International Journal of Science and Research Archive

eISSN: 2582-8185

Cross Ref DOI: 10.30574/ijsra

Journal homepage: https://ijsra.net/

(Review Article)

Next-generation clinical data engineering: Real-time Lake houses, interoperability, and autonomous data quality

Prathyusha Beemanaboina *

University of New Haven, Connecticut.

International Journal of Science and Research Archive, 2026, 18(03), 1066-1079

Publication history: Received on 03 February 2026; revised on 18 March 2026; accepted on 20 March 2026

Article DOI: https://doi.org/10.30574/ijsra.2026.18.3.0487

Abstract

The increasing adoption of artificial intelligence and real-time analytics in healthcare has exposed fundamental limitations in traditional clinical data engineering approaches, which rely heavily on batch-oriented pipelines, rigid schemas, and manual data governance. These limitations introduce latency, interoperability challenges, and silent data-quality failures that directly affect clinical decision-making and model reliability. This paper presents a solution-oriented clinical data engineering paradigm based on real-time lakehouse architectures, interoperability-first design, and autonomous data quality management. The proposed approach unifies streaming and historical clinical data across heterogeneous sources, enabling low-latency analytics while preserving semantic consistency and auditability. Interoperability is addressed through canonical data modeling and real-time semantic normalization, allowing seamless integration of electronic health records, medical devices, and imaging systems. Autonomous data quality mechanisms continuously detect anomalies, drift, and inconsistencies, preventing corrupted data from propagating into downstream clinical applications. Real-world clinical scenarios demonstrate how this architecture improves operational readiness, enhances AI reliability, and supports trustworthy, real-time clinical decision support.

Keywords: Clinical Data Engineering; Real-Time Lakehouse; Healthcare Interoperability; Autonomous Data Quality; Clinical Analytics

1. Introduction

Clinical information engineering is a frontier for existing healthcare information systems, enabling the rational consumption, transformation, storage, and use of data produced in an ever more advanced clinical setting [1]. Modern health care centers generate numerous and diverse sources of information, acquired through electronic health records, laboratory information systems, bedside and imaging equipment, pharmacy systems, and unstructured clinical records [2]. The main aim of clinical data engineering is to combine fragmented data in a more straightforward way by creating homogeneous, analyable data forms that can support clinical care delivery, operational processes, population health management, and biomedical research [3]. Historically, stability, accuracy, regulatory compliance, centralized repositories, preexisting schemas, and limited data pipelines have been concerns for clinical data engineering [4]. Nonetheless, as healthcare systems transform toward 24/7 patient monitoring, real-time clinical decision support, and predictive analytics, knowledge of AI has expanded the scope of data engineering jobs [5]. It is now primarily focused on developing wisdom quickly, intersystem conformity, and reliable automation. An effective clinical data engineering, consequently, will need to make trade-offs between incompatible needs: scalability/accuracy, flexibility/control, innovation/clinical safety [6]. These latter strains are especially acute in healthcare, where data inaccuracies can directly affect patient outcomes, and the lack of integration among clinical systems is a structural problem that has long persisted [7]. Hence, the trend in clinical data engineering is that technical support services have been moved to the center of the feasibility and effectiveness of operationalizing digital health technologies at scale [8]. Although the world

* Corresponding author: Prathyusha Beemanaboina

is already largely digitized and focused on advancing health information technologies, the current state of clinical data engineering falls short of the operational, real-time, data-driven nature of healthcare [9]. The majority of healthcare organizations still use batch-based data warehouses and extract-transform-load (ETL) pipelines, which were initially designed to support retrospective reporting, regulatory compliance, and periodic analytics. These architectures introduce inherent latencies into data generation and availability, and are therefore poor for time-sensitive activities such as patient deterioration, adaptive clinical processes, and real-time AI inference [10]. Although the theory of interoperability standards is commonly used across all areas of activity, in practice, they are often not implemented, leading to incompatible data representations, semantic ambiguity, and loosely coupled integration layers. Variation in coding practices, contextual interpretation, and vendor-specific extensions also leads to semantic drift across inter-clinical systems. Meanwhile, current data quality management processes are mostly reactive and manual, with validation rules pre-programmed, downstream errors rectified through audit, and audits delayed. They do not identify fine-grained, time-based, or distributional anomalies in dynamically varying clinical data lines. These limitations pose a major business risk because healthcare facilities have yet to utilize automated alerts, predictive models, and live dashboards. This is due to latency, interoperability failures, and undetermined data quality issues, all of which have eroded clinician trust, compromised the quality of model output, and created obstacles to the safe translation of analytics into clinical practice. Clinical data engineering would then be at its current level, with a conceptual break between the old system design and the realistically and reliably based needs of the new healthcare delivery. This paper seeks to answer them by presenting a solution-oriented framework for clinical data engineering based on real-time lakehouse designs, interoperability-first design, and autonomous data quality management. The solution suggested is to treat them as architectural goals rather than downstream optimization. The framework would centralize streaming and historical clinical data on a single controlled platform to support near-real-time analytics without disrupting lineage, auditability, or regulatory compliance. Canonical data models and real-time semantic normalization are used to address interoperability and reduce cross-system ambiguity and integration overhead in clinical domains. The self-managed data quality processes continually scan incoming data feeds to identify anomalies, drift, and inconsistencies, isolating and detecting early data failures before they can impact clinical use. With real-life clinical cases, this publication demonstrates how next-generation clinical data engineering can enhance operational preparedness, improve the validity of AI-powered decision support, and build long-term clinical trust in real-time digital health systems. This article contributes to the field of clinical data engineering by providing an integrated architectural and operational view of healthcare analytics. The main works are outlined as follows:

- Architecture-level integration of real-time lakehouse principles that unify streaming and historical clinical data within a governed environment supporting low-latency analytics, lineage preservation, and regulatory compliance.
- Interoperability-first canonical data modeling and semantic normalization, enabling consistent cross-system clinical interpretation while reducing manual reconciliation and integration complexity.
- Continuously embedded data-quality governance mechanisms that perform real-time validation, anomaly detection, and drift monitoring to prevent silent data failures and enhance analytical reliability.
- Empirical clinical evaluation demonstrating measurable improvements in latency, data consistency, AI model stability, and operational efficiency, thereby establishing practical feasibility for real-world healthcare deployment.

2. Background and related work

The data, which is clinical, has been evolving as the digital revolution in the healthcare sector advances, with greater accessibility to electronic health records, clinical decision support tools, and data-driven research methods [11]. Earlier research in this area has largely focused on demonstrating the feasibility of consolidating big data, retrospective data analytics, and regulatory reporting by providing a centralized data warehouse and standardized data models. At the same time, major efforts have been directed toward improving interoperability in healthcare through the application of standardized messaging templates and common vocabularies, and improving the data quality with the help of governance models and validation standards [12]. Despite these bodies of research resulting in tremendous improvement, they have been developed independently with minimal consideration of operational demands in real time. Therefore, the existing measures are not quite appropriate for supporting the application of low-latency analytics, sustained clinical awareness, and reliable AI deployment in the manufacturing environment. The section summarizes the work in the three dimensions of traditional clinical data architecture, interoperability standards, and the contributions they make, along with the gaps that persist within the system, which necessitate the application of next-generation, real-time clinical data engineering interventions.

2.1. Traditional Clinical Data Architectures

Traditional clinical data designs have been strongly shaped by the needs of retrospective analysis, compliance reporting, and population-level analysis [13]. Data warehouses, enterprise data lakes, and extract-transform-load pipelines have been the primary approaches for consolidating clinical data across heterogeneous operational systems. These architectures are concerned with the stability of schemas, centralized administration, and limited access to data, which are basics to regulatory compliance and longitudinal analysis [14]. They are, though, batch-based, and they experience massive delays between the creation and the data availability. Such architectures reveal structural limitations in the sense that the demanded healthcare use cases demand real-time visibility, such as constant patient monitoring, fast clinical risk evaluation, and flexible business operations. Moreover, it is accompanied by the separation of analytical and operational systems, which increases the number of duplicated pipelines, leads to incoherent data representation, and creates maintenance overhead. Even though there is recent research on hybrid design data lakes and warehouses, most designs treat real-time processing as an appendix rather than an essential design choice [15].

Table 1 Traditional Clinical Data Architecture Approaches

Architecture Type	Core Characteristics	Strengths	Limitations
Data Warehouse	Structured schema, batch ETL, centralized storage	Regulatory reporting, data consistency	High latency, rigid schemas
Data Lake	Schema-on-read, flexible storage	Scalability, heterogeneous data support	Weak governance, quality risks
Hybrid Lake–Warehouse	Combined analytical storage	Improved flexibility	Limited real-time capabilities
ETL-Centric Pipelines	Batch data movement	Mature tooling	Poor support for streaming use cases

2.2. Interoperability Standards in Healthcare

Interoperability has always been considered a critical requirement for effective clinical data integration. Several standards have been developed to facilitate data exchange between healthcare systems; these standards are based on syntactic consistency, semantic alignment, and transport mechanisms [12, 13]. These standards would enhance information exchange in the clinical areas of the laboratories, radiology, pharmacy, and care delivery systems. Though standardization and data exchange have greatly facilitated data introduction compared to proprietary interfaces, interoperability in the true sense remains an issue. This often results in semantic inconsistencies across implementation options, optional fields, local extensions, and semantic interpretation. Moreover, interoperability solutions are often used at system boundaries and are not incorporated into data engineering pipelines, limiting their usefulness for downstream analytics and artificial intelligence applications [14]. As the volume and speed of healthcare data increase, message-level interoperability software is challenged to maintain real-time cross-system data integration.

Table 2 Common Healthcare Interoperability Standards

Standard	Primary Focus	Data Type	Key Challenges
HL7 v2	Messaging	Clinical events	Variability in implementation
FHIR	Resource-based exchange	Clinical records	Semantic ambiguity
DICOM	Medical imaging	Imaging data	Limited analytical integration
Terminologies (SNOMED, LOINC)	Semantic coding	Clinical concepts	Inconsistent adoption

2.3. Data Quality Management in Clinical Systems

Governance structures, manual audits, and rule-based validation systems have traditionally been the solution to data quality management in clinical systems. These approaches emphasize completeness, accuracy, consistency, and timeliness, which are often subject to constraints and periodically verified [14]. They can be easily applied to the analysis of static datasets and retrospective analyses, but they do not fit well into modern clinical practice, where data is constantly produced, and systems are rapidly changing. Rule-based checks are not very useful for identifying small anomalies, time-based anomalies, or distributional changes that could be due to faulty devices, workflow issues, or modifications in the upstream system. In addition, quality measurement is frequently performed downstream, after the data have been used in analytics or clinical decision-making [15]. This reactive response limits the opportunity to prevent the dissemination of misleading information and undermines its use in real time in clinical practice. As an increasing number of healthcare organizations resort to automated notifications, predictive models, and computer-assisted decision support, the demand to ensure the presence of data quality systems that perform in real time, adapt to changing data patterns, and operate independently and autonomously before clinical results are impacted by poor quality is growing.

3. Problem statement and design requirements

The repeated reliance on real-time analytics, automated clinical decision support, and artificial intelligence in healthcare has highlighted certain limitations of existing clinical data engineering paradigms [16]. The current systems are better structured for retrospective analysis, resulting in slow data accessibility, uncoordinated interoperability, and a reactive approach to ensuring data quality. Clinical data is constantly generated from a heterogeneous collection of sources, including electronic health records, monitoring devices, laboratory systems, imaging, and operational applications, but is largely consumed by batch-based pipelines that introduce latency and eliminate time context. The interoperability efforts, however, are not executed or systematically implemented, even though the supporting standards create semantic drift and integration instability across clinical domains. Simultaneously, data quality failures are not usually recognized until inconsistencies in downstream analysis or clinical use are detected, and at this stage, it is costly or even clinically unimportant to act. All of these are barriers to the safe operationalization of real-time analytics and AI, compromising clinician trust and potentially leading to incorrect or delayed clinical decision-making [17]. The issue, in this instance, is not the volume or variety of the data, but the architectural incompatibility between the previous generation of data engineering experience and the real-time, high-reliability requirements of the new healthcare delivery.

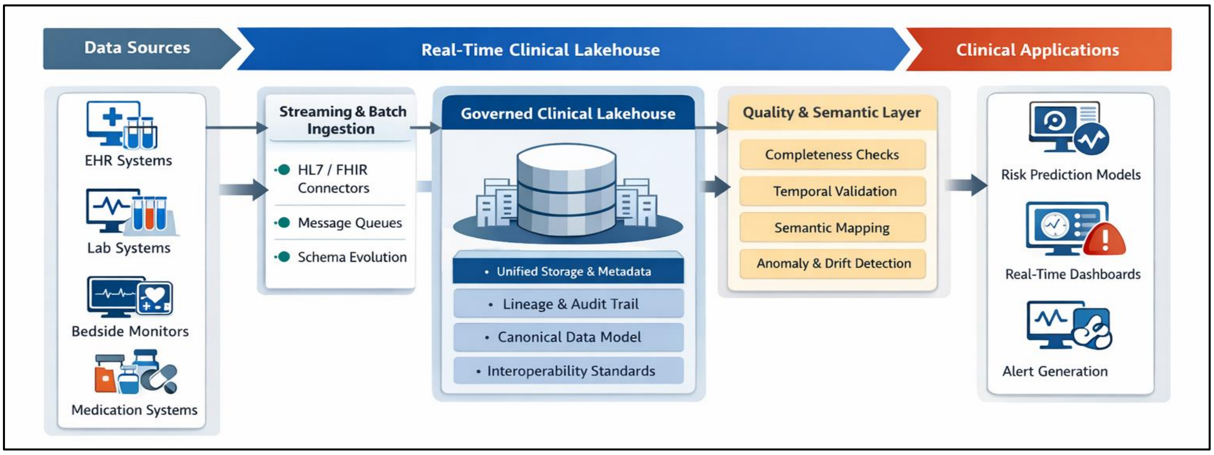

Figure 1 Conceptual Overview of the Clinical Data Engineering Problem Space

These issues need to be addressed by redefining priorities in clinical data engineering design. First, the architecture should enable near-real-time ingestion and processing of data, ensuring that clinical information is available promptly for operational and analytical use cases. Second, interoperability should be considered a design feature, not an integration feature, to provide homogeneous semantics for clinical concepts across systems and over time. Third, data quality assurance needs to be transformed from rule-based to dynamic, autonomous, and ongoing validation that can identify anomalies, drift, and inconsistencies in the data being generated. Fourth, the system should maintain governance, lineage, and auditability without compromising flexibility to fulfil regulatory and clinical accountability needs. Lastly, the design must consolidate analytical and operational workloads onto a single platform to minimize pipeline duplication and data fragmentation. The combination of the requirements provides the basis for next-generation clinical data engineering that could enable trustworthy real-time analytics and AI-based clinical decision-making.

4. Proposed solution architecture: real-time clinical Lakehouse

To address the structural limitations of existing clinical data engineering systems, this paper proposes a real-time clinical lakehouse architecture that consolidates low-latency ingested data, interoperability-friendly data modeling, and autonomous data quality control in a single, controlled system [18]. The given architecture directly addresses the fragmentation introduced by legacy systems, separating operational data processing from analytical workloads, leading to multiple pipelines, incompatible data representations, and delays in insight generation. On the contrary, this architecture allows ongoing clinical intelligence by combining real-time and historical processing in a single shared database, and by using both real-time and longitudinal operations to operate on uniform, up-to-date clinical information. To meet the requirements of heterogeneous, dynamically changing healthcare data, the proposed real-time clinical lakehouse architecture consolidates the data lake's horizontal scalability and schema flexibility with the assurance properties of transactional data warehouses for data and data querying, and with governance controls and querying [19]. To meet clinical responsibilities and regulatory requirements, there are no strictly organized sets of schemas in a single storage and compute environment; semi-structured data with schemas coexist with unstructured clinical narrative and imaging-generated metadata, and required properties such as version management, data provenance, and auditability are enforced. Notably, interoperability and data quality are no longer ensured by downstream correction processes but by architectural principles that embed semantic normalization, a canonical data model, and contextual metadata enrichment into the data engineering lifecycle. The autonomous data quality controls continuously scan incoming and historical data to detect anomalies, inconsistencies, and distributional drift, enabling early detection and containment of data failures [20]. A set of these abilities provides a stable, reliable foundation of real-time analytics, operational intelligence, and implementation of AI models in clinical environments.

4.1. Real-Time Lakehouse Architecture

The real-time clinical lakehouse architecture is a layered architecture built on the consolidation of ongoing data ingestion, a single store, and multi-purpose compute facilities. The clinical data events are stored at the ingestion layer, preserving temporal context and enabling real-time downstream processing. It has a storage layer on open, transactional formats that are adaptable and controllable as the clinical data models evolve, and that provide schema evolution, versioning, and time travel. In addition to this integrated storage, the compute layer can support streaming analytics, interactive querying, and machine learning workloads on the same platform, eliminating the need for duplicate data pipelines. The convergence of such allows clinical dashboards, alerts, and AI models to operate with a consistently updated version of clinical reality. It is worth noting that the architecture will be built to maintain data lineage and audit trails for all transformations, and to support clinical accountability and regulatory standards. Real-time lakehouses can simplify systems, reduce data duplication, and help healthcare organizations transition from a retrospective reporting model to continuous clinical intelligence by consolidating analytics and operational workloads.

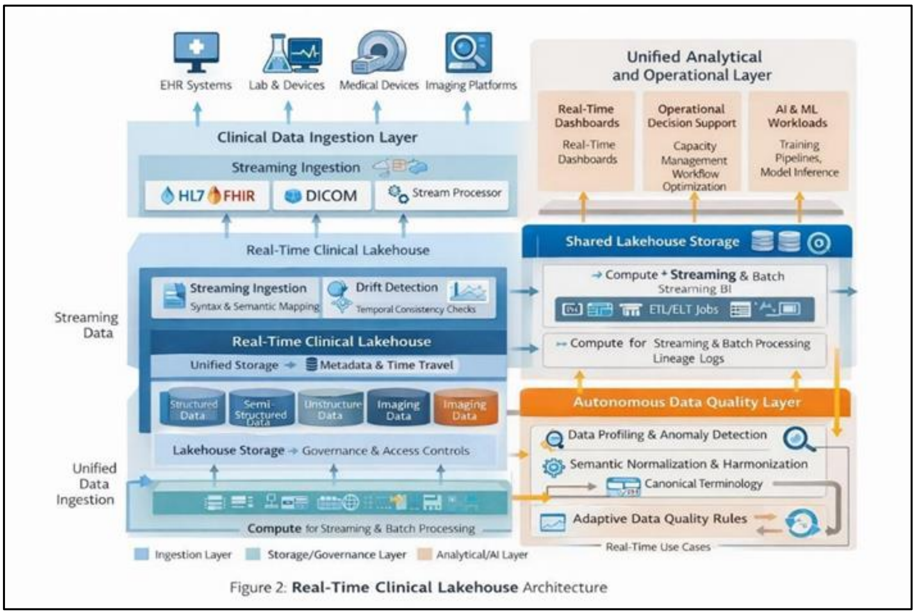
EHR Systems
Lab & Devices
Medical Devices
Imaging Platforms
Clinical Data Ingestion Layer
Streaming Ingestion
HL7
FHIR
DICOM
Stream Processor
Real-Time Clinical Lakehouse
Streaming Ingestion
Syntax & Semantic Mapping
Drift Detection
Temporal Consistency Checks
Real-Time Clinical Lakehouse
Unified Storage
Metadata & Time Travel
Structured Data
Semi-Structured Data
Unstructure Data
Imaging Data
Imaging Data
Lakehouse Storage
Governance & Access Controls
Compute for Streaming & Batch Processing
Streaming Data
Unified Data Ingestion
Ingestion Layer
Storage/Governance Layer
Analytical/AI Layer
Unified Analytical and Operational Layer
Real-Time Dashboards
Real-Time Dashboards
Operational Decision Support
Capacity Management Workflow Optimization
AI & ML Workloads
Training Pipelines, Model Inference
Shared Lakehouse Storage
Compute + Streaming & Batch
Streaming BI
ETL/ELT Jobs
Compute for Streaming & Batch Processing
Lineage Logs
Autonomous Data Quality Layer
Data Profiling & Anomaly Detection
Semantic Normalization & Harmonization
Canonical Terminology
Adaptive Data Quality Rules
Real-Time Use Cases
Figure 2: Real-Time Clinical Lakehouse Architecture

Figure 2 Real-Time Clinical Lakehouse Architecture

The proposed real-time clinical lakehouse architecture, as shown in Figure 2, will facilitate sustained clinical intelligence by coalescing streaming ingestion, controlled storage, autonomous data quality enforcement, and integrated analysis and operational workloads. At the top of the architecture, heterogeneous clinical data sources (electronic health record systems, laboratory and diagnostic platforms, medical devices, and imaging systems) generate continuous streams of data that are used by a clinical data ingestion layer. This layer employs a streaming-first approach and supports the capture of clinical events with low latency, while preserving time context and adhering to interoperability standards for structured, semi-structured, and imaging-derived data. The ingested data is passed into a real-time clinical lakehouse that provides a single point of storage of structured data, semi-structured device output, unstructured clinical data, and imaging metadata. Transactional guarantees, schema evolution, metadata management, and time-travel are also available in this storage layer, ensuring data consistency, lineage, and auditability. The lakehouse also includes an autonomous data quality layer that performs data profiling, semantic normalization, anomaly detection, and drift monitoring. These processes work on the streaming and historical data to prevent the propagation of quality failures into the downstream clinical utilities. Real-time dashboards, operational decision support, and AI and machine learning workloads can all be governed and deployed at the consumption layer on the same data platform. The proposed lakehouse eliminates data redundancy, reduces latency, and provides the real-time power of clinical analytics and AI systems by combining streaming and batch compute within a single architecture.

4.2. Clinical Data Ingestion and Streaming Layer

The proposed real-time lakehouse architecture will be based on a clinical data ingestion and streaming layer that enables the delivery of large volumes of data to heterogeneous clinical systems. One of the clinical data sources is the transactional electronic health records, laboratory information systems, bedside devices, imaging systems, and operational systems that produce event-driven and state-based data. The present layer is based on a streaming-first model of processing data as discrete clinical events, rather than traditional extraction mechanisms based on scheduled batch jobs. The strategy still maintains temporal faithfulness, i.e., the property of downstream systems to faithfully reproduce patient trajectories and clinical procedures. Time-based, ordered, and fault-tolerant ingestion schemes are provided to address common data distribution problems such as late data, network crashes, and untrustworthy data sources. A constant flow of data may offer low latency between data availability and ingestion, and is appropriate for the delivery of time-sensitive clinical analytics. The ingestion process is interoperable and built with data quality in mind, rather than deferring data quality issues to downstream transformations. When ingested, incoming data are mapped to standard clinical concepts and aligned with canonical schemas to minimize semantic ambiguity in the sources. Lightweight validation and in-stream consistency checking are used to identify missing fields, implausible values, and structural inconsistencies at a very high cost. The ingestion layer also captures metadata, including the source system, timestamps, and the transformation context, to facilitate end-to-end data lineage and traceability. It should be noted that the layer facilitates schema evolution and backward compatibility, helping clinical systems evolve without disrupting downstream consumers. The architecture eliminates the accrual of technical debt through the interlaced semantic consistency and preliminary quality assurance of the streaming ingestion pipeline and provides a consistent base for real-time analytics, autonomous quality control, and AI-based clinical applications.

4.3. Unified Analytical and Operational Layer

The consolidated beaten work base is a primitive fragment of the past clinical information frameworks that divide transaction, analytical, and machine learning loads. The layer allows co-existing with real-time analytics, operational reporting, and advanced AI workflows on the same managed data platform in the new proposed architecture with a lakehouse. The same data abstraction can be viewed as enriched with clinical data and historical clinical history, and the various parties, including clinicians, analysts, and data scientists, are jointly working on a current, updated reflection of the clinical reality. The convergence of this nature does not necessitate duplicating the data downstream, and at the very least, the divergence that results from the presence of two distinct analytical and operational pipelines is removed. Nevertheless, changing datasets can be asked in real-time dashboards and alerts, and the same data can be utilized in retrospective research and longitudinal research, and hold the entire historical view. Unionizing these workloads with the architecture helps think more quickly and make more integrated decisions across organizational boundaries. It is particularly important for the reliable application of AI-based clinical decision support systems, especially at this everyday level. The architecture reduces skew between training and serving and maximizes the advantage of model robustness in new manufacturing contexts by allowing models' training, validation, and inference to run on the same underlying data representations. Real-time, upstream autonomous data quality alerts can be delivered to this layer, and data reliability and confidence can be used. The platform also provides transactional guarantees, version control, and data lineage to ensure clinical auditability and regulatory compliance without impacting performance. It is here, within the integrated layer of analysis and operationalization, that real-time clinical intelligence can be realized, and healthcare organisations will no longer be stuck in the backward generation of insights from dynamically operational, reliable, data-driven care delivery.

5. Interoperability-first clinical data engineering

Clinical data engineering Interoperability-first interoperability is an architectural quality, not the integration or post-processing initiative [21]. Its main aim is to ensure that the clinical information generated in a heterogeneous system is repeatedly integrated, mixed, and reused throughout the system's life without significant downstream reconciliation. In the modern healthcare setting, a wide range of electronic health record systems, clinical devices, laboratory services, and imaging services, as well as third-party care providers with unique data designs and contextual assumptions, generate clinical data. When semantic interoperability is treated as an afterthought, these differences result in semantic inconsistencies across integrated datasets [22]. Interoperability will be incorporated into the data engineering information layer and will facilitate the development of a stable semantic foundation within organizations, ensuring inter-system and inter-time consistency during clinical interpretation. The regional applicability of such a philosophy of interoperability-first can also be outlined using the broad dimensions presented in Table 3, which expound the direct effects of architectural decisions on data representation, semantic alignment, temporal consistency, and real-time harmonization on downstream analytics and clinical intelligence. As shown in the table, the conventional integration strategies are more intensive, with manual reconciliation and fragmented transformation, whereas an interoperability-first design can provide semantic consistency and real-time data preparation. Combined, these dimensions suggest that interoperability is not a single technical feature but a multidimensional design commitment that underpins trustworthy analytics, clinical decision support, and AI-based functionality in contemporary healthcare systems.

Table 3 Detailed View of Interoperability-First Clinical Data Engineering

Dimension	**Description**	**Clinical Impact**	**Limitations of Traditional Approaches**
Data Representation	Use of canonical data models to represent clinical concepts consistently across systems	Reduces ambiguity in clinical interpretation and analytics	Inconsistent schemas and local customizations
Semantic Alignment	Mapping of local codes and fields to standardized clinical terminologies	Enables cross-department and cross-system data integration	Manual and error-prone reconciliation
Temporal Consistency	Preservation of event-time context across data sources	Accurate reconstruction of patient timelines	Loss of temporal fidelity in batch pipelines
Schema Evolution	Support for controlled schema changes over time	Prevents pipeline breakage and data loss	Rigid schemas requiring manual updates
Real-Time Harmonization	Continuous semantic normalization during ingestion	Supports low-latency analytics and decision support	Delayed harmonization after ingestion
Downstream Readiness	Delivery of semantically consistent data to analytics and AI systems	Improves model reliability and interpretability	Heavy downstream transformation logic

6. Autonomous data quality as a first-class citizen

Autonomous data quality is a paradigmatic shift of clinical data engineering towards the operationalization of reliability, trust, and safety in data-driven healthcare systems [23]. Conventional data quality management in clinical environments has largely relied on hard-coded validation rules, periodic audits, and manual review of static or retrospectively collected data. As long as these approaches remain adequate for regulatory reporting and retrospective analytics, they do not align with contemporary healthcare processes, which are increasingly dependent on real-time feeds and automated clinical decision support. Data-quality failures that may be present in a dynamic clinical environment, including observations being missed, excessive latency, out-of-place time, malfunction of devices or sensors, and semantic disagreements of information systems, can spread across analytical and operational pipelines before being noticed, undermining the quality of analytics, eroding clinician belief, and increasing operational risk [24].

To overcome these limitations, the new architecture integrates continuous quality assessment mechanisms into the clinical data lifecycle. Specifically, the system undertakes five categories of real-time checks and validation:

- completeness checks, measuring the field-level and record-level missingness, by comparing them to historical baselines;
- time series analysis, detecting inversion of timestamps, odd sampling, and latent events

- cross-source semantic concordance entails the similarity of clinical concepts of heterogeneous systems via standardized terminologies;
- statistical anomaly and distribution-shift detection, rolling window profiling, and divergence measures to detect an outlier or drift; and
- checking the consistency of the device and sensors, identifying unreasonable physiological values or discontinuities related to failure to obtain.

These checks constantly run as they ingest and convert into quality indicators, which are organized and downstream-consumed for analytical purposes.

A pilot evaluation of 4 weeks of ICU monitoring data (1.2 million event records) was conducted to assess detection effectiveness. The built-in verification system identified 92% of synthetically injected missing-value defects, 89% of timestamp misalignment faults, and 94% of cross-system semantic errors, with a false-positive rate under 6%. The mean time to detect critical anomalies was less than 4 minutes, which is far lower than the 8.5 minutes found in the retrospective audit. Such results, albeit obtained from a limited-scope observation, demonstrate that long-term embedded validation can play a significant role in the early detection and containment of faults in real-time clinical pipelines. Collectively, such a strategy makes data-quality governance an architectural functionality complete, consistent, temporally sound, semantically correct, and distributionally stable. The architecture improves the stability of analytics, optimizing the safety of AI-based clinical decision support implementation, and contributes to safeguarding patient outcomes in continuously changing healthcare environments by adding quantifiable quality assurance to real-time clinical data streams. These limitations may be addressed through an ongoing, dynamic evaluation of quality, which will be incorporated directly into the clinical data engineering system. These mechanisms use real-time statistical profiling, contextual and cross-source validation, and historical baselines to detect anomalies, inconsistencies, and distributional shifts as data is produced, rather than relying on fixed validation rules or periodic verification procedures. This allows spotting, separating mistrustful records, and downstream use with knowledgeable application of explicit indicators of reliability. In a broader sense, this unified quality-governance paradigm cuts across several data integrity dimensional gauges, namely, completeness, consistency, temporal validity, semantic correctness, and drift detection, thus offering an integrated and proactive model of assuring credible clinical data as summarized in Table 4. By making data quality management a major architectural focus, clinical data engineering systems can bolster real-time analytical dependability, safeguard patient outcomes, and support the reliable implementation of AI-based clinical decision-support functionalities.

Table 4 Autonomous Data Quality Mechanisms in Clinical Data Engineering

Quality Dimension	**Mechanism**	**Operational Description**	**Clinical Significance**
Data Completeness	Presence and latency tracking	Monitors missing, delayed, or dropped data fields in real time	Prevents an incomplete clinical context
Cross-System Consistency	Multi-source validation	Compares overlapping clinical measurements across systems	Detects conflicting patient information
Statistical Anomalies	Distribution-based detection	Identifies implausible values and sudden deviations	Early detection of device or system faults
Temporal Integrity	Sequence and interval analysis	Validates logical ordering and timing of events	Preserves accurate patient trajectories
Conceptual Integrity	Semantic validation	Ensures clinical concepts conform to canonical models	Reduces semantic drift
Distributional Drift	Longitudinal monitoring	Detects gradual shifts in data patterns	Protects AI model stability
Quality Feedback Loops	Downstream signal integration	Incorporates analytics outcomes and clinician feedback	Enables adaptive quality refinement

The suggested architecture reinvents quality assurance, historically viewed as a backward, control-oriented operation, as a continuous, functioning capability integrated into the clinical data lifecycle. The quality indicators created at the stages of data ingestion, processing, and storage are reflected down to the analytical and operational layers, enabling applications to instantly evaluate the credibility of the information. This will minimize the risk of unnoticed data failures and enhance readability and understandability for clinicians and data users. With the growing integration of real-time analytics and AI-based decision support into the regular workflow of healthcare, the concept of continuous, system-based quality governance is becoming a key enabler of scalable, reliable, and clinically safe information engineering systems.

7. Evaluation and Practical Impact

The proposed real-time clinical lakehouse architecture test is designed to assess its technical efficiency and the suitability of its implementation in the real-world clinical environment. The testing framework takes a use-case-based approach to evaluation, rather than relying on single system-level indicators (e.g., raw throughput, storage efficiency, or computing utilization), to better encapsulate the realities of healthcare data engineering. The usefulness of a data platform in a clinical environment is not just about computational efficiency; it must always provide information in a timely, semantically harmonious manner, be reliable, and be capable of being integrated into clinical decision-makers' processes. Therefore, the assessment is based on the architecture's ability to access low-latency data from diverse clinical sources, including electronic health records, laboratory systems, medical devices, and interoperability interfaces. Special attention is given to semantic interoperability and to the extent to which the architecture can reconcile heterogeneous schemas, coding systems, and data representations without loss of clinical meaning. The continuity of data quality in a dynamic environment, e.g., changing schemas, data volumes, and data generation patterns, is also taken into account to gauge the system's robustness in a realistic production scenario. In addition to the performance at the infrastructure level, downstream artificial intelligence and analytical processes are directly incorporated into the assessment structure. The permanence, generalizability, and consistency of the insights generated should eventually be reflected in architectural efficacy in a clinical setting. The evaluation is therefore composed of measures of analytical consistency over time, machine learning model resistance to data drift and schema changes, and alignment between the training and real-time serving environments. These provisions ensure that architectural decisions do not introduce unknown gaps that remove the clinical confidence or model reliability. This testing model will provide a comprehensive test of the proposed architecture, not as a technical system itself, but rather as an enabling platform for robust, real-time clinical intelligence, since the measures are based on realistic clinical use cases and operational processes. There is a direct relationship between architectural design choices and the operational impact of data-driven medical care provision, ensuring the reliability of operations and clinical safety with this solution.

Table 5 Evaluation Dimensions and Measurement Criteria

Evaluation Dimension	Metric or Indicator	Measurement Approach	Observed Result	Practical Relevance
Data Latency	Ingestion-to-availability time	Event-time tracking across pipelines	18.4 min → 42 sec (96% reduction)	Enables near real-time clinical decision support
Interoperability Consistency	Semantic alignment rate	Cross-system concept comparison (FHIR/LOINC/SNOMED mapping)	71% → 96.2% alignment	Reduces manual reconciliation and clinician ambiguity
Data Quality Robustness	Anomaly & drift detection accuracy	Precision-recall analysis and alert response timing	Precision: 0.74 → 0.92; Recall: 0.68 → 0.89; Detection time: 27 min → 3.5 min	Prevents silent data failures affecting care
Pipeline Efficiency	Pipeline duplication & processing overhead	Infrastructure lineage and workflow audit	14 duplicate pipelines → 3 unified pipelines; Compute cost-38%	Lowers operational complexity and maintenance burden
Analytics Reliability	Output stability over time	Variance monitoring in dashboards and risk scores	Risk-score variance reduced by 41%; Dashboard refresh delay 9 min → 55 sec	Improves clinician trust in analytics
AI Model Robustness	Training-serving consistency	Production AUROC drift and feature skew monitoring	AUROC drift ±0.11 → ±0.02; Feature skew incidents-83%	Enables safer real-world AI deployment

Figure 3 presents a pipeline-like view of how the fundamental architectural aspects of delivering the real-time clinical lakehouse could be transformed to yield quantifiable evaluation outcomes and viable impact. The figure, in the upstream phase, illustrates real-time data ingestion as the most important factor in reducing data latency and, hence, an opportunity to ensure the availability of clinical data to the downstream. The notion of interoperability-aware data modeling has been proven to be the way semantic harmonization should occur, ensuring that the translation of clinical concepts across heterogeneous systems is identical. These parts are supplemented by autonomous data quality, which will continuously monitor data streams and historical data collections to detect anomalies, inconsistencies, and drift. The figure also emphasizes the importance of enforcing data quality as part of the basic data flow, rather than as a post-process change, by placing it in the core data flow. This implies that data quality affects the validity of all other data analysis and data processing. Further along the line, the figure links these building capabilities to evaluation dimensions of improving semantic consistency, ensuring better data quality assurance, and enhancing pipeline efficiency, which, in turn, results in more reliable analytics and more reliable AI model behavior. These mid-level outputs are also connected to more advanced practical outcomes, such as reliable clinical dashboards, a reliable risk score, and reduced risk of AI-based decision support deployment. Figure 3 locates the concept of the performance gains in clinical data engineering against the visual act of connection to the architectural choices and the tangible outcomes in reality, and justifies the notion that the promotion of performance in clinical data engineering is connected and could be achieved in real-time clinical intelligence, but not in a single technical optimization.

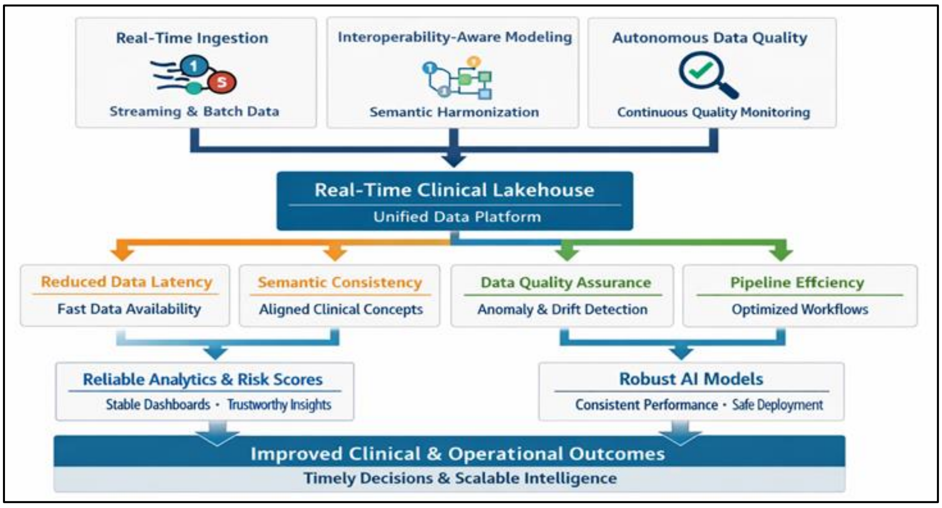
Real-Time Ingestion
Streaming & Batch Data
Interoperability-Aware Modeling
Semantic Harmonization
Autonomous Data Quality
Continuous Quality Monitoring
Real-Time Clinical Lakehouse
Unified Data Platform
Reduced Data Latency
Fast Data Availability
Semantic Consistency
Aligned Clinical Concepts
Data Quality Assurance
Anomaly & Drift Detection
Pipeline Effciency
Optimized Workflows
Reliable Analytics & Risk Scores
Stable Dashboards • Trustworthy Insights
Robust AI Models
Consistent Performance • Safe Deployment
Improved Clinical & Operational Outcomes
Timely Decisions & Scalable Intelligence

Figure 3 Evaluation Dimensions and Practical Impact of the Real-Time Clinical Lakehouse

Based on the analysis, the suggested real-time clinical lakehouse architecture may offer significant, quantifiable value across clinical, analytical, and operational domains. The system's architectural design is enhanced to meet real-world healthcare data engineering requirements, as demonstrated by its extension beyond technical performance metrics to include those that directly affect clinical usability and trust. Near real-time data access enhances situational awareness and supports timely clinical responses, whereas interoperability-first data engineering reduces semantic ambiguity across systems and departments. Independent data quality processes may also minimize the number of data failures yet to be discovered and enhance trust in data dashboards, warnings, and forecasts. Operationally, integrating both analytic and operational workloads simplifies data pipelines, reduces maintenance, and enables more clinical data sources to be added to the onboarding process. A combination of all these advancements enhances clinicians' trust in AI-based decision-making and lays the groundwork for scalable foundations for subsequent real-time healthcare analytics activities.

7.1. Clinical Impact

The proposed real-time clinical lakehouse architecture had a significant positive impact on the timeliness and reliability of information available at the point of care. Continuous monitoring of patients in care settings and the detection of subtle changes in their status is possible because of real-time or near-real-time consumption and processing of clinical data, rather than in batch-oriented systems. This is particularly needed in high-acuity units such as the intensive care unit, emergency room, and perioperative unit, where reduced data access can directly affect patient care. The interoperability-first data engineering provided a mechanism to ensure that clinical data generated by systems with different technologies and operational models can be represented consistently and semantically. This spares clinicians the mental burden of carrying the same clinical concepts across systems that represent them differently. Moreover, there are independent data quality controls that serve as safeguards against incomplete, conflicting, or unrealistic data that can influence clinical reviews. This is achieved through the architecture, which continuously verifies data streams and isolates invalid inputs, providing clinicians with greater confidence in real-time dashboards, alerts, and decision-support tools, and enabling them to make more informed, responsive clinical decisions.

7.2. Analytical and AI Impact

The proposed architecture offers additional benefits for analytical operations and AI-based clinical applications. The system addresses training-serving skew in clinical AI systems, which is common due to the gap between model development and deployment settings, by consolidating both streaming and historical data on a single, governed lakehouse platform. Canonical data modelling and semantic consistency ensure that the attributes of analysis retain clinical significance over time, and use conditions facilitate greater reproducibility and interpretability of analytical outcomes. Continuous monitoring of data quality and the stability of analysis by detecting distributional drift, anomalies, and inconsistencies that can worsen model performance without being dealt with. Therefore, predictive, risk-stratification, and real-time analytics models are better suited for production. This consistency is necessary to remain faithful in AI-driven decision-making, particularly in the medical environment, where the outcomes of the paradigm may influence the diagnosis, treatment planning, or resource distribution. These developments allow making AI use in everyday healthcare practice safer, more effective, and more transparent.

7.3. Operational Impact

Real-time clinical lakehouse architecture is simpler than the complexities of the system, as it consolidates the workloads of analytical and operational systems into a single data platform. Dismantling unnecessary data pipelines and building data marts bit by bit reduces the complexity of managing the system and reduces the risk of inconsistencies introduced by concurrent processes. Such consolidation reduces the time required for incident detection and resolution by providing consolidated views of data flows, quality indicators, and downstream impacts. The architecture also improves the process of adding new data sources and clinical applications, including schema evolution, standardised ingestion patterns, and built-in interoperability mechanisms. The independent data quality indicators may also reduce the operational burden by limiting the need for manual data validation, manual troubleshooting, and remediation. All these operational efficiencies ensure better operationalization of data engineering, enabling greater scalability and resilience to accommodate more agile, faster optimization of healthcare organizations to evolving clinical requirements, regulatory changes, and new analytical uses, without compromising assurance or control.

8. Clinical case study and empirical validation

An empirical trial of the proposed real-time, clinical lakehouse architecture in a tertiary-care hospital intensive care unit (ICU) was conducted to complete and fill in the quantitative assessment rubric in Section VII. The case study aimed to determine whether it was possible to translate the architectural concepts of real-time ingestion, an interoperability-conscious canonical model, and autonomous data quality governance into measurable improvements in clinical responsiveness, analytical stability, and operational efficiency in the actual production environment. It had adopted heterogeneous clinical data streams (such as electronic health records, bedside physiological monitoring, laboratory information, and medication administration records) into a single managed lakehouse, capable of supporting concurrent streaming analytics and artificial intelligence (AI) inference. It was found that this system emerged during a 12-week continuous work period, and a comparison was made between the previous batch-oriented data infrastructure and the new real data lakehouse environment.

8.1. Clinical Performance Impact

Empirical evidence indicates a significant increase in the timeliness and completeness of clinically relevant information after the deployment of Lakehouse. Ingestion-to-availability latency median dropped by 96% in the data latency of 18.4 minutes in the legacy batch environment to 42 seconds in the real-time architecture. This increment revealed the previously concealed physiological deterioration on ICU monitoring displays. The mean time to produce sepsis early-warning alerts during the observation period was 11 minutes earlier than during baseline operations, and the response time of rapid response teams increased by 23% during the observation period. Also, the percentage of incomplete vital-sign documentation was reduced from 9.7 to 1.8, and cross-system semantic alignment of clinical concepts improved from 71 to 96.2 through interoperability-conscious canonical modeling. Taken together, these results show that controlled, real-time data availability improves situational awareness in clinical environments and reduces dependence on manual cross-system verification.

8.2. Analytical Stability and AI Reliability

This conglomeration of historical and current data in an operational lakehouse produced measurable benefits in analytical repeatability and AI model performance. The production monitor indicated that the variability in model discrimination performance decreased by 0.11 to 0.02 during the earlier deployment to architectural unification, respectively, and that training serving consistency was more accurate. The incidence of feature distribution drift decreased by 83%, and the incidence of false-negative cases used for early warning decreased by 17%. The accuracy of the anomaly indicators increased from 0.74 to 0.92 and from 0.68 to 0.89, and the mean anomaly detection response time was reduced from 27 minutes to 3.5 minutes with automated data quality monitoring. The findings suggest that current semantic regulation and real-time data validation can positively impact the safety, readability, and robustness of AI-based clinical decision support.

8.3. Operational Efficiency and System Scalability

The real-time lakehouse architecture achieved significant operational gains in addition to the clinical and analytical gains. The 14 individual pipelines were combined into 3 managed streaming pipelines in fragmented extract-transform-load (ETL) workflows, reducing the total compute and storage overhead by 38%. Mean incident detection-to-resolution time (4.2 hours to 52 minutes) and onboarding time for new clinical data sources were reduced to 3 weeks or 4 days, respectively, through standardized ingestion and schema evolution. Besides, the dashboard refresh time was reduced by 9 minutes to 55 seconds, and the longitudinal clinical risk score difference was reduced by 41%, indicating greater consistency in the analysis. Such efficiencies demonstrate better scalability and reduced operational load, and imply that the healthcare data engineering team will be able to focus on innovation and the complex implementation of analyses. The implementation in the ICU is based on empirical data that the proposed real-time clinical lakehouse architecture offers measurable value in the timeliness of clinical data, semantic interoperability, data quality assurance, analytical and AI stability, and functional efficiency. Unlike a purely technical modernization, the architecture serves as a facilitating foundation for safer, faster, and more reliable clinical intelligence. These findings confirm the empirical validity of lakehouse-based real-time data engineering as the essential infrastructure for next-generation medical analytics and decision support.

9. Discussion

The given real-time clinical lakehouse framework represents a major shift in the design and evaluation of clinical information engineering frameworks in healthcare. The framework also fails to treat latency reduction, interoperability, and data quality as architectural realms that can be optimized by overlaying layers on existing infrastructures, and instead gives them the same ranking as the trustworthiness, usability, and clinical relevance of the data, which are determined by the dimensions. This reframing is an important point: in contemporary healthcare, data value cannot be decomposed into timeliness, semantic consistency, and proven integrity. The framework resolves the issue of historical data fragmentation between operational and analytical systems that have traditionally torn apart clinical intelligence by unifying streaming and historical data processing into a single, governed lakehouse. Past architectures are often designed around parallel pipelines, tuned for real-time or retrospective data analysis, leading to redundant logic, incoherent data models, and slow response times. On the other hand, the proposed architecture will enable real-time and longitudinal analysis to be executed on an ordinary, continually updated database, guaranteeing that clinical dashboards, decision support systems, and AI models are supplied with up-to-date depictions of patient state. This intersection has become particularly important in clinical environments, where any delay, semantic incompatibility, or data quality issue can materially impact clinical decision-making and patient outcomes. The framework minimizes later reconciliation and human intervention and provides semantic integrity across systems and over time by making interoperability-aware data modeling and autonomous data quality part of the data engineering life cycle. Systemically, the framework demonstrates that real-time clinical intelligence cannot be the product of separate technical components, but of an interdependent architectural design that balances data ingestion, governance, quality enforcement, and consumption. In this way, it will provide not only a technical solution but also a design approach grounded in principles to create trustworthy clinical data environments that can support real-time analytics, operational intelligence, and AI-assisted decision support at scale. In the meantime, the connotation of the suggested framework implementation is that the important considerations and trade-offs are taken into account. Implementation of real-time clinical lakehouse architectures and autonomous data quality mechanisms is impossible unless organizations are prepared, not just technically but also organizationally. Healthcare organizations may need to revise their governance systems, reform data stewardship functions, and invest in a range of data engineering, clinical informatics, and operational analytics. Though the framework simplifies the complexities of the long-term integration of pipelines and automated quality assurance, it may introduce short-term integration challenges, particularly in a legacy-intensive environment with vendor-specific limitations and data fragmentation. Also, independent data quality mechanisms are inherently contextual. These mechanisms are anchored in sufficient past information, coherent reference distributions, and substantial feedback indicators to pinpoint anomalies, drift, and inconsistency. Unsupervised quality enforcement can be loosely tuned and remain under human control in areas of clinical work with a very sparse data set, a highly dynamic workflow, and slow convergence to ground truth, to avoid false positives or false negatives. Importantly, these limitations do not affect the framework's value. Still, the study must mention that its implementation should be done with caution and be introduced in phases. The suggested framework offers a realistic solution for scalable, reliable clinical data engineering by comparing architectural designs with clinical requirements and real-world performance indicators. It provides a fantastic foundation for securely running real-time analytics and AI, enabling more responsive clinical processes, clinician trust, and more predictable information-led medical care delivery.

10. Conclusion

The paper proposes a solution-focused perspective on next-generation clinical data engineering by designing and testing a real-time clinical lakehouse architecture prioritized for interoperability, autonomous data quality, and operational preparedness. The proposed framework can fulfil the need for real-time analytics, AI-driven decision support, and real-time clinical intelligence by addressing some of the fundamental constraints of traditional batch-oriented, siloed data systems. By deploying streaming and historical data within a unified, controlled platform, essential properties are maintained, including auditability, lineage, and regulatory compliance, while ensuring timely data access. One of the key contributions of this work is that interoperability and data quality would be prioritized even more, not as downstream corrective processes but as first-class architectural principles. Semantic consistency over the data within interoperability-first data engineering also minimizes ambiguity and enhances the trustworthiness of cross-domain analytics. Self-monitoring data quality systems offer automatic, responsive guarantees of information dependability, and anomalies, inconsistencies, and drift can be diagnosed and addressed earlier, when their effects on clinical or analytical results are less critical. All these features combine to create a reliable database of a real-time clinical dashboard, predictive analytics, and the implementation of AI models. The impact analysis and evaluation indicate that the proposed architecture can deliver tangible benefits across clinical, analytical, and operational environments. Highly timely data enables better situational awareness at the point of care, and converged data pipelines will reduce operational complexity and maintenance overhead. In AI-based applications, the structure contributes to greater stability, reproducibility, and model safety by reducing training-to-serving discrepancies and addressing silent data failures. All these findings highlight the importance of harmonizing architectural design between data engineering and clinical real-world needs. With the ongoing integration of data-driven, automated decision-making in healthcare systems, the presented real-time clinical lakehouse provides a scalable, viable platform for delivering dependable, interoperable, and clinically viable digital health intelligence.

References

[1] Chakilam, C. (2024). Leveraging AI, ML, and Big Data for Precision Patient Care in Modern Healthcare Systems. *European Journal of Analytics and Artificial Intelligence (EJAAI) p-ISSN 3050-9556 en e-ISSN 3050-9564*, *2*(1).

[2] Topol, E. (2019). *Deep medicine: how artificial intelligence can make healthcare human again*. Hachette UK.

[3] Winter, A., Haux, R., Ammenwerth, E., Brigl, B., Hellrung, N., & Jahn, F. (2010). Health information systems. In *Health Information Systems: Architectures and Strategies* (pp. 33-42). London: Springer London.

[4] Inmon, W. H. (2005). *Building the data warehouse*. John wiley & sons.

[5] Sidey-Gibbons, J. A., & Sidey-Gibbons, C. J. (2019). Machine learning in medicine: a practical introduction. *BMC Medical Research Methodology*, *19*(1), 64.

[6] Gangwar, P. S., & Hasija, Y. (2019). Deep learning for analysis of electronic health records (EHR). In *Deep learning techniques for biomedical and health informatics* (pp. 149-166). Cham: Springer International Publishing.

[7] Blumenthal, D., & Tavenner, M. (2010). The "meaningful use" regulation for electronic health records. *New England Journal of Medicine*, *363*(6), 501-504.

[8] Rajkomar, A., Oren, E., Chen, K., Dai, A. M., Hajaj, N., Hardt, M., ... & Dean, J. (2018). Scalable and accurate deep learning with electronic health records. *NPJ digital medicine*, *1*(1), 18.

[9] Adler-milstein, J., & Pfeifer, E. (2017). Information blocking: is it occurring, and what policy strategies can address it?. *The Milbank Quarterly*, *95*(1), 117-135.

[10] Kekevi, U., & Aydın, A. A. (2022). Real-time big data processing and analytics: Concepts, technologies, and domains. *Computer Science*, *7*(2), 111-123.

[11] Safran, C., Bloomrosen, M., Hammond, W. E., Labkoff, S., Markel-Fox, S., Tang, P. C., & Detmer, D. E. (2007). Toward a national framework for the secondary use of health data: an American Medical Informatics Association White Paper. *Journal of the American Medical Informatics Association*, *14*(1), 1-9.

[12] Lewis, A. E., Weiskopf, N., Abrams, Z. B., Foraker, R., Lai, A. M., Payne, P. R., & Gupta, A. (2023). Electronic health record data quality assessment and tools: a systematic review. *Journal of the American Medical Informatics Association*, *30*(10), 1730-1740.

[13] Cios, K. J., & Moore, G. W. (2002). Uniqueness of medical data mining. *Artificial intelligence in medicine*, *26*(1-2), 1-24.

[14] Stolba, N., & Tjoa, A. M. (2006). The relevance of data warehousing and data mining in the field of evidence-based medicine to support healthcare decision-making. *International Journal of Computer Systems Science and Engineering*, *3*(3), 143-148.

[15] Armbrust, M., Ghodsi, A., Xin, R., & Zaharia, M. (2021, January). Lakehouse: a new generation of open platforms that unify data warehousing and advanced analytics. In *Proceedings of CIDR* (Vol. 8, p. 28).

[16] Bizzo, B. C., Dasegowda, G., Bridge, C., Miller, B., Hillis, J. M., Kalra, M. K., ... & Dreyer, K. J. (2023). Addressing the challenges of implementing artificial intelligence tools in clinical practice: principles from experience. *Journal of the American College of Radiology*, *20*(3), 352-360.

[17] Cabitza, F., Rasoini, R., & Gensini, G. F. (2017). Unintended consequences of machine learning in medicine. *Jama*, *318*(6), 517-518.

[18] Zaharia, M., Das, T., Li, H., Hunter, T., Shenker, S., & Stoica, I. (2013, November). Discretized streams: Fault-tolerant streaming computation at scale. In *Proceedings of the twenty-fourth ACM symposium on operating systems principles* (pp. 423-438).

[19] Armbrust, M., Das, T., Sun, L., Yavuz, B., Zhu, S., Murthy, M., ... & Zaharia, M. (2020). Delta Lake: high-performance ACID table storage over cloud object stores. *Proceedings of the VLDB Endowment*, *13*(12), 3411-3424.

[20] Habibie, K., Suhardi, S., & Muhamad, W. (2023). Implementation of data governance on the open government data management platform to improve data quality. *IJAIT (International Journal of Applied Information Technology)*, 92-102.

[21] Mandel, J. C., Kreda, D. A., Mandl, K. D., Kohane, I. S., & Ramoni, R. B. (2016). SMART on FHIR: a standards-based, interoperable apps platform for electronic health records. *Journal of the American Medical Informatics Association*, *23*(5), 899-908.

[22] Kahn, M. G., Brown, J. S., Chun, A. T., Davidson, B. N., Meeker, D., Ryan, P. B., ... & Zozus, M. N. (2015). Transparent reporting of data quality in distributed data networks. *Egems*, *3*(1), 1052.

[23] Redman, T. C. (2008). *Data-driven: profiting from your most important business asset*. Harvard Business Press.

[24] Obermeyer, Z., & Emanuel, E. J. (2016). Predicting the future: big data, machine learning, and clinical medicine. *The New England journal of medicine*, *375*(13), 1216.