Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

An EMOI-Based Object-Level Event Governance Framework for Multimodal Healthcare Data Integration and Clinical Research

Yueyue Li
First Affiliated Hospital of Army Medical University
Liu Yang
yangliu0623@tmmu.edu.cn
First Affiliated Hospital of Army Medical University
Fei Wang
First Affiliated Hospital of Army Medical University
Xiaofei Li
First Affiliated Hospital of Army Medical University

Research Article

Keywords: Multimodal medical data, data governance, EMOI, lake-warehouse integration, health and medical big data

Posted Date: November 1st, 2026

DOI: https://doi.org/10.21203/rs.3.rs-10810768/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Abstract

Background

To address critical bottlenecks in hospital-wide multimodal medical data management, specifically, inadequate data association, semantic heterogeneity, and low utilization efficiency—this study proposes a multimodal medical data governance framework tailored for clinical applications.

Methods

Utilizing 72 TB of real-world medical data from a tertiary hospital, we designed a layered, lakehouse-based governance architecture. Central to this framework is the Enterprise Multimodal Object Identifier (EMOI), a mechanism built upon the Enterprise Master Patient Index (EMPI) that establishes an object-level event association chain linking the “encounter–examination–report–diagnosis–treatment” workflow. By integrating a nine-step closed-loop governance process, the framework facilitates data cleansing, standard mapping, identity unification, object association, privacy protection, and normalized data asset management.

Results

The implementation of this platform generated 336 million standardized data records, achieving an EMPI consistency of 86.7% and an EMOI association rate of nearly 100%. Furthermore, data consistency, accuracy, and timeliness reached 99.63%, 97.78%, and 95.70%, respectively.

Conclusions

These validated outcomes have been successfully applied to support clinical research cohorts, 360-degree patient profiling, intelligent clinical decision support, and operational analytics. Ultimately, the EMOI-based object-level event governance mechanism effectively resolves the fragmentation of multimodal medical data, enabling full-lifecycle management from data integration and quality governance to downstream research utilization.

Scoping review registration:

As a scoping review drawing on retrospective institutional operational and clinical data, this study does not report any findings from medical intervention trials with human participants.

Background

In recent years, national strategies under the **Healthy China Initiative** have accelerated the development of health information technology. The *Healthy China 2030 Planning Outline* identifies health information technology as a strategic priority. In 2025, the World Health Assembly extended the *Global Strategy on Digital Health* through 2027, underscoring data interoperability, security, and ethics as shared goals for intelligent healthcare systems worldwide [1–3]. This extension highlights the enduring priority of health data governance [4, 5]. Furthermore, data protection regulations have been comprehensively implemented globally, including the EU's GDPR, the US's HIPAA, and China's *Data Security Law* and *Personal Information Protection Law*. Against this regulatory backdrop, achieving high-value health data circulation under secure and compliant conditions has become a fundamental challenge that smart hospital development must address [6].

Years of healthcare informatization in large tertiary general hospitals have led to the accumulation of extensive multimodal healthcare data resources across diverse clinical information systems, including Hospital Information Systems (HIS), Electronic Medical Records (EMR), Picture Archiving and Communication Systems (PACS), Laboratory Information Systems (LIS), and nursing information systems. These resources encompass heterogeneous data modalities such as clinical text, medical imaging, laboratory measurements, and electrophysiological signals [7–9]. Although interoperability standards based on HL7 Fast Healthcare Interoperability Resources (FHIR) [10] have enabled data aggregation at the physical level, the focus of hospital data governance has evolved from simple interoperability toward comprehensive data asset management. A central challenge in clinical data analytics is the effective integration and reconstruction of fragmented electronic health records across multiple modalities and time points [11]. EMPI frameworks unify patient identities across heterogeneous information systems through record-matching algorithms, thereby addressing the question of "**to whom does this data belong**". However, clinical research scenarios require a more granular understanding of "**during which episode of care was this data generated,**" together with object-level associations across multimodal records. Current EMPI-based approaches are unable to distinguish CT images, laboratory results, pathology findings, and treatment records generated during different clinical encounters for the same patient, making it difficult to establish a complete association chain of multimodal records corresponding to a single clinical event [12, 13]. Previous studies have shown that patient identity-based linkage strategies connect only approximately 43% of encounter records across heterogeneous systems, whereas encounter-centric linkage strategies can achieve complete coverage. These findings highlight the inherent limitations of patient identity identifiers in clinical research applications and identify them as a critical bottleneck restricting the efficient utilization of multimodal healthcare data for scientific research [14].

To address the aforementioned challenges, this study proposes an EMOI mechanism based on a real-world dataset comprising 72 TB of multimodal medical data from a tertiary hospital. Building upon the EMPI, the proposed approach establishes an object-level event association framework linking "encounter–examination–procedure execution–report–diagnostic decision-making," thereby endowing fragmented multimodal healthcare data with a coherent spatiotemporal semantic structure.

Furthermore, this study develops a "layered lakehouse-integrated governance framework" [15, 16] and a "nine-step closed-loop governance system" that enable unified management of multimodal healthcare data through standardized data acquisition, hybrid storage architecture, multimodal data fusion, and a security framework integrating "hierarchical protection with privacy-preserving computing" [17, 18]. The governance outcomes generated under this framework provide robust support for institutional clinical research platforms and intelligent clinical applications, offering a reproducible engineering paradigm for the construction of healthcare big data integration platforms and the enhancement of clinical research efficiency.

Methods

Study Setting and Data Sources

This study was conducted in a tertiary general hospital with approximately 1,200 operational beds. The study environment encompasses multiple heterogeneous clinical information systems, including HIS, EMR, PACS, LIS, and nursing information systems, forming a multimodal data repository spanning five major domains: clinical care, medical imaging, scientific research, hospital operations, and biobank resources (Table 1).

Table 1
Overview of multimodal healthcare data sources and governance characteristics

Data domain	Representative data	Data modality	Storage format	Primary standard / ontology	Targeted governance entity
Structured clinical data	Diagnoses, medications, procedures, laboratory results, examination reports	Structured	Relational tables, CSV	ICD-10, LOINC, HL7 FHIR	Encounter, diagnosis, treatment
Clinical documents	Admission notes, discharge summaries, operative reports, consultation notes	Semi-structured / Unstructured	XML, PDF, RTF, TXT	HL7 CDA, NLP terminology standards	Clinical event
Medical imaging	CT, MRI, CR, PET-CT	Image	DICOM	DICOM	Imaging examination
Physiological signals	ECG, EEG, bedside monitoring waveforms	Time-series signal	EDF, EDF+, XML	IEEE 11073 (where applicable)	Physiological observation
Operational data	Outpatient visits, inpatient statistics, billing records, service indicators	Structured	Tables, numerical records	Hospital management standards	Operational events
Research data	Research cohorts, study projects, omics datasets	Structured / Semi-structured	Tables, text	Research metadata schema	Research object
Biobank data	Specimen images, sample metadata, traceability information	Image + Metadata	JSON, image files	MIABIS (or institutional standard)	Biospecimen object

Note: CSV, comma-separated values; ICD-10, International Classification of Diseases 10th Revision; EMOI, Enterprise Multimodal Object Identifier; DICOM, Digital Imaging and Communications in Medicine; HL7 FHIR, Health Level Seven Fast Healthcare Interoperability Resources; LOINC, Logical Observation Identifiers Names and Codes.

The data collection period extended from January 2003 to February 2026 and included more than 1.7 million patients, comprising over 317 million records with a total data volume of 72 TB. Structured data accounted for approximately 2 TB (2.9%), whereas unstructured data comprised approximately 70 TB (97.1%), representing a highly complex heterogeneous data governance scenario.

All data included in this study were anonymized and de-identified prior to analysis. Data processing and analysis were performed on a privacy-preserving computing platform certified to China's Classified Protection Level □ standard. The study protocol was reviewed and approved by the institutional Medical Ethics Committee.

Layered Lakehouse-based Architecture and Governance Mechanism

Based on the spatiotemporal object indexing mechanism of the **EMOI**, an efficient integrated governance framework founded on a **five-tier progressive hybrid lakehouse architecture** was established (Fig. 1).

At the **data source layer (cross-system interoperability layer)**, standardized decoupling of more than 30 heterogeneous clinical information systems, including HIS, PACS, and LIS, was achieved in accordance with the Integrating the Healthcare Enterprise (IHE) framework using HL7 v2/v3/FHIR standard interfaces, JDBC connectivity, and database replication technologies. To address the sensitivity of healthcare data, a BERT-based sensitive information recognition module was embedded within the data acquisition pipeline to automatically identify 14 categories of protected health information, including patient identifiers and contact information, thereby providing a secure foundation for subsequent data governance.

During the **raw data persistence stage**, a homogeneous data layer (raw data persistence layer) was constructed based on the TiDB distributed database. Leveraging the strong consistency guarantees provided by the Raft consensus protocol, the platform ensures reliable data storage and high system availability while supporting the horizontal scalability and unified management of terabyte-scale multimodal healthcare data. This architecture effectively satisfies the high-concurrency access requirements of clinical and scientific research applications.

During the evolution of the **core homogeneous layer (multimodal feature fusion layer)**, the governance paradigm shifts from physical data storage to semantic association. Centered on the spatiotemporal indexing capability of the **EMOI**, the system performs multimodal data fusion by integrating knowledge graphs with Key–Value transformation strategies to convert free-text electronic medical records into structured representations.

A cross-attention-based alignment strategy is employed to project imaging spatial features extracted by **ResNet-50**, including lesion region-of-interest (ROI) characteristics such as density, morphology, and anatomical location, together with report text embeddings generated by a **Transformer encoder**, into a unified high-dimensional feature space. This approach enables semantic alignment between **pixel-level lesion features** and **terminology-level clinical descriptions**. Furthermore, the temporal evolution trajectories of multimodal feature vectors are quantified using the **Cosine Similarity** algorithm, establishing close associations among **imaging changes, disease progression narratives, and final diagnostic outcomes**. Consequently, fragmented clinical records are transformed into research objects with complete causal and temporal logical chains (Fig. 2).

During the transformation of data into standardized data assets, a **nine-step closed-loop quality control framework** was established. In the **“data cleansing–deduplication–mapping”** stage, an extensible rule repository was developed based on predefined medical business rules to automatically identify and correct outliers and logical inconsistencies, and missing values. In addition, **K-nearest neighbors (KNN)-based outlier detection** and **generative adversarial network (GAN)-based missing value imputation** were introduced to improve anomalous data detection and enhance the longitudinal completeness of patient disease trajectories.

During the **“format conversion–EMPI construction–EMOI construction”** stage, the governance framework adopted **ICD-10** for hierarchical disease classification and **LOINC** for standardized laboratory and examination coding as its underlying ontologies, while achieving deep semantic alignment with multiple industry standards. A weighted probabilistic matching algorithm was employed to construct the **EMPI**, enabling unified patient identity management across heterogeneous information systems. Building upon the EMPI, the proposed **EMOI** innovatively establishes object-level event association cohorts linking **patient–encounter–examination/imaging–diagnosis/treatment**, thereby enabling fine-grained multimodal clinical data integration, as outlined in the nine-step workflow (Fig. 3).

During the **“Separation–De-identification–Normalization”** stage, a defense-in-depth framework based on the "Onion Model" was established (Fig. 4). The system automatically identifies **Protected Health Information (PHI)** [19] and implements a five-level field classification management strategy, with sensitive data stored in isolated environments to reduce the risk of information leakage. Supported by the **SM4 national cryptographic algorithm** and a privacy-preserving computing framework characterized by “**models move while data remain local**” [20], the framework prevents cross-domain transfer of raw data and enables secure sharing of model parameters, thereby providing a foundation for cross-institutional collaborative research and federated analytics (Table 2).

Table 2
Security classification and governance strategy for multimodal healthcare data

Security level	Data category	Representative attributes	Governance strategy	Privacy protection mechanism
Level 1	Basic demographic information	Age, sex, region, education level	Restricted access	Symmetric encryption (SM4) + role-based access control (RBAC)
Level 2	Direct identifying information	Name, national identity number (ID), passport number, medical insurance ID	Identity isolation	SM4 encryption + pseudonymization
Level 3	Quasi-identifying information	Occupation, employer name, workplace address, zip code	Context-aware access control	Dynamic de-identification
Level 4	Clinical health information	Diagnoses, laboratory results, medications, operative records, pathology reports	EMOI-based object association before de-identification	Dynamic masking + controlled access
Level 5	Sensitive personal information	Genetic data, religious beliefs, criminal records, substance abuse history, infectious disease status	Maximum protection	Full-lifecycle encryption + privacy-preserving computation (e.g., MPC/FL)

Note: SM4 is a certified commercial symmetric block cipher standard in China, equivalent to AES (Advanced Encryption Standard).

Evaluation Metrics

This study established an evaluation framework encompassing **data quality, governance efficiency, and application effectiveness** by referring to the **International Society for Quality in Health Care (ISQua) Data Quality Evaluation Framework**, the **Healthcare Information and Management Systems Society (HIMSS)**, and the **Analytics Maturity Assessment Model (AMAM)** (Table 3).

Table 3
The evaluation framework for the proposed multimodal data governance system

Evaluation dimension	Evaluation metric	Definition	Measurement method	Reference standard
Data quality	Core field compliance rate	Percentage of mandatory fields satisfying governance rules	Random sampling and automated validation	ISO 8000
	Cross-system consistency	Consistency of identical entities across heterogeneous systems	EMPI-based record matching	ISO 8000
	Data freshness	Time interval between data generation and synchronization	Timestamp comparison	DQAF
	EMOI association rate	Percentage of multimodal records successfully linked to complete clinical events	EMOI event chain validation	Proposed method
Governance efficiency	Mean time to insight (MTTI)	Average time from data request to research-ready cohort generation	System log analysis	Internal benchmark
	Cross-system integration cycle	Average duration required for heterogeneous system integration	Interface monitoring logs	Operational statistics
Research support	Research data service response time	Average latency of cohort retrieval services	Platform monitoring	Service-level agreement
	PHI de-identification accuracy	Accuracy of automatic PHI identification and masking	Manual verification and sampling	HIPAA
	multimodal semantic coverage	Coverage of cross-modal event associations generated by EMOI	Knowledge graph validation	Proposed method

Note: DQAF, Data Quality Assessment Framework; EMOI, Enterprise Multimodal Object Identifier; EMPI, Enterprise Master Patient Index; HIPAA, Health Insurance Portability and Accountability Act; ISO, International Organization for Standardization; MTTI, Mean Time to Insight; PHI, Protected Health Information.

The **data quality** dimension adopted the core subset of the **Data Quality Assessment Framework (DQAF)**, including key indicators such as error rates of core clinical variables, cross-system data consistency, data update timeliness, and compliance rates of core data fields in accordance with the **ISO 8000** standard.

The **governance efficiency** dimension introduced the **MTTI** as a key performance metric to quantify the average duration from the initiation of a raw data request to the generation of a structured research cohort. This metric encompasses both the cross-system integration cycle and the time required for research data preparation and organization.

The **research support** dimension was evaluated using indicators including the response time for scientific research data services and the reduction in data preparation time, thereby reflecting the preliminary effectiveness of the proposed framework in supporting institutional clinical research applications.

Results

Data Asset Scale and EMOI-Based Association Performance

An engineering-oriented quantitative evaluation was conducted to assess the implementation of the proposed layered lakehouse-integrated governance framework in a tertiary hospital. The platform successfully integrated **19 categories of medical data**, establishing **81 standardized data tables** comprising **1,479 data fields**. The integrated data assets encompass the core clinical modules of hospital practice, including medical record front pages, laboratory tests, pathology, patient encounters, medical orders, healthcare expenditures, nursing records, surgical and anesthesia records, diagnostic examinations, and biobank specimens.

The **EMPI weighted probabilistic matching algorithm** consolidated **1.891 million** pre-merging patient identifiers into **1.639 million** unique patient identity records, eliminating approximately **251,000** duplicate identifiers and achieving a **deduplication rate of 13.3%**. The resulting **EMPI consistency rate reached 86.7%**, while the **EMOI association rate achieved 100%**, enabling the successful construction of a comprehensive knowledge graph linking **patients, examinations, medical images, and diagnoses**.

Approximately **105 million** raw records were initially ingested into the homogeneous data layer. Through multimodal semantic mapping based on the combined **EMPI–EMOI** framework in the core homogeneous layer, these records were expanded into **337 million event-associated linked records**. Following governance and standardization, the data were ultimately reconstructed into **336 million high-quality standardized records** within the data asset layer, establishing a fully traceable lifecycle spanning **data acquisition, aggregation, standardization, and assetization**.

Data Quality and Multidimensional Governance Effectiveness

To evaluate the effectiveness of the proposed nine-step closed-loop governance framework in improving data governance efficiency, data quality was assessed across five dimensions. The results demonstrated a **data consistency of 99.63%**, **accuracy of 97.78%**, and **timeliness of 95.70%**, all of which reached or

exceeded the corresponding industry benchmarks. The **standardization rate was 79.09%**, slightly above the typical industry range. Although the **overall compliance rate was 88.73%**, the average compliance rate for core clinical modules, including diagnostic examinations, laboratory tests, and medical orders, reached **92.63%**. The overall mean score across the five quality dimensions was **82.21%**.

With respect to interface decoupling and system interoperability, the adoption of **HL7/ FHIR-compliant interfaces** and a unified data integration mechanism reduced the average cross-system integration cycle from **9–14 days to 1–2 days**, substantially improving data acquisition and integration efficiency (Fig. 5).

Security and Compliance Framework

Supported by the platform's security governance architecture, refined management and compliant utilization of data assets were achieved. Based on sensitivity levels, data were classified into five categories: **Level 1 (public), Level 2 (management and research), Level 3 (service recipient notification), Level 4 (personalized services and management), and Level 5 (special disease diagnosis and treatment)**. PHI fields, including 14 categories such as names, identification numbers, and contact information, were accessible under role-based authorization within Levels 1–3, with an automatic classification and identification accuracy of **96.4%** for sensitive PHI fields. For core data classified as Levels 4–5, the coverage of encrypted storage using the SM4 national cryptographic algorithm and dynamic data masking approached 100%. In cross-institutional collaboration scenarios, the privacy-preserving joint research model based on a secure computation framework has completed preliminary technical validation. The full lifecycle management of data storage and processing complies with the controlled security and privacy requirements stipulated by the **Data Security Law of the People's Republic of China**, the **Personal Information Protection Law of the People's Republic of China**, and the **Health Insurance Portability and Accountability Act (HIPAA)**, thereby providing a compliant foundation for secure data sharing and collaborative research.

Effectiveness of Clinical Research Support

The establishment of a standardized multimodal data asset repository provides high-value data resources for **real-world research (RWR)** in clinical settings. Leveraging the **nine-step closed-loop governance framework** and the multimodal feature processing pipeline, the preparation time for research cohorts—including cohort inclusion and exclusion, dynamic de-identification, data standardization, and **EMOI-based event association**—was reduced from an average of **30 days** under the conventional manual data extraction workflow to **2–3 days**.

This framework effectively addresses the challenge of cross-modal temporal alignment among longitudinal electronic medical record narratives, terabyte-scale **DICOM** image series, and high-frequency physiological signals, reducing the acquisition cycle for complex multimodal research cohorts from the **monthly** scale to the **daily** scale.

The institutional clinical research big data retrieval platform maintains a stable query response time of **less than 20 seconds** and serves approximately **100 user sessions per month**, providing researchers with integrated capabilities for data retrieval, cohort construction, and phenotypic analysis. In addition, the platform offers **artificial intelligence-assisted application programming interfaces (APIs)** that support data mining and intelligent analytics, substantially improving both the efficiency and regulatory compliance of clinical research data acquisition.

Clinical Applications and Operational Management Outcomes

At the **clinical application level**, the **360-degree patient view** constructed based on the **Clinical Data Repository (CDR) integrates 25 categories of longitudinal clinical data** spanning the entire continuum of care and has cumulatively supported nearly **100,000** comprehensive clinical data aggregation tasks. The **specialty-specific intelligent view**, equipped with decision-support components such as disease risk early warning, intelligent recognition, and side-by-side lesion comparison, provides clinicians with services including medical report interpretation, medical record quality control, and intelligent question-answering. Within four months of deployment, the system was used approximately **42,000 times**, demonstrating the effective translation of multimodal data governance achievements into intelligent clinical applications.

At the **hospital operational management level**, the operational decision support system covers **eight administrative departments**, incorporates **more than 260 operational performance indicators together** with **58 specialized indicators across three medical technical departments**, and supports a **four-level analytical framework** encompassing hospital-level, departmental-level, operational personnel-level, and detailed business-level analyses. Combined with **mobile business intelligence (BI)** applications and real-time executive dashboards, the platform facilitates the transition of hospital operations management toward a **real-time, data-driven, and fine-grained management paradigm**.

Conclusions

Practical implementation demonstrated that the compliance rate of key data fields reached **≥ 99.99%**, while cross-system data consistency achieved **99.6%**. Leveraging the standardized multimodal data assets generated by the proposed framework, the time required for research cohort acquisition was reduced from the **monthly** scale to the **daily** scale, while simultaneously **providing** direct support for institutional intelligent clinical applications. These findings demonstrate that **object-level data governance for research-oriented healthcare data not only enhances the efficiency of clinical research but also creates substantial value for clinical decision support and hospital operational management**. The benefits of the proposed framework are primarily reflected in improvements to research cohort extraction, eligibility screening, and data field cleansing. However, systematic evaluation in downstream applications, such as **multimodal artificial intelligence model development** and **disease-specific database construction**, remains limited. Furthermore, although the **GAN-based missing value imputation**

module could be further optimized for data fields with **very low missing rates (< 10%)**, the overall implementation provides compelling evidence for the feasibility, practicality, and scalability of the proposed methodology and engineering framework, offering a valuable reference for multimodal healthcare data governance in real-world clinical settings.

Abbreviations

AMAM Adoption Model for Analytics Maturity

API Application Programming Interface

BI Business Intelligence

CDR Clinical Data Repository

DICOM Digital Imaging and Communications in Medicine

DQAF Data Quality Assessment Framework

EMOI Enterprise Multimodal Object Identifier

EMP Enterprise Master Patient Index

EMR Electronic Medical Record

GAN Generative Adversarial Network

HIMSS Healthcare Information and Management Systems Society

HIPAA Health Insurance Portability and Accountability Act

HIS Hospital Information System

HL7 FHIR Health Level Seven Fast Healthcare Interoperability Resources

IHE Integrating the Healthcare Enterprise

ISQua International Society for Quality in Health Care

KNN K-Nearest Neighbors

LIS Laboratory Information System

LOINC Logical Observation Identifiers Names and Codes

MTTI Mean Time to Insight

PACS Picture Archiving and Communication System

PHI Protected Health Information

ROI Region of Interest

SM4 Commercial Cryptographic Algorithm Standard (China)

Declarations

Ethics approval and consent to participate

This study was conducted in accordance with the principles of the Declaration of Helsinki. This study was approved by the Institutional Review Board of the Jiangbei Campus of the First Affiliated Hospital of Army Medical University (Approval No. ER2026KY046). The need for informed consent was waived by the Jiangbei Campus of the First Affiliated Hospital of Army Medical University due to the retrospective nature of the study and the use of fully de-identified patient data.

Consent for publication

Not applicable.

Competing interests

The authors declare that they have no competing interests.

Authors' information

[1] Department of Information, The Jiangbei Campus, The First Affiliated Hospital of Army Medical University, Chongqing, China.

[2] Artificial Intelligence and Big Data Center, The First Affiliated Hospital of Army Medical University, Chongqing, China.

Corresponding author: Liu Yang (Email: yangliu0623@tmmu.edu.cn).

Funding

This study was supported by the Jiangbei Campus of the First Affiliated Hospital of Army Medical University's Innovation Talent Cultivation Fund Project for 2024 (Grant No. 2025Y49).

Author Contribution

YYL, LY, FW, and XFL: conceptualization. YYL: methodology, data curation, formal analysis, writing – original draft. LY and FW: supervision, validation, writing – review and editing. XFL: visualization, data curation. All authors read and approved the final manuscript.

Acknowledgements

Not applicable

Data Availability

The datasets generated and/or analysed during the current study are not publicly available due to hospital data security policies, patient privacy regulations, and institutional data governance restrictions. However, de-identified data may be available from the corresponding author upon reasonable request and with approval from the relevant institutional authorities.

References

1. Vorisek CN, Lehne M, Klopfenstein SAI, Mayer PJ, Bartschke A, Haese T, et al. Fast healthcare interoperability resources (FHIR) for interoperability in health research: Systematic review. JMIR Med Inf. 2022;10(7):e35724.
2. Moody J, Johnson AEW, Ghassemi M, et al. MIMIC-IV on FHIR: Converting a decade of in-patient data into an exchangeable, interoperable format. J Am Med Inf Assoc. 2025;32(9):1701–10.
3. Wang MY, Li SY, Zheng T, et al. Design, development, and application. JMIR Med Inf. 2022;10(4):e36481. Big data health care platform with multisource heterogeneous data integration and massive high-dimensional data governance for large hospitals.
4. Mallappallil M, Sabu J, Gruessner A, Salifu M. A review of big data and medical research. SAGE Open Med. 2020;8:1–10.
5. Liu S, Wang X, Hou Y, Li G, Wang H, Xu H, et al. Multimodal data matters: Language model pre-training over structured and unstructured electronic health records. IEEE J Biomed Health Inf. 2022;27(1):504–14.
6. Sharma H, Pundir S, Deepak A, Mayuri K, Dwivedi SP, Kumar N. Multi-modal data fusion using transfer learning in big data analytics for healthcare. In: 2023 International Conference on Artificial Intelligence for Innovations in Healthcare Industries (ICAIIHI). IEEE; 2023. pp. 1–7.
7. Khan IH, Javaid M. Big data applications in the medical field: A literature review. J Ind Integr Manag. 2021;6(1):53–69.

8. Zhang Y, Sheng M, Liu X, Wang R. A heterogeneous multi-modal medical data fusion framework supporting hybrid data exploration. Health Inf Sci Syst. 2022;10(1):22.
9. Wang B, Hao XB, Yan XY, Lai JK, Jin FF, Liao XW, et al. Evaluation of the clinical application effect of eSource record tools for clinical research. BMC Med Inf Decis Mak. 2022;22(1):98.
10. Osamika D, Adebayo O, Oluwaseun T. A critical review of health data interoperability standards: FHIR, HL7, and beyond. World Sci News. 2025;203:194–233.
11. Davila MA 3rd, Sholle ET, Fuld X, Israel ML, Cole CL, Campion TR Jr. Linking patient encounters across primary and ancillary electronic health record systems: A comparison of two approaches. ACI Open. 2024;8(1):e43–8.
12. Grannis SJ, Williams JL, Kasthuri S, Murray M, Xu H. Evaluation of real-world referential and probabilistic patient matching to advance patient identification strategy. J Am Med Inf Assoc. 2022;29(8):1409–15.
13. AlSaad R, Abd-alrazaq A, Boughorbel S, Ahmed A, Renault MA, Damseh R, et al. Multimodal large language models in health care: Applications, challenges, and future outlook. J Med Internet Res. 2024;26:e59505.
14. Petri V, Tran A, Baranton E, Haas H, Freyssinet E, Vrzáková H. Clinical data integration and processing challenges in healthcare caused by contemporary software design. Digit Health. 2025;11:20552076251374233.
15. Armbrust M, Ghodsi A, Xin R, Zaharia M, Lakehouse. A new generation of open platforms that unify data warehousing and advanced analytics. In: Proceedings of the 2021 Conference on Innovative Data Systems Research (CIDR ‘21); 2021.
16. Yamashita T, Shibata SI, Takada A, Hotta T, Izukura R, Kang D, et al. A governance framework for medical code standardization to enhance multi-institutional data quality. BMC Med Inf Decis Mak. 2026;26(1):99.
17. Poudel P, Shrestha P, Amgain S, Shrestha YR, Gyawali P, Bhattarai B. CAR-MFL: Cross-modal augmentation by retrieval for multimodal federated learning with missing modalities. Medical Image Computing and Computer Assisted Intervention – MICCAI 2024. Cham: Springer Nature Switzerland; 2024. pp. 102–12.
18. Satti FA, Ali T, Hussain J, Khan WA, Khattak AM, Lee S. Ubiquitous Health Profile (UHPr): A big data curation platform for supporting health data interoperability. Computing. 2020;102(11):2409–44.
19. Yang X, Chen A, PourNejatian N, Shin HC, Smith KE, Parisien C, et al. A large language model for electronic health records. NPJ Digit Med. 2022;5(1):194.
20. Rieke N, Hancox J, Li W, Milletarì F, Roth HR, Albarqouni S, et al. The future of digital health with federated learning. NPJ Digit Med. 2020;3:119.

Figures

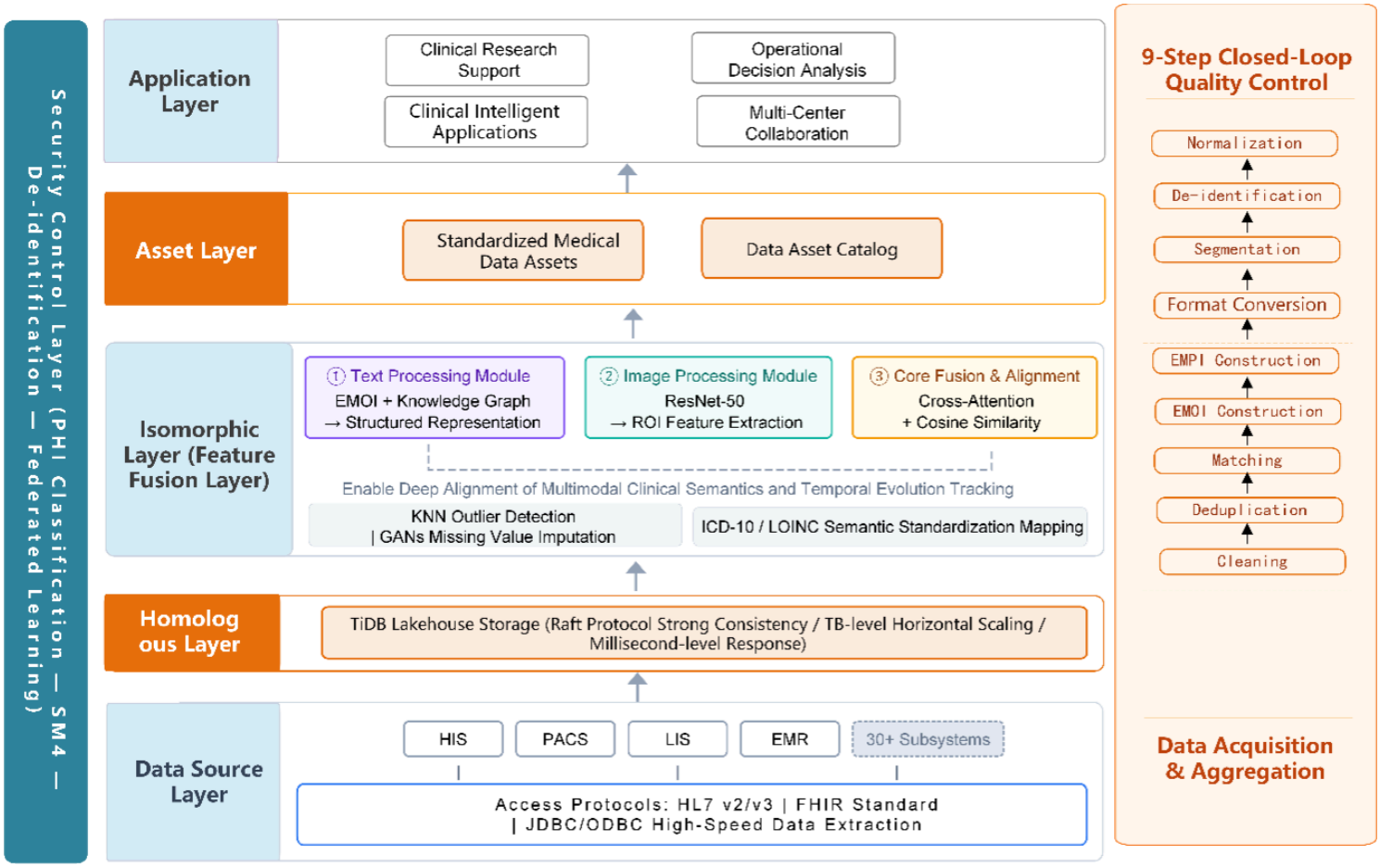


Figure 1

Layered lake-warehouse integration architecture

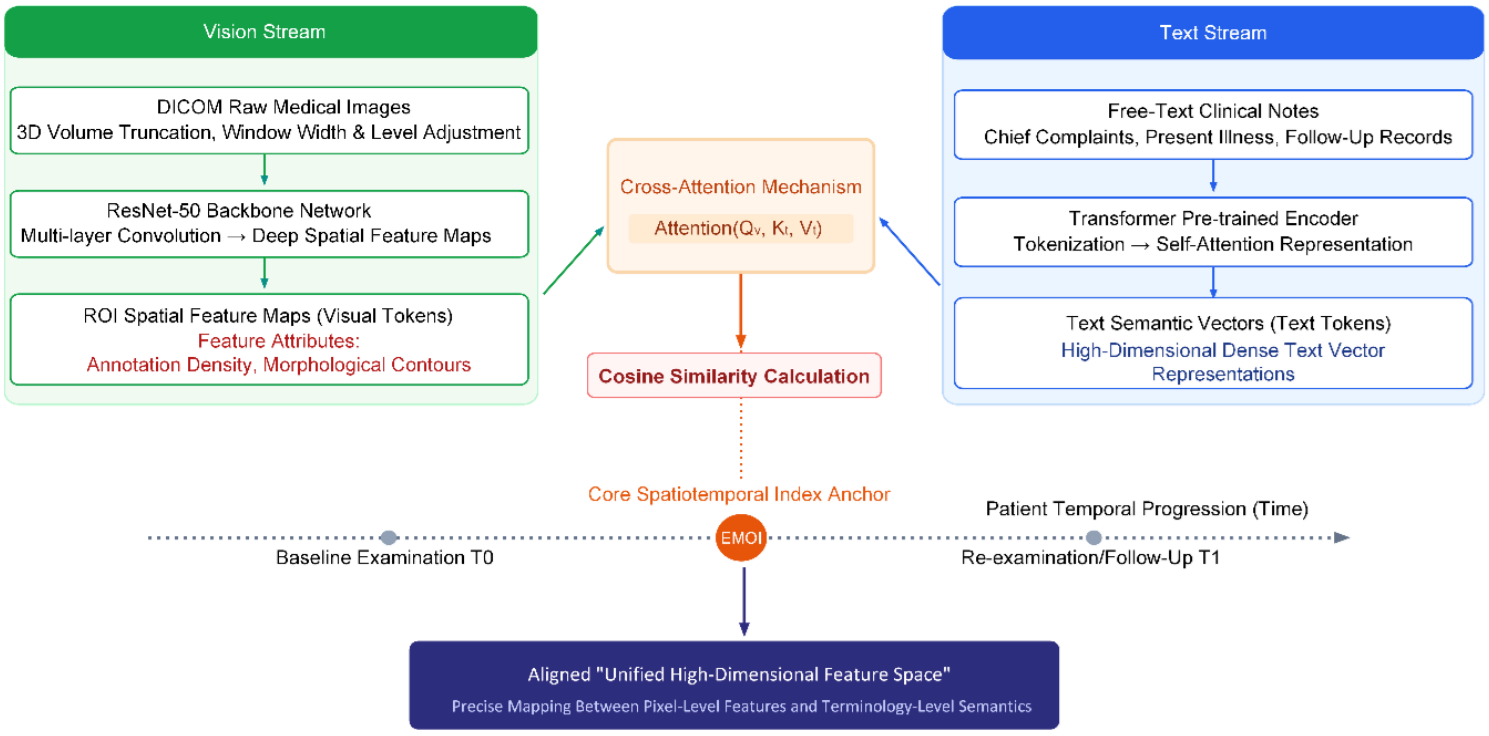


Figure 2

Multimodal feature fusion method

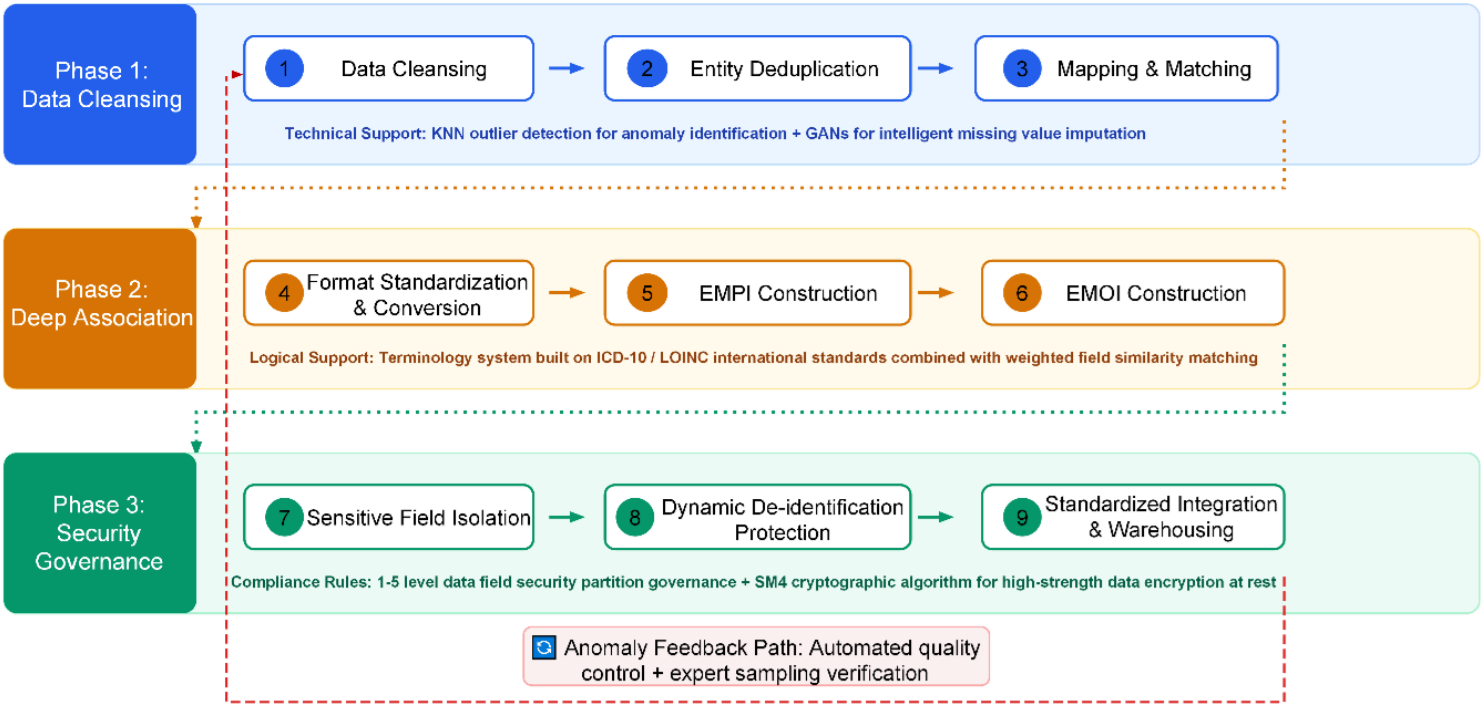


Figure 3

Nine-step closed-loop quality control process

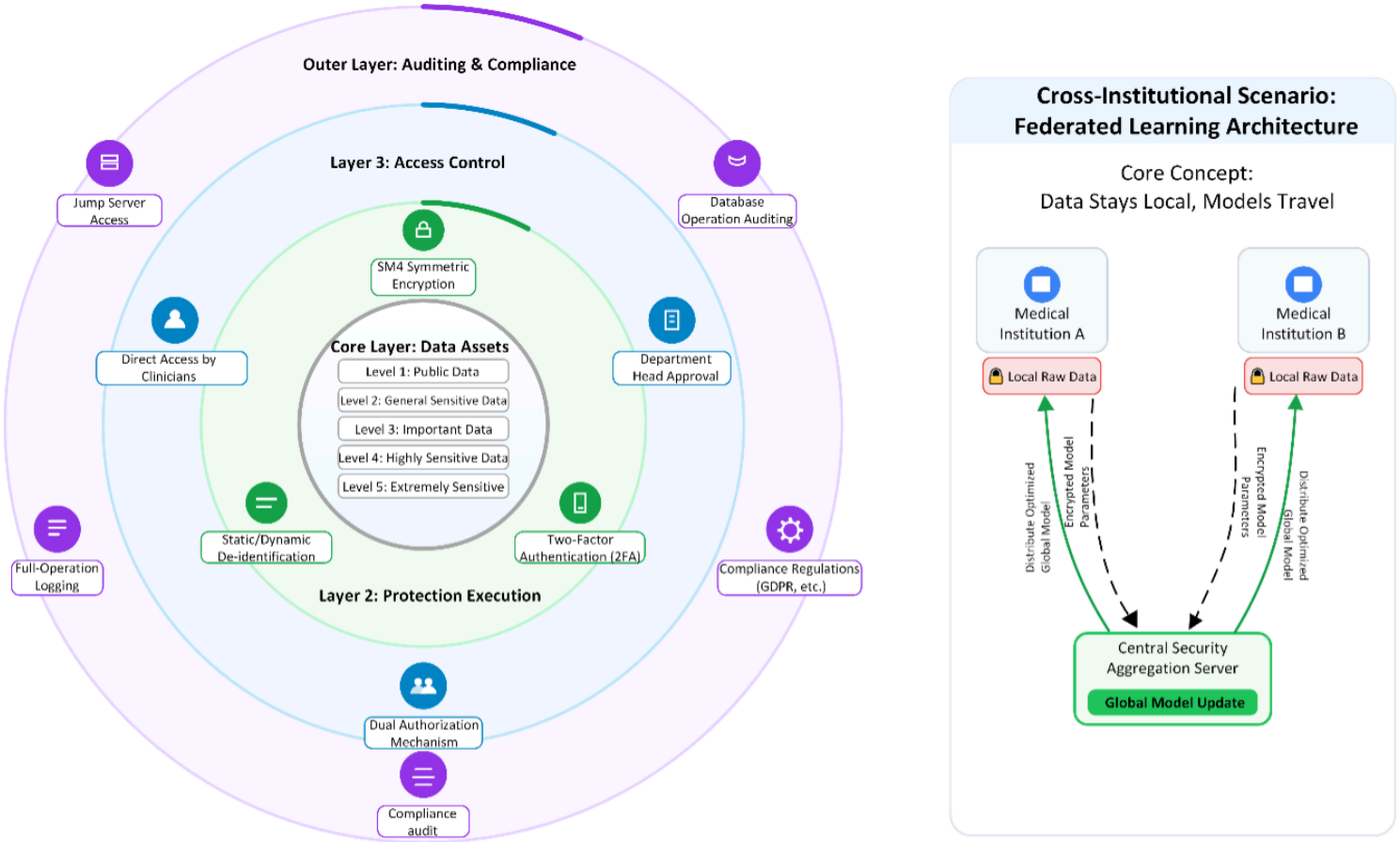


Figure 4

Medical data security and collaboration architecture based on the onion model

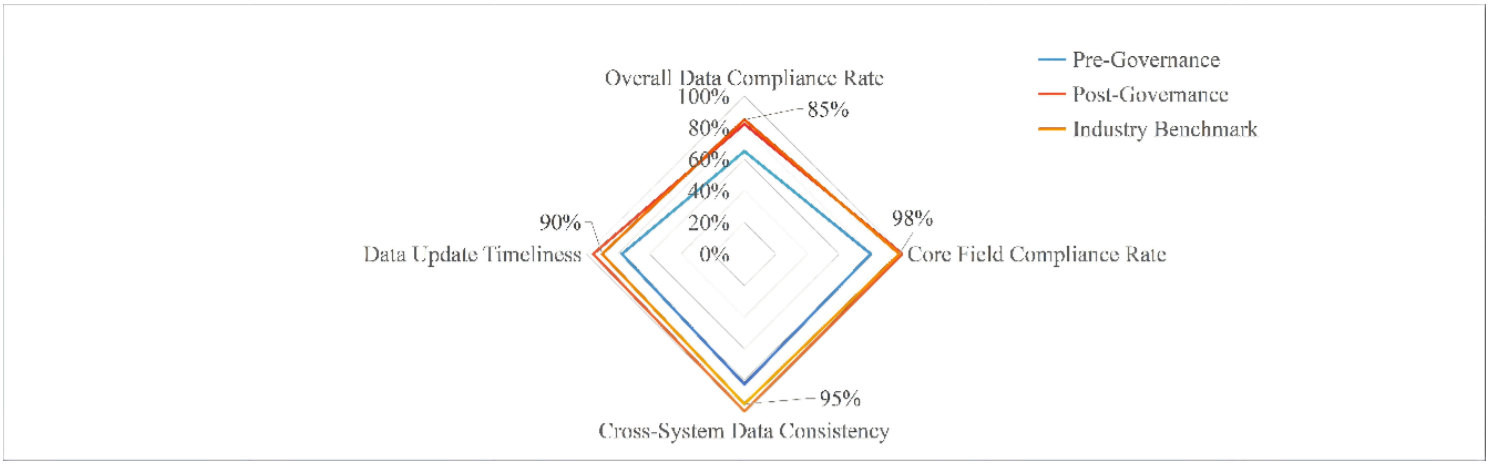


Figure 5

Radar chart comparing data quality governance before and after implementation