Sarcouncil Journal of Engineering and Computer Sciences

ISSN(Online): 2945-3585

Volume- 05| Issue- 03| 2026

Research Article **Received:** 20-01-2026| **Accepted:** 17-02-2026 | **Published:** 06-03-2026

Big Data Pipelines for HL7/FHIR Interoperability

Bhanuvardhan Nune

JNTU Kakinada, Andhra Pradesh, India

Abstract: The increasing digitization of the healthcare industry and the need for interoperability have revealed the necessity of scalable data integration systems. HL7 and FHIR standards have been central throughout the process of facilitating semantically enriched and structured data interchange across different health information systems. However, the emergence of Big Data in healthcare introduces new concerns in terms of volume, velocity, and variety and demands the implementation of more sophisticated data pipelines.This paper provides a summary of the architecture and implementation of Big Data pipelines specifically adapted for HL7/FHIR interoperability, including the contributions of AI, modular design, federated systems, and open-source tools to ensure semantic consistency and operational scalability. Oncology case studies demonstrate the advantages of these pipelines in improving the quality, security, and usability of data in clinical and research environments. Compared to clinical and research case studies, real-world oncology hospital systems further demonstrate the benefits of these pipelines. The paper also presents examples of integrating FHIR with data lakes, CDA conversion processes, and AI-driven interoperability frameworks of the future, providing general insight into the current state of current implementations and their further evolution.

Keywords: FHIR, HL7, Interoperability, Big Data Pipelines.

INTRODUCTION

Health systems have been computerized, and this has come with both opportunities and challenges. Interoperability among various health information systems with different types of data and standards is one of the most pressing requirements, since the former commonly utilizes the latter. HL7 (Health Level Seven) and FHIR (Fast Healthcare Interoperability Resources) implementations have become more or less standard, and they are designed to enable interoperability between dissimilar healthcare IT systems. The amount of data involved in healthcare and its complexity in recent years have necessitated the use of Big Data technologies in these interoperability efforts. As such, HL7/FHIR interoperability pipelines have played a pivotal role in ensuring scalable, efficient, and accurate health information exchange.

This paper focuses on the current trends and developments in Big Data pipelines to increase interoperability in accordance with HL7 and FHIR standards. It discusses the interaction between these technologies to address semantic, structural, and syntactic anomalies in health data, especially in relation to real-world applications and system designs that are consistent with present data engineering principles.

Understanding HL7 and FHIR Interoperability Standards

The HL7 family of standards has been a key component in healthcare data exchange since the 1980s. The issues with HL7 V2/V3, even though they were popular, included syntactic ambiguity and a lack of semantic clarity. FHIR was designed to cope with these challenges using a modular and web-friendly framework reliant on RESTful APIs, with data formats such as JSON/XML and standardized data models (Osamika, D.). Not only can FHIR be adapted to modern web development models with much greater ease, but it also provides semantically rich interoperability to ensure that meaningful information exchange is promoted instead of simple message relay.

FHIR has a design that enables various types of health applications, including electronic health records (EHR), personal health records, mobile health apps, and health data analytics platforms. The transition from HL7 V2/V3 to FHIR signifies a greater paradigm shift within the healthcare interoperability ecosystem and highlights the necessity of moving toward scalable and flexible architectures such as Big Data pipelines.

Role of AI in FHIR-Based Data Standardization

Clinical data transformation and standardization into FHIR formats are becoming more automated with the help of Large Language Models (LLMs) and Artificial Intelligence (AI). These tools, in particular, can be helpful in converting unstructured clinical narratives into structured FHIR resources. The recent adoption of transformer-based LLMs has demonstrated a high level of accuracy in enhancing clinical documentation and translating it into FHIR-compliant information, thereby reducing human

Copyright © 2022 The Author(s): This work is licensed under a Creative Commons Attribution- NonCommercial-NoDerivatives 4.0 (CC BY-NC-ND 4.0) International License

***Corresponding Author:** Bhanuvardhan Nune

DOI- https://doi.org/10.5281/zenodo.19166059

effort and potential errors (Riquelme, A. *et al.*, 2025).

The semantic interoperability gap can also be bridged using AI-assisted data standardization, as it can reconcile discrepancies among different coding systems, medical terminologies, and data structures. These systems can ascertain the context and meaning of clinical text that cannot be readily interpreted using rule-based or fixed mapping systems. This integration significantly accelerates the transformation of heterogeneous data sources into standardized, machine-readable formats required for downstream analytics and decision-making processes.

Modular FHIR-Driven Transformation Pipelines

The real-world healthcare setting is characterized by the presence of legacy systems, various data formats, and semantic inconsistencies. Modular transformation pipelines offer a solution here since they enable the decomposition of complex interoperability tasks into manageable steps. The overall structure of a FHIR-based pipeline typically includes data ingestion, preprocessing, standardization, validation, and finally integration with target systems (Marfoglia, A. *et al.*, 2025).

Modular pipelines enhance the flexibility of data processing processes and their scalability, enabling organizations to satisfy emerging standards and technologies without necessarily redesigning their current systems. They are commonly used in conjunction with distributed processing engines such as Apache Spark or Flink to process the high volume, variety, and velocity of medical information. FHIR profile validation tools, ontology mapping, and code translation services (e.g., SNOMED CT to LOINC) assist in achieving semantic interoperability.

The modular framework allows validation, enrichment, or anonymization modules to be added without affecting the remaining parts of the pipeline. This design also makes monitoring and logging easier, which is important for ensuring compliance with health data regulations.

Case Study: Oncology Data Transformation with FHIR

The necessity of standardization is even higher in oncology, where multiple modes of data are generated by diagnostics, therapies, and post-treatment follow-up. A FHIR-based pipeline implementation in an actual oncology setting was conducted to convert raw data from different sources into datasets that can be utilized in research. As shown in the case study, FHIR resources could be used to document end-to-end clinical processes such as patient encounters, lab results, radiology reports, and treatment plans (Carbonaro, A. *et al.*, 2025).

Distinct oncology terminologies and data models were incorporated into the pipeline through domain-specific FHIR profiles. In addition, the system used automated data validation schemes and human-in-the-loop data review to ensure data accuracy and clinical relevance.

The success of these implementations indicates the maturity of FHIR-based interoperability solutions, as well as their flexibility in addressing domain-specific requirements. The ability to integrate structured and unstructured data and the capacity to scale in processing make these pipelines a significant component of modern clinical informatics.

Table 1: Key Components of a Modular FHIR-Based Big Data Pipeline

Component	Description
Data Ingestion	Collects data from EHRs, medical devices, labs, etc.
Preprocessing	Cleanses and normalizes raw data formats
FHIR Mapping	Transforms data into FHIR resource structures
Semantic Alignment	Ensures consistency in medical terminology and coding systems
Validation	Applies FHIR schema checks and rule-based validation
Integration Layer	Connects with clinical apps, research platforms, or decision-support tools
Monitoring & Logging	Captures metrics for performance, compliance, and traceability

Integrating AI with Data Security and FHIR

Even though AI can be used to enhance automation in FHIR-based pipelines, the technology raises new concerns regarding data governance and security. Patient privacy and system integrity are important concerns in AI-enhanced health data systems. Advanced encryption, access control schemes, and anonymization are actively introduced into pipeline architectures in order to mitigate these risks (Oke, F. *et al.*,).

Copyright © 2022 The Author(s): This work is licensed under a Creative Commons Attribution- NonCommercial-NoDerivatives 4.0 (CC BY-NC-ND 4.0) International License

Publisher: SARC Publisher

Abnormal data access patterns, potential breaches, and appropriate remedial measures can also be detected with the assistance of AI itself. The popularity of federated learning models is growing because they allow the joint training of AI algorithms across institutions without sharing raw data, thereby preserving privacy. These features are important for aligning Big Data pipelines with regulatory security requirements such as HIPAA and GDPR.

Moreover, AI tools can facilitate auditability, as they can generate explainable outputs and maintain records of transformation logic, data lineage, and model behavior. This creates a level of trust and transparency that is required in clinical applications, where decisions often depend on data provenance and reliability.

Spezi Pipeline and Workflow Optimization

Another innovative pipeline that utilizes FHIR-based Big Data is the Spezi pipeline, which provides an easier approach to digital health workflows by unifying data ingestion, transformation, and delivery into a single framework. The design of the pipeline is fully interoperable, with the pipeline built using containerized services and a plug-and-play architecture so that components can be reusable and modular (Bikia, V. *et al.*, 2025).

The Spezi architecture can automatically map source data into FHIR resources and includes connectors to external terminologies, databases, and cloud storage facilities. In addition, the pipeline supports real-time analytics and monitoring and enables the management of dynamic decision support and adaptive care pathways.

One of the key characteristics of the Spezi framework is its focus on reproducibility and traceability. To facilitate repeatable research and regulatory audits, every step of data transformation is documented and versioned. Such a level of operational transparency is particularly important in clinical research and public health monitoring systems.

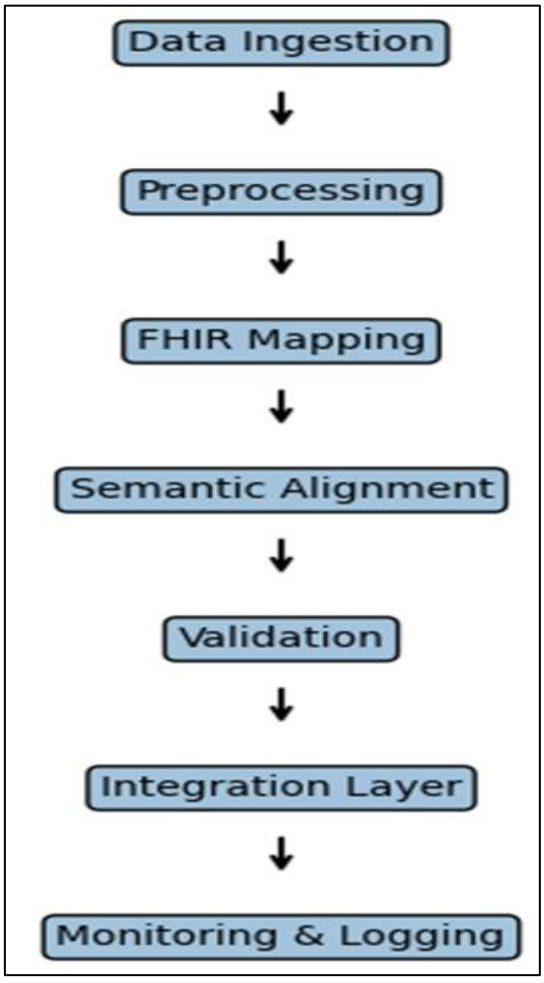


Figure 1: Spezi-Based FHIR Pipeline Architecture

(Source: Adapted from (Bikia, V. *et al.*, 2025)

Copyright © 2022 The Author(s): This work is licensed under a Creative Commons Attribution- NonCommercial-NoDerivatives 4.0 (CC BY-NC-ND 4.0) International License

Publisher: SARC Publisher

Federated Frameworks for Cross-Institutional Data Exchange

Variability in health IT systems across institutions is generally an obstacle to data sharing. Federated interoperability frameworks offer a solution because they enable decentralized data exchange while allowing local control of data to remain with each institution. A federated architecture implemented with FHIR allows secure and real-time transfer of health data among different hospital systems without requiring central data repositories (Adelusi, B. S. *et al.*, 2025).

These frameworks use standard FHIR profiles and metadata registries to ensure consistency in resource formats and meanings. Identity management and token-based authentication systems are federated to enable participating entities to obtain secure access control. The model is highly scalable and robust and can be deployed in national-level health information networks and collaborative clinical research programs.

Moreover, federated systems can be expanded with distributed ledger technologies (e.g., blockchain) to ensure accountability and immutability of data transactions. This also enhances stakeholder trust while maintaining data sovereignty.

Multi-Layered Architectures for Semantic Interoperability

The convergence of cloud-native technologies has enabled the creation of multi-layered architectures that support semantic interoperability at a large scale. These architectures typically consist of data ingestion layers, semantic enrichment modules, validation modules, and presentation modules, which are coordinated through container-based technologies and microservices. With such designs, healthcare organizations can deploy scalable and semantically aware Big Data pipelines (Pendyala, S. K.).

Semantic interoperability is achieved through the standardized application of ontologies, vocabularies, and standardized data models. Intermediate data representation formats such as RDF (Resource Description Framework) or OWL (Web Ontology Language) are typically used in such systems to translate between local data schema models and FHIR resources. This facilitated abstraction allows organizations to decouple data semantics from data design in order to add new data sources and technologies more easily without altering existing systems.

In addition, these architectures enable the integration of real-time analytics, machine learning frameworks, and clinical decision support systems with structured and semantically compatible information streams. These capabilities are particularly useful in intensive care units (ICUs), emergency rooms, and remote monitoring systems where timely data sharing may have a direct impact on patient outcomes.

Data Lakes and Interoperability Challenges

Healthcare data lakes, developed as repositories for both structured and unstructured data at scale, are increasingly being utilized to support advanced analytics and machine learning applications. However, they may be ineffective due to poor data interoperability. This challenge can be addressed by introducing FHIR standards into the data lake architecture, as it enables the use of a common data storage and access model (Anand, S. 2025).

With Snowflake-based architectures, for example, querying FHIR-transformed datasets across tenants can be performed while maintaining schema consistency. These architectures rely on semantic layers formed through data virtualization and metadata management techniques that operate on raw data. Using FHIR as a mid-level schema enables cross-platform querying and analytics, allowing insights to be retrieved from data generated by various hospitals, departments, or vendors.

Limitations of this design include schema drift, vocabulary differences, and the need to synchronize raw and curated data. These challenges can be mitigated through automated schema validation, ontology alignment tools, and AI-based data mapping tools that help simplify the interoperability process.

Copyright © 2022 The Author(s): This work is licensed under a Creative Commons Attribution- NonCommercial-NoDerivatives 4.0 (CC BY-NC-ND 4.0) International License

Publisher: SARC Publisher

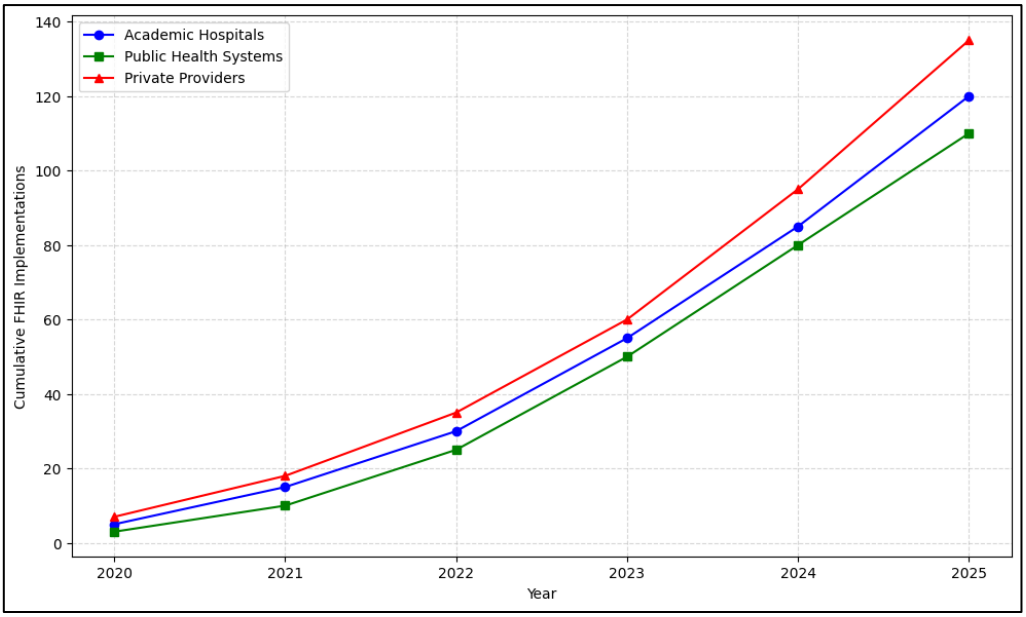


Figure 2: Growth of FHIR Adoption in Big Data Pipelines (2020–2025)

(Simulated data based on cumulative implementation trends across major institutions)
Note: Data simulated based on aggregated trends reported in (Bikia, V. *et al.*, 2025; Marfoglia, A. *et al.*, 2025), and (Carbonaro, A. *et al.*, 2025; Manam, 2025).

Description: This graph illustrates the exponential growth in the adoption of FHIR-compliant pipelines across academic hospitals, public health systems, and private healthcare providers. Between 2020 and 2025, the implementation of FHIR-based pipelines has increased significantly, reflecting a broader shift towards standards-based, semantically interoperable data infrastructures (Nellutla, 2022).

Open-Source Implementations and Vendor Neutrality

Interoperability frameworks cannot be effective through standards alone, but also require tools and ecosystems that support those standards. Open-source implementations play a crucial role because they provide transparent, customizable, and community-tested solutions for FHIR and HL7 integration. These tools reduce the likelihood of vendor lock-in and promote wider adoption by lowering the entry barriers for small institutions and start-ups (Kapitan, D. *et al.*, 2025).

Open-source FHIR servers such as HAPI FHIR and transformation engines such as Mirth Connect can be used as robust platforms to build, test, and deploy FHIR-conformant applications. They are also extensible and modular, which makes them suitable for pipelines used in Big Data environments. Moreover, community-based governance ensures continuous updates, security patches, and versioning in accordance with evolving standards.

The use of non-proprietary components also facilitates interoperability testing and verification in diverse environments such as academic, clinical, and regulatory settings. This collaborative ecosystem contributes to the development of the HL7/FHIR ecosystem and ensures that innovation is not restricted to proprietary systems alone.

Extraction Techniques for HL7 CDA Documents

Despite the fact that FHIR is the newer health data interoperability standard, the vast majority of healthcare facilities continue to utilize legacy data formats such as HL7 Clinical Document Architecture (CDA). Converting CDA documents into FHIR-conformant resources is a complex but necessary task to achieve full interoperability among health systems (Gollapudi, 2023).

CDA extraction functions are repeatable and typically involve parsing XML documents, recognizing coded entries, and mapping them to relevant FHIR resources such as Patient, Observation, Encounter, and MedicationRequest. These processes must take into account nested structures, hierarchies, and language ambiguities present in CDA templates (Talvik, H. A. *et al.*, 2025).

The outputs of data extraction algorithms are frequently based on a combination of rule-based and AI-assisted mapping, which enhances

Copyright © 2022 The Author(s): This work is licensed under a Creative Commons Attribution- NonCommercial-NoDerivatives 4.0 (CC BY-NC-ND 4.0) International License

Publisher: SARC Publisher

precision and reduces human intervention (Hassan et al., 2025). Quality assurance systems are also implemented to validate the integrity of the transformation and ensure that no clinical information is lost or distorted. These automated extraction processes reduce the time and cost required for FHIR migration and therefore should be included in any Big Data interoperability program.

Future Directions and Emerging Trends

HL7/FHIR interoperability is rapidly transforming alongside Big Data pipelines and is undergoing several emerging trends that are shaping its future course. One such trend is the convergence of interoperability with real-time streaming analytics, in which data is processed as it is generated and integrated directly with decision-support systems. Another trend is the growing use of federated learning to train AI models across institutions without compromising patient privacy and data sovereignty.

Quantum computing is an emerging technology that is being researched for its potential to enhance complex clinical simulations and analytics within FHIR-compliant data repositories. The concept of digital twins in healthcare systems is also gaining attention, as it enables predictive analytics and personalized medicine through interoperable data streams.

Finally, regulatory frameworks are increasingly emphasizing the use of open standards and open architectures that encourage vendors and institutions to invest in standards-based pipelines. Government regulations are likely to drive global standardization initiatives, such as those outlined in the U.S. 21st Century Cures Act and the European Health Data Space (EHDS) project.

CONCLUSION

Big Data pipelines are transforming healthcare interoperability by providing semantically aligned, secure, and scalable data exchange. These pipelines are used to consolidate heterogeneous data sources across institutional and national boundaries through the application of HL7 and FHIR standards. The adoption of FHIR-based systems is expanding, with AI, modular architectures, federated models, and open-source tools enabling solutions to long-standing challenges related to data standardization, security, and semantic consistency.

These technologies are not only making health data management more efficient but are also establishing the foundation for advanced clinical research, personalized medicine, and AI-based healthcare solutions. As healthcare systems continue to move toward digitization and decentralization, the adoption of FHIR-based Big Data pipelines will remain central to achieving true interoperability.

REFERENCES

1. Osamika, D., Adelusi, B. S., Kelvin-Agwu, M. T. C., Mustapha, A. Y., Forkuo, A. Y., and Ikhalea, N. “A Critical Review of Health Data Interoperability Standards: FHIR, HL7, and Beyond.”
2. Riquelme, A., Costa, P., and Martinez, C. “Large Language Models for Automating Clinical Data Standardization: HL7 FHIR Use Case.” *arXiv Preprint* (2025): arXiv:2507.03067.
3. Marfoglia, A., Nardini, F., Arcobelli, V. A., Moscato, S., Mellone, S., and Carbonaro, A. “Towards Real-World Clinical Data Standardization: A Modular FHIR-Driven Transformation Pipeline to Enhance Semantic Interoperability in Healthcare.” *Computers in Biology and Medicine* 187 (2025): 109745.
4. Carbonaro, A., Giorgetti, L., Ridolfi, L., Pasolini, R., Pagliarani, A., Cavallucci, M., et al. “From Raw Data to Research-Ready: A FHIR-Based Transformation Pipeline in a Real-World Oncology Setting.” *Computers in Biology and Medicine* 197 (2025): 111051.
5. Oke, F., Bolaji, O., Dopamu, O., Olatunji, A. P., Ibiyeye, A. O., and Stephen, O. “Revolutionizing Electronic Health Records Data Security and Interoperability: Harnessing Artificial Intelligence with FHIR, Fast Healthcare Interoperability.” *Authorea Preprints*.
6. Bikia, V., Schmiedmayer, P., Zahedivash, A., Aalami, L., Rao, A., Ravi, V., et al. “Spezi Data Pipeline: Streamlining FHIR-Based Interoperable Digital Health Data Workflows.” *arXiv Preprint* (2025): arXiv:2509.14296.
7. Adelusi, B. S., Osamika, D., Kelvin-Agwu, M. C., Yetunde, A., Mustapha, A. Y. F., and Ikhalea, N. “A Federated Interoperability Framework for Seamless Health Data Exchange Using FHIR Standards Across Multi-Hospital Systems.” (2025).
8. Pendyala, S. K. “Cloud-Driven Data Engineering: Multi-Layered Architecture for Semantic Interoperability in Healthcare.”
9. Anand, S. “Interoperability Challenges in Healthcare Data Lakes: A Snowflake-Based

Copyright © 2022 The Author(s): This work is licensed under a Creative Commons Attribution- NonCommercial-NoDerivatives 4.0 (CC BY-NC-ND 4.0) International License

Publisher: SARC Publisher

Approach." *International Journal of Emerging Trends in Computer Science and Information Technology* 6.1 (2025): 111–123.
10. Kapitan, D., Heddema, F., Dekker, A., Sieswerda, M., Verhoeff, B. J., and Berg, M. "Data Interoperability in Context: The Importance of Open-Source Implementations When Choosing Open Standards." *Journal of Medical Internet Research* 27 (2025): e66616.
11. Talvik, H. A., Oja, M., Tamm, S., Mooses, K., Särg, D., Loo, M., et al. "Repeatable Process for Extracting Health Data from HL7 CDA Documents." *Journal of Biomedical Informatics* 161 (2025): 104765.
12. Manam, K. B. "An Exploration of Gender Biases and Discrimination Manifesting on AI/ML Development Team Dynamics." Doctoral dissertation, University of the Cumberlands (2025).
13. Nellutla, N. "Secure DevSecOps Workflows for Medical IoT Device Integration in Smart Hospitals." *International Journal of AI, BigData, Computational and Management Studies* 3, no. 1 (2022): 114–122. DOI:10.63282/3050-9416.IJAIBDCMS-V3I1P113.
14. Gollapudi, R. "Operational Drift and Risk-Bounded Decision-Making in Production Database Systems." *Journal of International Crisis and Risk Communication Research* (2023): 132–147. DOI: 10.63278/jicrcr.vi.3762.
15. Hassan, S. Z., Alang, K., Nellutla, N., Balasubramanian, S. A., and Morusu, V. "GovernAI: Policy-Driven Model Governance for Dynamic and Multi-Tenant AI Systems." *Proceedings of the 2025 International Conference on Information, Implementation, and Innovation in Technology (I2ITCON)* (2025): 1–8. DOI: 10.1109/I2ITCON65200.2025.11210544.

Source of support: Nil; **Conflict of interest:** Nil.

Cite this article as:
Nune, B. " Big Data Pipelines for HL7/FHIR Interoperability." *Sarcouncil Journal of Engineering and Computer Sciences* 5.3 (2026): pp 62-68.

Copyright © 2022 The Author(s): This work is licensed under a Creative Commons Attribution- NonCommercial-NoDerivatives 4.0 (CC BY-NC-ND 4.0) International License

Publisher: SARC Publisher