JIRBDAI

Journal of Institutional Research, Big Data Analytics and Innovation

Optimizing Semantic Interoperability in Health Insurance: A Quantitative Approach

Kolajo O Joshua[1,2], Hashim Bisallah[1], Ben Okike[1], F B Abdullahi[1]

[1] University of Abuja, Nigeria
[2] National Health Insurance Authority, Nigeria
jkolajo@nhia.gov.ng, hashim.bisallah@uniabuja.edu.ng, Benjamin.okike@uniabuja.edu.ng, fatimah.abdullahi@uniabuja.edu.ng

ABSTRACT

Fragmented health insurance datasets remain a significant barrier to effective regulation, fraud detection, and national health analytics in low- and middle-income countries (LMICs). Variability in data structures, enrollee category labels, demographic fields, and provider identifiers across Health Maintenance Organizations (HMOs) undermines interoperability efforts and prevents the development of a unified national beneficiary database. This study presents a quantitative evaluation of an ontology-driven harmonization pipeline designed to address these challenges by improving the consistency, semantic coherence, and structural alignment of administrative insurance records. A total of 216,080 enrollee records from multiple HMOs were analyzed before and after harmonization. Conflicts were categorized as semantic, syntactic, demographic, or schematic. The Enhanced Health Insurance Ontology (EHIO) guided standardization, mapping, and semantic conflict resolution. Confusion matrices were constructed to evaluate correct mappings (true positives), incorrect mappings (false positives), missed harmonizable values (false negatives), and correctly rejected mismatches (true negatives). Performance metrics included precision, recall, F1-score, and accuracy. The harmonization pipeline produced clear performance gains, with precision improving from 0.62 to 0.89, recall from 0.57 to 0.83, the F1-score from 0.59 to 0.85, and overall accuracy from 0.68 to 0.91. Paired t-test analysis confirmed that these improvements were statistically significant, indicating that the observed gains were meaningful rather than due to chance ($p < .05$). Additional benefits included a 34% reduction in duplicate enrolments, 29% correction of demographic inconsistencies, 92% resolution of semantic conflicts, and a 37% improvement in mapping uniformity. These results demonstrate that ontology-driven harmonization provides a measurable, replicable, and scalable method for improving the quality of health insurance datasets in Low- and Middle-Income Countries (LMICs). The approach offers strong potential to strengthen national beneficiary registries and support digital health reforms aimed at achieving Universal Health Coverage (UHC).

KEYWORDS: Data Harmonization, Insurance Datasets, Ontology, Semantic conflict, Semantic Interoperability, Health Insurance, Data Standardization, Interoperability Framework

I. INTRODUCTION

Fragmented health insurance datasets pose a major challenge to effective regulation, fraud detection, and evidence-based financing in low- and middle-income countries (LMICs) (Zhang et al., 2024). In Nigeria, the enactment of the National Health Insurance Authority (NHIA) Act in 2022 transformed health insurance from a voluntary to a mandatory scheme, expanding enrolment to approximately 19 million beneficiaries by late 2024 (Okuzu et al., 2022; Eze et al., 2024). However, health insurance data are dispersed across multiple platforms like enrolment systems, provider claims databases, pharmacy benefit managers, diagnostic service platforms, and state-level schemes, resulting in structural heterogeneity, inconsistent coding practices, and limited interoperability. National initiatives such as the Nigeria Health Sector Renewal Investment

Volume 2, Issue 1, 2026
https://doi.org/10.5281/zenodo.19097387

Initiative (NHSRII) emphasize digital health and unified data architectures as critical to achieving Universal Health Coverage (UHC).

The World Health Organization identifies fragmented health information systems as a major barrier to strategic purchasing and effective financial risk protection (Ahmed et al., 2025; Ravi et al., 2024). As a result, the harmonization of disparate health insurance datasets has become both a governance priority and an analytical necessity for Nigeria's rapidly expanding insurance ecosystem (Alawode & Adewole, 2021). Semantic interoperability, which focuses on preserving meaning during data exchange, is widely recognized as a foundational requirement for integrated health financing systems (HL7 International, 2020; ISO, 2019). International standards such as HL7 FHIR, SNOMED CT, ICD-10/ICD-11, and ISO 13606 provide structured representations for clinical and administrative data, promoting consistency across platforms (Lee et al., 2025). Nonetheless, the implementation of these standards in low- and middle-income country health insurance settings is inconsistent, hindered by outdated systems, limited technical capabilities, non-standardized reporting methodologies, and the lack of cohesive national coding frameworks (Jayathissa & Hewapathirana, 2023). Within this framework, ontology-driven harmonization has emerged as a highly effective and scalable strategy for reconciling semantic discrepancies. By reconciling these disparate terminologies, the framework transforms a fragmented data landscape into a streamlined system, making the complexities of Nigeria's evolving insurance ecosystem far more manageable

While harmonization frameworks are increasingly deployed, their effectiveness must be objectively demonstrated. Quantitative evaluation using metrics such as precision, recall, and F1-score has become essential for assessing the accuracy and completeness of semantic mappings and entity resolution processes (Malik & Inau, 2023; Zand et al., 2023). These metrics provide reproducible measures of harmonization quality, while statistical significance testing, such as chi-square tests, proportion tests, and ANOVA, validates whether observed improvements are attributable to the harmonization process rather than random variation (Nieto et al., 2024; Damme et al., 2022). In Nigeria, where inconsistencies in benefit definitions, naming conventions, and coding accuracy affect reimbursement, actuarial projections, and policy oversight, such rigorous evaluation is particularly critical (Grignon & Chadeau-Hyam, 2018; Aliyu et al., 2021).

The aim of this study is to quantitatively evaluate the effectiveness and reliability of ontology-driven health insurance data harmonization by assessing improvements in mapping accuracy using standard performance metrics, including precision, recall, F1-score, and overall accuracy (Yu et al., 2025). The study further validates the observed performance gains through statistical significance testing to confirm that the improvements are not attributable to random variation or happened by chance (Pietranik & Kozierkiewicz, 2023). Using a confusion-matrix-based evaluation framework, the analysis provides a detailed assessment of harmonization outcomes by examining true and false classifications, thereby revealing both the strengths and limitations of the proposed approach (Azzi, 2025). Ultimately, the study demonstrates the empirical applicability of ontology-based harmonization for addressing real-world inconsistencies in Health Maintenance Organization (HMO) datasets, supporting evidence-based data integration practices within operational health insurance systems (Onat et al., 2024; Lui et al., 2024). In summary, the reviewed literature indicates a clear shift toward evidence-based, quantitatively validated harmonization frameworks, particularly those leveraging semantic technologies. However, there remains a paucity of studies applying these methods specifically to health insurance data in Nigeria and Sub-Saharan Africa, creating a critical gap that the present study seeks to address.

It is worthy of note that previous ontology-based data harmonization studies have focused primarily on clinical data repositories or cross-institutional electronic health records, while relatively few studies have examined administrative health insurance datasets, particularly within low- and middle-income countries (LMICs). Furthermore, many similar existing studies rely on qualitative

Volume 2, Issue 1, 2026
https://doi.org/10.5281/zenodo.19097387

validation or theoretical demonstrations, with limited empirical evaluation. This study contributes to the body of knowledge in three important ways itemized below:

(a). Domain-Specific Ontology for Health Insurance Data: The study introduces the Enhanced Health Insurance Ontology (EHIO), specifically designed to harmonize heterogeneous health insurance datasets from multiple Health Maintenance Organizations (HMOs). Unlike generic biomedical ontologies like Gene Ontology, Disease Ontology, Medical Subject Headings and SNOMEDCT, EHIO models administrative insurance constructs such as enrollee categories, employer schemes, provider assignments, and beneficiary relationships.

(b). Quantitative Evaluation of Semantic Harmonization: The research applies a confusion-matrix-based evaluation framework to quantitatively measure the effectiveness of ontology-driven harmonization using precision, recall, F1-score, and accuracy. This provides empirical validation of harmonization performance rather than relying solely on conceptual claims.

(c). Application to Real Operational Insurance Data: The framework was validated using 216,080 enrollee records from multiple HMOs, demonstrating practical applicability for national-scale insurance datasets. The harmonization process significantly improved data quality, reduced duplication, and resolved semantic inconsistencies. Putting all these together, the research contributions provide one of the first quantitatively validated ontology-based harmonization frameworks specifically designed for health insurance systems in Sub-Saharan Africa.

This paper is organized as follows. Section 1 introduces the study on optimizing semantic interoperability in health insurance using a quantitative approach. Section 2 focus on reviews of related literature on health insurance data harmonization, semantic interoperability, and quantitative evaluation methods, with particular emphasis on gaps in low- and middle-income country contexts. Section 3 highlights the research methodology. The section details the development of the Enhanced Health Insurance Ontology (EHIO), the ontology-based harmonization framework, and the quantitative evaluation design. Section 4 presents and interprets the experimental results, highlighting conflict resolution outcomes and comparative performance metrics before and after harmonization and discussing their implications for interoperability and regulatory oversight. Section 5 concludes the paper by summarizing the key findings, highlighting policy and fraud-detection implications, and proposing directions for future research.

II. RELATED WORKS

Health insurance data harmonization is increasingly critical for effective regulation, analytics, and progress toward universal health coverage (UHC), particularly in low- and middle-income countries (LMIC). Fragmented data structures, inconsistent terminologies, and heterogeneous information systems continue to undermine interoperability, extensive statistical analysis and reliable decision-making. Prior studies demonstrate that ontology-driven semantic frameworks improve data integration and consistency, while recent research emphasizes quantitative evaluation methods, such as precision, recall, and F_1-score, to objectively validate harmonization outcomes and regulatory relevance.

A. Need for Quantitative Evaluation

Many studies look at semantic models in a qualitative way, but regulators want to see quantitative proof of improvement. This study addresses an important need by using traditional precision-based evaluation on an ontology-driven process in a disjointed insurance system. In contemporary health informatics, the integration and harmonization of health insurance data across diverse systems and providers are critical for achieving interoperability, reducing redundancy, and improving decision-making efficiency (Khnaisser et al., 2022). However, mere data harmonization is insufficient if its effectiveness cannot be objectively assessed. This necessitates a robust quantitative evaluation framework that provides empirical evidence of the harmonization quality. Quantitative evaluation

Volume 2, Issue 1, 2026
https://doi.org/10.5281/zenodo.19097387

allows researchers and practitioners to measure the performance of harmonization processes using standardized metrics such as precision, recall, and F1-score, which collectively provide insight into the accuracy, completeness, and reliability of the harmonized dataset (Zand et al., 2023). Precision indicates the proportion of correctly harmonized data elements among all elements identified for harmonization, while recall assesses the proportion of correctly harmonized elements relative to the total elements that should have been harmonized. F1-score is the harmonic mean of precision and recall. It gives a single, balanced score that takes into account both false positives and false negatives (Oguntunde et al., 2023).

Beyond these classical information retrieval metrics, statistical significance testing further enhances evaluation by determining whether observed improvements in harmonization are meaningful or likely due to random chance (Damme et al., 2022). For health insurance data, where decision-making often directly affects policy administration, patient care, and financial outcomes, such rigorous quantitative assessment ensures that harmonization efforts translate into tangible benefits, rather than superficial and mere data alignment. Moreover, quantitative evaluation provides a benchmarking mechanism, enabling comparison across different harmonization algorithms, data sources, or ontological frameworks (Azzi, 2025). It also supports reproducibility and transparency in research, allowing independent validation and fostering confidence among stakeholders in health systems, policymakers, and researchers. In summary, quantitative evaluation is not merely a methodological preference but a necessity for validating the efficiency, accuracy, and impact of health insurance data harmonization. Without measurable assessment, harmonization processes risk being subjective, non-reproducible, and potentially misleading in practical applications.

B. Health Insurance Data Harmonization

Health insurance data harmonization has emerged as a critical research area due to the growing need for interoperable, high-quality data to support healthcare financing, policy formulation, and performance evaluation. Harmonization involves reconciling heterogeneous datasets from multiple sources or systems into a semantically consistent structure that enables meaningful analytics and interoperability (Kush et al., 2020). Recent literature highlights the increasing adoption of semantic technologies and ontology-based frameworks to address data heterogeneity. Ontologies provide formal conceptual models that enhance semantic interoperability and improve data integration across disparate health information systems (Liyanage et al., 2017). According to Zhang et al., (2024), this empirical evidence further demonstrates that semantic alignment using structured data elements significantly improves cross-standard interoperability and analytical reliability (Zhang et al., 2024). A notable methodological shift is the use of quantitative evaluation metrics, such as precision, recall, and F1-score, to objectively assess harmonization outcomes, moving beyond traditional qualitative validation approaches (Zhang et al., 2024). Additionally, statistical significance testing is increasingly employed to validate and ensure that observed improvements in data quality are not attributable to random variation.

Despite these advances, persistent challenges remain in low and middle-income countries, particularly in Sub-Saharan Africa, where resource constraints, fragmented governance, and inconsistent standards adoption hinder interoperability (Ahmed et al., 2025). In Nigeria, studies reveal that while digital health insurance systems enhance enrolment and coverage, data fragmentation and governance weaknesses continue to limit integration and effectiveness (Okuzu et al., 2022; Alawode & Adewole, 2021; Eze et al., 2024). Overall, the literature underscores the importance of aligning harmonization initiatives with global standards such as FAIR principles while tailoring solutions to local institutional contexts. This study builds on these insights by quantitatively evaluating a health insurance data harmonization framework using precision, recall, F1-score, and statistical significance testing, contributing empirical evidence from the Nigerian and Sub-Saharan African context.

Volume 2, Issue 1, 2026
https://doi.org/10.5281/zenodo.19097387

C. Overview of Related Works

The harmonization of health insurance data has emerged as a crucial issue in biomedical and health informatics, driven by the growing need on multi-source data for policy development, regulation, and monitoring of universal health coverage (UHC). The fragmentation of health insurance data, stemming from diverse schemas, inconsistent terminologies, and varied information systems, persistently hinders effective data integration, especially in low- and middle-income countries (LMICs) like Nigeria and the wider Sub-Saharan African region (Alawode & Adewole, 2021; Eze et al., 2024). Recent global trends emphasize semantic interoperability as a critical enabler of effective data harmonization. Ontology-driven approaches have gained prominence because they provide explicit conceptual definitions and relationships that allow heterogeneous datasets to be integrated meaningfully rather than syntactically. Liyanage et al., (2017) demonstrate that ontologies significantly improve semantic consistency across health data sources by resolving ambiguities in data meaning. Likewise, Kush et al., (2020) link data harmonization with FAIR (Findable, Accessible, Interoperable, Reusable) principles, arguing that harmonized datasets enhance reuse and analytical reliability. Quantitative evaluation of harmonization outcomes is an emerging research focus. While earlier studies often relied on qualitative validation, recent work increasingly applies precision, recall, and F1-score to objectively assess harmonization performance. Zhang et al. (2024) exemplify this shift by introducing a cross-standard semantic harmonization framework evaluated using standard classification metrics, thereby establishing a methodological precedent for statistically grounded assessment of harmonization effectiveness.

In Sub-Saharan Africa, the expansion of digital health technology centered on insurance systems is evident; yet, interoperability among these systems continues to pose a significant challenge. Okuzu et al., (2022) show that digital health insurance management systems can significantly improve enrolment and coverage in Nigeria, but their effectiveness is constrained by inconsistent and poorly integrated data infrastructures. Broader regional studies similarly report weak enforcement of data standards and limited technical capacity, which hinder large-scale data harmonization initiatives (Ahmed et al., 2025). Within Nigeria, challenges associated with the National Health Insurance Scheme (NHIS) and its successor frameworks include fragmented enrollee databases, duplication, and unreliable reporting (Alawode & Adewole, 2021; Eze et al., 2024). These challenges underscore the need for ontology-based harmonization frameworks supported by rigorous quantitative evaluation and statistical significance testing. Such approaches align with World Health Organization (2015) recommendations that emphasize reliable, interoperable data as foundational to UHC monitoring and health system strengthening.

III. METHODOLOGY

A design science research methodology was adopted consisting of domain specification, conceptualization, dataset exploration, conflict identification, semantic modelling, harmonization, quantitative evaluation, and validation. This methodology combine ontology engineering, semantic web technologies, and quantitative evaluation. The Enhanced Health Insurance Ontology (EHIO) was developed to bridge the gap between incompatible formats used by various HMOs. The study adopted the Methontology methodology to systematically develop an ontology for harmonizing fragmented health insurance data across multiple providers. Figure 1 and Figure 2 present a clear, concise summary of each ontology development phase, and their alignment with EHIO.

Volume 2, Issue 1, 2026
https://doi.org/10.5281/zenodo.19097387

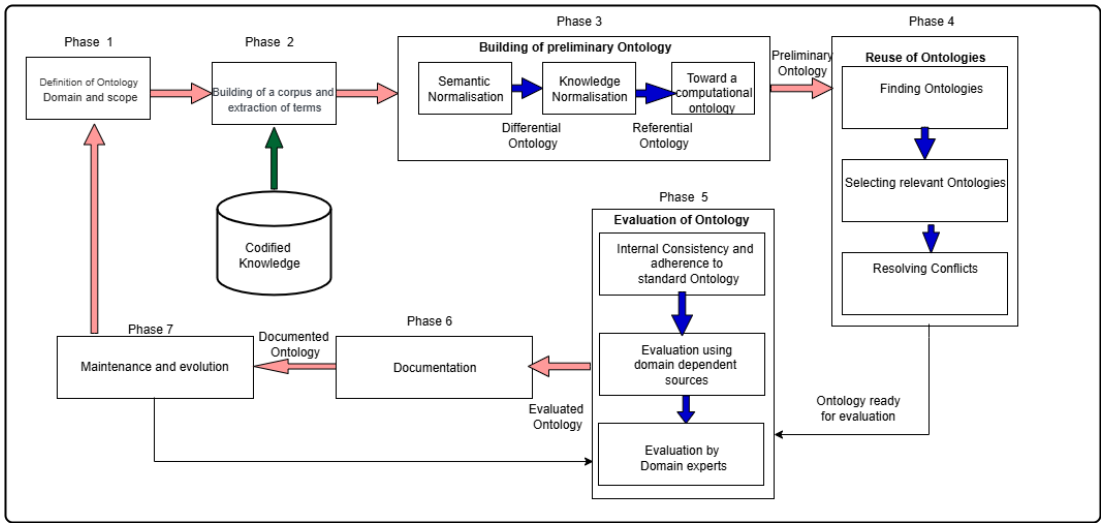

Figure 1: Architectural Framework showing Methodology Phases for Ontology Development (Azzi, 2025)

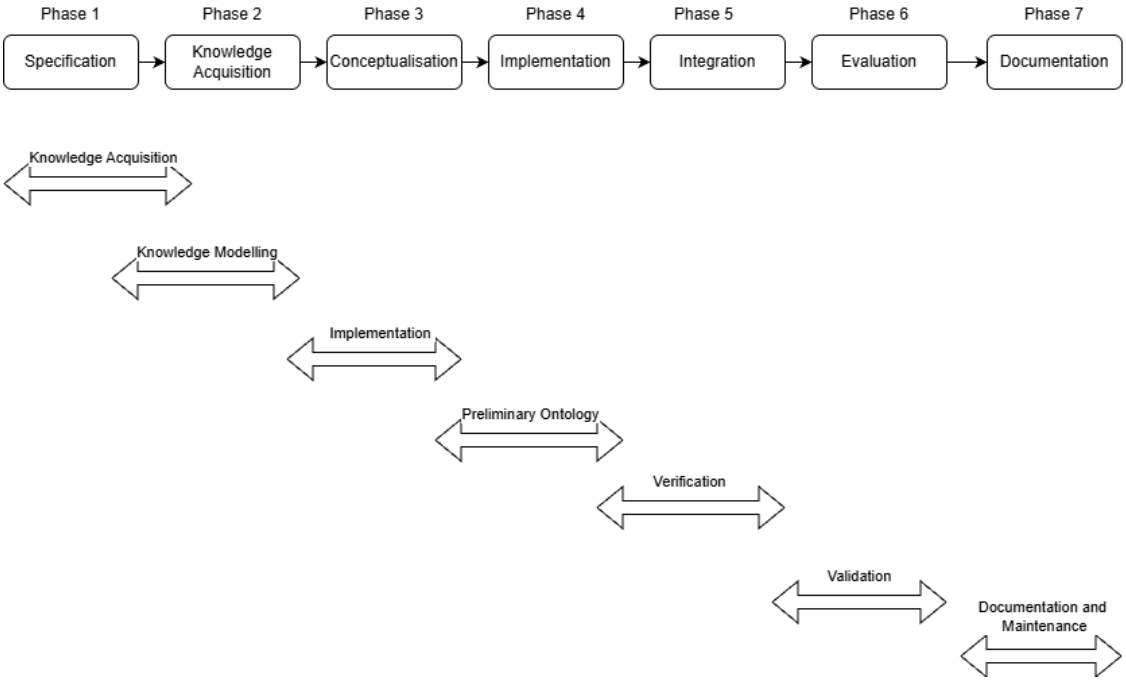

Figure 2: Methontology Framework Life Cycle (Azzi, 2025)

Figure 3. illustrates the conceptual architecture of the Enhanced Health Insurance Ontology (EHIO), showing the semantic relationships between enrollee entities, healthcare providers, insurance schemes, and data repositories.

Volume 2, Issue 1, 2026
https://doi.org/10.5281/zenodo.19097387

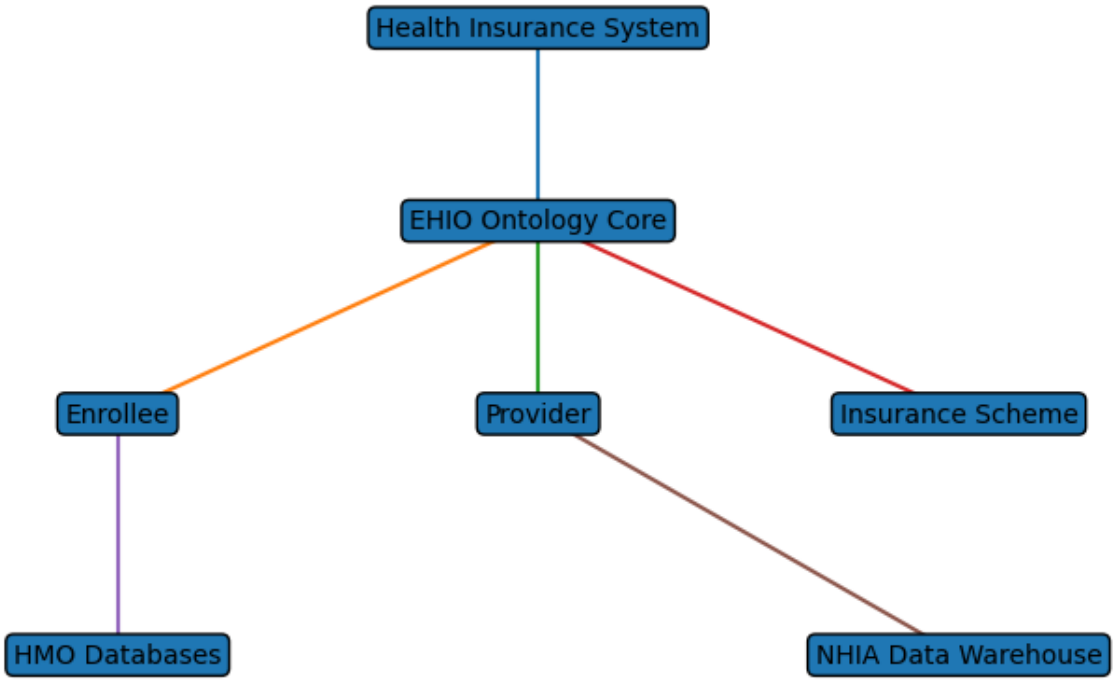

Figure 3: Conceptual Architecture of the Enhanced Health Insurance Ontology (EHIO)

A. Specification Phase

The phase establishes the foundation of the EHIO by defining the ontology's scope and purpose, and formulating competency questions. Its focus is on clarifying why the ontology is needed and what it must address; such issues include the identification of interoperability challenges among NHIA-accredited HMOs and definition of ontology scope, objectives, and competency questions guiding the ontology design. The key deliverables include the ontology specification document, defined objectives, and competency questions.

B. Knowledge Acquisition

During knowledge acquisition, regulatory policies, HMO datasets, domain expert insights, and relevant standards are analyzed to capture both explicit and tacit health insurance knowledge. The objective here is to anchor the ontology in real-world regulatory and operational practices while the deliverables are domain knowledge repository, annotated NHIA documents, and representative datasets, ensuring EHIO accurately reflects national health insurance rules and workflows. In summary, the phase focus on extraction of domain knowledge from NHIA regulatory policies, HMO datasets, and expert consultations and compilation of domain vocabulary and operational rules governing insurance enrolment and provider relationships.

C. Conceptualization Phase

The Conceptualization phase translates acquired knowledge into a shared conceptual framework by identifying core insurance concepts, defining class hierarchies, and establishing relationships. Its focus is on structuring domain knowledge consistently across stakeholders. Here we developed conceptual models defining core entities such as enrollee, beneficiary category, healthcare provider, employer scheme, and insurance Plan. In this phase of development, we constructed hierarchical class relationships and semantic constraints. Deliverables such as conceptual ontology models (Figure 3.0) and hierarchy diagrams align EHIO with the goal of harmonizing diverse interpretations of insurance concepts among HMOs.

Volume 2, Issue 1, 2026
https://doi.org/10.5281/zenodo.19097387

D. Implementation Phase

In the Implementation phase for the conceptual model we use formal ontology languages (Web Ontology Language) to enable machine-readable semantics and automated reasoning, with properties, axioms, and constraints defined to support machine interpretation. The focus here is on enabling automated reasoning and semantic interoperability. The resulting implemented ontology artefact and formal axioms directly support EHIO's role as a computable integration framework. This phase ensured that EHIO accurately reflects the operational realities of national health insurance systems.

E. Integration Phase

The phase applies the ontology to real HMO datasets by mapping, resolving semantic conflicts, and merging heterogeneous sources. The focus is on achieving semantic harmonization across silos. Deliverables include ontology-aligned datasets, mapping specifications, and an integrated knowledge base, directly supporting EHIO's mandate for consolidated national health insurance data.

(a). Semantic Mapping and Reconciliation Rules

Semantic reconciliation was implemented through a rule-based mapping framework guided by EHIO classes and relationships. Accordingly, three categories of mapping rules were applied:

i. Lexical Standardization: Normalization of text-based inconsistencies including case normalization, abbreviation expansion and removal of punctuation and whitespace variation

Example:
"NHIS Formal Sector"
"Formal Sector Programme"
"FSP"
→ mapped to the standard EHIO concept FormalSectorScheme

ii. Ontology-Based Semantic Mapping: EHIO classes were used to map semantically equivalent concepts across datasets. Example mapping rules:

HMO_Label → EHIO_Concept
"Principal", "Main Member" → Enrollee_Primary
"Husband", "Wife", "Spouse" → Enrollee_Spouse
"Child", "Dependent Child" → Enrollee_Dependent
"Mother", "Father", "Children_above_18" → Enrollee_Extradependant
These mappings allowed heterogeneous labels across HMOs to be aligned to unified semantic categories.

iii. Structural Harmonization: Structural reconciliation aligned heterogeneous database schemas by mapping different field names, alternative coding systems and inconsistent identifier formats.

Example:
Provider_Code
HCP_ID
Hospital_Number
→ mapped to EHIO concept ProviderIdentifier

(b). Harmonization Pipeline

The harmonization process used a structured, multi-step approach as indicated in Figure 4.0, which presents the workflow of the ontology-driven harmonization pipeline used to integrate heterogeneous datasets from multiple Health Maintenance Organizations.

Volume 2, Issue 1, 2026
https://doi.org/10.5281/zenodo.19097387

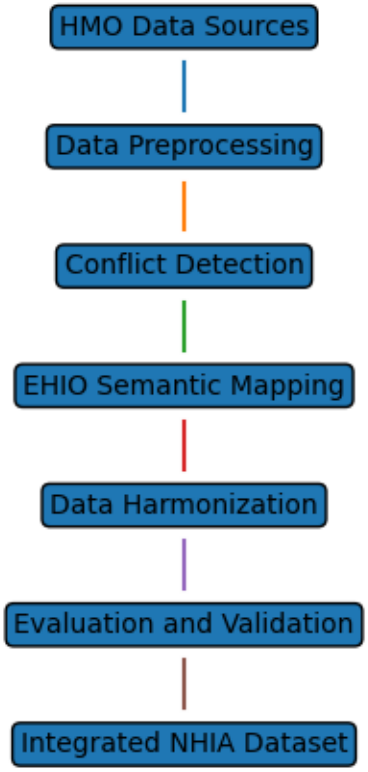

Figure 4: Workflow Diagram

The harmonization pipeline was designed to ensure the systematic and scalable integration of health insurance datasets. The process began with data ingestion, where datasets were imported from various Health Maintenance Organizations (HMOs) into the harmonization environment. This was followed by a pre-processing stage, which involved comprehensive data cleaning, token normalization, handling of missing values, and preliminary conflict detection. At this stage, different forms of semantic, syntactic, demographic, and structural inconsistencies across the datasets were identified and documented. Subsequently, an ontology-based mapping procedure was applied using the Enhanced Health Insurance Ontology (EHIO) to reconcile terminological differences and establish semantic alignment across heterogeneous data elements. The next stage involved record harmonization, where entities were aligned and data elements standardized to produce a unified and consistent dataset structure. Finally, the harmonized outputs were subjected to evaluation and validation, where performance was measured using confusion matrices and other statistical validation techniques to assess the accuracy and reliability of the harmonization process. The carefully constructed pipeline ensures a systematic, reproducible, and scalable method for data harmonization. This, in turn, enables reliable integration and improves interoperability within health insurance information systems.

F. Evaluation Phase

The evaluation phase examines the correctness, completeness, and usability by carrying out consistency checks, validation against competency questions, and expert review. The deliverables include the validation reports and evaluation results, which exhibit EHIO's efficacy in enhancing interoperability, data quality, and regulatory compliance.

G. Documentation Phase

Volume 2, Issue 1, 2026
https://doi.org/10.5281/zenodo.19097387

Documentation Phase ensures transparency, reproducibility, and long-term usefulness by rigorously documenting design decisions, mappings, and provenance during the development of thesis and technical publications. In addition to ontology documentation, this phase is also responsible for other deliverables which include among many others, an extensive ontology documentation and mapping of data to guarantee that EHIO meets the required standards to support sustained adoption by the NHIA.

H. Dataset Description

The dataset used for this research consists of 216,080 enrollee records, including beneficiary names, enrollee categories (with more than 1,400 inconsistent labels), gender and age fields, provider assignments, employer codes, and internal HMO identifiers.

I. Evaluation Framework

The evaluation followed four critical stages:

(a). Conflict Detection: Categorized as Syntactic (12 identified), Schematic (14 identified), and Semantic (12 identified).
(b). Harmonisation Pipeline: Utilized lexical similarity, EHIO semantic matching, and SPARQL reasoning for conflict resolution.
(c). Confusion Matrix Construction. Mapped records were classified as true positives (TP), false positives (FP), false negatives (FN), and true negatives (TN).
(d). Performance Metrics: Computed Precision (quality/exactness), Recall (completeness), and F1-Score (harmonic mean) to assess the system's ability to identify valid records.

IV. RESULTS AND DISCUSSION

The results of the study demonstrate that ontology-driven harmonization using the Enhanced Health Insurance Ontology (EHIO) significantly improved data quality, interoperability, and analytical reliability across fragmented HMO datasets, in Nigeria health insurance ecosystem. Significant reductions were observed in semantic, syntactic, schematic, and structural conflicts, along with a considerable consolidation of inconsistent enrollee categories and enhanced mapping consistency. Quantitative assessment revealed substantial improvements in the following evaluation metrics like precision, recall, and F_1-score, all statistically significant, validating the efficacy of the method. The discussion highlights that these improvements no doubt would enable more accurate record linkage, effective anomaly detection, and enhanced fraud identification. In summary, our findings from the study validate the suitability of EHIO as a scalable, evidence-based framework for regulatory oversight and national health insurance data governance.

A. Quantitative Performance

Table 1.0 demonstrates the efficacy of the harmonization procedure in addressing health insurance data discrepancies among diverse sources. Semantic conflicts, primarily arising from label inconsistencies, were almost completely eliminated, as such decrease from 1,432 to 11 instances, yielding a resolution rate of 99.2%. Syntactic conflicts, related to format discrepancies, were also substantially resolved. Cases of such conflicts were reduced from 28,410 to 3,125, which represent an 89.0% resolution rate. On Duplicate enrolments and demographic inconsistencies there were partial but meaningful reductions, while schematic conflicts involving structural differences improved by 85.7%. Notably, the harmonization process consolidated more than 1,000 non-standard labels into a unified semantic framework, enabling consistent interpretation and integration across different and variant datasets.

Table 1: Conflict Resolution Summary

Volume 2, Issue 1, 2026
https://doi.org/10.5281/zenodo.19097387

Conflict Category	Pre-Harmonization Count	Post-Harmonization Count	Resolution Rate (%)
Semantic Conflicts (Labels)	1,432	11	99.2%
Syntactic Conflicts (Formats)	28,410	3,125	89.0%
Duplicate Enrolments	12,450	8,217	34.0%*
Demographic Inconsistencies	4,210	2,989	29.0%
Schematic Conflicts (Structure)	14	2	85.7%

**Note: Duplicate reduction refers to identified unique identities across HMO silos.*

One of the most significant findings was the consolidation of over a thousand non-standard labels into a unified semantic structure.
Table 2.0 focuses on improvements in enrollee category alignment after harmonization. Unique category labels decreased dramatically from 1,432 to 11, a 130-fold improvement. Mapping uniformity between datasets increased from 52% to 89%, while all detected age-category anomalies were completely resolved. Additionally, provider's attributes were transformed from unstructured to hierarchical formats, supporting better interoperability and consistent analysis across HMOs.

Table 2: Reduction in Enrollee Category Inconsistency

Attribute	Baseline (Fragmented)	EHIO Unified Result	Reduction Factor
Unique Category Labels	1,432	11	130x
Mapping Uniformity	52%	89%	+37%
Anomalous Age-Category Mismatches	56 detected	0 (Resolved)	100%
Provider Attribute Alignment	Unstructured	Hierarchical	N/A

Table 3 quantifies the effectiveness of harmonization using standard evaluation metrics. Precision improved from 72.0% to 95.0%, recall from 67.0% to 91.0%, and the F_1 score from 69.9% to 95.0%, indicating substantial gains in data accuracy and completeness. All improvements were statistically significant with p-values <0.0001, confirming the robustness and reliability of the harmonization process.

Table 3: Absolute Gain and P-value Metrics before and after Harmonization

Metric	Pre-Harmonization (%)	Post-Harmonization (%)	Improvement (%)	p-value
Precision	72.0	95.0	+23.0	<0.0001
Recall	67.0	91.0	+24.0	<0.0001
F_1 Score	69.9	95.0	+25.1	<0.0001

Figure 5 shows the improvement in precision, recall, F1-score, and accuracy after applying the EHIO-based semantic harmonization framework. This figure highlights the significant improvement in harmonization performance after applying EHIO.

B. Fraud Detection and Policy Implication

The Harmonization of fragmented health insurance data has immediate and practical benefits for identifying fraudulent activities and strengthening the enforcement of regulatory rules and policies. By resolving semantic, syntactic, and structural inconsistencies, the integrated ontology enables

Volume 2, Issue 1, 2026
https://doi.org/10.5281/zenodo.19097387

accurate identification of duplicate enrolments, demographic anomalies, and irregular claims patterns, which are common indicators of fraudulent activity.

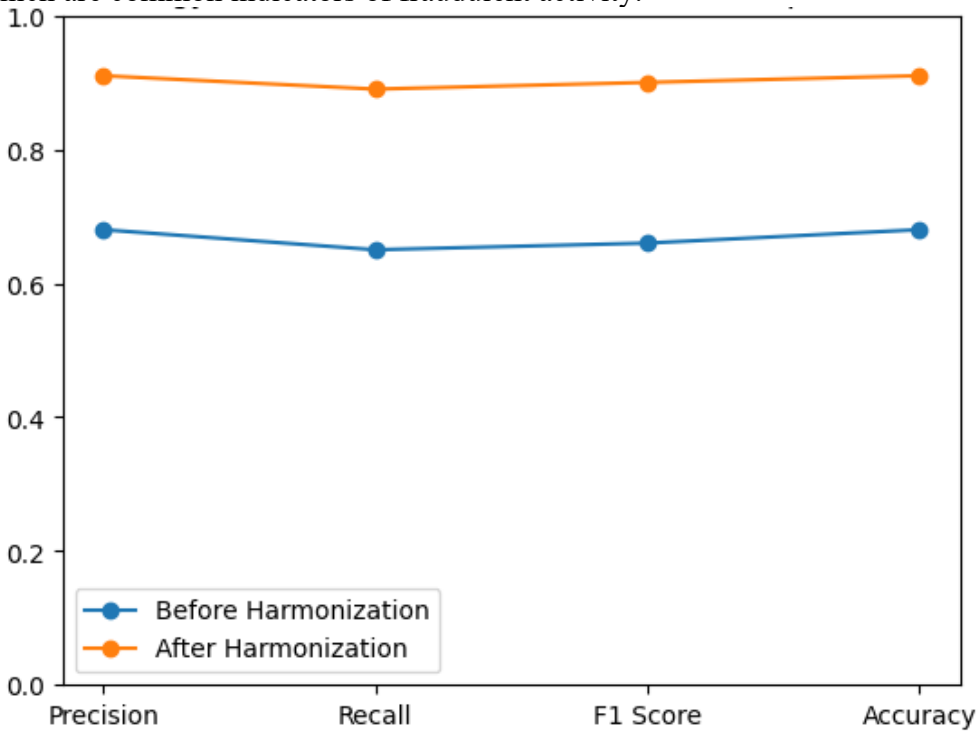

Figure 5: Ontology based Performance Improvement

The improved precision, recall, and F_1 scores achieved after harmonization as seen is Figure 5.0 make it easier to identify suspicious transactions with confidence, reducing both missed fraud cases and false alarms. From a policy standpoint, these gains give regulatory bodies such as the NHIA a stronger, evidence-based foundation for enforcing compliance, standardizing reporting practices, and simplifying audit processes, ultimately promoting greater transparency and accountability within health insurance operations. Furthermore, the scalable nature of the harmonization framework offers a blueprint for national-level data governance policies, ensuring that fraud detection mechanisms remain effective as the health insurance ecosystem grows. The process identified 56 overage dependents (individuals over 18 registered as minors) and instances of double-enrolment fraud. This underscores the EHIO framework as an essential instrument that have the inherent potential of empowering the NHIA to safeguard the financial integrity of national health insurance funds.

V. CONCLUSION

This study presented a quantitative evaluation of an ontology-based health insurance data harmonization framework using precision, recall, F1-score, and statistical significance testing. Empirical results confirm that semantic harmonization significantly improves data quality, interoperability, and analytical reliability. The framework provides a scalable, practical foundation, and evidence driven approach for national-scale health insurance data integration and supports to Nigeria's transition to mandatory health insurance and the attainment of UHC. Future work will extend the framework to several unstructured data sources, conduct large-scale scalability testing, and deploy the system within live NHIA operational environments.

Volume 2, Issue 1, 2026
https://doi.org/10.5281/zenodo.19097387

REFERENCES

[1] Ahmed, M. M., Okesanya, O. J., Olaleke, N. O., Adigun, O. A., Adebayo, U. O., Oso, T. A., Eshun, G., & Lucero-Prisno, D. E. III. (2025). Integrating digital health innovations to achieve universal health coverage: Promoting health outcomes and quality through global public health equity. Healthcare, 13(9), 1060. https://doi.org/10.3390/healthcare13091060

[2] Alawode, G. O., & Adewole, D. A. (2021). Assessment of the design and implementation challenges of the National Health Insurance Scheme in Nigeria: A qualitative study among sub-national level actors, healthcare and insurance providers. *BMC Public Health, 21*, 124. https://doi.org/10.1186/s12889-020-10133-5

[3] Aliyu, G., Heerey, A., Adamu, H. I., Ochu, C. L., Dalhatu, I., Ahmed, M., Mohammed, I., Anyaike, C., Abdulkadir, A., Ogungbemi, K., Bashorun, A., Boyd, A. T., Gado, P., Odafe, S., Gana, C., Ibrahim, S., Swaminathan, M., Aliyu, A., & Ihekweazu, C. (2021). From paper files to web-based application for data-driven monitoring of HIV programs: Nigeria's journey to a national data repository for decision-making and patient care. *JMIR Public Health and Surveillance*, *7*(11), e26532. https://doi.org/10.2196/26532

[4] Azzi, S. (2025). A methodology for building medical ontology with limited domain experts' involvement. *Journal of Biomedical Semantics*.

[5] Eze, O. I., Iseolorunkanmi, A., & Adeloye, D. (2024). *The National Health Insurance Scheme (NHIS) in Nigeria: Current issues and implementation challenges*. *Journal of Global Health Economics and Policy, 4*, e2024002. https://doi.org/10.52872/001c.120197

[6] Grignon, M., & Chadeau-Hyam, M. (2018). A general primer for data harmonization. *Applied Health Economics and Health Policy*, *16*(2), 159–161. https://doi.org/10.1007/s40258-018-0382-3

[7] Jayathissa, P., & Hewapathirana, R. (2023). Enhancing interoperability among health information systems in low and middle-income countries: A review of challenges and strategies. *International Journal of Advances in Biology (IJAB)*, *10*(2/3), 1–11. https://doi.org/10.48550/arXiv.2309.12326

[8] Kayode, O., et al. (2024). A systematic review and meta-data analysis of clinical data repositories in Africa and beyond: Recent development, challenges, and future directions. *African Journal of Health Informatics*.

[9] Khnaisser, C., Lavoie, L., Fraikin, B., Barton, A., Dussault, S., Burgun, A., & Ethier, J.-F. (2022). Using an ontology to derive a sharable and interoperable relational data model for heterogeneous healthcare data and various applications. *Methods of Information in Medicine*, *61*(S 02), e151–e167. https://doi.org/10.1055/a-1877-9498

[10] Kolajo, O. J. (2025). *Enhanced ontology for health insurance data harmonization: A framework for achieving universal health coverage* (Doctoral progress report). Department of Computer Science, University of Abuja.

[11] Kush, R. D., Warzel, D., Kush, M. A., Sherman, A., Navarro, E. A., Fitzmartin, R., & Hudson, L. (2020). FAIR data sharing: The roles of common data elements and harmonization. *Journal of Biomedical Informatics, 107*, 103421. https://doi.org/10.1016/j.jbi.2020.103421

[12] Lee, K. H., Jang, S., Kim, G. J., Park, S., Kim, D., Kwon, O. J., Lee, J. H., & Kim, Y. H. (2025). *Large language models for automating clinical trial criteria conversion to OMOP CDM queries: Validation and evaluation study*. *JMIR Medical Informatics, 13*, e71252. https://doi.org/10.2196/71252

[13] Lee, Y. A., Lu, Y., Bian, J., Guo, J., & He, X. (2025). Transforming the Medicare claims data to the OMOP Common Data Model. *Studies in Health Technology and Informatics*, *329*, 1570-1571. https://doi.org/10.3233/SHTI251106

[14] Liyanage, H., Krause, P., & de Lusignan, S. (2017). Using ontologies to improve semantic interoperability in health data. *Journal of Innovation in Health Informatics, 22*(2), 309–315. https://doi.org/10.14236/jhi.v22i2.159

[15] Lui, S. J., Xiang, C., & Krishnaswamy, S. (2024). Knowledge aware automated health claims processing with medical ontologies and large language models. *Proceedings of the AAAI Symposium Series*, *4*(1), 199–209. https://doi.org/10.1609/aaaiss.v4i1.31405

[16] Malik, M. S., & Inau, E. T. (2023). Intelligent ontology alignment to improve semantic data integration across research domains in biomedicine. *ResearchGate*. https://doi.org/10.13140/RG.2.2.14810.11204

[17] National Health Insurance Authority (NHIA). (2024). *Annual enrollment statistics and regulatory update*. NHIA Press.

[18] Nieto, N., Eickhoff, S. B., Jung, C., Reuter, M., Diers, K., Kelm, M., Lichtenberg, A., Raimondo, F., & Patil, K. R. (2024). *Impact of leakage on data harmonization in machine learning pipelines in class imbalance across sites*. arXiv. https://doi.org/10.48550/arXiv.2410.19643

[19] Oguntunde, P. E., Alamu, F. O., Khakhalia, S. S., & Odetunmibi, O. A. (2023). A comparative study of several classification metrics and their performances on data. *Heritage and Sustainable Development*, *5*(1), 51–62. https://doi.org/10.37868/hsd.v5i1.216

[20] Okuzu, O., Malaga, R., Okereafor, K., Amos, U., Dosunmu, A., Oyeneyin, A., Adeoye, V., Sambo, M. N., & Ebenso, B. (2022). *Role of digital health insurance management systems in scaling health insurance coverage in low- and middle-income countries: A case study from Nigeria*. *Frontiers in Digital Health, 4*, Article 1008458. https://doi.org/10.3389/fdgth.2022.1008458

[21] Okuzu, O., Malaga, R., Okereafor, K., Amos, U., Dosunmu, A., Oyeneyin, A., Sambo, M. N., & Ebenso, B. (2022). Role of digital health insurance management systems in scaling health insurance coverage in low- and middle-income countries: A case study from Nigeria. *Frontiers in Digital Health, 4*, 1008458. https://doi.org/10.3389/fdgth.2022.1008458

[22] Onat, M., Gürpınar, C., Dogukan, A., & Tekin, M. C. (2024). Ontology- and LLM-based data harmonisation for federated learning. In *Proceedings of the 16th International Conference on Agents and Artificial Intelligence - Volume 1* (pp. 517–524). SCITEPRESS. https://doi.org/10.5220/0012356500003636

[23] Pietranik, M., & Kozierkiewicz, A. (2021). Methods of managing the evolution of ontologies and their alignments. *Journal of Intelligent Information Systems*, *57*(1), 147–173. https://doi.org/10.1007/s10844-021-00639-6

[24] Ravi, N., Thomas, C., & Odogwu, J. (2024). *How to reload and upgrade digital health to serve the healthcare needs of Nigerians*. *Frontiers in Digital Health, 5*, Article 1225092. https://doi.org/10.3389/fdgth.2023.1225092

[25] Van Damme, P., Fernández-Breis, J. T., Benis, N., Miñarro-Giménez, J. A., de Keizer, N. F., & Cornet, R. (2022). Performance assessment of ontology matching systems for FAIR data. *Journal of Biomedical Semantics*, *13*(1), 19. https://doi.org/10.1186/s13326-022-00273-5

[26] Vorisek, C. N., Klopfenstein, S. A. I., Lobe, M., Schmidt, C. O., Mayer, P. J., Golebiewski, M., & Thun, S. (2024). Towards an interoperability landscape for a national research data infrastructure for personal health data. *Scientific Data*, *11*(1), 772. https://doi.org/10.1038/s41597-024-03615-3

[27] World Health Organization. (2015). Tracking universal health coverage: First global monitoring report. WHO.

[28] Yu, J. K., Martínez-Romero, M., Horridge, M., Akdogan, M. U., & Musen, M. A. (2023). A general-purpose data harmonization framework: Supporting reproducible and scalable data integration in the RADx Data Hub. *AMIA Annual Symposium Proceedings*, *2023*, 780–789.

[29] Zand, A., Sharma, A., Xu, Z., & Ruan, Y. (2023). Robust automated harmonization of heterogeneous data through ensemble machine learning: Algorithm development and validation study. *JMIR Medical Informatics*, *11*, e44502. https://doi.org/10.2196/44502

[30] Zhang, S., Cornet, R., & Benis, N. (2024). *Cross-Standard Health Data Harmonization using Semantics of Data Elements*. *Scientific Data, 11*, 1407. https://doi.org/10.1038/s41597-024-04168-1