Computers in Biology and Medicine 197 (2025) 111051

Contents lists available at ScienceDirect

Computers in Biology and Medicine

journal homepage: www.elsevier.com/locate/compbiomed

From raw data to research-ready: A FHIR-based transformation pipeline in a real-world oncology setting

Antonella Carbonaro [a,*], Luca Giorgetti [a], Lorenzo Ridolfi [b], Roberto Pasolini [b], Andrea Pagliarani [b], Martina Cavallucci [c], Alice Andalò [c], Livia Del Gaudio [d], Paolo De Angelis [c], Roberto Vespignani [e], Nicola Gentili [c]

[a] Department of Computer Science and Engineering - DISI, University of Bologna, Via dell'Università 50, 47521, Cesena, Italy
[b] Aivolution Srl, Dogana, San Marino
[c] Data Unit, IRCCS Istituto Romagnolo per lo Studio dei Tumori (IRST) "Dino Amadori", 47014, Meldola, Italy
[d] Department of Engineering "Enzo Ferrari" - University of Modena and Reggio Emilia, Modena, Italy
[e] Servizio Informatico, IRCCS Istituto Romagnolo per lo Studio dei Tumori (IRST) "Dino Amadori", 47014, Meldola, Italy

ARTICLE INFO

Keywords:
Health informatics
Data conversion
Semantic data management
FHIR
Knowledge graph
Oncology

ABSTRACT

The exponential growth of healthcare data, driven by advancements in medical research and digital health technologies, has underscored the critical need for interoperability and standardization. However, the heterogeneous nature of real-world clinical data poses significant challenges to ensuring seamless data exchange and secondary use for research purposes. These challenges include syntactic inconsistencies (e.g., variable use of terminologies like ICD-10 vs SNOMED CT), semantic mismatches (e.g., differing conceptualizations of disease staging across institutions), and structural fragmentation (e.g., laboratory results encoded in free text rather than structured fields).

Fast Healthcare Interoperability Resources (FHIR) has emerged as a leading standard for structuring and harmonizing healthcare data, enabling integration across diverse systems. This work presents a FHIR-based transformation pipeline that leverages Resource Description Framework (RDF) to convert raw, conceptually heterogeneous oncology data into research-ready, semantically enriched datasets. By representing FHIR resources as RDF graphs, our approach enables semantic interoperability, enhances data linkage across heterogeneous sources, and supports automated reasoning through ontology-based queries and inference mechanisms. The pipeline employs a templated conversion strategy, allowing for the declarative definition of mappings that enable domain experts to focus on the data model.

In Cancer Virtual Lab, we applied this methodology to a real-world oncology dataset comprising 36,335 anonymized patient records, successfully converting 1,093,705 clinical records into 1,151,559 distinct RDF-based FHIR resource types. The process incorporated syntactic and semantic validation, along with expert review, to ensure technical correctness and clinical relevance. Our results demonstrate the feasibility of semantically integrating oncology data using FHIR and RDF, fostering machine-readable, interoperable knowledge representation. This enriched representation supports data quality monitoring and improvement, data harmonization, longitudinal analysis, advanced analytics, and AI-driven decision support, promoting large-scale secondary use.

1. Introduction

The healthcare landscape is characterized by continuous evolution driven by advancements in medical research and technological innovation, resulting in an outstanding growth of clinical data, representing approximately 30% of the world's data volume [1]. This increase is further accelerated by the involvement of major tech companies in the $8.3 trillion healthcare industry. Managing and sharing this information across institutions has become a critical challenge for progress in the sector.

Effective data reuse requires interoperable, standardized, and semantically consistent data management strategies—key to the digital health transformation, providing a basis for efficient communication, prospective data curation, and the secondary use of real-world data

* Corresponding author.
E-mail address: antonella.carbonaro@unibo.it (A. Carbonaro).

https://doi.org/10.1016/j.compbiomed.2025.111051
Received 16 May 2025; Received in revised form 30 July 2025; Accepted 2 September 2025

in medical data science–essential elements for well-informed decision-making [2]. Its economic potential is exemplified by the European Health Data Space (EHDS), which aims to create a unified health data ecosystem. By improving data access in healthcare and facilitating secondary use for research and innovation, the EHDS is projected to save the EU €11 billion over ten years [3].

Although the U.S. mandates Fast Healthcare Interoperability Resources (FHIR) adoption through the 21st Century Cures Act, many other countries still lack clear regulatory frameworks [4]. This has resulted in individual hospitals and software companies adopting custom data formats and interfaces to prioritize short-term gains, thereby hindering interoperability. To overcome this challenge, FHIR, developed by Health Level Seven International (HL7), has emerged as a prominent standard for achieving interoperability in healthcare systems [5]. FHIR conceptualizes healthcare data as modular resources with explicit semantics, offering flexibility and extensibility. This data standard design supports incremental adoption, promoting scalability, cost-effectiveness, and seamless integration into healthcare infrastructures. Moreover, the FHIR specification can be tailored to suit specific contexts of use through its FHIR profiles mechanism. This feature enables the creation of customized resource definitions by applying constraints and extensions to the base resource, promoting standardized implementations that cater to unique requirements effectively. A concrete application of these principles is demonstrated in our previous work, where we developed a FHIR-standardized dataset describing the rehabilitation pathway of trans-femoral amputation patients. This initiative emphasized the value of structured, interoperable data representations for reusability and secondary analysis in clinical research [6].

Building on this foundation, this study presents a modular and semantically validated transformation pipeline based on FHIR and RDF, designed to support real-world oncology research. Our approach addresses the limitations of existing solutions by emphasizing technical correctness, reusability, and semantic alignment with external knowledge sources. The pipeline was developed within the framework of the Cancer Virtual Lab (CVL) project, which aims to support oncology research through the semantic integration of clinical data and scientific evidence.

Our approach employs a templated conversion strategy, enabling the definition of mappings in a standardized and declarative manner, thus empowering users without technical expertise to concentrate on the data model. In this study, we present a transformation pipeline designed to convert large-scale oncology data into semantically enriched FHIR resources. We demonstrate its feasibility by applying it to real-world clinical records from IRST IRCCS (Istituto Scientifico Romagnolo per lo Studio e la Cura dei Tumori) "Dino Amadori" in Meldola, Italy, a recognized cancer research and treatment institute.

The transformation process was evaluated through both syntactic and semantic validation, with the involvement of domain experts to ensure clinical relevance and correctness.

The successful conversion demonstrates the capabilities of the adopted transformation process to enhance semantic interoperability and facilitate the reuse of real-world data. This work highlights the feasibility of semantically transforming real-world oncology data into standardized FHIR/RDF representations through a modular and reusable pipeline. While it demonstrates high semantic fidelity and supports advanced reasoning, its current application remains limited to a single institution and a dataset of deceased patients. These constraints, along with the need for broader validation and performance benchmarking, are discussed in detail alongside the observed benefits of semantic enrichment and data quality enhancement emerging from the transformation process.

The remainder of this manuscript is organized as follows. Section 1 introduces the clinical and technical context of the work, outlining the motivation for adopting FHIR-based RDF representations to ensure semantic interoperability in cancer data integration.

Section 2 contains related works, Section 3 presents the architecture of the transformation pipeline, emphasizing its modular design principles-maintainability, incremental development, reusability, and scalability-supported by Docker-based containerization. Section 4 details the semantic validation strategy that underpins the transformation process, including the use of ontologies, RDF Schema,[1] and OWL[2] constraints to ensure structural and logical consistency of the generated RDF graphs. Section 5 reports the outcomes of the pipeline validation, showcasing quantitative metrics, ontology quality assessments, and experimental insights into data completeness and refinement. Finally, Section 6 discusses the methodological contributions and current limitations. Finally, Section 7 outlines directions for future extensions.

2. Related works

Recent advancements have underscored FHIR's critical role in ensuring transparency and reproducibility in health data pipelines, particularly for AI-driven applications. Namli et al. introduced a scalable data preparation pipeline using a declarative language to extract AI-ready datasets from EHR data while ensuring traceability and compliance with emerging regulations like the AI Act [7]. This approach highlights the importance of transparent data transformation processes in creating reproducible AI solutions, addressing challenges such as harmonizing hierarchical FHIR data into AI-compatible formats and handling missing or irregularly sampled clinical data.

Emerging applications highlight FHIR's adaptability in tackling healthcare interoperability challenges. A scoping review also emphasizes FHIR's expanding role in chronic disease management, including cancer, cardiovascular diseases, and diabetes, within digital health ecosystems [8]. A representative example is the beHEALTHIER platform [9], which leverages a microservices-based architecture to manage and analyze large-scale healthcare data, including social and environmental determinants, to inform public health policy—demonstrating how standardized data pipelines can be effectively applied beyond oncology. Additionally, FHIR supports modular healthcare platforms that integrate diverse data sources to create standardized Digital Twins (DTs) of patients [10]. These DTs provide semantic-rich foundations for advanced healthcare applications, ensuring data integrity while advancing patient-centric care.

Nowadays, most FHIR projects focus on developing FHIR-based data infrastructure and pipelines designed to provide standardized data for easy sharing and clinical analysis. These platforms' key features should prioritize incorporating tools facilitating semi-automated data integration and anonymization [11]. Building on these trends, Kouremenou et al. proposed a standard-driven methodology for data modeling that emphasizes the systematic analysis of core FHIR resources and elements to promote interoperability across diverse domains, including beyond healthcare, and lays the groundwork for developing practical implementation guides enriched with AI capabilities [12]. Turki et al. [6] have recently underscored the value of using the FHIR RDF specification and developing health data as knowledge graphs in RDF[3] format to create structured electronic health records, demonstrating how this approach facilitates the representation of clinical information, the semantic validation of biomedical data, and the application of advanced reasoning techniques through SPARQL and machine learning.

In cancer clinical trials research, FHIR-based data pipelines facilitate the automatic population of case report forms (CRFs). For instance, an Electronic Data Capture framework was developed to model colorectal cancer trials, demonstrating the feasibility of leveraging electronic health record (EHR) data for real-world trials [13]. Another study

[1] https://www.w3.org/TR/rdf-schema/.
[2] https://www.w3.org/OWL/.
[3] https://www.w3.org/RDF/.

Table 1
Overview of the functional scope and methodological differences across reviewed studies, highlighting FHIR support, use of graph-based and semantic technologies, mapping strategies, validation mechanisms, AI integration, and application domains.

Study	FHIR support	Graph-based	Declarative mapping	Semantic validation	AI integration	Application domain
[7]	✓		✓		✓	General EHR/AI
[9,16]	✓				✓	General EHR, Ovarian cancer
[10]	✓	✓	✓	✓		Cancer survivorship
[12]	✓		✓			General EHR
[6]	✓	✓		✓		Knowledge graph
[13,15]	✓		✓			Colorectal cancer
[14]	✓	✓	✓		✓	Prediction of Primary Cancers
[17]	✓					Cancer survivorship
[18,19]	✓		✓	✓		Cancer Interoperability, Phenotypes
CVL	✓	✓	✓	✓	✓	Oncology, real-world data

explored cancer classification and prediction using genetic and phenotypic data extracted from oncology reports and EHRs. A network-based infrastructure, integrating FHIR and the Resource Description Framework (RDF), was employed to enhance predictive modeling through machine learning and deep learning techniques [14]. Similarly, a colorectal cancer study utilized data extraction from CRFs, mapping relevant elements to the FHIR cancer profile [15]. For ovarian cancer, an interactive analytics platform, Shiny FHIR, was implemented, incorporating FHIR resources and R-based statistical tools to enable dynamic data analysis [16]. Another initiative, the Cancer Survivorship Interoperable Data Elements model, mapped survivorship-related data elements to FHIR resources, promoting interoperability and secondary use of patient data. This approach demonstrated the advantages of combining FHIR-based models with machine learning while enabling seamless integration with FHIR-based EHRs [17]. Moreover, a harmonized FHIR-based data model was developed to align with German cancer registry requirements, comparing XML-based representations with the extended model of the German Cancer Consortium. The study illustrated how multiple healthcare domains can be effectively structured using FHIR, enabling seamless integration with other standards through embedded mapping annotations [18]. Additionally, another FHIR-based model was proposed for integrating genetic data from heterogeneous sources, supporting Phenome-Wide Association Studies across institutions. This model enabled genotype-phenotype association identification and validation through literature-based evidence [19].

These studies collectively highlight the potential of FHIR-based frameworks in oncology for structuring, integrating, and analyzing heterogeneous clinical and genetic datasets, fostering interoperability and enhancing predictive analytics. Despite progress in this direction, existing solutions face limitations, particularly in terms of portability, code reuse, and maintenance [20]. Moreover, several approaches reveal challenges arising from manual conversion and validation processes and the use of internal mapping languages that hinder the sharing and publishing of mappings [21].

To provide a clearer understanding of the functional scope and methodological differences between the reviewed studies, Table 1 summarizes the key characteristics of each approach, including support for the FHIR standards, use of semantic technologies, mapping strategies, validation mechanisms, AI integration, and application domain.

3. Transformation pipeline design and architecture

The proposed conversion pipeline adopts a modular architecture to ensure semantic interoperability, scalability, and maintainability when transforming conceptually heterogeneous oncology data into RDF-based FHIR resources. This design supports the seamless transformation of raw, heterogeneous oncology data into RDF-based FHIR resources that are semantically enriched and interoperable. Each module encapsulates a specific functionality within the transformation workflow, ensuring that the system remains adaptable, robust, and extensible in complex clinical data environments. The design principles and architectural choices are aligned with previous work on modular transformation pipelines leveraging standard mapping specifications such as the FHIR Mapping Language, which demonstrated the pipeline's generalizability and effectiveness in real-world clinical scenarios beyond oncology contexts [21], [22]. Modular architecture improves maintainability by isolating components and minimizing interdependencies, thus confining the impact of modifications to individual modules, an essential feature in the context of evolving clinical data models. Its decomposition into autonomous units supports incremental development and promotes effective collaboration among domain experts, developers, and ontologists. Furthermore, the design emphasizes reusability, with core modules, such as the RDF template engine and ontology-based validation tools, easily adaptable to diverse clinical domains and datasets. Finally, clear separation of concerns improves testing and debugging, enabling precise issue identification and facilitating the rigorous validation of both transformation logic and semantic integrity.

To support this modular architecture, the pipeline employs Docker containerization. Docker was chosen because of its ability to encapsulate each module within isolated, reproducible environments, ensuring consistency across different systems and deployment scenarios. This isolation aligns with the modular design by allowing independent development, testing, and deployment of each component without cross-module interference. Moreover, Docker enhances portability and scalability, enabling the pipeline to run seamlessly across local, institutional, and cloud infrastructures, without imposing limitations on storage or computational resources as the CVL project grows.

The following sections, as shown in Figs. 1 and 2 provide a detailed walk-through of each stage in the transformation pipeline, Input, Refinement, Mapping, Validation, Graph Deployment, describing how raw clinical data are ingested, semantically enriched, structurally aligned with FHIR and domain ontologies, validated for consistency and completeness, deployed into a triple store, and finally augmented through rule-based and ontology-driven inferencing.

The use of CSV at the input stage enabled a clean handoff between the institutional data unit, responsible for data extraction and initial cleaning, and the semantic transformation logic developed downstream. This separation facilitated clear responsibilities and simplified debugging. The adoption of FHIR-compliant JSON as an intermediate format served a dual purpose: on one hand, it ensured alignment with the FHIR standard and supported the possibility of integrating the pipeline with FHIR-based infrastructures in the future; on the other, it provided a structured, serializable, and easily validated representation suitable for intermediate inspection and reusability.

The output was transformed into RDF in Turtle (TTL) format, which offers both machine interpretability and human readability, and is natively supported by the majority of RDF triplestore implementations. These formats are not merely serialization choices, but enable independent validation, partial execution, and error isolation. Persisting these intermediate representations also supports resilience and traceability, allowing the pipeline to recover from partial failures and facilitating inspection during development.

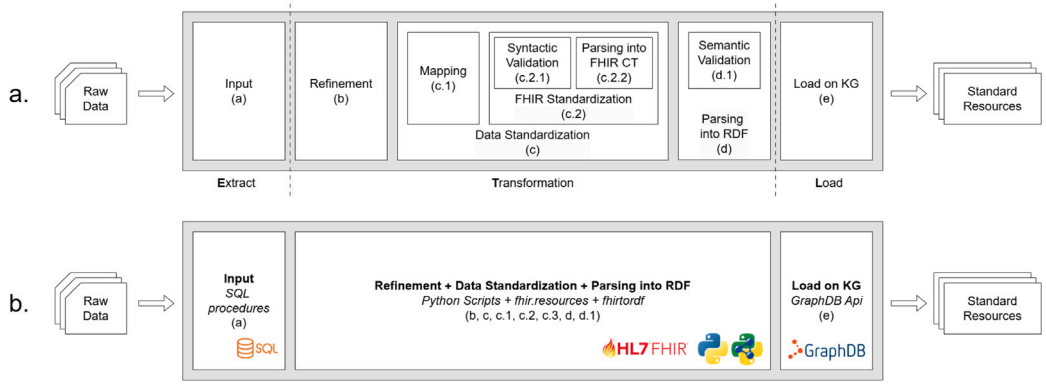

Fig. 1. (a) Modular representation of the pipeline architecture, organized according to the Extract, Transform, and Load (ETL) process steps. (b) Implementation details of the developed transformation pipeline.

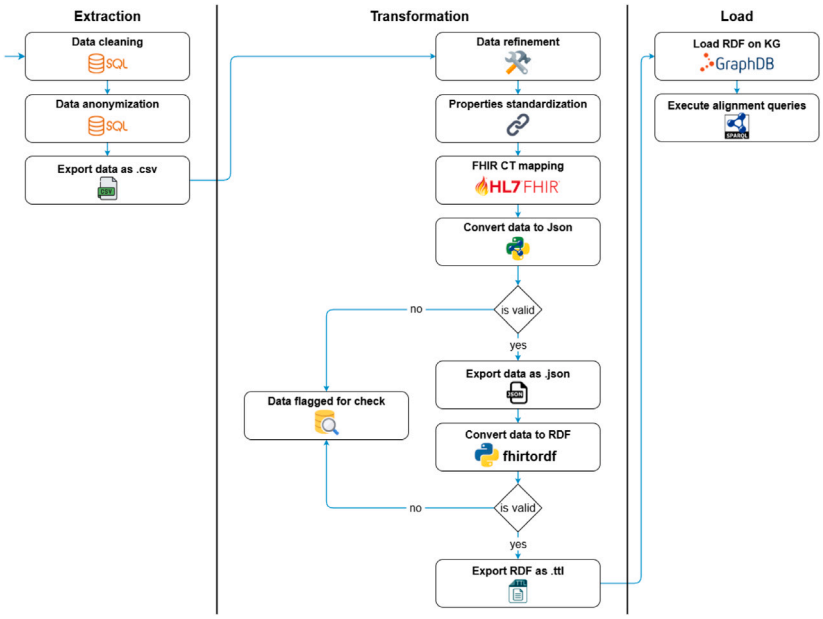

Fig. 2. UML flow diagram representing pipeline steps.

3.1. Input

The data are extracted from the hospital's electronic medical records using an ETL (Extract, Transform, Load) process. This process is implemented through structured SQL queries executed within a business intelligence environment, specifically using Qlik Sense.[4] During the data processing phase, several tasks are carried out to ensure data quality and compliance with ethical standards. The CVL project has been approved by the IRST Scientific Board on 2024-02-13. These include the dataset anonymization to protect privacy, the cleaning of datasets to remove inconsistencies and erroneous entries, and the standardization of variables to ensure uniformity across records and facilitate subsequent analyses.

The dataset, sourced from a single institutional system with standardized workflows and unified coding, is inherently cleaner than those from heterogeneous sources. Nonetheless, we addressed issues like missing values, non-standard encodings, and inconsistent formats through a multi-step preprocessing approach inspired by recent advances in healthcare data quality assurance. While we did not implement the full multi-layered cleaning stack by Mavrogiorgos et al. [23], our curated EMR context allowed for a streamlined process with high semantic fidelity. Unlike more generalized frameworks such as the ontology-driven approach by Kiourtis et al. [24], which targets multi-format inputs, our focus was on aligning structured tabular clinical data with the FHIR model. Still, our semantic preprocessing shares conceptual similarities with ontology-driven containerization, pointing to promising future convergence. Importantly, drawing from a single standardized source reduced risks of selection or measurement bias, and we intentionally avoided imputation or rebalancing techniques to preserve the real-world distribution and ensure reliable downstream analytics.

For the anonymization process, a statistical noise addition technique is applied, ensuring continuity of the noise across related data variables for each entity. This method was selected to balance data utility with privacy protection, minimizing the risk of patient re-identification and reducing the likelihood of individual recognition during data analysis.

The cleaning process involves a detailed inspection of each variable, assessing the types of values originating from the raw data in the electronic medical records. The quality of the information is evaluated, and fields with missing dates or missing essential variables are excluded. The standardization process contributes to harmonizing

[4] https://www.qlik.com/it-it/products/qlik-sense.

categorical values, such as those related to TNM classification[5] and staging. The diverse clinical concepts processed during this phase, such as anamnesis, diagnosis, ESAS scores,[6] adverse events, procedures, and drug administrations, are detailed in the tables presented in Appendix B of the supplementary material.

3.2. Refinement

The refinement phase of the transformation pipeline, as shown in Fig. 1a, is crucial in ensuring that raw data obtained from the input phase is well prepared for subsequent mapping and integration.

While the Input phase focuses on the extraction and initial cleaning of historical data within the institute's ETL environment, the Refinement phase works on the exported CSV files to enforce semantic normalization, prepare complex fields, and assign consistent FHIR-compliant identifiers across all related entities. Unlike the Input phase, which is more tailored to the initial transformation of legacy datasets, the Refinement step is designed to support future updates and real-time data streams by enabling repeatable, structured enrichment and identifier management prior to mapping.

This step involves several sub-steps, including reading input CSV files, verifying data coming from the input phase, cleaning unsupported formats when necessary, preprocessing complex or language-specific fields (e.g., TNM or lifestyle habits), replacing local patient identifiers with consistent FHIR identifiers, and updating bindings between patients and associated resources.

The first step in refinement is the ingestion of CSV files, which serve as the primary format for incoming clinical and research data. These files originate from the input phase, which collects this data from different sources such as hospital records, medical reports and laboratory analysis. Given the heterogeneity of these data sources, a fundamental task is to verify whether the data has been sufficiently cleaned by previous procedures. For instance, when preparing data for compliance with the FHIR standard, specific mandatory fields must be present. In case essential fields, such as diagnosis dates or identifier codes, are missing, the row containing the missing information is excluded from the next steps and flagged for correction. Fields containing unsupported characters, instead, are cleaned and passed to the next step.

Another component of refinement is the preprocessing of complex fields, where raw clinical information is transformed into structured, analysable formats. This step primarily involves two types of records: TNM and Physiological Anamnesis. TNM (Tumor-Node-Metastasis) records (staging data obtained from the previous step) are initially presented in a condensed format (e.g., *"T2N1M1"*), but are split into separate values for the T, N, and M classifiers to facilitate easier querying and improve interoperability with clinical decision support systems. In the case of Physiological Anamnesis, both smoking attitude and alcohol consumption are presented in a string format ("*Sì*", "*No*"), which is language-dependent and specific to Italian. Both fields are converted into a boolean format, less error-prone and well serializable.

An essential aspect of refinement is the generation of a new, anonymized patient identifier and the systematic replacement of the original identifier across all related resources. This process ensures both data privacy and traceability, preserving internal consistency while safeguarding patient privacy in compliance with ethical and legal regulations. By ensuring that all connected resources correctly reference the updated identifier, the pipeline maintains the integrity of patient records without exposing personally identifiable information.

[5] https://www.ncbi.nlm.nih.gov/books/NBK553187/.

[6] https://pmc.ncbi.nlm.nih.gov/articles/PMC5337174/.

3.3. Mapping

The mapping phase is responsible for converting raw tabular data into structured, semantically enriched representations that adhere to the FHIR standard. This process ensures that extracted clinical and research data can be standardized, interoperable, and machine-readable, facilitating integration with external healthcare systems, knowledge bases and ontologies.

The mapping phase consists of two primary transformations:

1. CSV to FHIR JSON: structuring raw data into FHIR-compliant JSON format, ensuring alignment with external ontologies and healthcare standards.
2. FHIR JSON to FHIR RDF: further transforming structured records into the FHIR RDF (Resource Description Framework), enabling advanced semantic queries, linked data integration, and inferencing.

3.3.1. CSV to FHIR JSON: Structuring data

The first step in mapping involves converting flat, tabular CSV data into FHIR-compliant JSON structures. CSV files, while widely used for data exchange, lack relationships and semantic annotations, making them difficult to integrate with complex healthcare models.

To address this, the ETL pipeline employs structured mapping rules that align CSV fields with corresponding FHIR resource attributes, ensuring that each data element is mapped to an appropriate standard structure.

3.3.1.1. Structuring data with FHIR. To ensure data standardization and interoperability, the proposed pipeline, as displayed in Fig. 1a at step c.2, maps every entity to a specific FHIR R4 concept, resource classes that encapsulate various aspects of the patient, including their clinical history, examinations, and other relevant information that supports oncological patient care. The key FHIR concept included in the Cancer Virtual Lab project can be found in Table 2.

In the FHIR standard, certain concepts may correspond to multiple distinct raw clinical entities, introducing ambiguity in semantic interpretation. A representative case of this complexity can be observed in the use of the `Condition` and `Observation` resources.

For the `Condition` resource, differentiation between types of clinical phenomena — such as diagnoses and adverse event effects — is achieved through the `Condition.category` field. Specifically, a category value of "encounter-diagnosis" denotes a formal diagnosis made during an encounter, whereas "problem-list-item" is typically used to capture issues or unintended effects relative to adverse events. This distinction enables semantic disambiguation while using the same resource type.

A similar pattern emerges with the `Observation` resource, which is versatile and capable of representing a wide range of clinical measurements, assessments, and findings. To support disambiguation of the varied semantic roles that `Observation` may play, FHIR provides fields such as `Observation.category` and `Observation.note`. These fields in combination — as shown in Table 3 — are often used to infer the clinical intent or context of the observation,

This flexible schema, while powerful, necessitates careful interpretation of categorical and supporting metadata fields to accurately distinguish between observational contexts.

3.3.1.2. Mapping with external terminologies. To enhance the robustness and standardization of the mapped data, the transformation process incorporates external biomedical terminologies, as described into step c.1 of Fig. 1a. These terminologies offer standardized codes and attributes that enable the consistent interpretation and processing of clinical data across different systems. Table 4 shows the utilized terminologies and the mapped information.

Table 2
Core FHIR resources and their associated clinical content as modeled within the Cancer Virtual Lab project.

FHIR concept	Contained information
`Patient`	Basic information about an individual receiving healthcare or other health-related services.
`Condition`	Diagnostic information and adverse effects of an administered ATC-classified medication that have reached a level of clinical concern.
`Observation`	Measurements and simple assertions made about a patient, such as biomarkers, genetic variants, laboratory test results, physical and physiological assessments, ECOG performance status, ESAS evaluations, and TNM staging classifications.
`CarePlan`	Therapy based on a specified diagnosis.
`Medication administration`	Administration of a specific active principle prescribed as part of a therapy.
`Medication`	Administered active principle.
`AdverseEvent`	An unintended adverse event that may occur following medication administration.
`ChargeItem`	Clinical service delivered by a healthcare organization and accessed by a patient as part of their care pathway.
`Organization`	A healthcare organization or one of its departments.

Table 3
FHIR `Observation` disambiguation through `Observation.category` and `Observation.note` fields.

Observation	Observation.category	Observation.note
Biomarker	Laboratory	Marker
Variant	Laboratory	Variant
Laboratory test	Laboratory	Laboratory
Physiological history	Social-history	
ECOG performance status	Survey	ECOG performance status (SNOMED CT)
ESAS (Edmonton Symptom Assessment System)	Survey	Edmonton Symptom Assessment System (SNOMED CT)
Body measurement	Vital-signs	
Therapeutic response	Therapy	
TNM classification		SNOMED CT Category codes

Table 4
Standard clinical terminologies used in the Cancer Virtual Lab project, including ontology references and the categories of information mapped to each terminology.

Ontology name	URL	Mapped information
Anatomical Therapeutic Chemical Classification (ATC)	https://bioportal.bioontology.org/ontologies/ATC	• Administered medication
Human Genome Organisation (HUGO) Gene Nomenclature	https://bioportal.bioontology.org/ontologies/HGNC	• Variant gene names
International Classification of Diseases, Version 10 (ICD10)	https://bioportal.bioontology.org/ontologies/ICD10	• Cancer sites
Logical Observation Identifier Names and Codes (LOINC)	https://bioportal.bioontology.org/ontologies/LOINC	• Marker • Variant • Laboratory test • Patient gender
Medical Dictionary for Regulatory Activities Terminology (MedDRA)	https://bioportal.bioontology.org/ontologies/MEDDRA	• Adverse event effect grades
National Cancer Institute Thesaurus (NCIT)	https://bioportal.bioontology.org/ontologies/NCIT	• Administered marker value • TNM classification • Therapy lines and settings
SNOMED CT (SNOMEDCT)	https://bioportal.bioontology.org/ontologies/SNOMEDCT	• Diagnosis stage • ECOG values • Clinical service • TNM classification • ESAS symptom assessment findings • Drinker and smoker patient • Body measurement types

Consistently with recent evidence showing that SNOMED CT,[7] LOINC,[8] ICD-9/10,[9] UMLS,[10] and RxNorm[11] are among the most commonly adopted terminologies in FHIR-based studies [25], our transformation pipeline relies on these widely used and standardized sources to ensure maximum interoperability and alignment with current best practices. When integrating FHIR resources with external terminologies, such as SNOMED CT, LOINC, or NCIT,[12] it is often necessary to bind data elements to standardized value sets to ensure semantic interoperability. However, a challenge arises because existing terminologies do not always provide comprehensive coverage for all required data values, particularly in domain-specific values.

For example, Cancer Virtual Lab data includes *Lines of Therapy* beyond the 8th line, specifically the 9th and 10th, whereas the NCIT framework currently only provides mappings for lines of therapy from the 1st through the 8th. Similarly, concepts such as the *Advanced Therapy Setting* are represented in the CVL data but lack corresponding mappings within NCIT. Another case is the classification of stages within *Salvage Therapy*: while *Salvage Therapy* itself is recognized and mapped in NCIT, its individual stages are not yet specified. To address this gap, as we can see in Section 3.6, we introduce an intermediary class that accommodates both bound and unbound values. In this model, the class may define fields that are *optionally* bound to a terminology, allowing values to be either selected from a controlled vocabulary or local codes when no appropriate standard term exists. This hybrid approach enables flexibility in data representation while preserving data standardization where available.

3.3.1.3. JSON file generation. Upon completion of the binding process — both to internal FHIR resource structures and to external standardized terminologies — the primary stage of the mapping process ends in the generation of individual FHIR-compliant JSON files for each clinical entity. To guarantee adherence to the FHIR specification, the pipeline makes use of `fhir:resources` and `fhir` Python libraries. These tools provide programmatic access to the FHIR meta-model, which includes a detailed schema for all resource types. This ensures that the mapping is not only structurally correct but also semantically accurate, embedding clinical information within defined FHIR resource classes.

Although the pipeline is designed to process data in-memory to maximize efficiency and reduce I/O operations, serialization of the JSON output is performed as a backup strategy: persisting intermediate representations to disk enhances data recoverability and reproducibility of the transformation process. When dealing with large datasets, especially when managing frequent observations such as laboratory tests, physiological history or symptom assessment, the generation process incorporates a chunking mechanism. This strategy partitions the data into smaller subsets, which are processed sequentially. Chunking is fundamental in optimizing memory usage, reducing processing overhead, and enhancing the scalability of the ETL pipeline during high-volume data transformations. Each persisted JSON file contains an array of FHIR JSON nodes, like in Listing A.1.

3.3.2. FHIR JSON to FHIR RDF: Enabling semantic reasoning and linked data

Once the data has been structured in FHIR JSON, the next transformation (refer to step d of Fig. 1a) consists of converting it into FHIR RDF. RDF is a graph-based format that enables semantic reasoning, knowledge graph integration, and linked-data queries. This transformation is particularly useful in medical research, where complex relationships can be inferred dynamically.

3.3.2.1. FHIR identifier management. An important aspect of this transformation is managing FHIR resource identifiers. Each FHIR resource (e.g., a patient, observation, or medication record) must have a unique and persistent identifier to maintain referential integrity within the RDF dataset. Depending on the need to preserve links between different entities, the pipeline handles identifiers in two different ways. In the preferred scenario, identifiers are specified during the creation of FHIR resources, a practice that allows the pipeline to preserve continuity and traceability across different stages of data processing. This method ensures that entities remain logically consistent and semantically linked. In cases where a predefined identifier is not specified, the pipeline assigns a generic, randomly generated identifier. While this identifier does not carry intrinsic semantic value, it guarantees that each resource remains distinct and addressable within the RDF representation.

This approach ensures that the resulting semantic graph maintains internal consistency and completeness, even in the face of heterogeneous, sparse, or partially anonymized input datasets.

3.3.2.2. TTL file generation. The outcome of the FHIR JSON to RDF transformation phase is a Turtle file (.ttl) containing the RDF-encoded representation of the dataset. This format serves as an important element for enabling semantic interoperability, facilitating tasks like ontology-based reasoning and integration with external linked data sources. The Turtle syntax[13] is selected due to its compactness, readability, and wide adoption within the Semantic Web community.[14] Turtle provides a human-readable yet syntactically structured format that simplifies both manual inspection and debugging during development and validation phases. Its structure allows for concise representation of RDF triples, especially when dealing with nested nodes, which are common in the FHIR resources model. Moreover, Turtle is natively supported by most RDF processing frameworks and triplestores, facilitating seamless ingestion and being supported by most knowledge graph platforms.

The generation of Turtle (.ttl) files is accomplished through a Python-based toolchain, which ensures accurate translation of FHIR resources into RDF representations. Specifically, the process relies on two key libraries:

- `fhirtordf`, which performs parsing, validation, and transformation of FHIR JSON files into RDF triples. This library guarantees that the resulting semantic structure remains faithful to both the FHIR data model and RDF syntax conventions.
- `rdflib`, a widely adopted Python framework for working with RDF data, is employed to instantiate a virtual RDF graph for each processed JSON resource. It also handles the serialization of these graphs into Turtle files, ensuring compatibility with SPARQL engines and semantic repositories.

Each persisted Turtle file contains a set of triples for each entity, like in Listing A.2.

This combination of tools allows for a scalable, standards-compliant conversion pipeline, which paves the way to the loading step, as described in Section 3.5.

3.4. Validation

The validation mechanism within our RDF conversion pipeline is coupled with the semantic structure of the FHIR specification. As specified in Section 3.3.2, the `fhirtordf` library is responsible for the transformation of FHIR JSON resources into RDF graphs by invoking the `FHIRResource` class, which in turn employs a preloaded FHIR R4 metadata vocabulary (`fhir.ttl`) represented as an RDF Graph. This ontology serves as a formal specification of FHIR classes, properties, and datatypes, and is fundamental to the validation process.

[7] https://www.snomed.org/.
[8] https://loinc.org/.
[9] https://icd.who.int/browse10/2019/en.
[10] https://www.nlm.nih.gov/research/umls/index.html.
[11] https://www.nlm.nih.gov/research/umls/rxnorm/index.html.
[12] https://ncithesaurus.nci.nih.gov/.
[13] https://www.w3.org/TR/turtle/.
[14] https://www.w3.org/2001/sw/wiki/Main_Page.

For each resource instance, a `FHIRMetaVocEntry` is instantiated based on its *resourceType*, providing a typed interface to query and validate the resource's structure. Validation occurs through several mechanisms. First, the presence of the resourceType is checked against the ontology to ensure it corresponds to a recognized FHIR class. Subsequently, each JSON field is mapped to its corresponding RDF predicate using the ontology, with RDFS.domain and RDFS.range constraints applied to validate the proper use of properties. The system then verifies the RDF type of each predicate, ensuring consistency with the expected OWL types and RDF Schema semantics. Finally, primitive data types are validated against the FHIR model to determine the appropriate xsd:datatype, which supports both default RDF literals and FHIR-specific type mappings.

In case of inconsistencies, such as undefined properties, incorrect types, or malformed structures, exceptions are raised during the conversion phase, stopping the process and preventing invalid triples from being serialized. Additionally, the vocabulary-based validation supports inheritance and subclass relationships, allowing the system to resolve predicates and constraints that are indirectly inherited from superclass definitions in the FHIR ontology. This enables accurate modeling of complex FHIR resources with nested components and extensions.

By being based on the OWL-based FHIR ontology, our pipeline ensures semantic integrity and syntactic correctness of the RDF output, laying a robust foundation for interoperability, inferencing, and integration within Linked Data ecosystems.

3.5. Graph deployment

The last step of the CVL transinformation pipeline, as described in Fig. 1a at step e, consists in deploying data in a knowledge graph. All RDF files in Turtle format generated in the previous steps, along with the Turtle file defining the ontology that models the domain of the CVL project, need to be uploaded into a triple store. This process ensures that both the domain-specific schema and instance data are available for querying and reasoning within the same semantic environment. The chosen platform for the Cancer Virtual Lab project is Ontotext GraphDB.[15] GraphDB has been selected due to its robust support for standard semantic web technologies, including RDF, RDFS, OWL, and SPARQL,[16] ensuring full compatibility with the structured data generated throughout the pipeline. It provides a comprehensive REST API, which enables the automation of the bulk upload of RDF data directly from Python scripts, thereby streamlining the integration process into the pipeline. An important feature of GraphDB is its integrated reasoning engine, which allows for the inference of implicit knowledge based on the ontology's axioms, enriching the dataset. Additionally, GraphDB offers several SPARQL plugins that extend its native query functionalities. Among them, the Time Functions Extension (OFN) plugin and the Provenance plugin were particularly valuable. The OFN plugin extends SPARQL by providing advanced time-based query operations, such as `ofn:asDays(?duration)` and `ofn:years-from-duration(?duration)`, where `?duration` represents the difference between two dates. These functions are particularly useful for handling temporal data, which is common in oncology research, enabling complex date-based calculations. Meanwhile, the Provenance plugin facilitates the tracing and verification of reasoning steps, ensuring transparency and the accuracy of the inference process. Thus, GraphDB provides a scalable, standards-compliant, and feature-rich environment that fully supports the semantic interoperability and advanced reasoning needs of the Cancer Virtual Lab project.

[15] https://graphdb.ontotext.com/documentation/11.0/.

[16] https://www.w3.org/TR/rdf-sparql-query/.

3.5.1. Publishing pipeline

After the Turtle files are successfully uploaded, a set of SPARQL INSERT queries is automatically executed to align the FHIR resources with the corresponding concepts defined in the ontology. This alignment ensures that the resulting RDF graph conforms to the domain model and supports semantically enriched queries. In particular, these INSERT operations not only establish class and property assertions but also add human-readable labels (rdfs:label) for each FHIR resource, improving data interpretability and enabling effective visualization within the triplestore interface. Additionally, the SPARQL updates are used to bridge semantic gaps by explicitly linking FHIR resources to concepts from external ontologies that could not be directly encoded within the FHIR structure. This allows for a richer and more expressive representation of the data, extending its semantic coverage beyond the constraints of the original FHIR model. As detailed in Section 3.6, this automated post-processing serves as a practical and necessary workaround for the limitations of standard entailment rules, which are not fully compatible with the nested structure and dotted property paths characteristic of FHIR-based RDF. By materializing the necessary semantic relationships, the system compensates for the lack of native reasoning support and also resolves issues related to data visualization, since FHIR data, due to their complex triple patterns, cannot be directly displayed within the triplestore.

3.6. Reasoning

An initial objective of the CVL project was to exploit entailment rules to automatically infer additional semantic relationships from the RDF data, to enrich the dataset and enable more expressive and ontology-driven queries. However, during the implementation phase, it became clear that r-entailment rules—defined by Horst [26] and natively supported by the chosen triplestore are not compatible with the current data model.

This limitation depends on the structural complexity of FHIR-encoded RDF, where entities are frequently represented using deeply nested constructs and property paths expressed through dotted notation (e.g., `fhir:Observation.component.valueString`, as shown in the alignment query in Listing A.3). While this encoding reflects the FHIR specification, it presents significant challenges for rule engines based on r-entailment, which are designed to operate on simpler triple patterns. As a result, standard entailment mechanisms are unable to parse or match such complex path expressions, making them inadequate for reasoning over RDF graphs structured according to the FHIR model. To address this issue, a complementary solution was adopted in the form of automated SPARQL INSERT queries. These queries are executed immediately after the RDF data and ontologies are loaded into the triplestore and serve to explicitly assert the triples that entailment would otherwise generate. The INSERT operations enrich the existing dataset by adding semantic annotations, such as `rdf:type` declarations or property alignments, that establish a direct connection between the FHIR resources and the corresponding concepts in the domain ontology. By the added `rdf:type` and properties, the system overcomes the limitations of rule-based entailment and ensures that semantic alignment is both accurate and operational within the same RDF graph.

Nevertheless, GraphDB's built-in reasoner was employed in conjunction with a custom ruleset to perform additional reasoning tasks. In particular, this setup enabled the inference of the `DrinkerPatient` and `SmokerPatient` classes based on physiological anamnesis data. Furthermore, the reasoner was effectively used to apply standard OWL-based entailments, including constructs such as `owl:equivalentClass`, `owl:subClassOf`, `owl:subPropertyOf`, `owl:inverseOf`, as well as `rdfs:domain` and `rdfs:range`.

This hybrid reasoning strategy — combining rule-based reasoning with automated SPARQL assertions — allowed the system to maintain reasoning capabilities that not only ensure semantic coherence but also support the enrichment of knowledge available to researchers and systems of ontology-aware analyses.

Table 5
Summary of ontology quality metrics used to assess the CVL ontology across semantic accuracy dimension.

Metric title	Description	Value
Missing annotation labels	Verifies if instances (CVL classes) lack labels (rdfs:label).	1.0
White space in annotation	Checks if labels are empty or consist only of white spaces.	1.0
Datatype consistency	Ensures that the datatype of values follows the correct format for xsd types.	1.0
Functional property violation	Verifies that a subject has at most one object for each functional property.	1.0
Inverse functional property violation	Ensures each object has at most one subject for each inverse functional property.	1.0
Domain constraint compliance	Ensures subjects of a property belong to the declared domain class.	1.0
Range constraint compliance	Ensures objects of a property belong to the declared range class.	1.0

3.7. Software and library choices

The software architecture was designed for portability, standards compliance, and maintainability. Each pipeline module is containerized using Docker, enabling reproducible deployment across local, institutional, and cloud environments. We used open-source components compatible with HL7 FHIR and semantic web standards. Specifically, the `fhir.resources` library was selected for modeling and validating FHIR resources in Python due to its broad version support and built-in validation. For semantic enrichment, `fhirtordf` was used to convert FHIR JSON into RDF, producing an rdflib graph for flexible RDF manipulation and serialization. GraphDB was chosen as the triple-store for its SPARQL 1.1 compliance, reasoning support, and plugins for temporal and provenance-aware queries, key in oncology research. Its adherence to W3C standards ensures long-term interoperability and avoids vendor lock-in.

4. Semantic validation strategy

In this study, we present a set of comprehensive metrics designed to evaluate the quality and consistency of the CVL ontology within the context of semantic accuracy, consistency, conciseness, and defined properties and classes compliance. These metrics are important for ensuring that our ontology adheres to formal and semantic constraints, improving its reliability.

The metrics used in this study are based on KGHeartBeat [27], a community-shared open-source knowledge graph quality assessment tool. KGHeartBeat allows for the quality analysis of a wide range of freely available knowledge graphs, including those registered on the LOD cloud and DataHub. We have implemented their queries, transitioning from pseudocode to SPARQL queries to assess the ontology's quality.

The *Semantic Accuracy* section includes metrics such as the presence of empty or white space-only annotation labels, ensuring that instances are appropriately labeled. Furthermore, we verify the consistency of data types within the ontology, ensuring that the values conform to their expected types as per the ontology's design. We also assess whether functional and inverse functional properties are correctly implemented, maintaining logical integrity by ensuring that subjects and objects follow the constraints imposed by the ontology.

The *Consistency* metrics focus on detecting logical errors, such as entities being assigned to disjoint classes or misplaced classes and properties. These checks guarantee that the ontology does not violate fundamental logical rules. Additionally, we evaluate the ontology for the use of deprecated classes or properties, ensuring that all elements are up-to-date and aligned with the latest ontology standards. Additionally, we examine whether undefined classes or properties are being used, which would indicate a lack of clarity in the ontology structure.

In the *Conciseness* section, we assess both intensional and extensional conciseness. Intensional conciseness addresses the redundancy of properties defined with the same domain and range, while extensional conciseness identifies redundant properties that may produce the same effects through a concatenation of properties. These metrics are helpful to remove duplications from the ontology and improve its overall efficiency. *Defined Classes Compliance* and *Defined Properties Compliance* metrics are employed to ensure that all defined classes and properties in the ontology are being used effectively, with no unused elements lingering in the ontology. By evaluating the ratio of unused defined classes and properties, we can ensure that the ontology is both efficient and fully leveraged.

The results of these metrics provide a comprehensive evaluation of the ontology's quality and are summarized in Table 5, Tables 6 and 7. This assessment serves as a valuable tool for ontology developers, helping them to identify potential issues early and improve the overall structure and usability of the ontology.

The results of the quality assessment, demonstrate the high semantic and structural integrity of the developed ontology. Every metric yields a perfect score of 1, indicating full compliance with the expected semantic, logical, and structural standards. Specifically, the absence of missing or whitespace-only annotation labels confirms the ontology's semantic completeness and attention to detail. Data type consistency and the lack of violations of functional and inverse functional property constraints highlight strong logical soundness. Similarly, complete adherence to domain and range constraints, coupled with the absence of entities assigned to disjoint classes or the use of deprecated or undefined classes and properties, ensures the structural integrity and correctness of the ontology.

From the perspective of conciseness, described in Table 7, the ontology exhibits no intensional redundancy, reflecting an efficient and optimized design. Among the evaluated metrics, *Extensional Conciseness* yielded a score of 0.97, indicating a minor degree of redundancy in property paths with equivalent semantic implications. Upon closer inspection, this result stems from the coexistence of two distinct relationship paths in the knowledge graph: `cvl:Patient` → `cvl:Diagnosis` and `cvl:Patient` → `cvl:Therapy` → `cvl:Diagnosis`.

Although these paths converge semantically by ultimately linking patients to diagnoses, their structural differentiation is intentional and domain-justified. In oncology data modeling, direct associations between a patient and a diagnosis, as well as indirect associations mediated through therapeutic interventions, represent semantically distinct contexts. Maintaining both paths within the ontology preserves critical information about the provenance and clinical significance of each link.

The *Defined Classes Compliance* metric returned a score of 0.96, indicating that 3 out of 86 declared classes (namely `cvl:Variant`, `cvl:LaboratoryTest`, and `cvl:Service`) are currently unused

Table 6
Summary of ontology quality metrics used to assess the CVL ontology across consistency dimension.

Metric title	Description	Value
Entities as members of disjoint classes	Verifies if entities are assigned to disjoint classes, which is logically incorrect.	1.0
Misplaced classes	Checks if classes are used as properties, which is semantically incorrect.	1.0
Misplaced properties	Verifies if properties are used as rdf:type values, which is incorrect.	1.0
Use of members of deprecated classes or properties	Measures the use of deprecated classes or properties.	1.0
Undefined classes	Verifies if there are classes used without being defined (rdf:type without owl:Class).	1.0
Undefined properties	Verifies if there are properties used but not defined as rdf:Property.	1.0

Table 7
Summary of ontology quality metrics used to assess the CVL ontology across conciseness dimension.

Metric title	Description	Value
Intensional conciseness	Identifies duplicate properties with the same domain and range.	1.0
Extensional conciseness	Identifies redundant properties with the same effect, even in compound paths.	0.97
Defined classes compliance	Verifies how many declared classes are never used in any instances.	0.96
Defined properties compliance	Verifies how many defined properties are never used in any triples.	0.62

in any instance data. This design decision is deliberate and reflects a forward-compatible modeling strategy. These classes were explicitly included in the ontology to anticipate future expansions of the dataset and to ensure the structural readiness of the knowledge graph for the upcoming integration of genomic data, laboratory diagnostics, and service-level metadata. The *Defined Properties Compliance* metric yielded a score of 0.62, reflecting that 31 out of 120 declared `owl:DatatypeProperties` are not currently instantiated in any triple. This outcome stems from the intentional inclusion of properties designed to support future data integration efforts. Specifically, these unused properties are associated with the newly defined but not yet populated classes `cvl:Variant`, `cvl:LaboratoryTest`, and `cvl:Service`, which have been introduced into the ontology to accommodate forthcoming data extensions. The presence of these properties ensures schema completeness and semantic readiness, enabling a seamless and semantically consistent evolution of the graph as new clinical and molecular data are incorporated.

Overall, these results attest to the high semantic and technical quality of the ontology, positioning it as a reliable and robust foundation for advanced knowledge graph applications, ontology-driven reasoning, and semantic interoperability in complex data ecosystems.

5. Pipeline validation

The evaluation of the transformation pipeline was conducted by analyzing entity-level correspondence across the three main stages of the process: raw input data, intermediate FHIR representations, and final CVL-aligned RDF entities. As detailed in Table 8, each of the 16 clinical entities, ranging from `Patient` and `Condition` to more granular constructs such as `Observation` and `Procedure`, was tracked through its transformation trajectory. The results demonstrate a perfect one-to-one mapping across all stages: the number of FHIR resources and CVL entities generated precisely matches the initial count of raw entities for each category. This outcome confirms the structural and semantic integrity of the pipeline, indicating that no entity loss, duplication, or semantic drift occurred during the conversion. The fidelity of the transformation ensures that the resulting knowledge graph preserves the granularity and completeness of the source data, establishing a robust foundation for downstream reasoning, integration, and analytical tasks.

Beyond ensuring structural fidelity across the transformation stages, the pipeline also served as a diagnostic tool for assessing data completeness and consistency. During the mapping of raw entities to standardized FHIR and CVL representations, the process revealed previously unregistered clinical instances, particularly in the domain of pharmacological treatments. Specifically, several administered drugs were found to be absent from the initial reference database, either because they represented newly approved medications or because existing pharmaceutical codes had been updated over time. These discrepancies, which became apparent through the normalization and validation steps, enabled a systematic revision and augmentation of the original drug registry. As a result, the curated database expanded from an initial 152 entries to 176 entries, representing a 116% increase and significantly enhancing its coverage. This outcome highlights the bidirectional value of the transformation pipeline: not only does it standardize and encode the source data into semantically rich formats, but it also acts as a quality control mechanism that supports the continuous evolution and refinement of foundational data assets.

In addition to the currently mapped clinical entities, we have proactively implemented transformation logic for three additional classes of entities: *Laboratory Tests*, *Healthcare Services*, and *Genetic Variants*. Although these entities are not yet utilized in the current version of the transformation pipeline, their early integration into the data model reflects a forward-compatible design strategy. By formalizing their structure and embedding them within both the FHIR and CVL semantic layers, we have laid the groundwork for future data ingestion and alignment.

6. Discussion

The starting point of this work was a clinical database, from which tabular data in CSV format were extracted through an ETL process. These CSV files contained clinical records without explicit relationships between entities. The initial request from domain experts was to provide structure and semantic meaning to these records, enabling data integration and reuse in research. This challenge led to the design of a semantically enriched transformation pipeline based on FHIR and RDF, capable of turning flat, disconnected data into a coherent knowledge graph.

Several insights emerged during the development and implementation of the pipeline. First, the ETL phase required substantial manual effort to extract, clean, and restructure data from the EMR source

Table 8
Entity-level correspondence across transformation stages. Each clinical entity was traced from raw input to FHIR and CVL RDF representations.

Raw entity		FHIR resource		CVL entity		Mapping
Entity	#	Concept	#	CVL Class	#	%
Physiological history	3.869	Observation	3.869	cvl:Physiological history	3.869	100,00
Active substance	176	Medication	176	cvl:ActiveSubstance	176	100,00
Diagnosis	19.464	Condition	19.464	cvl:Diagnosis	19.464	100,00
ECOG status	141.134	Observation	141.134	cvl:ECOGStatus	141.134	100,00
ESAS Symptom assessment	225.148	Observation	225.148	cvl:Symptom assessment	225.148	100,00
Adverse event	37.741	AdverseEvent	37.741	cvl:AdverseEvent	37.741	100,00
Adverse event effect	37.741	Condition	37.741	cvl:AdverseEvent effect	37.741	100,00
Biomarker	5.171	Observation	5.171	cvl:Biomarker	5.171	100,00
Patient	36.335	Patient	36.335	cvl:Patient	36.335	100,00
Body measurement	214.590	Observation	214.590	cvl:Body measurement	214.590	100,00
Department	48	Organization	48	cvl:Department	48	100,00
Response	73.892	Observation	73.892	cvl:Response	73.892	100,00
Administration	308.736	Medication administration	308.736	cvl:Administration	308.736	100,00
Therapy	20.220	CarePlan	20.220	cvl:Therapy	20.220	100,00
TNM	7.205	Observation	7.205	cvl:TNM	7.205	100,00
TNM Classification	20.089	Observation	20.089	cvl:TNM classification	20.089	100,00

systems. While the semantic modules were reusable and standardized, the ETL logic had to be tightly coupled with the local data environment, underscoring the importance of contextual adaptation.

Second, the success of the pipeline relied heavily on interdisciplinary collaboration. Clinical experts, data engineers, and ontologists worked closely to define mappings, validate semantic correctness, and ensure the clinical relevance of the transformed data. This collaborative model was essential for navigating the complexity of real-world oncology data and aligning technical implementation with medical understanding.

Despite its strengths, the pipeline was tested and validated on data from a single institution, and only for deceased patients due to institutional privacy constraints. This limits the dataset's diversity and raises generalizability concerns. While the architecture is modular and reusable, key components, particularly the ETL layer, were adapted to the local data environment. Broader adoption will therefore require regulatory adjustments, patient consent mechanisms, and multi-institutional validation to ensure portability and confirm scalability.

7. Conclusions and future directions

The implementation of the FHIR-based transformation pipeline within the Cancer Virtual Lab project has demonstrated the feasibility of converting large-scale, heterogeneous oncology data into standardized, semantically enriched RDF representations. The modular architecture supports scalable integration, semantic coherence, and reuse across clinical and research domains. Our results show successful transformation of over one million clinical records from more than 36,000 patients into interoperable FHIR resources, with integrated syntactic and semantic validation processes and expert review ensuring technical correctness and clinical relevance. The pipeline also enabled detection of incomplete or outdated data, highlighting its role in data harmonization and quality improvement, as well as institutional data governance. Using RDF facilitates ontology-based reasoning, advanced querying, and AI-driven analytics, with GraphDB supporting temporal and provenance-aware queries critical for oncology research. Although currently focused on core clinical entities, the pipeline is designed to support future expansion to additional data classes such as laboratory observations and genetic variants. Technically validated within a single institution, the pipeline's containerized, modular design supports scalability and deployment across multiple sites. The RDF graph enables both SPARQL and natural language queries, enhancing data accessibility, and serves as a basis for knowledge graph embeddings useful in biomedical informatics.

Future work includes multi-center validation and extension to other healthcare domains, leveraging the extensibility of the FHIR standard and alignment with FAIR principles and initiatives like the European Health Data Space. Looking ahead, we plan to develop semi-automated terminology alignment tools to reduce manual mapping effort and accelerate pipeline configuration. Current literature on this topic is still limited, and automation will be key to achieving scalable semantic interoperability. This pipeline addresses diverse stakeholder needs, from clinicians and researchers to data scientists and IT professionals. While this work has been validated within a single institution, the pipeline is already technically equipped to scale across multiple clinical sites. Its modular and containerized architecture, along with the clear separation between local ETL logic and reusable semantic modules, supports deployment in heterogeneous environments. Future work will include empirical validation in multi-center settings to assess portability and performance.

In conclusion, this work offers a replicable strategy to bridge raw clinical data and semantically interoperable knowledge graphs, validated in a real-world oncology context, thus laying the groundwork for secondary data use and cross-institutional research aligned with evolving digital health ecosystems.

CRediT authorship contribution statement

Antonella Carbonaro: Writing – review & editing, Writing – original draft, Supervision, Methodology, Funding acquisition, Formal analysis, Conceptualization. **Luca Giorgetti:** Writing – original draft, Software, Resources. **Lorenzo Ridolfi:** Software. **Roberto Pasolini:** Software. **Andrea Pagliarani:** Supervision, Resources, Investigation, Funding acquisition, Conceptualization. **Martina Cavallucci:** Validation, Software, Resources, Data curation. **Alice Andalò:** Validation, Software, Data curation. **Livia Del Gaudio:** Data curation. **Paolo De Angelis:** Software. **Roberto Vespignani:** Software, Investigation. **Nicola Gentili:** Writing – review & editing, Supervision, Project administration, Funding acquisition, Formal analysis, Conceptualization.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This research was co-funded by the Italian Complementary National Plan PNC-I.1 “Research initiatives for innovative technologies and pathways in the health and welfare sector” D.D. 931 of 06/06/2022, “DARE - DigitAl lifelong pRevEntion” initiative, code PNC0000002, CUP: B53C22006450001. This work was partially supported by IFAB International Foundation Big Data and Artificial Intelligence for Human Development, Call for Projects 2022–2023. Research Board (Medical-Scientific Comittee) Seduta del CMS-BOARD si svolge il 13/02/2024.

We gratefully acknowledge Andrea Zaccheroni and Nicola Caroli from IT service, IRCCS Istituto Romagnolo per lo Studio dei Tumori (IRST) “Dino Amadori”, 47014, Meldola, Italy for the System Administrator and security architecture support.

Appendix. Code examples

```
{
    "resourceType": "Observation",
    "status": "registered",
    "category": [
      {
        "coding": [
          {
            "system": "http://terminology.hl7.org/CodeSystem/
                observation-category",
            "code": "survey",
            "display": "Survey"
          }
        ]
      },
      {
        "coding": [
          {
            "system": "http://snomed.info/sct",
            "code": "423740007",
            "display": "ECOG performance status"
          }
        ]
      }
    ],
    "code": {
      "coding": [
        {
          "system": "http://snomed.info/sct",
          "code": "425389002",
          "display": "ECOG performance status - grade 0"
        }
      ]
    },
    "subject": {
      "reference": "Patient/24754"
    },
    "effectiveDateTime": "2012-06-30"
}
```

Listing A.1: Example structure of a persisted FHIR-compliant JSON file generated by the pipeline. The file contains an array of FHIR resource instances, illustrating the outcome of the mapping process with chunked serialization for scalable processing

```
@prefix fhir: <http://hl7.org/fhir/> .
@prefix sct: <http://snomed.info/id/> .
@prefix xsd: <http://www.w3.org/2001/XMLSchema#> .

<http://hl7.org/fhir/Observation/01c14ba1-4bf8-41f8-a0f3-1
     cea5b8d8a67> a fhir:Observation ;
    fhir:nodeRole fhir:treeRoot ;
    fhir:Observation.category [
        fhir:index "0"^^xsd:integer ;
        fhir:CodeableConcept.coding [
            a <http://terminology.hl7.org/CodeSystem/
                observation-category/survey> ;
            fhir:index "0"^^xsd:integer ;
            fhir:Coding.code [
                fhir:value "survey"
            ] ;
            fhir:Coding.display [
                fhir:value "Survey"
            ] ;
            fhir:Coding.system [
                fhir:value "http://terminology.hl7.org/
                    CodeSystem/observation-category"
            ]
        ]
    ],
        [
        fhir:index "1"^^xsd:integer ;
        fhir:CodeableConcept.coding [
            a sct:423740007 ;
            fhir:index "0"^^xsd:integer ;
            fhir:Coding.code [
                fhir:value "423740007"
            ] ;
            fhir:Coding.display [
                fhir:value "ECOG performance status"
            ] ;
            fhir:Coding.system [
                fhir:value "http://snomed.info/sct"
            ]
        ]
    ] ;
    fhir:Observation.code [
        fhir:CodeableConcept.coding [
            a sct:422894000 ;
            fhir:index "0"^^xsd:integer ;
            fhir:Coding.code [
                fhir:value "425389002"
            ] ;
            fhir:Coding.display [
                fhir:value "ECOG performance status - grade 0"
            ] ;
            fhir:Coding.system [
                fhir:value "http://snomed.info/sct"
            ]
        ]
    ] ;
    fhir:Observation.effectiveDateTime [
        fhir:value "2012-06-30"^^xsd:date
    ] ;
    fhir:Observation.status [
        fhir:value "registered"
    ] ;
    fhir:Observation.subject [
        fhir:link <http://hl7.org/fhir/Patient/6> ;
        fhir:Reference.reference [
            fhir:value "Patient/24754"
        ]
    ] ;
    fhir:Resource.id [
        fhir:value "01c14ba1-4bf8-41f8-a0f3-1cea5b8d8a67"
    ] .
```

Listing A.2: Example of a Turtle (.ttl) file generated from FHIR JSON resources. The file encodes RDF triples corresponding to clinical entities

```
PREFIX fhir: <http://hl7.org/fhir/>
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
PREFIX cvl: <http://irst.emr.it/cvl/>
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>

INSERT {
    ?observation rdf:type cvl:Response ;
        rdfs:label ?label ;
        cvl:response_code ?codeUri ;
        cvl:response_value ?value ;
        cvl:response_date ?effectiveDateTime .
}
WHERE {
    ?observation a fhir:Observation ;
        # Filter only responses
        fhir:Observation.category/fhir:CodeableConcept.coding/
            fhir:Coding.code/fhir:value "therapy" ;
        fhir:Observation.effectiveDateTime/fhir:value ?
            effectiveDateTime ;
        fhir:Observation.component/fhir:Observation.component.
            code/fhir:CodeableConcept.coding/fhir:Coding.
            system/fhir:value ?codeSystem ;
        fhir:Observation.component/fhir:Observation.component.
            code/fhir:CodeableConcept.coding/fhir:Coding.code/
            fhir:value ?codeValue ;
        fhir:Observation.component/fhir:Observation.component.
            code/fhir:CodeableConcept.coding/fhir:Coding.
            display/fhir:value ?codeDisplay ;
        fhir:Observation.component/fhir:Observation.component.
            valueString/fhir:value ?value ;
        fhir:Resource.id/fhir:value ?resourceId .

    # Creation of the label
    BIND(CONCAT("Response: ", ?codeDisplay, " - Date: ", STR(?
        effectiveDateTime)) AS ?label)
    BIND(IRI(CONCAT(STR(?codeSystem), ?codeValue)) AS ?codeUri)
}
```

Listing A.3: Example of an ontology alignment query using dotted FHIR RDF property paths.

Appendix B. Structured representation of clinical concepts and source data fields

See Tables B.9–B.15.

Table B.9
List of fields per IRST entity. The number of rows indicates the dataset size for each entity.

IRST entity (*rows*)	Fields	
Physiological history (*3.869*)	• Patient ID • Alcohol consumption (Y/N) • Smoke (Y/N)	
Active substance (*176*)	• ATC code • ATC description	
Diagnosis (*19.464*)	• Diagnosis ID • Patient ID • Site description • Site ICD10 code • Diagnosis date	• Clinical stage • Pathological stage • Clinical TNM value • Pathological TNM value
ECOG status (*141.134*)	• Patient ID • ECOG value • ECOG date	
Symptom assessment (*225.148*)	• Patient ID • Fatigue • Depression • Pain • Nausea • Anxiety	• Drowsiness • Loss of appetite • Malaise • Breathing difficulty • ESAS assessment date • ESAS assessment time
Adverse event (*37.741*)	• Administration ID • Toxicity description • Toxicity grade description • Toxicity grade description • Administration date	
Laboratory test (*0*)	• Patient ID • Laboratory exam code • Laboratory exam description • Laboratory result value	• Reference range minimum • Reference range maximum • Unit of measurement • Laboratory exam date
Biomarker (*5.171*)	• Patient ID • Marker description • Operation date • Marker value	
Patient (*36.335*)	• Patient ID • Gender • Birth date • Death date	
Body measurement (*214.590*)	• Patient ID • Body measurement date • Weight • Height	
Service (*0*)	• Patient ID • Department code • Service code • Price • Quantity	• Category 1 code • Category 2 code • Category 3 code • Service date
Department (*48*)	• Department code • Department name	
Therapy response (*73.892*)	• Patient ID • Response value • Response date	
Administration (*308.736*)	• Therapy ID • Administration ID • ATC code • Administration date	
Therapy (*20.220*)	• Therapy ID • Patient ID • Line of therapy • Therapy setting • Therapy site ICD10 code	• First administration date • Last administration date
Variant (*0*)	• Patient ID • Gene • Variant • Variant 2	• Amino acid change • Varianti classification • Genetic exam date

Table B.10
Diagnosis stage values and associated number of instances.

Diagnosis stage	Number of instances
NULL	16,720
0	14
I	330
IA	33
IB	63
IC	3
II	134
IIA	335
IIB	229
IIC	9
III	129
IIIA	169
IIIB	188
IIIC	222
IV	873
IVA	7
IVB	3
IVC	3

Table B.11
Therapy setting values and associated number of instances.

Therapy setting	Number of instances
NULL	155
Adjuvant therapy	1284
Advanced therapy	16,465
Concomitant with radiotherapy	22
Consolidation therapy	12
Induction therapy	775
Maintenance therapy	100
Neoadjuvant therapy	675
Salvage therapy 1	387
Salvage therapy 2	172
Salvage therapy 3	104
Salvage therapy 4	39
Salvage therapy 5	16
Salvage therapy 6	6
Salvage therapy 7	5
Salvage therapy 8	2
Salvage therapy 9	0
Salvage therapy 10	1

Table B.12
Line of therapy values and associated number of instances.

Line of therapy	Number of instances
NULL	3890
Line of therapy 1	7541
Line of therapy 2	4191
Line of therapy 3	2289
Line of therapy 4	1130
Line of therapy 5	586
Line of therapy 6	287
Line of therapy 7	148
Line of therapy 8	85
Line of therapy 9	50
Line of therapy 10	26

Table B.13
Tumor size (T) values and associated clinical and pathological instances.

T value	Number of clinical instances	Number of pathological instances
NULL	35,683	35,537
T0	5	11
T1	86	158
T1a	5	65
T1b	14	149
T1c	51	249
T2	567	621

(*continued on next page*)

Table B.13 (*continued*).

T value	Number of clinical instances	Number of pathological instances
T2a	20	55
T2b	19	46
T2c	9	9
T3	938	858
T3a	15	129
T3b	30	112
T3c	11	40
T4	785	413
T4a	38	95
T4b	93	86
T4c	6	2
T4d	21	3
Tis	4	18
TX	528	272

Table B.14
Spread to regional lymph nodes (N) values and associated clinical and pathological instances.

N value	Number of clinical instances	Number of pathological instances
NULL	35,704	35,629
N0	450	1188
N1	683	737
N1a	10	121
N1b	15	32
N1c	0	5
N2	911	408
N2a	13	62
N2b	12	29
N2c	4	4
N3	386	172
N3a	1	58
N3b	4	10
N3c	13	5
NX	222	468

Table B.15
Presence of distant metastasis (M) values and associated clinical and pathological instances.

M Value	Number of clinical Instances	Number of pathological instances
NULL	35,327	35,599
M0	1047	2089
M1	2273	310
M1a	9	24
M1b	19	17
M1c	31	27
MX	222	310

References

[1] J. Thomason, Data, digital worlds, and the avatarization of health care, Glob. Heal. J. 8 (1) (2024) 1–3.

[2] S.T. Liaw, Digital public health: quality, interoperability and capability maturity, in: Oxford Research Encyclopedia of Global Public Health, 2024.

[3] Q & A on the European Health Data Space, URL: https://ec.europa.eu/commission/presscorner/detail/en/qanda_24_2251.

[4] T. Benson, G. Grieve, Why interoperability is hard, in: T. Benson, G. Grieve (Eds.), *Principles of Health Interoperability: FHIR, HL7 and SNOMED CT*, Springer, Cham, 2021, pp. 21–40.

[5] V. Quintero, C. Chevel, P. Sanmartin-Mendoza, Analysis on the interoperability of health information systems, in: 2024 IEEE Technology and Engineering Management Society, pp. 1–6.

[6] H. Turki, et al., FHIR RDF—Why the world needs structured electronic health records, J. Biomed. Inf. 136 (2022) 104253.

[7] T. Namli, et al., A scalable and transparent data pipeline for AI-enabled health data ecosystems, Front. Med. 11 (2024) 1393123.

[8] R. Gazzarata, et al., HL7 fast healthcare interoperability resources (HL7 FHIR) in digital healthcare ecosystems for chronic disease management: Scoping review, Int. J. Med. Inf. 189 (2024) 105507.

[9] A. Mavrogiorgou, et al., beHEALTHIER: A microservices platform for analyzing and exploiting healthcare data, in: 2021 IEEE 34th International Symposium on Computer-Based Medical Systems, CBMS, IEEE, pp. 283–288.

[10] A. Marfoglia, et al., CONNECTED: A knowledge graph-driven platform for clinical data harmonization, in: PerCom Workshops 2025, pp. 1–6.

[11] J. Gehrmann, et al., What prevents us from reusing medical real-world data in research, Sci. Data 10 (1) (2023) 459.

[12] E. Kouremenou, A. Kiourtis, D. Kyriazis, A data modeling process for achieving interoperability, in: International Conference on E-Health and Bioengineering, Springer Nature Switzerland, Cham, 2023, pp. 711–719.

[13] N. Zong, et al., Developing an FHIR-based computational pipeline for automatic population of case report forms for colorectal cancer clinical trials, JCO Clin Cancer Inf. 4 (2020) 201–209.

[14] N. Zong, et al., Leveraging genetic reports and electronic health records for the prediction of primary cancers, JMIR Med. Inf. 9 (5) (2021) e23586.

[15] N. Zong, et al., Modeling cancer clinical trials using HL7 FHIR: a case study with colorectal cancer data, Int. J. Med. Inf. 145 (2021) 104308.

[16] N. Hong, et al., Shiny FHIR: an integrated framework leveraging shiny R and HL7 FHIR, Stud. Health Technol. Inform. 245 (2017) 868–872.

[17] L. González-Castro, et al., CASIDE: a data model for interoperable cancer survivorship information based on FHIR, J. Biomed. Inf. 124 (2021) 103953.

[18] M. Lambarki, et al., Oncology on FHIR: a data model for distributed cancer research, Stud. Health Technol. Inform. 278 (2021) 203–210.

[19] N. Zong, et al., Developing a FHIR-based framework for phenome wide association studies, AMIA Jt Summits Transl. Sci. Proc. (2020) 750–759.

[20] P. Tabari, et al., State-of-the-art FHIR-based data model and structure implementations, JMIR Med. Inf. 12 (1) (2024) e58445.

[21] A. Marfoglia, et al., Towards real-world clinical data standardization: A modular FHIR-driven transformation pipeline, Comput. Biol. Med. 187 (2025) 109745.

[22] V.A. Arcobelli, et al., FHIR-standardized data collection on the clinical rehabilitation pathway of trans-femoral amputation patients, Sci. Data 11 (1) (2024) 806.

[23] K. Mavrogiorgos, A. Kiourtis, G. Manias, D. Kyriazis, A multi-layer approach for data cleaning in the healthcare domain, in: Proc. of the 2022 8th International Conference on Computing and Data Engineering, 2022, pp. 1–6.

[24] A. Kiourtis, A. Mavrogiorgou, G. Manias, D. Kyriazis, Ontology-driven data cleaning towards lossless data compression, in: Challenges of Trustable AI and Added-Value on Health, IOS Press, 2022, pp. 421–422.

[25] F. Amar, A. April, A. Abran, Electronic health record and semantic issues using FHIR: systematic mapping review, J. Med. Internet. Res. 26 (2024).

[26] H.J.T. Horst, Combining RDF and part of OWL with rules: Semantics, decidability, complexity, in: The Semantic Web – ISWC 2005, pp. 668–684.

[27] KGHeartBeat, URL: https://isislab-unisa.github.io/KGHeartBeat/.