International Journal of Emerging Trends in Computer Science and Information Technology
ISSN: 3050-9246 | https://doi.org/10.63282/3050-9246.IJETCSIT-V6I4P120
Eureka Vision Publication | Volume 6, Issue 4, 134-138, 2025

Original Article

Data Lake Governance – Establishing a Single Source of Truth in Healthcare Enterprises

Selvakumar Kalyanasundaram
Independent Researcher,Texas, USA.

Received On: 27/10/2025 ***Revised On: 19/11/2025*** ***Accepted On: 27/11/2025*** ***Published On: 05/12/2025***

Abstract - Healthcare organizations generate enormous volumes of multi-modal data electronic health records (EHR), pharmacy claims, medical claims, imaging, genomics, IoT sensor streams, and administrative data. However, fragmented systems prevent efficient data sharing, analytics, and decision-making. A well-governed healthcare data lake provides a scalable architecture to integrate structured and unstructured data while maintaining quality, security, and compliance. This paper proposes a comprehensive governance framework enabling a unified Single Source of Truth (SSOT) for healthcare enterprises. The framework integrates metadata management, data lineage, interoperability standards, AI-driven quality checks, and federated access controls. The proposed model ensures trustworthy, timely, and regulated data access for clinical, operational, and financial use cases including population health, pharmacy benefit optimization, risk scoring, and value-based care. The framework further incorporates ethical safeguards to mitigate AI bias, enforce algorithmic fairness, and ensure transparency and accountability in all automated governance decisions.

Keywords - Data Lake, Healthcare Analytics, Data Governance, Single Source of Truth, Interoperability, Data Quality, Metadata Management, HIPAA, Value-Based Care.

1. Introduction

Healthcare enterprises face exponential growth of heterogeneous datasets generated from clinical workflows, payer systems, provider networks, pharmacy operations, and patient engagement channels. Traditional data warehouses are rigid and highly schema-dependent, limiting the ingestion of complex data types such as medical images, unstructured clinician notes, HL7/FHIR streams, and remote patient monitoring (RPM) data. To address these challenges, healthcare organizations are adopting data lakes capable of storing petabyte-scale data in raw, curated, and consumer-ready zones. However, without proper governance, a data lake can degrade into a “data swamp” unstructured, untrustworthy, and unusable. This paper develops a structured governance architecture to establish a Single Source of Truth (SSOT) within healthcare enterprises, enabling unified clinical intelligence, cost containment initiatives, and regulatory compliance.

2. Background and Related Work

Recent studies emphasize the need for unified analytics environments in healthcare. Literature also highlights the limitations of isolated departmental data marts and warehouse architectures.

2.1. Healthcare Data Complexity

Prior researches stresses data heterogeneity. Healthcare data comes in many different formats, structures, and sources, which makes integration and analysis difficult. Medical claims (ICD-10, CPT, HCPCS codes), Pharmacy claims (NDC, dosage, refill info), Lab results and EHR vital signs data are in structured format.HL7 Messages, FHIR bundles (JSON/XML), Device data and insurance eligible files are in semi structured format. Doctor notes, patient messages, PDF and scanned reports, radiology images and pathology slides are in unstructured format. This shows that healthcare data is extremely diverse and inconsistent, making consistent analytics and governance challenging.

2.2. Existing Governance Approaches

In healthcare, governance frameworks are more complex because of HIPAA, HITRUST, PHI/PII protection, FDA/clinical audit requirements, Interoperability mandates (HL7, FHIR). Existing governance model focus on Master Data Management which ensures that key business entities have one clean, consistent definition across the enterprise, preventing duplication and conflicts. This can be achieved by matching, merging, and deduplication algorithms. The accuracy, completeness, timeliness, Validity and consistent of the data been maintained by using the tools like Talend DQ, Informatica DQ, Great Expectations. Few companies started using AI driven anomaly detectors. Through Metadata management and Data Cataloging, the healthcare data been classified based on sensitivity. For each dataset, data stewards/ owners (primary + secondary owners assigned. Last refresh timestamp been maintained along with lineage. Tools like Collibra, alation, Informatica EDC, Google Data Catalog, AWS Glue Data catalog, Azure Purview, etc., are some of the

common tools used for Metadata management. Data lineage and traceability are mandatory components of healthcare data governance because they ensure regulatory compliance, clinical accuracy, and operational transparency across the data lifecycle. Healthcare data moves through multiple systems like EHRs, claims processors, pharmacy systems, analytics platforms, and data lakes. As a result, they undergo numerous transformations that directly impact patient care, reimbursement, and reporting. Frameworks such as HIPAA, HITRUST, CMS RADV, and ONC Interoperability Rules require organizations to maintain verifiable provenance of all protected health information (PHI), including where the data originated, how it was transformed, and who accessed or modified it. Lineage enables auditability of clinical and financial calculations, supports risk-adjustment validation, ensures correct mapping of coding systems (ICD-10, CPT, NDC, LOINC), and allows rapid root-cause analysis when discrepancies arise. Without robust lineage, organizations cannot guarantee data integrity, validate analytics outputs, or meet regulatory obligations, ultimately compromising patient safety and undermining trust in enterprise data assets.

In healthcare environments, the protected health information (PHI), claims data, lab results, and medication histories must be tightly safeguarded. Role Based Access Control (RBAC) assigns permissions based on job roles rather than individuals. By standardizing permissions through predefined roles, RBAC reduces the risk of unauthorized disclosure, minimizes insider threats, supports HIPAA and HITRUST compliance. It simplifies audits by creating traceable, consistent access patterns. However, few frameworks address healthcare-specific regulations (HIPAA, HITRUST), multi-source interoperability, and cross-benefit integration (pharmacy + medical). This paper fills the gap by presenting a healthcare-centric, AI-enabled data lake governance architecture.

3. Data Lake Architecture in Healthcare

Before implementing the proposed governance framework, the organization operated a traditional data warehouse centric architecture built around an on-premises Datawarehouse environment. Governance was largely implicit and manual. Metadata was scattered across ETL job definitions, spreadsheets, and tribal knowledge held by individual teams. There was no centralized metadata catalog, limited business glossary coverage, and no end-to-end lineage visibility from source to report. Data quality rules were implemented ad hoc within ETL jobs, without formal trust scoring or standardized monitoring. Access control to PHI was implemented at the database and schema level, but without granular role-based policies tailored to modern least-privilege or Zero-Trust principles. As a result, the organization experienced duplicated logic, conflicting metrics across departments, long reconciliation cycles, and delayed regulatory reporting.

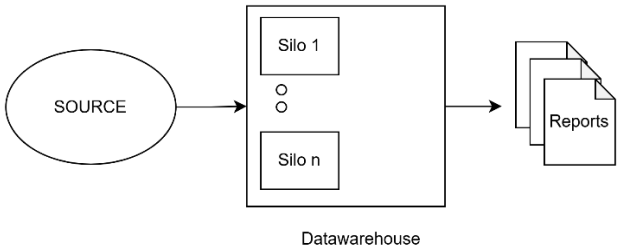

Figure 1. Enterprise Legacy Data pipeline

A modern healthcare data lake is architected as a multi-layered ecosystem, with each zone performing a distinct governance and transformation function to ensure data quality, compliance, and analytic readiness. The Raw Zone serves as the initial landing environment where ingested data is preserved in its native format. Few examples are HL7 v2 messages, FHIR bundles, flat files, JSON payloads, and DICOM imaging.

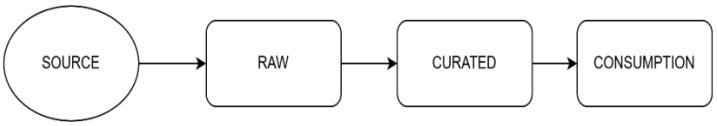

Figure 2. Proposed Enterprise Data pipeline

The Curated Zone applies rigorous data engineering and governance processes such as schema harmonization, cross-mapping of clinical coding systems (e.g., ICD to SNOMED, NDC to RxNorm), and automated de-identification pipelines that enforce HIPAA requirements. This zone also supports curation workflows that resolve inconsistencies and normalize semantic structures. The Curated Zone consolidates and integrates datasets to produce high-value analytic assets, including pharmacy-medical record linkages, patient-level longitudinal profiles, encounter-centric datasets, and enriched features incorporating social determinants of health (SDOH). Analytics and Consumption Zone provide governed, role-appropriate access for downstream use cases such as population-health dashboards, predictive risk scoring, fraud detection, and value-based care (VBC) insight generation. Together, these layers form a robust governance framework that transforms raw clinical data into trusted, actionable intelligence for healthcare decision-making.

4. Proposed Governance Framework

The proposed healthcare data governance model is built on eight foundational pillars designed to meet the complex regulatory, clinical, and operational needs of modern healthcare enterprises.

4.1. Metadata Management

Forms the backbone of discoverability and transparency, supported by a centralized metadata catalog that documents table and column definitions, data types, PHI/PII

classifications, domain ownership, and quality scorecards. Data can be logically grouped based on characteristics, usage, and context. Thereby they are classified as Data domains. The three level data domains are.

- L1 Domain - Enterprise domain which represent broad categories of data such as customer, clinical and finance.
- L2 Domain - Smaller areas within L1 such as Dental, Vision within Claims (L1)
- L3 Domain - More detailed datasets within L2 domains

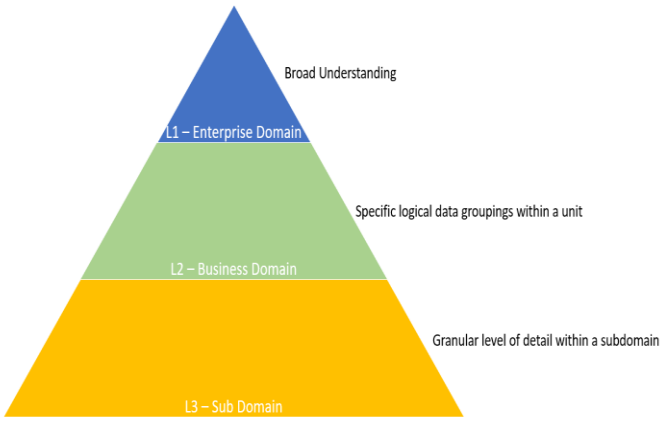

Figure 3. Enterprise Data Domain Hierarchy

This grouping establishes clear and comprehensive data domains across all levels to ensure the data is organized in a way that supports the business operations and objectives.

4.2. AI-enhanced scanning

Further enriches metadata through auto-schema detection, entity matching, and NLP-based tagging of clinician notes.

4.3. Data Lineage and Traceability

Provides full end-to-end visibility from source systems through ingestion, transformation and consumption and capturing the domain specific flows such as ICD-to-risk-score mappings and NDC-to-utilization analytics, with graph-based lineage enabling auditors to validate HIPAA and CMS compliance.

4.4. Interoperability and Standardization

Is maintained by adopting healthcare semantic standards including FHIR R4, RxNorm, LOINC, NPI, HL7/C-CDA, and X12 EDI. This ensures consistent meaning and unified analytics across payers, providers, pharmacies.

4.5. Data Quality Management

Integrates AI-driven rules to identify eligibility mismatches, duplicate claims, impossible clinical events, erroneous medication or diagnosis codes, outliers in cost/utilization trends, and temporal gaps in encounters, with each dataset assigned a dynamic 0–100 trust score.

4.6. Security, Privacy, and Compliance

Is reinforced through HIPAA-aligned access layers, attribute-based access control, data masking, tokenization, audit logging, and end-to-end encryption to protect PHI. Data is classified in four levels to dictate how the information is protected and encrypted. i.e., Restricted (SSN, HICN, etc.,) Confidential (Legal and Financial data), Proprietary (business artifact and Intellectual properties) and Public.

4.7. Master Data Management (MDM)

Ensures enterprise-wide consistency by establishing unified entities such as patient, provider, medication, and benefit masters. Federated Data Access and Zero-Trust Exchange support secure collaboration across payers, providers, pharmacies, TPAs, and ACOs, enforcing least-privilege access while enabling governed cross-organization data exchange. Finally.

4.8. AI-Assisted Governance

Strengthens reliability and automation by auto-classifying sensitive fields, detecting anomalies, reconstructing lineage, predicting quality degradation before SLA breaches, and auto-generating compliance documentation. Together, these pillars establish a comprehensive, scalable, and future-ready governance framework for healthcare data ecosystems.

5. Single Source of Truth (SSOT) Model

A Single Source of Truth (SSOT) ensures that all enterprise analytics, operational workflows, and decision support systems rely on the same validated and trusted datasets. At its core, the SSOT framework integrates several key components, including a Unified Patient Record that consolidates claims, clinical data, and social determinants of health (SDOH); a Golden Dataset Registry that designates authoritative data assets for enterprise use; and a Semantic Enterprise Data Model that harmonizes meaning across domains. A Certified Consumption Layer enables analysts and applications to access only vetted, quality-assured datasets, while version control for data which is like GitOps ensures transparency, reproducibility, and governance over dataset evolution. Together, these elements deliver significant benefits: reducing redundancy across systems, accelerating insight generation, supporting consistent clinical and financial decision-making, enhancing provider–payer coordination, and lowering overall operational costs.

6. Case Study: Healthcare Enterprise Implementation

6.1. Data source and Governance Scope

To evaluate the effectiveness of the proposed healthcare data governance framework, two high-impact enterprise domains were selected: Patient Data and Prescription

(Pharmacy) Data. These domains are among the most operationally critical and governance-sensitive due to their role in clinical decision-making, reimbursement, regulatory compliance, and patient safety.

Patient and prescription data were sourced from multiple enterprise systems to ensure comprehensive clinical and medication coverage. Patient data originated from EHR and EMR systems, eligibility and enrollment platforms, laboratory systems capturing clinical vitals, Social Determinants of Health (SDOH) data feeds, and provider directories with encounter records. Prescription data were obtained from pharmacy claims systems, National Drug Code (NDC) reference files, e-prescription platforms, medication adherence and refill datasets, and drug utilization and formulary management systems. These heterogeneous datasets were ingested through FHIR-enabled pipelines into the governed Raw, Curated, and Certified Consumption zones of the data lake architecture and were subjected to AI-assisted PHI classification, semantic code harmonization using standards such as ICD, SNOMED, RxNorm, NDC, and LOINC, Master Data Management (MDM), automated data quality validation, trust scoring with end-to-end lineage enforcement, and robust RBAC and Zero-Trust security controls to ensure privacy, accuracy, and regulatory compliance.

6.2. KPI Improvements After Governance Implementation

Table 1. Before vs After KPI Performance for Patient & Prescription Data

KPI	Before Governance	After Governance	Net Improvement
Data Ingestion Latency	4–8 hours (batch ETL across EHR & pharmacy systems)	~ 20 minutes (FHIR + streaming pipelines)	~ 95% reduction
Query Response Time	70-150 seconds (multi-mart joins across patient & Rx tables)	1–4 seconds (certified Big Query SSOT views)	~ 93% faster
Data Reconciliation Duration	7-14 business days per reporting cycle	1–2 business days	~ 88 % reduction
Data Accuracy Percentage	87–90%	98–99%	~ 9 % improvement
Duplicate Record Rate (Patient + Rx)	7–11%	< 0.8%	~ 90 % reduction

6.3. Comparative Study

Table 2. Comparative Study

Before Governance	After Governance
1. Siloed Teradata data marts 2. Manual reconciliation 3. Limited metadata documentation 4. No automated lineage 5. Inconsistent PHI access control 6. Delayed regulatory reporting	1. Unified Data Lake 2. Automated metadata catalog and Lineage 3. FHIR-based ingestion pipelines 4. AI-driven PHI detection 5. Centralized SSOT-certified datasets 6. Real-time audit-ready compliance

7. Challenges and Limitations

Despite its advantages, the governance framework introduces several challenges that healthcare enterprises must address. High initial implementation costs can be a barrier, particularly when establishing metadata systems, lineage tools, and security controls. The model also requires a highly skilled workforce, including data engineers, interoperability specialists, and privacy experts, to manage complex data flows. Multi-source coding systems further complicate operations, as mapping across ICD, SNOMED, LOINC, RxNorm, and proprietary payer codes demands meticulous semantic alignment. Vendor interoperability issues may arise when disparate systems fail to fully support standards such as FHIR or X12, leading to integration bottlenecks. Additionally, AI models embedded in governance processes require continuous monitoring and validation to prevent model drift, ensuring ongoing accuracy and regulatory compliance.

7.1. Ethical and Societal Implications of AI-Driven Data Governance

The integration of artificial intelligence into healthcare data governance introduces significant ethical responsibilities related to bias mitigation, algorithmic fairness, and model transparency. While AI-driven automation enhances scalability and efficiency in metadata classification, PHI detection, and data quality management, it also raises concerns regarding unintended discrimination, opaque decision-making, and potential regulatory misinterpretation.

7.1.1. AI Bias

AI models trained on historical healthcare data may inherit systemic biases related to race, gender, age, socioeconomic status, and access to care. In governance applications, such biases can manifest as unequal PHI detection accuracy, inconsistent data quality scoring across patient populations, or skewed anomaly detection in underserved groups. To mitigate this risk, the proposed framework mandates diverse training datasets, continuous bias audits, subgroup performance evaluation, and periodic revalidation of models using demographically representative samples. These controls help

ensure that governance automation does not exacerbate existing healthcare disparities.

7.1.2. Algorithmic Fairness

Algorithmic fairness is critical when AI systems influence access control, dataset certification, and regulatory compliance decisions. Biased trust scores or misclassification of sensitive attributes can indirectly affect patient inclusion in analytics, reimbursement modeling, and quality reporting. The framework enforces fairness through rule-based governance constraints layered above AI outputs, ensuring that automated decisions are systematically reviewed against HIPAA, CMS, and HITRUST requirements.

7.1.3. Model Transparency and Explainability

Transparency and interpretability are essential for regulatory compliance and organizational trust. Black-box AI models in governance pipelines may generate outcomes such as PHI classification, trust score assignment, or anomaly detection that are difficult to justify during audits. To address this, the framework integrates explainable AI (XAI) mechanisms, including SHAP and LIME, to provide feature-level attribution for governance decisions. All model actions are logged with audit-ready justifications, enabling healthcare compliance teams, regulators, and internal auditors to trace how specific outputs were produced. This ensures alignment with emerging regulatory expectations for algorithmic accountability in healthcare AI. Collectively, these ethical safeguards ensure that the proposed AI-driven governance architecture not only complies with technical and regulatory standards but also upholds principles of equity, transparency, and responsible AI deployment within healthcare data ecosystems.

8. Future Research Directions

Healthcare data governance is poised to evolve significantly, driven by the integration of advanced AI and distributed technologies. Future architectures will incorporate LLM-based semantic governance engines capable of interpreting clinical context, auto-classifying sensitive fields, and enforcing standards dynamically. Autonomous data quality pipelines will continuously detect and remediate errors without manual intervention, while reinforcement learning models will optimize data workflows by adapting transformation logic and resource allocation over time. Blockchain-enabled audit trails will introduce immutable, tamper-proof compliance records, strengthening trust and regulatory alignment. Additionally, real-time federated analytics across payer–provider ecosystems will support secure, zero-trust data collaboration, enabling population health insights and clinical decision-support without centralized data sharing. Together, these innovations will define the next generation of healthcare data governance.

9. Conclusion

A healthcare-specific data lake governance framework is essential to create a Single Source of Truth that supports clinical excellence, regulatory compliance, and advanced analytics. The proposed model integrates metadata management, interoperability standards, lineage, security, and AI-driven quality control. As healthcare shifts to value-based care and precision medicine, organizations must invest in robust governance to unlock reliable insights and reduce systemic inefficiencies.

References

[1] HL7 International, “FHIR Release 4,” 2021.
[2] CMS, “Risk Adjustment Data Validation,” 2020.
[3] IBM Healthcare, “AI in Data Governance,” 2023.
[4] Google Cloud Healthcare API Documentation, 2024.
[5] Khosla et al., “Interoperability in Healthcare Data Systems,” IEEE Access, 2022.
[6] HITRUST Alliance Framework, 2022.