



Towards real-world clinical data standardization: A modular FHIR-driven transformation pipeline to enhance semantic interoperability in healthcare

Alberto Marfaglia^a, Filippo Nardini^c, Valerio Antonio Arcobelli^b, Serena Moscato^b,
Sabato Mellone^b, Antonella Carbonaro^a*

^a Dept. of Computer Science and Engineering - DISI, University of Bologna, Via dell'Università 50 47521, Cesena, Italy

^b Dept. of Electrical, Electronic and Information Engineering "Guglielmo Marconi" - DEI, University of Bologna, Viale del Risorgimento 2 40136, Bologna, Italy

^c Dept. of Industrial Engineering - DIN, University of Bologna, Viale del Risorgimento 2 40136, Bologna, Italy

ARTICLE INFO

Keywords:

Health informatics
Standardization
Data conversion
Semantic technology
Software architecture

ABSTRACT

Background: Given the exponential increase in clinical data, which accounts for around 30% of global data volume, effective information management has become crucial to ensuring robust interoperability. This trend is further expedited by implementing consumer-oriented Internet of Things platforms, contributing to the growth of the \$8.3 trillion healthcare industry. These advancements, combined with challenges such as heterogeneous data formats and a lack of incentives, necessitate the development of pragmatic infrastructures and tools that harness contemporary clinical standards like Fast Healthcare Interoperability Resources (FHIR).

Objective: This study aims to present a modular conversion pipeline employing a templating strategy for translating clinical data into the FHIR model. Emphasis is placed on utilizing a standard mapping specification like FHIR Mapping Language. This ensures essential properties such as platform independence, portability, and code reusability.

Methods: The pipeline was developed incrementally, dividing its core functionalities into five modules: Input, Refinement, Mapping, Validation, and Export. These were subsequently validated by converting a dataset from a prosthetic fitting and rehabilitation center to demonstrate the approach's validity in a real-world data context.

Results: A total of 1962 hospital stay records of 1006 unique patients were converted successfully to 15 distinct types of FHIR resources. The successful conversion states the pipeline's effectiveness, additionally showcasing its capabilities for enhancing semantic interoperability and facilitating the reuse of real-world data.

Conclusion: Our approach emerges as a modular data conversion framework that addresses the limitations of existing solutions, making significant contributions to the creation of standardized, interoperable, and high-quality clinical datasets that serve as a foundation for further work.

1. Introduction

The healthcare landscape is marked by continuous evolution driven by advancements in medical research and technological innovation, resulting in an outstanding growth of clinical data — approximately 30% of the world's data volume [1]. This growth is further accelerated by the involvement of Big Tech companies in the USD 8.3 trillion healthcare industry. These companies have introduced consumer-focused Internet of Things (IoT) platforms and services, including Amazon Care, Apple's wearables, and Alphabet's artificial intelligence (AI)-driven precision medicine initiatives [1]. Additionally, emerging technologies such as the IoT and 5G are transforming healthcare, with IoT expected to contribute between USD 0.5–1.8 trillion (14%) of its total economic

value by 2030 [2]. The global wearable healthcare market, valued at USD 14.4 billion in 2020, is projected to grow at an annual rate of roughly 18.9% [1]. A significant portion of this growth is driven by home healthcare, which dominated the wearable medical device market in 2021, accounting for 54.5% of the total revenue [1] (see Table 1).

Managing and sharing this information effectively across diverse organizations has emerged as a pivotal challenge on which the progress of the healthcare industry depends [3–5]. To address this issue, it is essential to develop healthcare systems that are both semantically and structurally cohesive, enabling the seamless exchange of meaningful data. This necessitates the adoption of formal data descriptions, standardized vocabularies, and machine-readable formats [4]. Essentially,

* Corresponding author.

E-mail addresses: alberto.marfaglia2@unibo.it (A. Marfaglia), filippo.nardini3@unibo.it (F. Nardini), valerio.arcobelli2@unibo.it (V.A. Arcobelli), serena.moscato3@unibo.it (S. Moscato), sabato.mellone@unibo.it (S. Mellone), antonella.carbonaro@unibo.it (A. Carbonaro).

<https://doi.org/10.1016/j.combiomed.2025.109745>

Received 2 September 2024; Received in revised form 9 January 2025; Accepted 22 January 2025

0010-4825/© 2025 The Authors. Published by Elsevier Ltd. This is an open access article under the CC BY license (<http://creativecommons.org/licenses/by/4.0/>).

Table 1
Statement of significance.

Problem or Issue	Heterogeneous and proprietary clinical data formats hinder interoperability, complicating data reuse and undermining data semantics.
What is Already Known	Existing solutions provide conversion pipelines but lack comprehensive, portable, and reusable mapping strategies
What this Paper Adds	Presents a modular conversion pipeline utilizing templating and the FHIR Mapping Language, enhancing interoperability by providing a standardized approach for translating clinical data into the FHIR model. Addresses limitations of existing solutions, facilitating semantic preservation, platform independence, and code reusability.

interoperability emerges as the linchpin of the digital health revolution [3,6], providing a basis for efficient communication, prospective data curation, and the secondary use of real-world data in medical data science — essential elements for well-informed decision-making. 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 [7].

Despite the mandate in certain regions like the U.S. under the 21st Century Cures Act [8] requiring the use of the Fast Healthcare Interoperability Resources (FHIR) for health data exchange, many countries lack well-defined regulatory directives and incentives. This has resulted in individual hospitals and software companies adopting custom data formats and interfaces to prioritize short-term gains, thereby hindering interoperability [6]. To overcome this challenge, standardization is pivotal in achieving data harmonization and improving healthcare information availability, safety, and accuracy [3,9], as well as integrating heterogeneous data sources. In this context, FHIR, developed by Health Level Seven International (HL7), has emerged as a prominent standard for achieving interoperability in healthcare systems. 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 existing healthcare infrastructures [6]. Moreover, the FHIR specification can be tailored to suit specific contexts of use through its FHIR profiles mechanism.¹ This feature enables the creation of customized resource definitions by applying constraints and extensions to the base resource. As a result, it promotes standardized implementations that cater to unique requirements effectively. The recent surge in FHIR adoption can be attributed to its robust framework for exchanging electronic healthcare information and to regulatory measures implemented, particularly in countries like the U.S. [10].

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. [11] 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. 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. For example, platforms integrating Internet of Medical Things (IoMT) devices with 5G network slicing use FHIR to manage data heterogeneity and enable real-time data exchange [12]. A scoping review also emphasizes FHIR’s expanding role in chronic disease management, including cancer, cardiovascular diseases, and diabetes, within digital health ecosystems [13]. Additionally, FHIR supports modular healthcare platforms that integrate diverse

data sources to create standardized Digital Twins (DTs) of patients. These DTs provide semantic-rich foundations for advanced healthcare applications, ensuring data integrity while advancing patient-centric care [14].

Nowadays, the focus of most FHIR projects is on developing FHIR-based data infrastructure and pipelines designed to provide standardized data for easy sharing and clinical analysis [15]. The key features of these platforms should prioritize incorporating tools facilitating semi-automated data integration and anonymization [16]. Despite progress in this direction [17–20], existing solutions face limitations, particularly in terms of portability, code reuse, and maintenance.

Bennet et al. [21] converted the MIMIC-IV database into the FHIR data model to create a real-world FHIR store, an invaluable asset for research. However, the absence of a data sanitization phase and the use of internal mapping logic make the workflow challenging to share. Montazeri et al. [9] developed a minimum Data Set customized for Computerized Physician Order Entry (CPOE) in the cardiovascular field, integrated with the Electronic Health Record (EHR) system through the classification and mapping of data elements into FHIR. A notable constraint is the absence of digital tools for the mapping process, necessitating individual and manual verification of each resource. Sinaci et al. [22] defined and implemented a methodology to enable the FAIR-ification of real-world health data, allowing domain experts to transform existing health data into FHIR via user-friendly GUIs. However, the dependence on an internal mapping language and adopting FHIR profiles, which enforce constraints on resource data fields, impede widespread adoption. Furthermore, Williams et al. [17] developed an FHIR-based data harmonization pipeline to facilitate an AI-friendly representation of clinical data. Despite its effectiveness, the mapping process implemented in Python and the explicit representation of source database attributes make the mappings strictly dependent on low-level programming details, impeding code reuse. Ladas et al. [23] converted the German Corona Consensus (GECCO) dataset, which describes COVID-19-related patient data, into FHIR to enhance accessibility. Their approach relied on a web-based tool named openEHR-to-FHIR, designed to convert data from the openEHR model into an equivalent GECCO FHIR profile. However, the tool’s implementation, which relies on Java classes with regular expressions, limits its reusability for other use cases. Similarly, Soares et al. [24] developed the OMOP-to-Bulk FHIR tool to process data from an EHR stored in the OMOP Common Data Model (CDM) into FHIR resources. Their workflow involved custom SQL queries to extract data into an intermediate format, which was then transformed into FHIR resources using translation rules. This process was powered by the Whistle Engine, a tool developed by Google that supports FHIR-specific functions and transformations. While effective within the OMOP framework, this solution’s dependence on the OMOP CDM as the input data model greatly limits its broader adoption. Finally, Simon et al. [25] explored the use of a metadata repository (MDR) to streamline the conversion of data from primary clinical systems into standardized formats. Their methodology involved extracting metadata from hospital forms, defining mappings and transformations in the MDR, and using a REST API to automate the conversion process. Despite its potential, the approach depends heavily on custom scripts, which may hinder generalizability and demand significant effort to adapt to different scenarios.

¹ <https://build.fhir.org/profiling.html>

Table 2

Summary of Related Work: The table outlines the goals and challenges of prior works in the field, highlighting their limitations and how our approach overcomes these gaps to offer novel contributions.

Ref.	Goal	Challenges	Our Approach
[21]	Convert the MIMIC-IV database into the FHIR data model for a real-world FHIR store	Lack of data sanitization, Internal mapping logic	Elimination of internal mapping logic, standardized and shareable mapping process
[9]	Develop a minimum Data Set for CPOE in the cardiovascular field, integrated with the EHR system	No digital tools for mapping, Individual manual verification required for each resource	Modular, automated mapping system eliminating manual verification
[22]	Enable the FAIR-ification of health data using user-friendly GUIs	Dependence on internal mapping language, Enforcing of FHIR Profiles	Language-agnostic, templated approach for defining mappings
[17]	Develop an FHIR-based data harmonization pipeline for AI-friendly clinical data representation	Mapping process dependent on Python and explicit source database attributes, limiting code reuse	Portable, reusable framework using standardized mapping language
[23]	Convert COVID-19-related patient data (GECCO) from the openEHR model to FHIR profiles	Implementation based on Java classes and regular expressions, limiting reusability	Reusable, standardized mapping language avoiding reliance on specific implementations
[24]	Transform EHR data stored in OMOP CDM into FHIR resources	Dependence on the OMOP CDM as the input model, limiting adoption	Flexible framework supporting multiple input models for broader applicability
[25]	Automate the conversion of clinical data into standardized formats using an MDR	Reliance on custom scripts, requiring additional effort for generalization and maintenance	Standardized mapping language minimizing reliance on custom scripts

These approaches, summarized in Table 2, reveal challenges arising from manual conversion and validation processes and the use of internal mapping languages that hinder the sharing and publishing of mappings. We recognized these shortcomings and defined the essential functionalities necessary for an effective conversion pipeline. These features, detailed in Section 2, guided the development of our solution — a modular framework designed to translate data from various source formats into FHIR seamlessly.

Our approach employs a templating conversion strategy, implemented through the FHIR Mapping Language. This approach enables the definition of mappings in a standardized and declarative manner, empowering users without technical expertise to concentrate only on the data model, and avoiding low-level programming issues. Moreover, our mappings, based on a standard language definition, are language-agnostic and highly portable, facilitating seamless operation on any FHIR server.

We converted the MOTU dataset to validate our pipeline, a retrospective study evaluating safety against falls and effectiveness for balance and mobility outcomes among different prosthetic knees [26].

The remainder of this manuscript is organized as follows. Section 2 outlines the methodology used in our study, including the pipeline design, architecture, and implementation. In Section 3, we present the results of the data conversion and validation processes. Section 4 discusses the implications of our findings and potential avenues for future research. It also explores directions for further investigation. Finally, Section 5 summarizes the key contributions and provides concluding remarks.

2. Methods

Our work focuses on establishing a conversion pipeline that seamlessly translates data from diverse source formats into modern clinical standards, thereby enhancing healthcare interoperability. This section delineates the development strategy, the architectural framework, the implementation methodologies, and the tools employed in crafting the pipeline.

2.1. Pipeline design

2.1.1. Templating strategy for declarative mapping

Our conversion process relies on a templating strategy that articulates data mappings in a declarative manner. This approach decouples mapping definitions from a specific programming language, promoting flexibility and adaptability across various data models and formats.

2.1.2. Modular architecture for structural advantages

The conversion pipeline is structured into modular components, providing several architectural advantages crucial for robustness and scalability:

- **Maintainability:** The modular design prevents unwanted dependencies between components. This isolation ensures that any requirements alterations impact only the relevant module, minimizing unintended consequences.
- **Incremental Development:** Breaking tasks into distinct components allows for an incremental implementation workflow. This “divide and conquer” approach facilitates a more manageable and systematic development process.
- **Reusability:** These modular components are designed for reuse across various pipeline implementations. This feature significantly reduces the time needed to address new use cases.
- **Testing and Debugging:** The well-defined responsibilities of each module enable precise control over the software. This granularity simplifies the testing and debugging of smaller, individual components, streamlining the development, and troubleshooting process.

2.2. Pipeline architecture

The design and development of our pipeline are based on the Extract Transform Load (ETL) framework [27], and its functionalities were organized into sequential modules (Fig. 1).

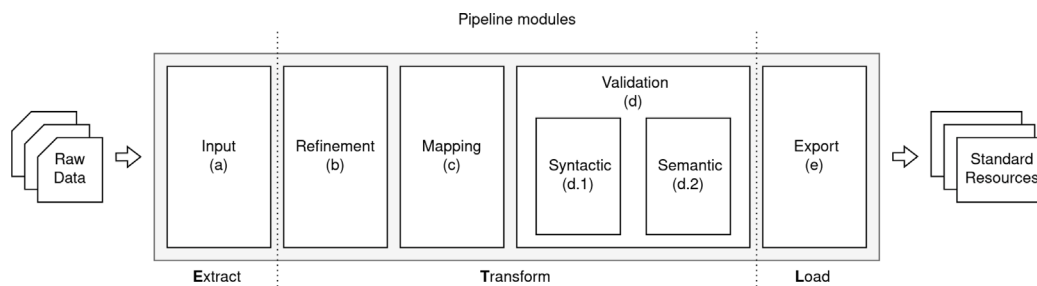


Fig. 1. Modules representation of the pipeline architecture, organized according to the Extract, Transform, and Load (ETL) process steps.

1. **Input:** This module represents the strategy to acquire input data from a specific source, with no restrictions on the input data model and format, as long as a proper API to gather data is provided (e.g., databases, platforms with HTTP API, files, etc.). Additionally, this module addresses data security and regulatory compliance, ensuring that data quality, integrity, and confidentiality are not compromised.
2. **Refinement:** This module is designed to preprocess and refine input data. Given the diverse formats and varying quality of the source datasets, this module ensures a uniform output. Moreover, it ensures consistency by eliminating missing data and structuring the information (e.g. using JSON format) to enable seamless conversion operations in subsequent steps. Specifically, the module preserves rows with non-essential null values and removes the ones where missing data are crucial for conveying the meaning of the information.
3. **Mapping:** This module takes the sanitized source data as input and performs the conversion to the target data model by employing a templating strategy.
4. **Validation:** It carries out all the procedures aimed at checking the conformity of output resources with the standard target data model. In case of undesirable outcomes, the module takes immediate action to notify the presence of invalid data. It also initiates any roll-back functions to maintain the integrity of the system's state.
5. **Publishing:** The final module executes the necessary logic to store the information in an accessible repository. It also provides essential functionalities, such as querying, exporting, and logging, which enable developers to debug results and design automated tests.

2.3. Pipeline implementation

Our pipeline was methodically constructed in stages, employing data extracted from a database and compliant with the HL7 FHIR standard for the output data model. The choice of HL7 FHIR was driven by its flexibility, scalability, and ability to integrate with modern healthcare systems seamlessly, ensuring compatibility and interoperability across diverse data formats and systems in the healthcare domain.

The *Input* (a) and *Refinement* (b) modules (Fig. 2) were implemented using Python scripts supported by data manipulation libraries. These scripts transform the source data into sanitized JSON documents for further processing. In our specific case, the developed modules do not include any security checks because the dataset to be converted was already anonymized and verified for quality and integrity at the source.

The *Mapping* (c) module was developed utilizing Matchbox, a versatile FHIR templating engine characterized by its offline transformation capabilities, allowing standalone microservice deployment, and providing seamless integration as a Java library. Matchbox transforms input JSON object into an FHIR Transaction,² a mechanism that empowers clients to send multiple interactions to a server to be processed

atomically. This approach adopts the FHIR Bundle Resource as the transport, serving as a container for a collection of FHIR resources. The transactional approach has proved invaluable to performing non-trivial operations, such as storing a collection of related resources, performing conditional creates or updates on existing ones, etc. These features ensure the flexibility of our pipeline to accommodate future evolutions and changing requirements. Matchbox leverages templates defined using the FHIR Mapping Language³ to efficiently transform data from any format to the FHIR data model. Specifically, these templates are parsed into Structure Maps, a detailed set of rules describing the relationships between different data structures, required to enable their automatic conversion. Both the incoming data structures and the resulting FHIR resources are represented by instances of Structure Definition – an FHIR resource that serves as a metamodel for describing schemas. To ensure the integrity of the process, Matchbox conducts syntactic validation (d.1) of the results using the built-in Structure Definitions of the FHIR data model. This step adds a quality control layer, aimed at verifying adherence of produced results with the FHIR specification.

Upon completion of the pipeline, the HAPI FHIR JPA Server,⁴ an open-source, self-contained solution, was deployed. This server implements storage and retrieval logic without additional coding and exposes a RESTful API. Its primary role involves semantic validation (d.2) of incoming resources concerning the official HL7 Structure Definitions before storing them in the database (e). Lastly, data from the server can be exported in various file formats such as XML, JSON, and FHIR-RDF.

2.4. Validation methodology

To rigorously assess the accuracy of our pipeline, we focused on converting the MOTU dataset to the FHIR model. MOTU originates from a retrospective observational study conducted at the Italian National Institute for on-the-Job Injuries (INAIL) Prosthesis Center in Budrio, Italy. The study was approved by the Ethics Committee “Area Vasta Emilia Centro” (ref. MOTU 18088, CE AVEC n. 380/2018/OSS/AUSLBO) and complies with the Declaration of Helsinki. It aims to evaluate safety against falls and effectiveness for balance and mobility outcomes among prosthetic knees. The dataset is provided in Comma Separated Values (CSV) format and composed of 1962 hospital stays of 1006 patients with unilateral trans-femoral amputation from January 2011 to May 2020. It includes five tables: Patient, HospitalStay, Fall, DrugHospitalStay, and ProstheticKnee. Patient data feature anonymous IDs, age, and gender, while HospitalStay records admission/discharge dates, amputation details, rehabilitation goals, and assessments on balance, mobility, and daily living activities through questionnaires and functional tests. The Fall table provides information about 146 falls experienced during hospitalizations, detailing prosthesis use and resulting injuries. The DrugHospitalStay table reports information about the medication intake for each hospital stay, listed by trade name and ATC (Anatomical Therapeutic Chemical) code, and counts 3035 instances.

³ <https://build.fhir.org/mapping-language.html>

⁴ <https://hapifhir.io/>

² <https://build.fhir.org/http.html#transaction>

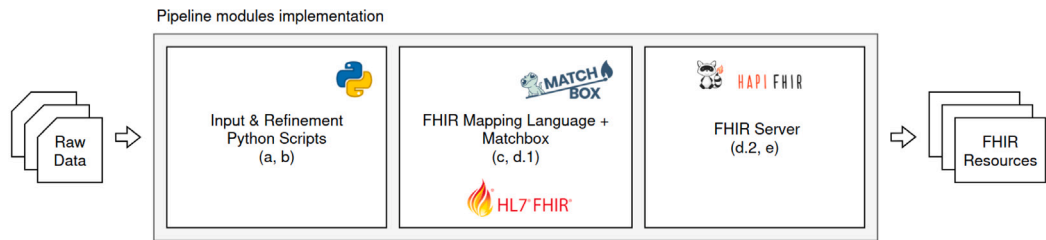


Fig. 2. A possible implementation of the pipeline. The identifiers (a, b), (c, d.1), and (d.2, e) refer to the modules of the pipeline architecture shown in Fig. 1.

Table 3
Demonstration queries used to validate the mapping results. It shows the MOTU tables and the corresponding FHIR resources involved in each query.

Query	MOTU Table(s)	FHIR Resources
Q1: How many patients who came for an initial supply return for a second hospitalization?	HospitalStay	Encounter
Q2: How many patients transition from the initial supply of a mechanical knee to a second supply of an electronic knee?	Knee	Patient DeviceRequest DeviceDefinition
Q3: How many patients use anxiolytics/antidepressants (ATC Code: N06A and N05A) related to the risk of falls?	Drug	Patient Medication MedicationRequest
Q4: Hospital stays from patients over 65 years old?	Knee Patient	Patient DeviceRequest
Q5: How many falls per type of prosthesis (AMK, FK, etc.)	HospitalStay Fall Knee	DeviceDefinition DeviceRequest Encounter AdverseEvent

The Knee table contains 40 different prosthetic knees, characterized by device details, and manufacturer specifics. To maintain transparency and reproducibility, the input dataset used in this study is publicly available on Zenodo [28].

As a strategy for assessing the completeness and accuracy of the conversion outcomes, we defined five sample queries (Table 3) encompassing a wide range of tables, columns, and data types within the source model. These queries were designed in collaboration with experts from the INAIL, one of Italy’s largest prosthetic centers, including a researcher with 30 years of experience and a physical therapist involved in rehabilitating patients with amputations. First, we employed the pandas library in Python to execute the queries on the original dataset, ensuring that the logic and criteria set by the domain experts were faithfully replicated. Subsequently, using the SPARQL querying language, we developed corresponding queries replicating the logic of the original ones. This process involved retrieving the FHIR resources in FHIR/RDF format from the HAPI FHIR server’s REST API and then uploading them to an Apache Jena Fuseki⁵ server, an easy to setup open-source SPARQL server that enables simple querying of data through a web interface. The primary goal of such an approach was to compare the resulting number of retrieved instances, aiming for identical outcomes on corresponding queries to confirm the success of the transformation.

To ensure high data quality in the resulting dataset, we adopted the Kahn framework [29], a systematic methodology for data quality assurance. The framework emphasizes three main categories: *Conformance*, which checks the data against syntax and structural constraints; *Completeness*, which ensures that there are no missing values; and *Plausibility*, which assesses the believability and truthfulness of the data values. Each category is further divided into subcategories, providing a detailed structure for quality evaluation (Table 4).

The framework distinguishes between two evaluation contexts: verification, which assesses how well the data aligns with local knowledge and expectations, and validation, which uses external knowledge and benchmarks to evaluate quality. Our study focuses on the verification

context, using the precise definitions of data quality of the Kahn framework in all categories to evaluate how effectively our methodology addresses these standards.

3. Results

The implemented pipeline successfully converted the MOTU dataset into FHIR-compliant resources [30]. The transformation process, leveraging the previously described modules, encompassed the following key steps (Fig. 3):

- **Configuration Setup (Step 0):** Upon deployment, the pipeline was configured by uploading Structure Definitions and Structure Maps obtained in collaboration with MOTU dataset experts [26].
- **Data Preprocessing (Step 1):** Raw data from the MOTU dataset [28] underwent a thorough preprocessing step by using data sanitization services. This step ensured data quality, consistency, and compatibility with subsequent pipeline stages.
- **Data Transformation (Step 2):** The CSV data was transformed into JSON objects to streamline processing by Matchbox. This step included renaming, typing, and filtering source data fields.
- **Mapping (Step 3):** Matchbox transformed each JSON object into an FHIR Transaction. Moreover, Matchbox performed a syntactic validation to ensure the correctness of input and output data schemas.
- **Bundles Upload (Step 4):** Generated Bundles were dispatched individually to the HAPI FHIR server to validate and store the enclosed resources. Placeholder IDs, temporary and randomly generated UUIDs, were used to link resources within the Bundle, facilitating the determination of corresponding concepts and preventing duplications.
- **Transaction Application (Step 5):** Upon arrival at the HAPI FHIR server, Bundles underwent a processing step with reference integrity checks, resolution of temporary UUIDs, and assessment of resource conformity to FHIR profiles using the server’s integrated FHIR Instance Validator module.⁶ This will use terminology services to validate codes, and StructureDefinitions to

⁵ <https://jena.apache.org/documentation/fuseki2/>

⁶ https://hapifhir.io/hapi-fhir/docs/validation/instance_validator.html

Table 4
Implementation strategies for the Kahn framework data quality assessments (DQA).

DQA	Description	Implementation
Value Conformance	Data values conform to allowed formats and values	Adopting an FHIR Server with strict validation features ensures that data values conform to the correct formats, relational constraints are respected, and required values are always present, thus guaranteeing completeness and data integrity.
Relational Conformance	Data values conform to relational constraints and no duplicate keys	
Completeness	Absence of data under a predefined threshold	
Computational Conformance	Computed values conform to programming specification	Computed values, like the number of medications, are not stored, as they can be derived dynamically using FHIRPath queries.
Uniqueness Plausibility	There aren't two or more resources that identify the same subject	By adopting FHIR transactions for incremental updates, the risk of creating duplicate resources for the same patient is eliminated.
Atemporal Plausibility	Data measures are coherent to a local knowledge	During the pre-processing phase, Python scripts with custom rules are used to validate input data, ensuring that data adheres to expected properties and is consistent with local knowledge.
Temporal Plausibility	Data measures are coherent with expected temporal properties	

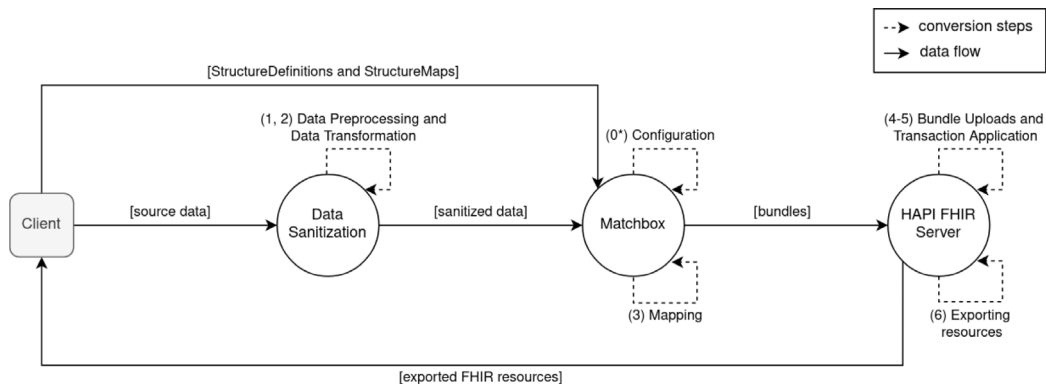


Fig. 3. Representation of the role of different services participating in the pipeline. (*) The setup action is performed only once, during the initial configuration on the pipeline.

validate semantics. Successful transactions modified the state of the server, enabling incremental updates to the same resource from subsequent Bundles. On the other hand, failed transactions triggered a complete rollback to the previous state, ensuring successful storage only upon error-free processing.

- **Exporting Resources (Step 6):** All stored resources within the HAPI FHIR server resulting from the conversion process were exported in NDJSON format using the server's built-in FHIR Bulk Data Access protocol. This operation organized FHIR resources by their type and segmented them into files. The resulting FHIR resources dataset can be accessed on Zenodo [31].

A concrete example of data workflow conversion is illustrated in Fig. 4. In Step 1, the initial phase entailed data sanitization, involving the removal of unused columns, and the conversion of dates into the ISO 8601 format. Afterward, in Step 2, the data is transformed into

JSON format, enabling its utilization through Matchbox (Fig. 4.b). Step 3 encompassed the mapping of the source data onto a Bundle comprising FHIR-compliant resources, each identified by temporary UUIDs, thereby enabling seamless cross-referencing (Fig. 4.c). Notably, the semantic richness of FHIR resources was enhanced by the integrating formal CodeSystems. Subsequently, the Bundle was transmitted to the HAPI FHIR server to execute the necessary transaction (Step 4). Finally, in Step 5, the resultant FHIR resources were individually stored in the HAPI FHIR server, and their temporary UUIDs were substituted with universal ones to enable their integration with the existing knowledge (Fig. 4.d).

The experiment was conducted on a system with an Intel Core i9-14900KF processor (32 cores) and 128 GB of RAM, running Ubuntu 24.04.1 LTS. The pipeline converted the entire MOTU dataset in 8 min and 52.467 s.

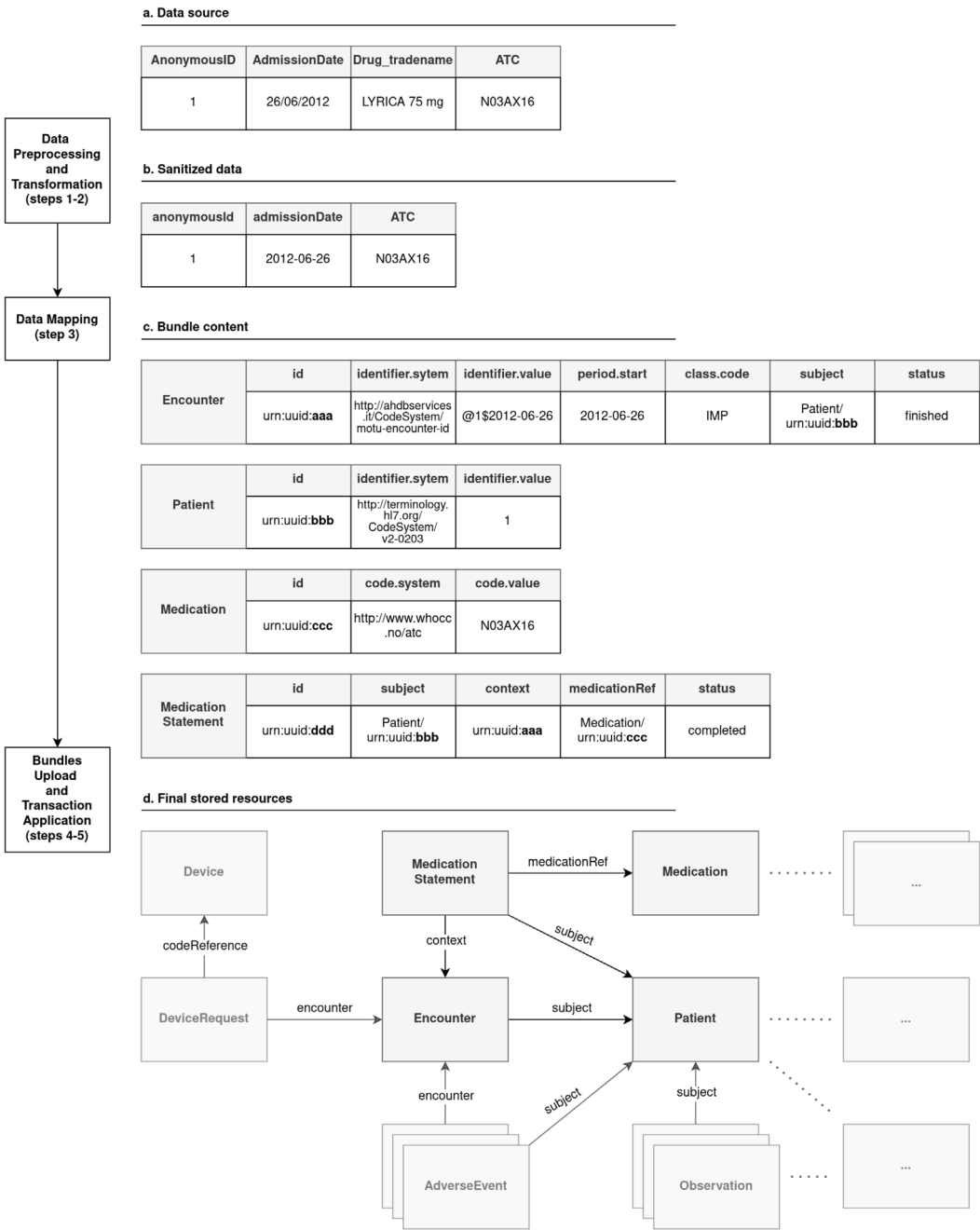


Fig. 4. Diagram showing data conversion workflow from MOTU DrugHospitalStay table to FHIR domain, bundling resources in step 3, and integrating them with prior knowledge in steps 4-5".

3.1. Implemented mappings

The pipeline conversion process involves 157 fields from the MOTU tables, mapped to 15 distinct types of FHIR resources (Table 5). A correspondence to FHIR variables was established for 155 fields, covering 98.7% of the dataset [30]. Of these, 143 fields (91.1%) were directly mapped to FHIR resource fields, while the remaining 12 (1.3%) fields were recognized as computable values through FHIRPath expressions. Several external ontologies, such as LOINC,⁷ SNOMED-CT,⁸ and Units of Measurement (UOM)⁹ ontology were adopted, to provide the results

with a precise, consistent, and machine-readable representation of clinical concepts. As the information from the source dataset is specific, a corresponding concept from standard vocabularies was sometimes missing. Different Code Systems were created to overcome this lack of coverage. Moreover, before initiating the pipeline, shared concepts between all the records, such as Questionnaires, and Organizations were manually created and uploaded to the FHIR server.

3.2. Validation

The validation process for resources derived from the MOTU dataset involved utilizing FHIRPath queries executed on the HAPI FHIR Server, and Python scripts for analyzing the original dataset. Initially, the validation encompassed counting the “root” resources, signifying overarching concepts uniquely identified by a primary key within the MOTU

⁷ <https://loinc.org/>
⁸ <https://www.snomed.org/>
⁹ <https://units-of-measurement.org/>

Table 5

Shows high-level mappings between MOTU tables and FHIR resources. These mappings establish a central reference point - Encounter, DeviceDefinition, Medication, AdverseEvent, and Patient - since it is referenced by all other resources derived from the same MOTU table's mapping process.

MOTU	FHIR	Total resources	Usage
HospitalStay	Encounter	1962	Models the Information about the hospitalization of a patient
	QuestionnaireResponse	23 544	Models the responses to standard questionnaires (Barthel, Morse, etc.)
	Account	1962	Outline the entity in charge of payment
	Coverage	1962	
	Observation	11 709	Describes the measurements and the analysis results performed during the hospitalization
	Procedure	1962	Defines an intervention of amputation (body part, cause, date)
	Condition	181	Models any patient's disorders, such as diabetes.
	DeviceRequest	1835	Model a request for knee prosthesis
	Careplan	1962	Describe the rehabilitation and the achieved goals
	Goal	1962	
Knee	DeviceDefinition	40	Defines the properties of a certain model of knee prosthesis
Drug	Medication	370	Describe a drug and its prescriptions
	MedicationStatement	3032	
Fall	AdverseEvent	146	Models a fall event associated with the hospitalization.
Patient	Patient	1006	Models the patient's details

Table 6

Comparison of distinct item counts in the original dataset and the corresponding FHIR “root” resources derived from mapping. Results are aligned, except for three MedicationStatement entries (*) excluded due to lacking ATC codes.

MOTU Table	FHIR Resource	No. MOTU	No. FHIR
HospitalStay	Encounter	1962	1962
Knee	DeviceDefinition	40	40
Drug	MedicationStatement	3035	3032*
Fall	AdverseEvent	146	146
Patient	Patient	1006	1006

Table 7

Summary of results from demonstration queries. The total count for query Q5 is 144, reflecting the exclusion of two rows from the HospitalStay table due to the absence of any reference to the prosthesis model, leading to their omission in the inner join operation.

Query	Python count	SPARQL count
Q1: How many patients who came for an initial supply return for a second hospitalization?	85	85
Q2: How many patients transition from the initial supply of a mechanical knee to a second supply of an electronic knee?	34	34
Q3: How many patients use anxiolytics/antidepressants (ATC Code: N06A and N05A) related to the risk of falls?	67	67
Q4: Hospital stays from patients over 65 years old	674	674
Q5: How many falls per type of prosthesis (AMK, FK, etc.)	AMK - 11	AMK - 11
	FK - 79	FK - 79
	LK - 26	LK - 26
	MPK - 28	MPK - 28

tables. This count was subsequently compared to the total number of rows within the tables, as outlined in Table 6. The consistent alignment of query counts across all assessments confirmed the accurate transformation of the core concepts of the MOTU domain. Subsequently, utilizing queries provided by domain experts (Table 3), we found consistent outcomes across both groups of queries (Table 7).

4. Discussion

We successfully designed and implemented a modular FHIR-based conversion pipeline capable of extracting, preprocessing, and standardizing real-world health data. Validation, performed on the MOTU dataset, ensures that the transformed data maintains accuracy and meaning. Consistent counts validated the preservation of core concepts, while clinician-driven queries affirmed the retention of semantic integrity and relational coherence. This comprehensive validation reinforces the trustworthiness of the transformed FHIR resources in capturing the essence of the MOTU domain. The resulting FHIR-conformant resources enhance interoperability, support high-throughput computation [32], and improve accessibility and readability for both humans and machines. This achievement is reinforced by the integration of standard ontologies, contributing to enriched semantic content and reduced

ambiguities. Furthermore, the generated resources are efficiently findable through the RESTful Search API – a critical component for realizing a scalable platform capable of handling extensive volumes of information. Our harmonization process yields output with a greater volume than the input, attributable to the inclusion of metadata, mandatory fields, and references to external ontologies or concepts within FHIR resources. Managing this expanded knowledge corpus necessitates the deployment of automated software infrastructures, as emphasized in prior research [16].

The implemented pipeline provides a versatile and scalable framework to support the FHIR modeling of custom data, including modules aimed at streamlining semi-automated data integration. The validation module removes the need for manual evaluation, cutting down time and reducing errors. This is pivotal for achieving semantic and structural interoperability in health information, identified as a significant research challenge in recent years [4,5]. As demonstrated with the MOTU dataset, this framework could play a crucial role in addressing a key aspect of digital health research — facilitating the comprehensive and meaningful reuse of real-world data. This capability is vital for supporting analyses in the field of Medical Data Science [16].

In the realm of health informatics, the showcased pipeline exhibits several remarkable strengths that distinguish it from existing

approaches. Primarily, its modular design ensures scalability and adaptability for a broad range of clinical studies and data sources, providing a versatile solution applicable across various scenarios. Notably, the adoption of Bundle transactions significantly enhances efficiency, enabling the processing of large datasets at speeds up to 50 times¹⁰ faster than individual operations. Furthermore, utilizing the RESTful interface provided by the FHIR server allows for parallel processing of bundles, further enhancing overall performance.

An essential feature of the framework is its automated validation mechanism, systematically ensuring the adherence of each FHIR resource to the FHIR model. This procedure, supported by Structure Definitions and Structure Maps, distinguishes our study from others that rely on manual validation processes [9]. Additionally, the adoption of a templating conversion strategy introduces a language-agnostic approach. This empowers users to concentrate exclusively on the mapping activity, irrespective of their technical proficiency. This eliminates the necessity for alterations to the framework code, allowing users to navigate the conversion process without requiring expertise in specific programming languages.

Unlike similar solutions such as the .NET-based FHIR Converter,¹¹ limited to handling specific source data types (HL7 version 2 and CDA), our pipeline boasts versatility by accommodating any input format. In addition, adopting the FHIR Mapping Language not only formalizes the conversion process but also makes it externally accessible and publishable. This distinctive feature promotes transparency in our mappings, a quality often lacking in comparable studies [17,22].

The primary limitation of our research lies in its validation, which was exclusively conducted using the MOTU dataset. To enhance the robustness of our solution, our objective is to broaden the validation scope to encompass a wider range of datasets. This step is crucial for unveiling potential challenges and obstacles while assessing the practical adaptability of our pipeline architecture across diverse scenarios. For instance, it should be noted that, for the MOTU dataset, creating FHIR profiles was not necessary to handle explicit constraints given by domain-specific rules or requirements. On the contrary, other datasets may require the definition of FHIR profiles. As an example, the US Core FHIR profile¹² provides its definitions for many FHIR resources, such as the Patient resource, which includes mandatory fields such as name and additional fields such as date of death, not natively mandatory in the basic Patient resource. The inclusion of FHIR profiles creation must be seamlessly integrated into our pipeline to map the structure and characteristics of the source information, avoiding any loss of context or meaning.

Other limitations arise from the current manual procedures for implementing ad-hoc input/refinement scripts and defining mappings. These tasks, inherently domain-specific, require substantial time and effort investment. A promising path for improvements involves exploring the implementation or integration of Natural Language Processing (NLP), as exemplified in Google's Healthcare Natural Language API.¹³ NLP tools can be used to automatically extract structured data from unstructured clinical documents [32] and to assist in creating mappings by suggesting the most suitable FHIR resources and external ontology codes to model specific concepts. In this regard, we are developing an automatic mapping tool that uses NLP to convert textual descriptions of variables into FHIR resources. By fine-tuning the model with FHIR documentation, the tool will act as a decision support system, suggesting outcomes that operators can review and approve. To enhance usability for non-technical users and streamline the process, these tools should feature a graphical user interface, ultimately saving time and simplifying the mapping workflow.

Finally, to enhance the efficiency of the overall post-hoc validation step, we are working on automation to streamline procedures. This involves integrating automated functionalities into a new pipeline module, aligned with DevOps principles. This module will orchestrate a series of tasks triggered by changes in mapping logic or source dataset. For instance, the generated FHIR resources will be converted directly into RDF format and integrated into a pre-deployed knowledge graph repository, like Apache Jena Fuseki. Afterward, predefined queries in configuration files will be analyzed against both input and output providing a comprehensive summary of results. Automating these tasks facilitates seamless coverage and ensures consistent application of semantic integrity tests. Importantly, this approach decouples the technical implementation from the semantics of the queries outlined by non-technical users, significantly enhancing the acceptance process and pipeline usability.

5. Conclusion

This paper presents an FHIR-based conversion pipeline designed to enhance the interoperability of real-world health data. The transition to a modern clinical standard, such as HL7 FHIR, represents a pivotal step towards unlocking latent value within heterogeneous data silos. Our solution, characterized by its modular architecture and templating mapping strategy based on the FHIR Mapping Language, offers a flexible and scalable approach to implementation, effectively addressing the complexities inherent in health data management. Furthermore, the employed templating approach ensures the formalization of mappings, providing an implementation-independent solution suitable for external publication. The effectiveness of our proposed solution underwent rigorous validation against the MOTU dataset. Through well-conceived queries on both the source dataset and the output generated by the pipeline, we demonstrated the pipeline's ability to seamlessly transform the MOTU dataset into FHIR without compromising accuracy or completeness.

Building on these achievements, future efforts will prioritize expanding validation across diverse datasets, encompassing variations in structure, quality, and domain-specific requirements. Incorporating FHIR profile creation, where needed, will ensure that domain constraints are met without losing context or meaning. To address the manual bottlenecks in mapping and refinement, we plan to integrate advanced automation techniques, including natural language processing, to support variable extraction and mapping processes. Developing intuitive graphical interfaces for these tools will empower non-technical users, fostering broader adoption and usability.

Additionally, enhancing the automation of post-hoc validation through DevOps-aligned modules will streamline workflows, ensuring consistency and semantic coherence across all outputs. The incorporation of automated RDF conversion and semantic integrity tests within knowledge graphs will further reinforce the reliability of the pipeline while decoupling technical complexities from user interactions.

By addressing these aspects, we aim to solidify the applicability and scalability of our FHIR-based conversion pipeline in real-world settings, effectively addressing prevailing challenges in health data interoperability.

CRedit authorship contribution statement

Alberto Marfaglia: Writing – review & editing, Writing – original draft, Software, Resources, Methodology, Data curation. **Filippo Nardini:** Writing – review & editing, Writing – original draft, Software, Resources, Methodology, Data curation. **Valerio Antonio Arcobelli:** Writing – review & editing, Writing – original draft, Software, Resources, Methodology, Data curation. **Serena Moscato:** Writing – review & editing, Writing – original draft, Software, Resources, Methodology, Data curation. **Sabato Mellone:** Writing – review & editing, Supervision, Project administration, Investigation, Funding acquisition, Formal analysis, Conceptualization. **Antonella Carbonaro:** Writing – review & editing, Writing – original draft, Supervision, Resources, Project

¹⁰ https://smilecdr.com/docs/fhir_standard/transactions.html

¹¹ <https://github.com/microsoft/FHIR-Converter>

¹² <http://hl7.org/fhir/us/core/>

¹³ <https://cloud.google.com/healthcare-api/docs/how-tos/nlp?hl=it>

administration, Methodology, Funding acquisition, Formal analysis, Conceptualization.

Code availability

The MOTU-to-FHIR Mapping Pipeline code is available on GitHub at <https://gitlab.com/almahealthdb/ahdb-mapping-service/>. This repository includes all the necessary scripts, configuration files, and guidelines for converting the MOTU CSV dataset into a FHIR dataset according to the HL7 FHIR standard (R4 version). It also offers a Docker Compose file configuring Matchbox and the HAPI FHIR Server, enabling automated pipeline deployment. For detailed instructions on running the pipeline, please refer to the README.md file in the repository.

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 article.

Acknowledgments

This research was co-funded by the 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, Italy, code PNC0000002, CUP, Italy: (B53C22006450001) and by the Italian National Institute for Insurance against Accidents at Work (INAIL), Italy within the MOTU++ project (PR19-PAI-P2). Authors express their gratitude to the AlmaHealthDB Team.

References

- [1] J. Thomason, Data, digital worlds, and the avatarization of health care, *Glob. Heal. J.* 8 (1) (2024) 1–3, <http://dx.doi.org/10.1016/j.glohj.2024.02.003>, URL <https://www.sciencedirect.com/science/article/pii/S2414644724000034>.
- [2] F.A. Seixas-Lopes, C. Lopes, M. Marques, C. Agostinho, R. Jardim-Goncalves, Musculoskeletal disorder (MSD) health data collection, personalized management and exchange using fast healthcare interoperability resources (FHIR), *Sensors* 24 (16) (2024) <http://dx.doi.org/10.3390/s24165175>, URL <https://www.mdpi.com/1424-8220/24/16/5175>.
- [3] M. Lehne, J. Sass, A. Essenwanger, J. Schepers, S. Thun, Why digital medicine depends on interoperability, *Npj Digit. Med.* 2 (1) (2019) 79, <http://dx.doi.org/10.1038/s41746-019-0158-1>, URL <https://www.nature.com/articles/s41746-019-0158-1>.
- [4] A. Chatterjee, N. Pahari, A. Prinze, ?H17 FHIR with SNOMED-CT to achieve semantic and structural interoperability in personal health data: A proof-of-concept study, *Sensors* 22 (10) (2022) 3756, <http://dx.doi.org/10.3390/s22103756>, Number: 10 Publisher: Multidisciplinary Digital Publishing Institute. URL <https://www.mdpi.com/1424-8220/22/10/3756>.
- [5] K.T. Kadakia, M.D. Howell, K.B. DeSalvo, Modernizing public health data systems: Lessons from the health information technology for economic and clinical health (HITECH) act, *JAMA* 326 (5) (2021) 385–386, <http://dx.doi.org/10.1001/jama.2021.12000>.
- [6] T. Benson, G. Grieve, Why interoperability is hard, in: T. Benson, G. Grieve (Eds.), *Principles of Health Interoperability: FHIR, HL7 and SNOMED CT*, in: Health Information Technology Standards, Springer International Publishing, Cham, 2021, pp. 21–40, http://dx.doi.org/10.1007/978-3-030-56883-2_2.
- [7] Q&A on the European Health Data Space. URL https://ec.europa.eu/commission/presscorner/detail/en/qanda_24_2251.
- [8] 21st century cures act: Interoperability, information blocking, and the ONC health IT certification program, 2020, URL <https://www.federalregister.gov/documents/2020/05/01/2020-07419/21st-century-cures-act-interoperability-information-blocking-and-the-onc-health-it-certification>.
- [9] M. Montazeri, R. Khajouei, A. Mahdavi, L. Ahmadian, Developing a minimum data set for cardiovascular computerized physician order entry (CPOE) and mapping the data set to FHIR: A multi-method approach, *J. Med. Syst.* 47 (1) (2023) 47, <http://dx.doi.org/10.1007/s10916-023-01943-2>.
- [10] C.N. Vorisek, M. Lehne, S.A.I. Klopstein, P.J. Mayer, A. Bartschke, T. Haese, S. Thun, Fast healthcare interoperability resources (FHIR) for interoperability in health research: Systematic review, *JMIR Med. Inform.* 10 (7) (2022) e35724, <http://dx.doi.org/10.2196/35724>, Company: JMIR Medical Informatics Distributor: JMIR Medical Informatics Institution: JMIR Medical Informatics Label: JMIR Medical Informatics Publisher: JMIR Publications Inc., Toronto, Canada. URL <https://medinform.jmir.org/2022/7/e35724>.
- [11] T. Namli, A. Anil Sinaci, S. Gönül, C.R. Herguido, P. Garcia-Canadilla, A.M. Muñoz, A.V. Esteve, G.B.L. Ertürkmen, A scalable and transparent data pipeline for AI-enabled health data ecosystems, *Front. Med.* 11 (2024) 1393123.
- [12] A. Mavrogiorgou, A. Kiourtis, M. Touloupou, E. Kapassa, D. Kyriazis, Internet of medical things (IoMT): Acquiring and transforming data into HL7 FHIR through 5G network slicing, *Emerg. Sci. J.* 3 (2) (2019) 64–77, <http://dx.doi.org/10.28991/esj-2019-01170>, Number: 2. URL <https://ijournalse.org/index.php/ESJ/article/view/131>.
- [13] R. Gazzarata, J. Almeida, L. Lindsköld, G. Cangioli, E. Gaeta, G. Fico, C.E. Chronaki, HL7 fast healthcare interoperability resources (HL7 FHIR) in digital healthcare ecosystems for chronic disease management: Scoping review, *Int. J. Med. Inform.* 189 (2024) 105507, <http://dx.doi.org/10.1016/j.ijmedinf.2024.105507>, URL <https://www.sciencedirect.com/science/article/pii/S1386505624001709>.
- [14] A. Carbonaro, A. Marfoglia, F. Nardini, S. Mellone, CONNECTED: leveraging digital twins and personal knowledge graphs in healthcare digitalization, *Front. Digit. Heal.* 5 (2023) <http://dx.doi.org/10.3389/fdgh.2023.1322428>, Publisher: Frontiers. URL <https://www.frontiersin.org/journals/digital-health/articles/10.3389/fdgh.2023.1322428/full>.
- [15] S.N. Duda, N. Kennedy, D. Conway, A.C. Cheng, V. Nguyen, T. Zayas-Cabán, P.A. Harris, HL7 FHIR-based tools and initiatives to support clinical research: a scoping review, *J. Am. Med. Inform. Assoc.* 29 (9) (2022) 1642–1653, <http://dx.doi.org/10.1093/jamia/ocac105>, URL <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC9382376/>.
- [16] J. Gehrmann, E. Herczog, S. Decker, O. Beyan, What prevents us from reusing medical real-world data in research, *Sci. Data* 10 (1) (2023) 459, <http://dx.doi.org/10.1038/s41597-023-02361-2>, Number: 1 Publisher: Nature Publishing Group. URL <https://www.nature.com/articles/s41597-023-02361-2>.
- [17] E. Williams, M. Kienast, E. Medawar, J. Reinelt, A. Merola, S.A.I. Klopstein, A.R. Flint, P. Heeren, A.-S. Poncette, F. Balzer, J. Beimes, P. von Büna, J. Chromik, B. Arnrich, N. Scherf, S. Niehaus, A standardized clinical data harmonization pipeline for scalable AI application deployment (FHIR-DHP): Validation and usability study, *JMIR Med. Inform.* 11 (2023) e43847, <http://dx.doi.org/10.2196/43847>.
- [18] G. Xiao, E. Pfaff, E. Prud'hommeaux, D. Booth, D.K. Sharma, N. Huo, Y. Yu, N. Zong, K.J. Ruddy, C.G. Chute, G. Jiang, FHIR-ontop-OMOP: Building clinical knowledge graphs in FHIR RDF with the OMOP common data model, *J. Biomed. Inform.* 134 (2022) 104201, <http://dx.doi.org/10.1016/j.jbi.2022.104201>, URL <https://www.sciencedirect.com/science/article/pii/S1532046422002064>.
- [19] A. Kiourtis, A. Mavrogiorgou, A. Menychtas, I. Maglogiannis, D. Kyriazis, Structurally mapping healthcare data to HL7 FHIR through ontology alignment, *J. Med. Syst.* 43 (3) (2019) 62, <http://dx.doi.org/10.1007/s10916-019-1183-y>.
- [20] G. Lichtner, B.S. Alper, C. Jurth, C. Spies, M. Boeker, J.J. Meerpohl, F. von Dinkel, Representation of evidence-based clinical practice guideline recommendations on FHIR, *J. Biomed. Inform.* 139 (2023) 104305, <http://dx.doi.org/10.1016/j.jbi.2023.104305>, URL <https://www.sciencedirect.com/science/article/pii/S1532046423000266>.
- [21] A.M. Bennett, H. Ulrich, P. van Damme, J. Wiedekopf, A.E.W. Johnson, MIMIC-IV on FHIR: converting a decade of in-patient data into an exchangeable, interoperable format, *J. Am. Med. Inform. Assoc.* 30 (4) (2023) 718–725, <http://dx.doi.org/10.1093/jamia/ocad002>.
- [22] A.A. Sinaci, M. Gencturk, H.A. Teoman, G.B. Laleci Erturkmen, C. Alvarez-Romero, A. Martinez-Garcia, B. Poblador-Plou, J. Carmona-Pérez, M. Löbe, C.L. Parra-Calderon, A data transformation methodology to create findable, accessible, interoperable, and reusable health data: Software design, development, and evaluation study, *J. Med. Internet Res.* 25 (2023) e42822, <http://dx.doi.org/10.2196/42822>.
- [23] N. Ladas, S. Franz, B. Haarbrandt, K.K. Sommer, S. Kohler, S. Ballout, J. Fiebeck, M. Marschollek, M. Gietzelt, openEHR-to-FHIR: Converting openEHR compositions to fast healthcare interoperability resources (FHIR) for the german corona consensus dataset (GECCO), in: *Informatics and Technology in Clinical Care and Public Health*, IOS Press, 2022, pp. 485–486.
- [24] A. Soares, S. Essaid, M.G. Kahn, Omop-to-bulk fhir: A tool to convert population level clinical data into standardized FHIR batch data, 2023.
- [25] F. Simon, J. Schladezky, S. Macke, T. ABLAßa, J. Ingenerf, K.-S. Ann-Kristin, Metadata driven integration of clinical data for secondary use in FHIR-a pilot study at the UKSH, *Stud. Health Technol. Inform.* 317 (2024) 146–151.
- [26] V.A. Arcobelli, S. Moscato, A. Marfoglia, F. Nardini, P. Randi, A. Davalli, A. Carbonaro, P. Palumbo, L. Chiari, S. Mellone, MOTU on FHIR: A preliminary strategy to enable interoperability for retrospective dataset standardization, in: 2023 IEEE EMBS Special Topic Conference on Data Science and Engineering in Healthcare, Medicine and Biology, 2023, pp. 81–82, <http://dx.doi.org/10.1109/IEEECONF58974.2023.10404816>, URL <https://ieeexplore.ieee.org/document/10404816>.

- [27] M.J. Denney, D.M. Long, M.G. Armistead, J.L. Anderson, B.N. Conway, Validating the extract, transform, load process used to populate a large clinical research database, *Int. J. Med. Inform.* 94 (2016) 271–274, <http://dx.doi.org/10.1016/j.ijmedinf.2016.07.009>, URL <https://www.sciencedirect.com/science/article/pii/S1386505616301708>.
- [28] V.A. Arcobelli, S. Moscato, P. Palumbo, A. Marfoggia, F. Nardini, P. Randi, A. Davalli, A. Carbonaro, L. Chiari, S. Mellone, MOTU data. MOTU on FHIR: A 10-year data collection on the clinical rehabilitation pathway of 1006 trans-femoral amputees, 2024, <http://dx.doi.org/10.5281/zenodo.10683153>, URL <https://zenodo.org/records/10683153>.
- [29] M.G. Kahn, T.J. Callahan, J. Barnard, A.E. Bauck, J. Brown, B.N. Davidson, H. Estiri, C. Goerg, E. Holve, S.G. Johnson, S.-T. Liaw, M. Hamilton-Lopez, D. Meeker, T.C. Ong, P. Ryan, N. Shang, N.G. Weiskopf, C. Weng, M.N. Zozus, L. Schilling, A harmonized data quality assessment terminology and framework for the secondary use of electronic health record data, *EGEMs* 4 (1) (2016) 1244, <http://dx.doi.org/10.13063/2327-9214.1244>, URL <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5051581/>.
- [30] V.A. Arcobelli, S. Moscato, P. Palumbo, A. Marfoggia, F. Nardini, P. Randi, A. Davalli, A. Carbonaro, L. Chiari, S. Mellone, FHIR-standardized data collection on the clinical rehabilitation pathway of trans-femoral amputation patients, *Sci. Data* 11 (1) (2024) 806, <http://dx.doi.org/10.1038/s41597-024-03593-6>, Publisher: Nature Publishing Group. URL <https://www.nature.com/articles/s41597-024-03593-6>.
- [31] V.A. Arcobelli, S. Moscato, P. Palumbo, A. Marfoggia, F. Nardini, P. Randi, A. Davalli, A. Carbonaro, L. Chiari, S. Mellone, FHIR MOTU data. MOTU on FHIR: A 10-year data collection on the clinical rehabilitation pathway of 1006 trans-femoral amputees, 2024, <http://dx.doi.org/10.5281/zenodo.10684280>, URL <https://zenodo.org/records/10684280>.
- [32] N. Hong, A. Wen, F. Shen, S. Sohn, C. Wang, H. Liu, G. Jiang, Developing a scalable FHIR-based clinical data normalization pipeline for standardizing and integrating unstructured and structured electronic health record data, *JAMIA Open* 2 (4) (2019) 570–579, <http://dx.doi.org/10.1093/jamiaopen/ooz056>.