4], [5], [6]

Patient Demographics
Age
Gender
Ethnicity
Country of Residence

Personal medical history
Genetic & Hereditary CRC Syndromes
• Family history of hereditary CRC syndromes • Lynch Syndrome (Hereditary Nonpolyposis Colorectal Cancer – HNPCC) • Familial Adenomatous Polyposis (FAP) • Polyposis Syndromes • MUTYH-Associated Polyposis (MAP)
Lifestyle and Behavioral Data
• Diet (fiber intake, consumption of red and processed meats, and overall nutritional status) • Physical activity • Obesity • Smoking • Alcohol consumption • Inflammatory Bowel Diseases • Screening and Prevention
Environmental Data
• Environmental Data: Exposure to Carcinogens (air pollution, hazardous chemicals, radiation), • Geographic Data • Occupational Factors (location and occupation of an individual)
Psychological Well-being
Emotional distress, Anxiety level, Depression Level, Stress Level, Coping Mechanisms, Social support,
Clinical Data
Symptoms: rectal bleeding, changes in bowel habits, abdominal pain, weight loss, iron deficiency anaemia, fatigue.
Duration of symptoms
Genetic Testing
Medical Conditions and Medications
Biopsy and pathology reports: data about tissue samples obtained from the colon or rectum, confirming the presence of colorectal cancer and detailing its histological type and grade.

Tumor staging: Data regarding the extent of the cancer, including the size of the tumor, lymph node involvement, and the presence of metastases.
Surgical reports: • *Operative Procedure*: detailed description of the surgical procedure performed, including the specific type of surgery (e.g., colectomy, proctectomy); • *Operative Findings*: information about the condition of the colorectal area during the surgery, unexpected or unusual findings encountered during the procedure. • *Specimens Removed*: Description of the colorectal specimens removed during the surgery. Documentation of lymph nodes removed for pathological examination. • *Intraoperative Complications* • *Postoperative Instructions*
Laboratory Data: quantitative and qualitative information about various biomarkers and substances related to colorectal cancer
Tumour Markers: Carcinoembryonic Antigen (CEA),
Blood Chemistry and Haematology Tests: Liver Function Tests (LFTs), Complete Blood Count (CBC)
Genetic and Molecular Tests: Microsatellite Instability (MSI) Testing, BRAF Mutation Testing
DNA Analysis: Tissue DNA Analysis
Diagnostic Data: • *Laboratory Tests:* blood tests, genetic tests, and DNA analyses (see above). • *Pathological Tests* • *Faecal Tests*: Faecal Occult Blood Test (FOBT), Stool DNA Test (FIT-DNA) • *Colonoscopy findings*, including the location and size of tumors or polyps. • *Biopsy results* confirming the presence of colorectal cancer and its histological type. • *Endorectal ultrasound findings* for rectal cancer cases. • *Imaging Studies*: such as CT scans, MRI scans, and PET scans.
Staging Data
• Stage of the cancer (e.g., TNM staging) based on the extent of tumor invasion, lymph node involvement, and metastasis. • Information about cancer in nearby structures or organs.
Histological Data
• Details about the tumor type and grade. • Information on the presence of specific markers or proteins (immunohistochemistry) in the cancer cells.

CRC Medications and Treatment Data
• Data related to cancer treatments, such as chemotherapy regimens, radiation therapy plans, targeted therapies, and immunotherapies. • Information about medications, e.g., Targeted Therapy medications: Cetuximab, Panitumumab, Bevacizumab, including dosages and administration schedules.

Collectively, these various types of data provide a comprehensive understanding of the patient's colorectal cancer disease profile and are crucial for early detection and effective management of colorectal cancer. They are critical for personalized, evidence-based care and decision-making in colorectal cancer diagnosis and eventual treatment (not part of ONCOSCREEN).

3. How Digital CRC Data can Help with Diagnosis

CRC data in section 2 play a vital role in the diagnosis of colorectal cancer by providing essential information to confirm the presence of cancer, determine its characteristics, and plan an appropriate treatment strategy. Colorectal cancer diagnosis involves a combination of clinical and diagnostic data to confirm the presence of the disease. Below we describe how typical CRC data are used in the diagnosis of colorectal cancer [7], [8], [9]:

1. **Clinical History and Physical Examination**: A patient's medical history, including risk factors and symptoms, is an essential part of the diagnostic process. Common symptoms may include rectal bleeding, changes in bowel habits, unexplained weight loss, and abdominal pain. A physical examination helps identify any palpable masses or signs of advanced disease.
2. **Precancerous Lesions - Colonoscopy and Polyp Detection**: During a colonoscopy, clinicians can identify and remove precancerous polyps. This not only aids in the prevention of cancer but also provides information about the patient's risk factors.
3. **Blood Tests**: Various blood tests may be conducted as part of the diagnosis and staging process. These can include tests to assess liver function and tumor markers like carcinoembryonic antigen (CEA), which can be elevated in colorectal cancer.
4. **Stool Tests**:
    - **Fecal Occult Blood Test (FOBT)**: FOBT is a non-invasive test that checks for the presence of blood in the stool, which may indicate the presence of colorectal cancer or other gastrointestinal issues.
    - **Stool DNA Test (FIT-DNA)**: This test looks for DNA changes in the stool that are associated with colorectal cancer.
5. **Confirmation of Cancer Presence**:
    - **Biopsy Results**: Tissue samples obtained during a biopsy, typically during a colonoscopy or surgery, are examined by a pathologist. The presence of cancer cells in these samples confirms the diagnosis of colorectal cancer.
6. **Cancer Type and Histology**:
    - **Histological Examination**: Biopsy samples allow pathologists to identify the specific type of colorectal cancer, such as adenocarcinoma, mucinous carcinoma, or other rare variants. This information is essential for treatment planning.

7. **Staging of the Cancer**:
    - **Imaging Studies**: Imaging techniques like CT scans, MRI scans, and PET scans provide information on the size and location of the tumor, involvement of nearby lymph nodes, and the presence of distant metastases. This staging helps determine the extent of the disease and guides treatment decisions.
8. **Identification of Genetic Mutations**:
    - **Genetic and Molecular Testing**: Some colorectal cancers are associated with specific genetic mutations, such as mutations in the BRAF gene or microsatellite instability (MSI). Testing for these genetic markers can influence treatment options and prognosis.
9. **Tumor Marker Levels**:
    - **Tumor Marker Tests**: Tests like carcinoembryonic antigen (CEA) measure the levels of specific substances in the blood. Elevated levels may indicate the presence of colorectal cancer, though they are not diagnostic on their own.

4. CRC Risk Assessment

Assessing colorectal cancer risk involves collecting and analyzing various types of data to determine an individual's likelihood of developing the disease. The data required for assessing colorectal cancer risk can be categorized into the following areas.

- **Genetic and Hereditary Risk Factors:**
    - Genetic testing data: Information on the presence of specific genetic mutations associated with colorectal cancer, such as Lynch syndrome (MSH2, MLH1, MSH6, PMS2) or FAP (APC).
    - Personal medical history: Information about the individual's past health conditions, surgeries, and gastrointestinal issues.
    - Family history of colorectal cancer: Details about whether close relatives (parents, siblings, children) have had colorectal cancer or other related conditions, such as Lynch syndrome or familial adenomatous polyposis (FAP).
- **Demographic Data:**
    - Age: Colorectal cancer risk increases with age.
    - Gender: Colorectal cancer affects both men and women, but the risk may differ between sexes.
    - Ethnicity: Some populations have a higher risk of colorectal cancer.
- **Lifestyle and Behavioral Factors**: Certain lifestyle and behavioral factors can influence colorectal cancer risk. These include:
    - Diet: A diet high in red and processed meats, low in fiber, and lacking in fruits and vegetables may increase risk.
    - Physical activity: A sedentary lifestyle can be a risk factor.

    - Smoking: Smoking is associated with a higher risk of colorectal cancer.
    - Alcohol consumption: Excessive alcohol intake is linked to an increased risk.
    - Lifestyle factors: Information on behaviors like smoking, alcohol consumption, physical activity, and diet.

- **Dietary and Nutritional Data**:
    - Information on dietary habits, such as fiber intake, consumption of red and processed meats, and overall nutritional status. A diet rich in fruits, vegetables, and whole grains may lower the risk of colorectal cancer.

- **Environmental Data**:
    - Exposure to Carcinogens: Environmental data encompass exposure to potential carcinogens, such as air pollution, hazardous chemicals, radiation, and contaminated water sources. Prolonged exposure to these agents can increase cancer risk.
    - Geographic and Occupational Factors: The location and occupation of an individual can also contribute to cancer risk. Some regions may have higher rates of cancer due to specific environmental factors, while certain jobs may involve exposure to carcinogens.

Assessing colorectal cancer risk is typically performed through a combination of these data sources and risk assessment models. The most common approach includes:

- *CRC Risk Assessment*: estimate an individual's risk based on personal and family history, demographics, and lifestyle factors.
- *Genetic Counseling and Testing*: When a strong family history of CRC or hereditary syndromes is present, genetic counseling and testing may be recommended to identify specific genetic mutations.
- *Screening and Surveillance Recommendations*: Based on the assessed risk, healthcare providers recommend appropriate screening and surveillance protocols. Higher-risk individuals may be advised to start screenings at an earlier age or undergo more frequent examinations.
- *Lifestyle and Behavioral Interventions*: For individuals at increased risk due to lifestyle factors, healthcare providers may recommend lifestyle changes such as smoking cessation, dietary modifications, and increased physical activity.

5. Federated Architecture for Integrating Disparate CRC Data Sources

Federated database architecture is a useful approach for integrating healthcare data sources, especially in scenarios where healthcare organizations, research institutions, and various healthcare systems need to collaborate and share data while keeping their data local and under their control [10], [11]. A federated database architecture allows multiple databases, often distributed across different locations or managed by different organizations, to work together seamlessly. This approach enables users to access and query data from multiple sources as if they were part of a single, unified database. Key ways in which federated database architecture can be used for integrating healthcare data sources are [10], [11]:

Data Discovery:

- **Metadata Management**: Federated databases rely on extensive metadata management. Metadata describes the characteristics of data sources, including schema information, data location, data quality, access controls, and transformation rules.
- **Data Catalogs**: Metadata about each data source is often stored in a data catalog. Data catalogs help users discover available data sources, understand their content, and determine which sources are relevant to their needs.

Data Transformation and Integration:

- **Schema Mapping and Transformation**: Federated database architectures often include tools for mapping and transforming data from different sources to a common schema. This ensures that data from diverse sources can be integrated effectively.
- **Data Quality and Consistency**: Data integration tools may perform data cleansing, deduplication, and enrichment to maintain data quality and consistency.

Data Aggregation:

- Healthcare organizations often have multiple data sources, including Electronic Health Records (EHRs), laboratory systems, imaging systems, and more. A federated database architecture allows these organizations to aggregate and query data from these disparate sources in a unified and coherent manner without physically centralizing the data.

Distributed Query Processing:

- **Distributed Queries**: The federated database system optimizes queries to be executed against distributed data sources. It intelligently routes queries to the appropriate sources, minimizing data transfer and processing overhead.

Reduced Data Redundancy:

- By avoiding the need to duplicate data across multiple systems, a federated architecture can help reduce data redundancy and the associated storage and maintenance costs.

Data Sharing and Collaboration:

- In a federated model, healthcare entities can securely share data with each other while maintaining data sovereignty. For example, a hospital can share specific patient records with a research institution without transferring the data, ensuring data privacy and security.

Scalability and Flexibility:

- Healthcare data needs can change over time. A federated architecture can adapt to evolving data requirements and incorporate new data sources or remove outdated ones without major disruptions.

Data Security and Privacy:

- Security and privacy are paramount in healthcare. A federated database architecture can enforce access controls, encryption, and auditing to protect sensitive patient data and ensure compliance with data privacy regulations such as HIPAA (Health Insurance Portability and Accountability Act).

AI Tools & Analytics:

- AI tools can access a wide range of healthcare data across different organizations for epidemiological studies, clinical trials, and outcomes analysis, without compromising data security or breaching data use agreements.

Advanced metadata is a powerful tool for improving integration, privacy, and security in federated databases. It enhances data discovery, mapping, and transformation, enforces privacy controls, and supports security measures such as encryption, access controls, and auditing. Additionally, it aids in data governance and data lifecycle management, ensuring that data is handled in a way that aligns with organizational goals and regulatory requirements.

5.1 Enhanced Federated Approach: Federated Data Mesh

The ONCOSCREEN data-sharing architecture is designed as a federation of domain-oriented data products. Each data product is owned and managed by a specific domain, e.g., demographic, dietary, and nutritional data sources, in Figure-1 and they expose standardized interfaces for other domain-specific data sources to access and use their data. This architecture emphasizes:

- **Decentralized Data Ownership:** The federated data mesh emphasizes a decentralized approach to data ownership and management. Instead of a central authority, data is treated as a product, and individual domain teams are responsible for the data within their domain.
- **Domain-Oriented Data Products:** The federated data mesh encourages the creation of domain-oriented data products, where each domain team is responsible for their data's quality, reliability, and accessibility. This aligns with the concept of treating data as a product [12].

Through the data fusion tool that will be developed within ONCOSCREEN, several other data, including patient measurements, patients' demographics, patients' behavioural data and data gathered from other ONCOSCREEN tools, will be collected. This variety of retrospective data will be fused with data related to genetic data or laboratory/diagnostic data to help estimate the risk of CRC through an interplay of factors, including behavioural, socio-economic, and environmental data, and diagnose patients with colorectal cancer based on many factors, such as the stage and location of the cancer, a patient's age and overall health, and the response to treatment.

For privacy reasons, the prospective data will not be stored in a central database as it is the case with the ETL approach. Rather it will be accessed and processed in the data sources where it belongs.

Task 2.4 follows an enhanced federated data management approach to provide a unified view of data from multiple edge data sources while respecting the autonomy of data sources without compromising data privacy and security. This is shown in Figure 1.

To preserve privacy for medical applications the edge data sources (nodes) in various hospitals, which may include such data, such as Lifestyle & Behavioural data, Demographic, Dietary & Nutritional data will not be directly interconnected to one another – as is the conventional federated approach - but rather to a logically centralised control authority called virtual data lake (ONCOSCREEN Data Lake in WP-4) which comes equipped with appropriate aggregation, transformation services and privacy/security mechanisms, e.g., encryption and role-based access.

The virtual data lake will include enriched schema and meta-data descriptions, associations between meta-data descriptions and references (links) to the edge data sources (nodes) to which the meta-data refer and the CRC Knowledge Model types (in Section 8) that will be used to capture, associate, store and query the meta-data.

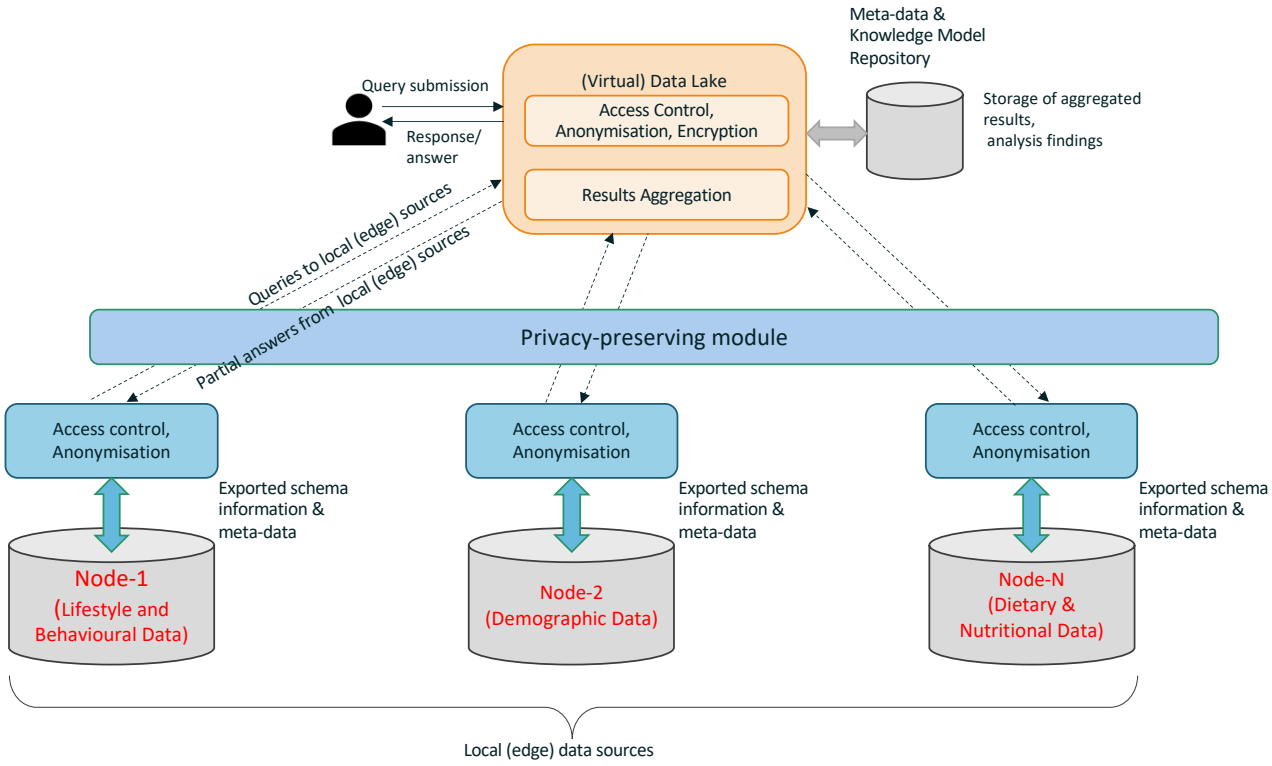

Figure 1: *Federate Data-Mesh Architecture.*

As Figure 1 shows, data will not migrate to the virtual data lake but they rather will be accessed and processed locally at the relevant database nodes, which will retain complete autonomy. Objective is to submit a query (user or tool-initiated request) at the level of the virtual data lake which will decomposed - based on information contained in the meta-data and Knowledge Model in Figure 1 - and will be distributed over the edge nodes in which the data resides. Subqueries will be executed locally with the partial answers aggregated and refined centrally and stored at the virtual data lake. This approach is very similar and conforms to the federated AI execution model, which could be used in the project.

5.2 Comparison of the Data Mesh and Federated Architecture for Retrospective Data Integration

Data Mesh and federated databases are related concepts but they approach the management and organization of data in different ways. Below we explore each concept to understand their similarities and differences.

Data Mesh:

- **Data Mesh** is an architectural paradigm and organizational approach for managing and scaling data across decentralized and autonomous domains within an organization.

Key Principles:

1. **Domain-Oriented Data Ownership:** Data is owned by the domain that understands it best. Each domain or business unit becomes responsible for its own data products.
2. **Data as a Product:** Treat data as a product with defined APIs and service-level agreements (SLAs). This includes designing and managing data products as standalone entities.

3. **Federated Computational Governance:** Instead of relying on a centralized data team, a federated approach to computational governance is adopted. Each domain has its own data teams responsible for the data within their domain.
4. **Self-Serve Data Infrastructure as a Platform (DIaaP):** Provide infrastructure tools and platforms that enable domains to build, deploy, and manage their own data products. This promotes autonomy and reduces dependencies on centralized teams.
5. **Product Thinking and Product Ownership:** Apply product thinking and product ownership to data, with cross-functional teams owning end-to-end responsibility for their data products.

Focus Areas:

- **Decentralization, Autonomy, and Scalability:** Data Mesh aims to address the challenges of scaling data in large organizations by decentralizing data ownership, promoting autonomy, and enabling scalability through domain-oriented data products.

Federated Databases:

- **Federated Databases** refer to a distributed database architecture where multiple databases, often geographically dispersed, are connected and treated as a single, unified database system.

Key Principles:

1. **Data Integration Across Disparate Systems:** Federated databases enable data integration and access across disparate database systems, potentially located in different physical locations or managed by different organizations.
2. **Virtualization of Data:** Data virtualization is a common technique in federated databases, where users can query and interact with data without being aware of its physical location.
3. **Distributed Query Processing:** Queries can be distributed across multiple databases, and the federated database system coordinates the retrieval and integration of results.
4. **Reduced Data Redundancy:** Federated databases aim to reduce data redundancy and provide a unified view of the data without the need for data replication.

Focus Areas:

- **Integration Across Heterogeneous Systems:** Federated databases are designed to provide a unified view of data across different systems, allowing organizations to leverage existing databases without the need for data migration.

Relationship:

- **Overlap in Decentralization:** Both Data Mesh and federated databases involve a degree of decentralization. In Data Mesh, decentralization is achieved through domain-oriented data ownership, while in federated databases, it is achieved through the integration of disparate database systems.
- **Autonomy and Scalability:** Both concepts aim to address issues of autonomy and scalability. Data Mesh achieves this through self-serve data infrastructure and domain-oriented teams, while federated databases achieve it by connecting and coordinating databases in a unified system.

Differences:

- **Scope and Purpose:** Data Mesh is a broader organizational and architectural paradigm that encompasses not only data storage and access but also ownership, governance, and the development of data products. Federated databases specifically focus on the integration of databases for unified data access.
- **Product Thinking:** Data Mesh places a strong emphasis on treating data as a product with dedicated product teams. Federated databases focus more on the technical aspects of integrating and querying data across distributed databases.

In summary, while both Data Mesh and federated databases involve decentralization and autonomy, Data Mesh is a more comprehensive organizational paradigm that addresses broader aspects of data management, including ownership, governance, and treating data as a product. Federated databases are more narrowly focused on providing a unified view of data across disparate database systems.

The approach used in ONCOSCREEN combines the product thinking approach of Data Mesh with the unified view of data across disparate database systems supported by federated databases.

6. Standards for Healthcare Data Integration

To achieve seamless data integration in the healthcare sector, various standards have been established. These standards help ensure that data from different sources and systems can be exchanged, shared, and used effectively. Here are some key standards for healthcare data integration:

HL7 (Health Level Seven): HL7 is a widely used set of standards for the exchange, integration, sharing, and retrieval of electronic health information. It includes different versions and profiles, with HL7 v2 and HL7 FHIR being the most prominent. FHIR (Fast Healthcare Interoperability Resources) is a modern and rapidly evolving standard designed for web-based and mobile applications, making it suitable for modern healthcare IT systems.

SNOMED CT (Systematized Nomenclature of Medicine - Clinical Terms): SNOMED CT is a comprehensive clinical terminology standard that provides a common language for healthcare data. It's used to encode clinical concepts, making it easier to standardize and exchange medical data.

Fast Healthcare Interoperability Resources (FHIR): The FHIR standard is designed to facilitate the exchange of healthcare information in a structured and standardized format. FHIR is designed to be highly flexible and extensible, allowing healthcare organizations and systems to create custom profiles and extensions for specific use cases, including colorectal cancer management. By adhering to FHIR standards, healthcare providers, electronic health record (EHR) systems, and healthcare applications can better exchange and integrate data related to colorectal cancer, leading to improved care coordination and patient outcomes. It can be used to handle data related to colorectal cancer in several ways:

1. **Patient Records**: FHIR allows for the representation of patient data, which includes information about their medical history, diagnoses, and treatments. For colorectal cancer, FHIR resources can store patient demographics, diagnostic information, lab results, and treatment records.
2. **Observations and Laboratory Results**: FHIR provides resources for observations and laboratory results. These can be used to represent data related to cancer screenings, such as colonoscopy results, fecal occult blood tests, and genetic testing results for hereditary colorectal cancer syndromes.

3. **Clinical Documents**: FHIR supports the exchange of clinical documents in standardized formats. This can include colonoscopy reports, pathology reports, and other clinical notes related to the diagnosis and management of colorectal cancer.
4. **Medications and Treatment Plans**: FHIR resources allow for the representation of medications and treatment plans. For colorectal cancer, this can include chemotherapy regimens, surgical procedures, and radiation therapy plans.
5. **Genomic Data**: Colorectal cancer is often associated with specific genetic mutations. FHIR can be used to exchange genomic data, including information about genetic mutations, genetic testing results, and personalized treatment plans based on genomics.
6. **Imaging and Diagnostic Reports**: FHIR can handle imaging studies and diagnostic reports, such as CT scans, MRI scans, and PET scans, which are often used in the diagnosis and staging of colorectal cancer.
7. **Care Coordination**: FHIR can support care coordination by enabling the exchange of care plans, referrals, and care team information for patients with colorectal cancer. This helps ensure that the patient's care is well-coordinated among different healthcare providers.
8. **Quality Measures**: FHIR can be used to capture and exchange data related to quality measures and performance metrics for colorectal cancer screening, diagnosis, and treatment. This is valuable for assessing and improving the quality of care.
9. **Patient Education and Engagement**: FHIR can also be used to provide patients with educational resources and engage them in their care. Patients with colorectal cancer can access educational materials and communicate with their healthcare providers through FHIR-enabled applications.

FHIR APIs (Application Programming Interfaces): In addition to FHIR's data exchange capabilities, FHIR also provides a set of RESTful APIs that enable real-time data access and interaction with healthcare data. These APIs are instrumental in modern healthcare application development.

The above standards play a vital role in achieving interoperability and data integration in healthcare. They enable healthcare systems, EHRs (Electronic Health Records), medical devices, and other healthcare entities to communicate and exchange data in a consistent and standardized manner, improving the quality of care and patient outcomes and their use will be considered in the context of retrospective data homogenization in T2.4.

7. Use of Advanced Meta-Data for CRC Data Source Integration

Advanced metadata plays a crucial role in improving integration, data autonomy, privacy, and security concerns mentioned in section-1. Metadata in healthcare applications refers to descriptive information about the data generated, collected, or stored within healthcare systems. It provides context and structure to the data, making it easier to manage, understand, and integrate data from heterogeneous sources.

Types of Metadata in Healthcare:

- **Descriptive Metadata:** This includes information like data source, creation date, author, patient ID, and data format.
- **Structural Metadata:** Describes the structure and organization of the data, such as database schema, data tables, and relationships.

- **Administrative Metadata:** Provides information about data ownership, access controls, and data management policies.
- **Technical Metadata:** Describes technical details like data encoding, data size, and data processing history.

In the context of federated databases, advanced metadata can address these concerns in the following ways:

1. **Integration**:
    - **Data Discovery**: Advanced metadata helps users discover and understand data sources within the federated database. It provides information about the location, content, and schema of distributed data sources, making it easier to identify relevant data for integration.
    - **Data Understanding:** Metadata helps healthcare professionals and data scientists understand the content and context of the data, which is crucial for accurate analysis and decision-making.
    - **Standardization:** Metadata can include standardized codes and vocabularies, making it easier to map and match data elements from different sources. For instance, using SNOMED CT or LOINC for clinical terminology (see section-6).
    - **Data Quality:** Metadata can be used to assess and maintain data quality by identifying errors, inconsistencies, or missing information.
    - **Data Mapping and Transformation**: Metadata can describe the transformation rules needed to integrate data from disparate sources. This includes mapping data attributes, defining data type conversions, and specifying how to join or relate data from different sources. Advanced metadata streamlines the integration process by providing clear instructions to data integration tools.
    - **Data Lineage**: Metadata tracks the lineage of data, showing where data comes from, how it's transformed, and where it's used. This helps maintain data quality and ensures that integrated data is traceable and auditable.
2. **Privacy**:
    - **Access Controls**: Advanced metadata can specify access controls and permissions for each data element. This ensures that only authorized individuals or systems can access sensitive information, helping to protect patient privacy in healthcare data, for example.
    - **Data Anonymization**: Metadata can include information about data anonymization techniques and policies. It can specify which data elements require anonymization and provide guidelines for protecting individuals' identities while still allowing meaningful analysis.
    - **Data Sensitivity Labels**: Metadata can tag data elements with sensitivity labels, indicating whether the data contains personal, sensitive, or confidential information. This labeling helps enforce privacy policies and ensures that data is handled appropriately.

3. **Security**:
    - **Encryption**: Metadata can describe the encryption mechanisms used to protect data during transfer and storage. This information is critical in ensuring that data remains secure throughout its lifecycle.
    - **Audit and Compliance**: Advanced metadata can store audit information, including logs of who accessed the data, when, and for what purpose. This supports security auditing and compliance with data protection regulations.
    - **Data Masking**: Metadata can specify data masking or obfuscation techniques used to protect sensitive data while still allowing some level of access for certain users or use cases.
    -

8. The ONCOSCREEN Knowledge Model

Advanced metadata can serve as a foundational component in the development of a knowledge model targeting colorectal cancer diagnosis. A knowledge model for cancer diagnosis typically relies on a wide range of data, medical knowledge, and patient information. Advanced metadata plays a crucial role in organizing, categorizing, and enriching this data, which is essential for building a comprehensive and effective knowledge model. This, in turn, contributes to the model's accuracy, reliability, and effectiveness in diagnosing colorectal cancer. Below we describe how advanced metadata in section-7 forms the basis of such a knowledge model.

Data Aggregation:

- Advanced metadata can describe the various data sources and types used in the CRC knowledge model, including medical records, pathology reports, imaging data, genetic information, and clinical studies. This metadata helps organize and consolidate diverse data into a unified system.

Patient and Disease Metadata:

- Metadata can describe patient-specific information, including demographics, medical history, family history, and lifestyle factors. It also specifies metadata related to colorectal cancer, such as tumor staging, histology, and biomarker data.

Terminology and Standards:

- Metadata can reference standardized medical terminologies, such as SNOMED CT or LOINC in section-6, and clinical guidelines like those from NCCN (National Comprehensive Cancer Network). This ensures that the model uses consistent and widely accepted terminology and standards.

Data Privacy and Security:

- Metadata includes information about how sensitive patient data is handled, following privacy regulations like HIPAA (Health Insurance Portability and Accountability Act). Metadata ensures that patient information is protected and compliant with relevant laws.

Explanation and Interpretability:

- Metadata can include documentation explaining the rationale behind data selection, model design, and the decision-making process for diagnosis. This is crucial for transparency and model interpretability.

Figure 2 illustrates how the Digital CRC Knowledge Model that we shall describe in the following subsection can be used to connect various tools and data sources. This figure shows the use of a CRC diagnostic tool that is interconnected to the three data sources in Figure 1, namely Lifestyle & Behavioural data, Demographic, Dietary & Nutritional databases.

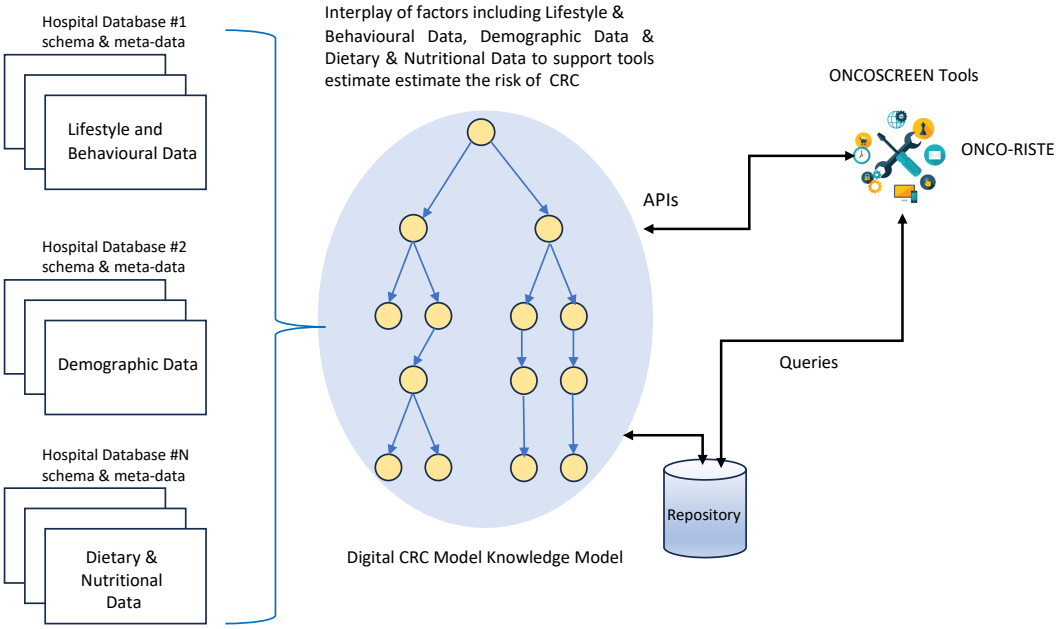

Figure 2: *The Digital CRC Knowledge Model Interplay.*

Figure 3 illustrates the comprehensive CRC Digital models and its associated knowledge parts (called medical blueprint frames or simply blueprints). The frames in Figure-3 correspond to the data types that we described in section-2. Collectively, these various types of data provide a comprehensive understanding of the patient's colorectal cancer disease profile, and is crucial for early detection and effective management of colorectal cancer. This is facilitated by traversing the interconnected blueprint types. This helps healthcare professionals and tools understand the content and context of the data, which is crucial for accurate analysis and decision-making. In this way, medical professionals or tools can for instance confirm presence of cancer, identify the specific type of colorectal cancer, determine the staging of cancer, identify mutations & tumor marker levels, prognosis & follow up.

9. Possible Use of Streaming Platforms for Retrospective Data

Apache Kafka is an open-source distributed streaming platform designed to handle real-time data feeds and provide a scalable, fault-tolerant, and highly available infrastructure for stream processing. Originally developed by LinkedIn, Kafka has become a popular technology in the world of data streaming and event-driven architectures.

A JDBC Kafka Connector is a specific type of Kafka Connector designed to connect Apache Kafka with relational databases using the Java Database Connectivity (JDBC) API. The purpose of this connector is to enable the integration of data between Kafka topics and relational databases, allowing for the ingestion or extraction of data in real-time. A JDBC Kafka Connector can help connect to relational databases, especially when dealing with retrospective medical data:

1. **Data Ingestion (Source Connector):**

When dealing with retrospective medical data, a JDBC source connector can be used to ingest data from a relational database into Kafka topics. This is particularly useful for capturing changes in the database over time or retrieving historical data.

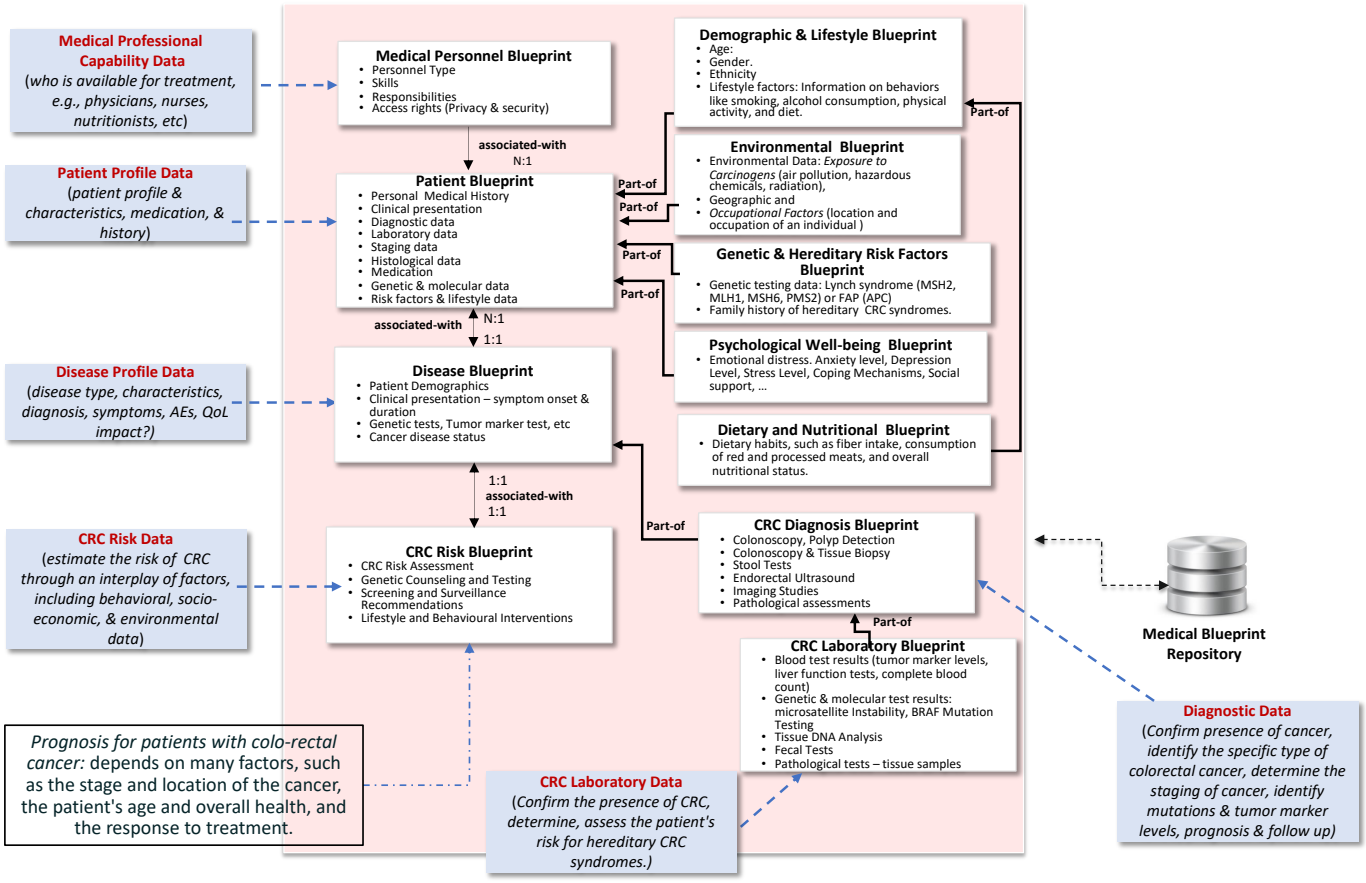

Figure 3: *Comprehensive CRC Digital models and its associated knowledge parts.*

2. **Real-time Streaming:**
    - The JDBC source connector continuously streams data changes (inserts, updates, deletes) from the relational database into Kafka topics. This ensures that any new or updated medical records are made available in real-time within the Kafka ecosystem.
3. **Schema Evolution:**
    - The JDBC Kafka Connector can handle schema evolution, allowing for changes in the structure of the data over time. This is important when dealing with retrospective data as the database schema may have evolved, and the connector can adapt to those changes.
4. **Efficient Data Movement:**
    - The JDBC Kafka Connector is designed for efficient and scalable data movement between Kafka and relational databases. It optimizes the process of capturing changes, serializing data, and publishing it to Kafka topics, as well as deserializing and writing data to databases.
5. **Connectivity and Configuration:**
    - The JDBC connector requires configuration settings such as database connection details, authentication credentials, and SQL queries or table configurations. This information is

specified in the connector configuration to establish connectivity between Kafka and the relational database.

In summary, a JDBC Kafka Connector could play a role in connecting Kafka with relational databases, making it possible to ingest, stream, and process retrospective medical data. By using this connector, you can create a real-time data pipeline that facilitates the integration of medical data stored in relational databases with Kafka, enabling applications and analytics to consume and analyze the data in a scalable and efficient manner.

Kafka is used mainly for streaming and masses of data in motion and it . However, it is not necessary to use Kafka for retrospective data as we will deal mainly with meta-data and queries which do not involve very large data sets. Using Kafka for such situations is not recommended.

Retrospective data in ONCOSCREEN is data at rest, which typically refers to data that is relatively static or unchanging over time. In the context of retrospective medical data, stable data might include information such as patient demographics, historical diagnoses, medical history, and other static patient information as explained earlier in section-2. This type of data doesn't frequently change and remains consistent across various points in time. As an example, patient demographics, such as name, date of birth, gender, and contact information, are stable data. Once recorded, these details are unlikely to change frequently. As Kafka involves handling a large volumes of data in motion and real-time streaming data, updates, or changes that occur during the capture, processing, or integration of data, it is not a suitable choice for retrospective data in ONCOSCREEN. In addition, Kafka introduces complexity due to its distributed nature and may require additional resources for setup, maintenance, and monitoring. As in ONCOSCREEN the data volume is consistently small and unlikely to grow significantly therefore a simpler messaging or data processing system as suggested in Figure 1 is easier to set up and manage.

REFERENCES

1 *Kimball, R., & Ross, M. (2013). The Data Warehouse Toolkit: The Definitive Guide to Dimensional Modeling (3rd ed.). Wiley. Doi: 978-1-118-73228-1.*

2 *Thantilage, R. D., Le-Khac, N. A., & Kechadi, M. T. (2023). Healthcare data security and privacy in Data Warehouse architectures. Informatics in Medicine Unlocked, 39.*

3 *American Cancer Society (2023) Cancer Facts & Figures,* https://www.cancer.org/content/dam/cancer-org/research/cancer-facts-and-statistics/annual-cancer-facts-and-figures/2023/2023-cancer-facts-and-figures.pdf.

4 *Scholefield J.H & Eng C., eds. (2014), Colorectal Cancer Diagnosis and Clinical Management. Wiley-Black-well. doi: 0.1002/9781118337929.*

5 *Beailieu, J.F. (2018) Colorectal Cancer: Methods and Protocols, Humana. doi: 10.1007/978-1-4939-7765-9.*

6 *The Surveillance, Epidemiology, and End Results SEER (2023). Cancer Stat Facts: Colorectal Cancer. https://seer.cancer.gov/statfacts/html/colorect.html.*

7 *Dimitriou N, Arandjelović O, Harrison DJ, Caie PD. (2018) A principled machine learning framework improves accuracy of stage II colorectal cancer prognosis. NPJ Digit Med. 2 (1):52. doi: 10.1038/s41746-018-0057.*

8 *Niazi MKK, Parwani AV, Gurcan MN. (2019) Digital pathology and artificial intelligence. Lancet Oncol. 2019 May;20(5). doi: 10.1016/S1470-2045(19)30154-8.*

9 *Nam S, Chong Y, Jung CK, Kwak TY, Lee JY, Park J, Rho MJ, Go H. Introduction to digital pathology and computer-aided pathology. Journal of Pathology & Translational Medicine 54(2):125-134.*

10 *Hallock H., et. al. (2021) Federated Networks for Distributed Analysis of Health Data. Frontiers in Public Health (9). doi: 10.3389/fpubh.2021.712569.*

11 *D. C. Nguyen, P. N. Pathirana, M. Ding and A. Seneviratne (2020). Blockchain and Edge Computing for Decentralized EMRs Sharing in Federated Healthcare. GLOBECOM 2020 - 2020 IEEE Global Communications Conference, Taipei, Taiwan, 1-6, doi: 10.1109/GLOBECOM42002.2020.9347951.*

12 *Dehghani Z. (2022) Data Mesh: Delivering Data-Driven Value at Scale. O'Reilly.*