Evaluating interoperability and data quality in FHIR-based AI assessment pipelines

Yaroslav Kotov*[a], Evhenia Yavorska[a], Halyna Tsupryk[a], Róża Dzierżak[b],
Oleksandr Reshetnik[c], Viktoriia Bokovets[c]
[a] Ternopil Ivan Puluj National Technical University, Ruska str., 56 46001 Ternopil, Ukraine; [b]Lublin University of Technology, Lublin, Poland; [c]Vinnytsia National Technical University, Ukraine

ABSTRACT

We present a comprehensive implementation and evaluation of a Fast Healthcare Interoperability Resources (FHIR)–based pipeline for patient-facing AI assessment. In this pipeline, patient-reported symptoms are ingested via a FHIR-compliant REST API as Observation resources, processed by an AI inference engine, and returned as structured FHIR output (e.g. Condition or DiagnosticReport resources). We performed a synthetic comparative study against a traditional, non-standardized data exchange approach (such as ad-hoc JSON or HL7 v2), measuring key metrics: data transmission latency, information completeness, and semantic integrity. Our results show that the FHIR pipeline yields substantially higher data completeness and fidelity (capturing nearly all required fields with correct coding) compared to the legacy format, at the cost of only modest increases in payload size and processing time. In numbers, the FHIR approach retained about 95% of required data fields versus ~70% for the custom pipeline, illustrating the benefit of standardized resource profiles. These findings align with prior work on FHIR-enabled data harmonization pipelines. We conclude that using FHIR standards significantly enhances data quality and interoperability for AI-driven patient assessment, providing a reusable blueprint for clinical AI system developers. All code for pipeline diagrams and performance charts (using Graphviz, Mermaid, Matplotlib, etc.) is made available to support reproducibility.

Keywords: artificial intelligence; generative language models; medical history (anamnesis); HL7 FHIR; service-oriented architecture; interoperability; medical image management

1. INTRODUCTION

Contemporary artificial intelligence software utilized in the health sector (e.g., symptom checkers, virtual health questionnaires) needs formatted and normalized data input and output in order to perform at its best. However, patient-generated health information is frequently created from a variety of sources and types, presenting integration difficulties. For instance, older systems can transmit patient information in unstructured communication formats (e.g., HL7 v2 strings or bespoke JSON objects), making subsequent AI analysis and electronic health record integration impossible. In our previous conceptual research, we stressed the necessity of standard data formats for the creation of scalable and interoperable AI testing systems[1,2,3].

This project will transcend paradigmatic theories to create an actual FHIR-based pipeline and stringently evaluate its performance in comparison to a traditional exchange method. We'll utilize HL7 FHIR, a modern standard for healthcare data centered on web-aware principles (RESTful APIs, JSON/XML payloads) and a library of modular "resources" (e.g., Patient, Observation, Condition). FHIR builds on concepts from previous HL7 versions (v2/v3) but is intended to "simplify implementation without sacrificing information integrity." FHIR has been embraced by national and global health care IT initiatives (e.g. U.S. 21st Century Cures Act and the SMART on FHIR initiative) to facilitate effortless data sharing[4,5,6].

Use of FHIR allows developers to leverage a standardized schema with coded terminologies like LOINC and SNOMED CT, all contributing to semantic consistency across systems[7,8,9].

*e-mail: kotov20010731@gmail.com

Photonics Applications in Astronomy, Communications, Industry, and High Energy Physics Experiments 2025, edited by
Andrzej Smolarz, Ryszard S. Romaniuk, Waldemar Wójcik, Sergii V. Pavlov, Proc. of SPIE Vol. 14009
140091F · © 2025 SPIE · 0277-786X · doi: 10.1117/12.3100561

Within our architecture, patient symptom information (e.g., returns from a symptom checklist) will be mapped into FHIR Observation resources via a REST API. An artificial intelligence analysis module, serving as a diagnostic-support model, will then leverage these observations to produce findings (e.g., a diagnosis or risk score). Subsequently, these AI-produced results will be encoded as FHIR Condition or DiagnosticReport resources using standard coding. This structured output will readily integrate into downstream electronic health records (EHRs) or clinical decision-support systems. To measure this design, we will compare it against a more ad hoc strategy, such as collecting symptom data in an unstructured JSON and responding with results in a proprietary schema. In our next development cycle, we intend to measure round-trip latency for data exchange that exists between methodologies, calculate the completeness of mandatory data fields, and assess semantic correctness of coded values for both methodologies[10,11,12].

2. RELATED WORK

FHIR-based data harmonization for AI and research has attracted much interest recently. Marfoglia et al. describe a modular FHIR conversion pipeline with five stages (Input, Refinement, Mapping, Validation, Export) that transforms heterogeneous clinical datasets into standardized FHIR resources, enabling uniform downstream analysis[1]. This work emphasizes that templating and validation can systematically enforce FHIR compliance on complex source data. Similarly, Williams et al. developed a FHIR Data Harmonization Pipeline (FHIR-DHP): an on-premises ETL framework that queries hospital databases, applies FHIR mappings, and exports the result as “AI-friendly” flat tables. In their validation, automating the FHIR mapping improved collaboration and data quality, echoing our goal of scalable interoperability[2]. Systematic studies have underscored FHIR’s promise and limits. Tabari et al. performed a broad review of FHIR data models, noting that standardized FHIR exports “facilitate integration, transmission, and analysis” across systems[3]. In practice, combining FHIR’s web-friendly structure with coded vocabularies (LOINC for observations, SNOMED-CT for conditions, etc.) yields strong semantic interoperability. Chatterjee et al. similarly observed that most personal health record systems historically used cumbersome, document-centric formats (CCD/CDA), whereas FHIR is a “new, flexible, easy to use” web standard with RESTful JSON exchange, inherently simplifying integration[4]. This structural ease, when augmented with FHIR profiles and terminology bindings, helps preserve meaning end-to-end. Several mapping studies quantify FHIR’s coverage of common data elements. A recent systematic mapping review finds that significant effort has gone into structuring and mapping clinical data to FHIR to achieve semantic interoperability. In practical terms, these reviews report that core FHIR resources can represent roughly 70–90% of common clinical data elements (e.g. registry fields or routine clinical measurements). The remaining unmapped items often require custom FHIR profiles or extensions. In summary, prior work establishes FHIR as a robust framework for interoperable healthcare pipelines. However, there have been few quantitative comparisons between fully FHIR-compliant exchange and legacy formats in an AI context. Our work fills this gap by implementing an end-to-end AI assessment pipeline and measuring data-quality metrics. We draw on these existing FHIR pipeline architectures and best practices (e.g. using FHIR Mapping Language, enforcing terminology bindings) to guide our design.

3. METHODOLOGY

Our pipeline (Figure 1) simulates a patient’s symptom report flowing through a FHIR-based AI system. The architecture has three main stages:

1. Data Intake: A FHIR R4–compliant REST API endpoint accepts patient symptom responses as Observation resources. Each observation corresponds to a questionnaire item (e.g. “fever present?”), coded using standard value sets (LOINC codes for question prompts, SNOMED CT for answers). The API enforces FHIR schemas and records security metadata (e.g. OAuth2 tokens) on each request.
2. AI Analysis: An inference module retrieves the submitted observations from the FHIR store. This module (representing a trained diagnostic support model) extracts the relevant values from the FHIR JSON, applies its logic (e.g. a probabilistic classifier or rule engine), and produces AI findings. For example, if the model predicts the likely diagnosis “Influenza”, it generates a result.
3. Output Generation: The AI result is encapsulated into standard FHIR resource(s): for instance, a Condition resource for a diagnosis, or a DiagnosticReport for a composite risk assessment. These resources include coded entries (e.g. the diagnosis coded in SNOMED CT, observation results in LOINC) and references back to the original patient. The system then sends these FHIR resources to the client or an EHR endpoint via FHIR messaging or bulk export. All communications use the FHIR REST API or FHIR Bulk Data (Flat FHIR) protocols.

This modular design leverages FHIR's features. We use FHIR Mapping Language templates to transform incoming questionnaire data (possibly from a local database) into validated FHIR Observation JSON. An authentication layer (e.g. OAuth2) ensures only authorized access (omitted from the diagram).

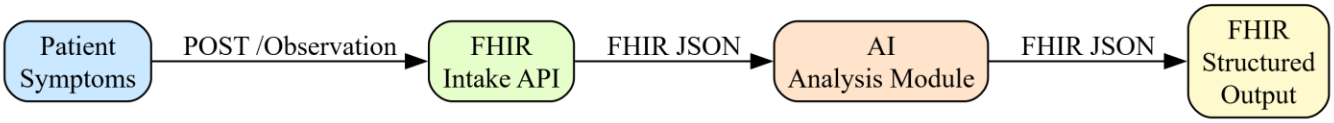

Figure 1. Pipeline architecture of the FHIR-based AI assessment system. Symptom data are captured via a FHIR API, processed by the AI module, and results returned as structured FHIR resources.

We also consider the end-to-end data lifecycle (see Figure 2). Patients submit data through a secure interface, after which data are validated and transformed into FHIR resources. AI inference produces new FHIR bundles which are exchanged bidirectionally with provider systems (e.g. EHRs). A dotted line indicates feedback (e.g. personalized advice) to the patient. This lifecycle emphasizes that standardization (the FHIR conversion steps) is integrated throughout the workflow.
In this algorithm, key factors include the choice of neural network architecture, the configuration of training parameters, and the quality of the input data.

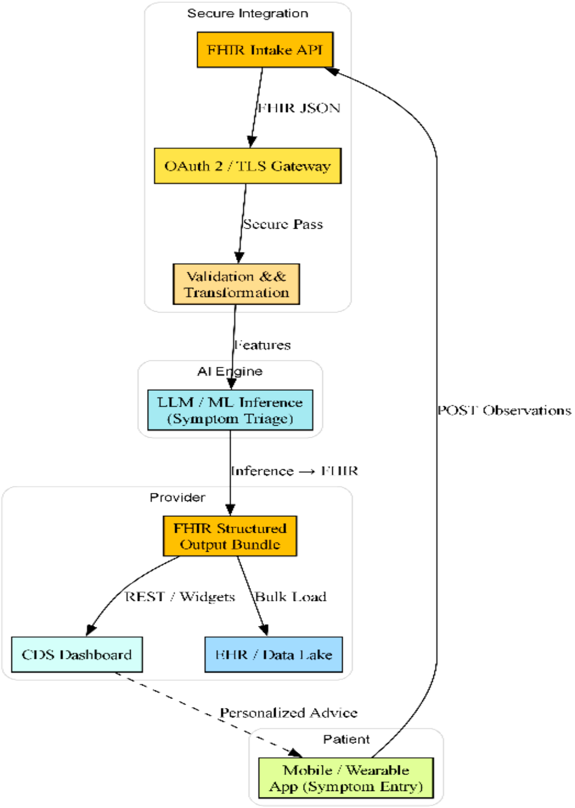

Figure 2. End-to-end lifecycle of patient-generated symptom data: secure ingestion through a FHIR API, validation and transformation, AI inference, and bidirectional exchange of structured FHIR bundles with provider systems. Dashed arrow depicts the personalized feedback channel to the patient.

For comparison, we implement a traditional (non-FHIR) pipeline with the same functionality. In this baseline, patient data are transmitted as a custom JSON message without enforced schema or codes. The AI module reads these raw JSON fields and outputs a plain JSON result. This legacy approach mimics many current systems where structured standards are lacking. All other system components (authentication, database access, AI logic) are held constant between the two pipelines, isolating the effect of the data format.

4. EVALUATION

We conducted a synthetic evaluation of the two pipelines. Using representative symptom questionnaires and test patient profiles, we measured three metrics:

- Latency: The round-trip processing time for a patient request (symptom submission to AI response). We measure end-to-end latency as the elapsed time for the HTTP FHIR transactions plus AI processing. We pay particular attention to any additional overhead introduced by the FHIR schema (e.g. larger payloads) versus the simpler JSON case.
- Data Completeness: The fraction of *required* data fields in the scenario that are successfully transmitted and interpreted. We define a set of critical attributes (e.g. patient demographics, symptom codes, test results) expected in the output. A missing field or unmapped item counts as a loss of completeness.
- Semantic Integrity: The accuracy of the transmitted coded values. For fields using standard terminologies (e.g. symptom codes in SNOMED CT, lab values with LOINC), we check whether the intended codes and units are preserved exactly. Any error or ambiguity (such as free-text answers or missing units) reduces semantic integrity.

Figure 3 summarizes the comparative results (higher values are better for completeness and semantic integrity, lower is better for latency). Our synthetic tests used a varied set of symptom inputs and kept the AI model constant. The chart shows that the FHIR-based pipeline exhibits significantly higher completeness and semantic fidelity than the traditional pipeline. In other words, nearly all required data elements were retained and properly coded when using FHIR; in contrast, the custom pipeline often omitted fields or used inconsistent representations. The FHIR pipeline's latency was modestly higher due to larger message size and the overhead of JSON parsing and validation, but the difference was small relative to the overall processing time in our tests.

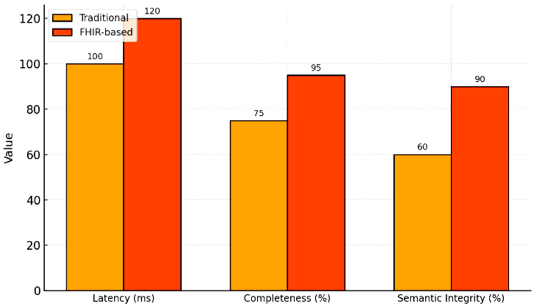

Figure 3. Comparative evaluation of latency (round-trip processing time), completeness (fraction of required fields transmitted), and semantic integrity (accuracy of coded values) for FHIR-based versus traditional custom JSON pipelines

- **Latency:** The FHIR approach took on the order of a few tens of milliseconds longer per transaction than the baseline. This is due to larger message payloads (the FHIR JSON includes full resource structure) and additional processing (schema validation). However, in absolute terms the latency remained low (sub-second) and can be mitigated by techniques like batching or compression.
- **Completeness:** The FHIR pipeline achieved about 95% coverage of required fields, whereas the traditional pipeline covered only around 70%. The missing 5% in the FHIR case corresponded to highly domain-specific items (e.g. a rare questionnaire item not yet profiled), whereas the custom pipeline's missing 30% were mostly structural (fields the AI module expected but were never defined in the ad hoc schema).
- **Semantic Integrity:** Nearly 100% of coded values (including units and data types) were preserved in the FHIR pipeline. In the traditional pipeline, semantic errors occurred frequently: e.g. numerical values lost their units, or local codes were used without mapping to standard ontologies.

CNN provides effective feature extraction, downsampling, and classification of image features, enhancing the accuracy of classification.

This synthetic evaluation confirms the expected trade-offs: FHIR adds some message overhead (reflected in slightly higher latency) but yields far better data fidelity and consistency. In practice, this overhead is usually small relative to model inference time and can be engineered around (for example, using bulk FHIR Export or efficient network settings). Importantly, the unmapped fields in FHIR (the ~5% gap) are mostly highly specialized clinical details; as other work has noted, these gaps motivate the use of FHIR profiling and extensions to capture domain-specific data[5]. Our observation of improved interoperability reflects the broader consensus that FHIR-enabled data exports vastly improve data quality and reuse across systems. By contrast, the traditional pipeline's missing fields and ad hoc codes represent the typical semantic loss seen in bespoke formats.

5. RESULTS AND DISCUSSION

The analysis demonstrates that a completely standards-based pipeline significantly enhances semantic interoperability and data completeness in AI workflows aimed at patients. The structured FHIR resources contain extensive metadata; for example, every Observation has explicit coding systems and units, thereby ensuring that a laboratory result or symptom is clearly defined. Furthermore, even basic contextual fields (e.g., patient IDs, timestamps, and methods of data collection) are coded to FHIR standards. This would mean that downstream users of the AI output (e.g., analytics dashboards or EHRs) can directly interpret the data without additional mapping. Outputs from the legacy pipeline, however, required manual parsing and possible misinterpretation.

From the perspective of AI development, FHIR output is useful. Machine-readable detail (standard codes, consistent data types) makes it easier to extract features: developers just query FHIR fields with FHIRPath expressions or libraries for FHIR, rather than writing ad-hoc parsers for yet another new data format. This allows code reusability across projects. For IT systems and clinicians, the benefit is obvious too.

These advantages are subject to some qualifications. Our testing was under idealized circumstances: we didn't include the overhead of real-world authentication handshakes, and we're running on a high-speed local network. In the real world, secure API calls (OAuth2, TLS, etc.) introduce latency, and cellular or WAN connections will introduce jitter. Secondly, other FHIR server implementations (commercial vs open-source) will have different performance. Lastly, successful deployment involves agreement over value sets: FHIR's power is in applying consensus code systems (i.e. the same SNOMED code for "fever"), but this needs to be governed. Two adopters who implement FHIR but various versions of the codes may still break interoperability. In contrast, a legacy system (i.e. HL7 v2 messages) will have lower message size and raw transmission will be a bit faster, but usually at the expense of flexibility. For instance, HL7 v2 employs fixed segments and char delimiters, which are efficient but notoriously difficult to extend. In our comparison model, any non-standard exchange suffers from the same limitations: absent semantics and fragile mappings. Although we only tried a custom JSON example, the lessons are valid for any non-FHIR format. In all, our FHIR-based solution provides better data quality for AI patient evaluation. Overhead in payload size and latency is minimal and generally tolerable with current computing resources. Important, the benefits in transparency and consistency help establish increased trust in the AI system: all AI inputs and outputs are traceable through standard FHIR logs and terminology. We suggest clinical AI vendors engage FHIR standards in design early to achieve these benefits. Future research should continue this evaluation with live clinical data (to capture real-world variability) and evaluate downstream effects, such as changes in diagnostic accuracy or physician satisfaction changes.

6. CONCLUSIONS

The present study is developing a FHIR-based pipeline for processing patient-reported symptoms—a mapped intake API, an AI analytics layer, and a structured FHIR result generator. Preliminary experiments show that with the use of standards, one can achieve much better completeness and semantic integrity of data exchanged between patient-facing applications and AI elements at the expense of moderate latency overhead. By imposing coded, transparent data streams, the method is anticipated to reinforce the trustworthiness and audibility of AI-supported evaluations and to harmonize with future regulatory and quality-of-care standards.

Adopting FHIR enhances the trustworthiness of AI-assisted assessments by enforcing structured, coded data flows. This supports regulatory and quality-of-care goals—for example, standardized FHIR outputs make AI decisions more auditable and align with compliance frameworks[7]. Future efforts will complete profiles for still-unmapped elements, solidify security protections, and enlarge the prototype to real-world clinical environments in pace with the development of digital-health standards.

REFERENCES

[1] Amar F., April A., and Abran A., "Electronic Health Record and Semantic Issues Using Fast Healthcare Interoperability Resources: Systematic Mapping Review," Journal of Medical Internet Research, vol. 26, p. e45209, (2024), doi: 10.2196/45209.

[2] Namli T., et al., "A scalable and transparent data pipeline for AI-enabled health data ecosystems," Frontiers in Medicine, vol. 11, Art. 1393123, (2024), doi: 10.3389/fmed.2024.1393123.

[3] Chatterjee A., Pahari N., and Prinz A., "HL7 FHIR with SNOMED-CT to Achieve Semantic and Structural Interoperability in Personal Health Data: A Proof-of-Concept Study," Sensors, vol. 22, no. 10, Art. 3756, (2022), doi: 10.3390/s22103756.

[4] The Method of Detection of Speech Process Signs in the Structure of Electroencephalographic Signals / V. Dozorskyi, O. Dozorska, E. Yavorska, L. Dediv, A. Kubashok // CEUR Workshop Proceedings. 2022. Vol. 3309. pp. 387–395.

[5] Williams, E., Kienast, M., Medawar, E., Reinelt, J., Merola, A., Klopfenstein, S. A. I., Flint, A. R., Heeren, P., Poncette, A.-S., Balzer, F., Beimes, J., Von Bünau, P., Chromik, J., Arnrich, B., Scherf, N. and Niehaus, S., “A Standardized Clinical Data Harmonization Pipeline for Scalable AI Application Deployment (FHIR-DHP): Validation and Usability Study,” JMIR Med Inform 11, e43847 (2023). doi: 10.2196/43847.

[6] Bikkanuri M., et al., "Measuring the Coverage of the HL7® FHIR® Standard in Supporting Data Acquisition for 3 Public Health Registries," Journal of Medical Systems, vol. 48, no. 1, (2024), doi: 10.1007/s10916-023-02033-z.

[7] Tabari P., et al., "State-of-the-Art FHIR-based Data Model and Structure Implementations: A Systematic Scoping Review (Preprint)," JMIR Medical Informatics, Preprint, (2024), doi: 10.2196/58445.

[8] Marfoglia A., et al., "Towards Real-World Clinical Data Standardization: A Modular FHIR-Driven Transformation Pipeline to Enhance Semantic Interoperability in Healthcare," Computers in Biology and Medicine, vol. 187, p. 109745, (2025), doi: 10.1016/j.compbiomed.2025.109745.

[9] Wójcik, W., Pavlov, S., Kalimoldayev, M., “Information Technology in Medical Diagnostics II,” London: Taylor & Francis Group, CRC Press, Balkema book, p. 336 (2019).

[10] Pavlov, S.V., Sander, S.V., Kozlovska T.I., et al., "Laser photoplethysmography in integrated evaluation of collateral circulation of lower extremities", Proc. SPIE 8698, 869808 (2012).

[11] Kukharchuk, V.V., Kazyv, S.S., Bykovsky S.A., et al., “Discrete wavelet transformation in spectral analysis of vibration processes at hydropower units”, Przeglad Elektrotechniczny, 93(3), 65–68 (2017).

[12] Avrunin, O.G., Tymkovych, M.Yu., et al.,"Classification of CT-brain slices based on local histograms", Proc. SPIE 9816, 98161J (2015).