Opening the Personal Gate between Technology and Health Care
M. Giacomini et al. (Eds.)

doi:10.3233/SHTI260393

Toward Interoperable Variable Definitions: A FHIR-Based Standardization Strategy for the METASTRA Project

Serena Moscato[a,b] Alberto Marfoglia[c], Valerio Antonio Arcobelli[a], Maria Rita Intagliata[d], Cristiana Griffoni[d], Giovanni Barbanti-Bròdano[d], Alessandro Gasbarrini[d], Antonella Carbonaro[c], Sabato Mellone[a,b]
[a] *Department of Electrical, Electronic, and Information Engineering – Guglielmo Marconi (DEI), University of Bologna, Bologna, Italy*
[b] *Health Sciences and Technologies – Interdepartmental Center for Industrial Research (CIRI-SDV), University of Bologna, Bologna, Italy*
[c] *Department of Computer Science and Engineering, University of Bologna, Italy*
[d] *Department of Spine Surgery, IRCCS Istituto Ortopedico Rizzoli, Bologna, Italy*
ORCiD ID: SeM: https://orcid.org/0000-0002-0538-650X, AM: https://orcid.org/0009-0000-5857-2376, VAA: https://orcid.org/0000-0002-1262-9899, CG: https://orcid.org/0000-0001-8344-7534, GBB: https://orcid.org/0000-0002-7075-7020, AG: https://orcid.org/0000-0002-9575-4061, AC: https://orcid.org/0000-0002-3890-4852, SaM: https://orcid.org/0000-0001-7688-0188

Abstract. The integration of data from multicenter clinical studies represents a key opportunity to enhance research quality. This potential can be further enhanced by standardizing variables and ensuring their semantic interoperability. In this work, we present the approach adopted, along with the preliminary results, to standardize a retrospective multicenter dataset collected within the METASTRA project, an EU H2022 initiative aimed at developing personalized strategies for patients with vertebral metastases. The dataset comprises 401 variables collected through electronic case report forms across four clinical centers. The proposed standardization strategy relies on mapping each variable to the most suitable HL7 FHIR resource and field, complemented by the use of SNOMED CT terminology. A modular transformation pipeline was applied to convert the raw data into FHIR resources. In this preliminary phase, we focused on a subset of 99 variables. Among these, 88% (87/99) were successfully standardized using nine FHIR resources and 177 SNOMED CT concepts. Validation queries confirmed full consistency between the original and standardized datasets, demonstrating the reliability of the process. This work contributes to creating a semantically coherent clinical knowledge base, enabling more effective data reuse and supporting evidence generation in multicenter clinical studies.

Keywords. FHIR, SNOMED-CT, Interoperability.

1. Introduction

Multicenter clinical studies offer the advantage of combining data from diverse sources, thereby increasing the representativeness of the study population and the heterogeneity of patterns that can be analyzed. The potential of such studies can be

further enhanced through harmonization and standardization of the collected variables, with the dual objective of ensuring semantic alignment across centers and enabling more interoperable and shareable datasets [1].

Interoperability requires complementary standards at different levels. At the structural level, for instance, HL7 Fast Healthcare Interoperability Resources (FHIR) [2] is an open standard for the exchange and representation of healthcare data [3]. At the semantic level, controlled vocabularies such as SNOMED CT provide consistent and, together with the structural level, both can enable machine-readable representations of clinical concepts and laboratory measurements, providing alignment by standardizing the meaning of clinical concepts.

In this paper, we present the data management strategy applied to a retrospective multicenter clinical study conducted involving four different clinical centers, along with the preliminary results on the standardization approach applied using HL7 FHIR as a standard framework and SNOMED CT as vocabulary.

2. Materials and Methods

2.1. Dataset

This work is part of the METASTRA project, aiming to develop personalized strategies for the treatment and care of cancer patients with bone metastases, through the harmonized integration of clinical and physiological knowledge from 15 partner institutions across EU.

The aim of the retrospective multicenter clinical study (Train-METASTRA) is to collect clinical information on patients with spine metastatic lesions to develop automatic algorithms to predict vertebra fractures, that will be validated with a prospective multicenter clinical study (Validate-METASTRA).

Each clinical center is asked to input data from the eligible patients through an electronic case report form (eCRF) developed in REDCap. The eCRF consists of 401 variables, each one labelled with a unique name defined by taking inspiration from the HL7 FHIR resource to which the variable will be mapped during standardization. The variables are collected through four different eCRF forms:

- First overview: a patient-centered form in which demographic and clinical information are collected;
- History of lesion: a spine metastatic lesion-centered form, collecting information regarding histological examination, stability, medical images, and the last follow up about the single lesion;
- Treatment of lesion: a spine metastatic lesion-centered form, collecting information regarding the therapeutic strategies each vertebra presenting a lesion underwent to;
- Last overview: a patient-centered form in which the last clinically relevant information available about the patient is inputted.

Here, we focused on the “First Overview” eCRF form, consisting in 99 variables and falling in the following domains:

- First visit information – 6 variables
- Demographic data – 5 variables
- Substance use – 3 variables
- Clinical assessment – 33 variables

- Blood test – 28 variables
- Medical therapy – 9 variables
- Oncological assessment – 15 variables

2.2. Standardization

This section describes the strategy adopted to standardize the collected data and the steps of a modular transformation pipeline [4], [5] used to produce and validate the FHIR version of the dataset.

- Data extraction and variable standardization
  - eCRFs were exported from the REDCap API in csv format
  - Each eCRF variable was mapped by two researchers to the most appropriate HL7 FHIR resource and specific field
  - SNOMED-CT codes were assigned to represent the corresponding clinical concepts
- Standardization pipeline steps
  - Configuration setup: upload of StructureDefinitions and StructureMaps defining METASTRA data structure and transformation rules
  - Data preprocessing: to ensure the quality and consistency of the raw CSV data
  - Data transformation: cleaned data are converted into JSON objects through field renaming, typing, and filtering
  - Mapping: Matchbox transforms JSON objects into FHIR Transactions, with syntactic validation ensuring schema correctness
  - Upload: Submissoin of FHIR Bundles to the HAPI FHIR server using temporary UUIDs to preserve internal links
  - Transaction application: integrity checks, semantic validation, and rollback on failure to ensure consistency
  - Export: release of validated FHIR resources in NDJSON format via the FHIR Bulk data Access protocol
- Consistency evaluation
  - Clinically meaningful validation queries were defined by clinical experts (M.R.I., C.G., G.B.B.).
  - Queries were executed on both CSV and FHIR datasets using Python.

3. Results

We preliminarily applied the standardization approach to the “First overview” data coming from 328 patients. Figure 1 shows a schematic representation of FHIR resources of the standardized dataset.

From the initial set of 99 variables of the “First Overview” form, we successfully mapped 88% (87 out of 99 variables) of them in 9 different FHIR resources, using a total of 177 SNOMED CT concepts, used to map not only variable definition, but also the response option to a single variable. Unmapped variables belong to the “Oncological assessment” domain, for which we have not found yet a proper standardization form for 12 out of 15 variables.

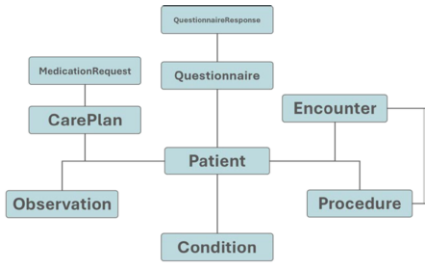

Figure 1 FHIR resources schema

Table 1 presents the number of variables mapped into different 9 FHIR resources, along with the number of instances for each resource.

Table 1 Mapping between FHIR resources and variables, and the number of instances for each resource

FHIR Resource	METASTRA variables	METASTRA instances
Encounter	2	666
Observation	26	2167
Condition	6	721
Questionnaire + QuestionnaireResponse	18	666
CarePlan + MedicationRequest	8	333
Patient	3	328
Procedure	3	15

Table 2 provides the number of instances found in response to a specific validation query in both the CSV and FHIR dataset. We found the same number of instances in both datasets, indicating that the standardization process did not result in any data loss.

Table 2. Validation queries

Query	METASTRA variables	METASTRA instances
Male smokers over 50 years with BMI over 25	9	9
Male under 50 years with prostate and neurological impaired bowel	0	0
Female under 50 years with breast cancer	23	23
Female over 50 years with breast cancer	81	81

4. Discussion

In this paper, we presented preliminary results regarding the standardization of a retrospective multicenter clinical study conducted within the METASTRA project involving patients with spine metastatic lesions.

As a preliminary result, we presented the standardization strategy for one out of the four forms collected during the retrospective study, successfully standardizing 88% of the variables. The remaining 12% of variables are still under discussion due to their domain-specific nature, particularly those related to concepts such as tumor markers and subtypes. These variables capture clinically complex concepts such as treatment intent, treatment modality and sequencing, and local disease courses. In HL7 FHIR, such information is distributed across multiple resources and often depends on contextual interpretation rather than explicit standardized elements, making an unambiguous mapping infeasible without loss of semantic fidelity. Additionally, the clinicians

involved in the study actively contributed to the development of the eCRF, thus providing the semantic foundation of the dataset, which guided the FHIR standardization.

The same conversion pipeline was previously applied to the MOTU dataset, a publicly-available clinical datasets on individuals with prosthetic knees prescription [6]. Comparison between the FHIR-based standardization of MOTU and METASTRA reveals complementary, domain-specific patterns: MOTU emphasizes functional assessments typical of orthopedics, whereas METASTRA focuses on diagnostic and disease-state resources aligned with oncology workflows. Despite these differences, both conversions employ the same FHIR core resources (Patient, Encounter, Observation, Condition), demonstrating sufficient flexibility to represent distinct clinical domains. Already widely used conversion tools like Mirth Connect, Cloverleaf, and Interfaceware Iguana, provide frameworks supported by extensive communities. Complementary template-based approaches such as Liquid offer lightweight, template-driven mechanisms, for example via the Microsoft FHIR Converter, to produce validated FHIR resources. Together, these tools enable workflow automation for healthcare standards; however, adoption can be hindered by upfront licensing, and the risk of vendor lock-in.

Beyond generating a standardized, machine-readable, and semantically enriched dataset suitable for analyses, this study validates the standardization strategy and demonstrates its adaptability across clinical research settings positioning it as an open, research-oriented solution, setting it apart from existing frameworks.

Acknowledgement

The research presented in this paper was conducted as part of several projects: METASTRA project, funded by EU H2022 program, grant ID 101080135); DARE - DigitAl lifelong pRevEntion, 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, code PNC0000002, CUP: B53C22006450001; MOTU++ project, funded by the Italian National Institute for Insurance against Accidents at Work (INAIL), grant ID PR19-PAI-P2.

References

[1] M. Pfeffer, M. Deneris, A. Shelley, P. Salcuni, and I. Altomare, “Utility of automated data transfer for cancer clinical trials and considerations for implementation,” *ESMO Real World Data Digit. Oncol.*, vol. 7, p. 100112, Mar. 2025, doi: 10.1016/j.esmorw.2025.100112.

[2] HL7 International, “HL7 FHIR Release 5.” https://www.hl7.org/fhir/ (accessed Oct. 13, 2025).

[3] R. Gazzarata *et al.*, “HL7 Fast Healthcare Interoperability Resources (HL7 FHIR) in digital healthcare ecosystems for chronic disease management: Scoping review,” *Int. J. Med. Inform.*, vol. 189, p. 105507, Sep. 2024, doi: 10.1016/j.ijmedinf.2024.105507.

[4] A. Marfoglia, F. Nardini, V. A. Arcobelli, S. Moscato, S. Mellone, and A. Carbonaro, “Towards real-world clinical data standardization: A modular FHIR-driven transformation pipeline to enhance semantic interoperability in healthcare,” *Comput. Biol. Med.*, vol. 187, p. 109745, Mar. 2025, doi: 10.1016/j.compbiomed.2025.109745.

[5] A. Carbonaro *et al.*, “From raw data to research-ready: A FHIR-based transformation pipeline in a real-world oncology setting,” *Comput. Biol. Med.*, vol. 197, p. 111051, Oct. 2025, doi: 10.1016/j.compbiomed.2025.111051.

[6] V. A. Arcobelli *et al.*, “FHIR-standardized data collection on the clinical rehabilitation pathway of trans-femoral amputation patients,” *Sci. Data*, vol. 11, no. 1, p. 806, Jul. 2024, doi: 10.1038/s41597-024-03593-6.