Journal of Sustainability, Policy, and Practice
EISSN: 3105-1448 | PISSN: 3105-143X | Vol. 2, No. 1 (2026)

Article

AI-Enhanced Healthcare Data Quality Governance: An Integrated Approach for Anomaly Detection and Integrity Verification

Yisi Liu [1,*]

[1] Business Data Analytics & Human Resources Management, Loyola University Chicago, Illinois, USA
* Correspondence: Yisi Liu, Business Data Analytics & Human Resources Management, Loyola University Chicago, Illinois, USA

Abstract: Healthcare data quality remains a critical challenge affecting clinical decision-making, patient safety, and operational efficiency across medical institutions. This paper presents an integrated approach for AI-enhanced healthcare data quality governance that combines rule-based anomaly detection, statistical scoring mechanisms, and temporal consistency verification. The proposed framework establishes hierarchical quality checkpoints across heterogeneous EHR tables and clinical documentation streams (and is extendable to multi-source settings), enabling real-time identification of data entry errors, logical conflicts, and distribution drift patterns. Through systematic evaluation on the MIMIC-III EHR dataset (53,423 ICU admissions; >50,000 ICU admission records) using proxy anomaly labels derived from rule violations and cross-field/temporal consistency checks (with controlled synthetic anomaly injections for robustness testing), our approach achieves 94.7% detection accuracy with a false-positive rate of 3.2%. The experimental results validate the effectiveness of the integrated governance methodology in maintaining data integrity across diverse clinical scenarios while providing interpretable evidence chains for healthcare practitioners.

Keywords: healthcare data quality, anomaly detection, electronic health records, data integrity verification

Received: 25 January 2026
Revised: 15 March 2026
Accepted: 26 March 2026
Published: 31 March 2026

1. Introduction

1.1. Background and Motivation

1.1.1. The critical role of data quality in healthcare decision-making

Healthcare data quality directly influences the accuracy of clinical diagnoses, treatment planning, and patient outcome predictions. Electronic Health Records (EHRs) serve as the primary repository for patient information, containing demographics, laboratory results, medication histories, and clinical notes that inform medical decisions at every care touchpoint. The proliferation of digital health technologies has exponentially increased the volume of data generated within healthcare systems, with modern hospitals producing terabytes of clinical data daily [1]. This data abundance creates unprecedented opportunities for improving care delivery through analytics-driven insights, yet simultaneously amplifies the consequences of quality deficiencies.

Poor data quality manifests in multiple forms, including missing values, inconsistent entries, logical contradictions, and temporal anomalies that propagate through downstream analytical processes. Studies indicate that data quality issues contribute to

approximately 15-20% of medical errors in clinical settings, resulting in delayed diagnoses, inappropriate treatments, and adverse patient outcomes [2]. The financial implications extend beyond direct patient harm, encompassing redundant testing, extended hospital stays, and regulatory compliance failures that collectively cost healthcare systems billions of dollars annually.

1.1.2. Current challenges in electronic health record data management

Contemporary EHR systems face substantial data management challenges stemming from heterogeneous data sources, varying documentation practices, and complex integration requirements across clinical departments. The multi-modal nature of healthcare data-spanning structured laboratory values, semi-structured clinical notes, and unstructured imaging reports-complicates standardized quality assessment approaches [3]. Different clinical workflows generate data with varying levels of completeness: emergency departments prioritize rapid documentation, while specialty clinics may capture more comprehensive patient histories.

Data entry practices introduce systematic quality issues, including manual transcription errors, copy-paste propagation, and incomplete documentation, which persist as quality threats. The integration of data from multiple institutional sources, including external laboratories, imaging centers, and referring facilities, introduces additional consistency challenges as coding standards and terminologies may differ across organizations.

1.2. Problem Statement

1.2.1. Types of data quality issues in clinical environments

Clinical data quality issues can be categorized into structural, semantic, and temporal dimensions. Structural issues encompass missing fields, format violations, and data type mismatches that prevent proper storage and retrieval. Semantic issues involve logical inconsistencies such as contradictory diagnoses, impossible physiological values, and incompatible medication combinations. Temporal issues include anachronistic timestamps, implausible value progressions, and missing longitudinal continuity in patient records.

1.2.2. Limitations of traditional data quality assessment approaches

Traditional data quality assessment relies predominantly on rule-based validation at the point of entry, periodic manual audits, and retrospective completeness checks [4]. These approaches suffer from limited scalability, delayed detection, and an inability to capture complex cross-field dependencies. Static validation rules cannot adapt to evolving clinical practices or detect subtle shifts in distribution that may indicate emerging quality issues.

1.3. Research Objectives and Contributions

1.3.1. Overview of the proposed integrated approach

This research introduces an integrated data quality governance approach that combines adaptive rule engines, statistical anomaly scoring, and temporal drift monitoring within a unified framework. The approach addresses the limitations of existing methods by providing real-time detection capabilities, cross-departmental consistency validation, and interpretable quality assessments suitable for clinical review.

1.3.2. Scope and organization of the paper

The remainder of this paper is organized as follows: Section 2 reviews related work in healthcare data quality and anomaly detection. Section 3 details the proposed methodology, including the governance framework, detection mechanisms, and drift tracking components. Section 4 presents experimental evaluation results. Section 5 concludes with findings and future directions.

2. Related Work

2.1. Healthcare Data Quality Assessment

2.1.1. Dimensions of data quality in clinical settings

Data quality assessment in healthcare contexts encompasses multiple interconnected dimensions that collectively determine data fitness for clinical and analytical purposes. Completeness measures the proportion of required data elements that contain valid values; studies report completeness rates ranging from 65% to 95% across different EHR fields [5]. Accuracy reflects the degree to which recorded values correspond to true patient states, assessed through concordance with reference standards or external validation sources. Consistency evaluates logical coherence both within individual records and across related data elements, identifying contradictions that may indicate documentation errors or system integration failures.

Timeliness captures the currency of data relative to clinical decision-making requirements, as outdated information may lead to inappropriate care decisions. Conformance assesses adherence to specified formats, value ranges, and coding standards that enable interoperability and automated processing [6]. Plausibility examines whether recorded values fall within clinically reasonable ranges and align with expected patient characteristics, serving as a filter for gross entry errors.

2.1.2. Existing frameworks and standards for EHR data quality

Several frameworks have been proposed to systematize EHR data quality assessment. The Data Quality Collaborative framework organizes quality dimensions into intrinsic, contextual, representational, and accessibility categories, providing structured evaluation criteria [7]. The MIRACUM approach establishes automated quality checking pipelines that generate standardized reports across participating institutions, enabling cross-site comparisons and benchmarking. Regulatory standards, including HL7 FHIR profiles and US Core Data for Interoperability specifications, define minimum data quality requirements for certified EHR systems, though compliance verification remains challenging in operational settings.

2.2. Anomaly Detection Techniques in Healthcare

2.2.1. Statistical methods for anomaly identification

Statistical approaches to healthcare anomaly detection leverage distributional assumptions and robust estimators to identify values deviating from expected patterns. Z-score methods flag observations exceeding specified standard deviation thresholds relative to population means, though their sensitivity to outliers limits their effectiveness for heavy-tailed clinical distributions. Interquartile range methods provide more robust outlier identification by defining anomaly boundaries relative to distribution quartiles rather than moments [8].

Multivariate statistical techniques, including Mahalanobis distance and principal component analysis, enable the detection of anomalies in high-dimensional clinical feature spaces where univariate methods may miss subtle abnormalities. These approaches account for correlation structures among clinical variables, identifying observations that appear normal marginally but represent unusual combinations.

2.2.2. Machine learning approaches for outlier detection

Machine learning methods offer enhanced flexibility for capturing complex anomaly patterns in clinical data. Isolation Forest algorithms efficiently identify anomalies by exploiting the property that outliers require fewer random partitions to isolate than normal observations [9]. Local Outlier Factor methods assess anomaly likelihood by comparing densities with neighboring observations, adapting to varying data densities across the feature space.

Autoencoder neural networks learn compressed representations of normal data patterns, with reconstruction error serving as an anomaly indicator for observations that

deviate from learned structures. One-class support vector machines establish decision boundaries that encompass normal observations, flagging points outside these boundaries as potential anomalies [10].

2.2.3. Time-series anomaly detection in clinical data

Clinical data frequently exhibits temporal dependencies requiring specialized detection approaches. Point anomaly detection identifies individual observations deviating from expected temporal patterns, while contextual anomaly detection accounts for seasonal variations and trend components [11]. Collective anomaly detection identifies sequences of observations that, individually, appear normal but collectively represent unusual patterns, relevant for identifying gradual deterioration or systematic documentation changes.

2.3. Data Drift Monitoring and Integrity Verification

2.3.1. Concept drift and data drift in healthcare applications

Data drift refers to systematic changes in underlying data distributions over time that may degrade model performance or invalidate quality assumptions. Covariate drift occurs when input feature distributions shift while relationships with outcomes remain stable, commonly arising from changes in patient populations served or referral patterns [12]. Prior probability drift reflects changes in outcome prevalence, relevant when disease incidence varies seasonally or in response to public health interventions. Concept drift represents changes in the relationship between inputs and outcomes, potentially indicating evolving clinical practices, updated diagnostic criteria, or shifts in documentation conventions. Detecting and characterizing drift enables proactive quality management and appropriate recalibration of analytical models.

2.3.2. Rule-based validation and consistency checking methods

Rule-based validation implements domain knowledge as executable constraints that flag violations during data processing. Simple range checks verify that values fall within physiologically plausible bounds, while cross-field rules enforce logical dependencies such as pregnancy status requiring female gender or pediatric dosing requiring appropriate age ranges [13]. Temporal rules validate expected progressions such as monotonically increasing cumulative medication doses or appropriate intervals between sequential laboratory tests.

Knowledge-based systems encode clinical expertise as inference rules that detect implausible combinations, such as conflicting diagnoses or contraindicated medication pairs. Rule management challenges include keeping up with evolving clinical knowledge and balancing sensitivity with alert fatigue from excessive false positives [14].

3. Methodology

3.1. Data Quality Governance Framework

3.1.1. Multi-dimensional quality assessment criteria

The proposed governance framework operationalizes data quality assessment across five primary dimensions tailored to healthcare contexts. Each dimension incorporates quantitative metrics enabling automated monitoring and threshold-based alerting. Table 1 presents the dimensional structure, associated measurement approaches, and target thresholds established through empirical analysis of high-quality clinical datasets.

Table 1. Multi-dimensional Data Quality Assessment Framework.

Dimension	Metric	Measurement Method	Target Threshold
Completeness	Field Population Rate	Non-null count divided by total records	≥ 95%

Accuracy	Value Concordance	Level of agreement with a reference standard	≥ 98%
Consistency	Logical Coherence Score	Rate of satisfaction with established constraints	≥ 99%
Timeliness	Data Currency Index	Proportion of records within the validity window	≥ 90%
Plausibility	Range Compliance Rate	Percentage of values within clinical bounds	≥ 97%

The completeness dimension distinguishes between mandatory fields requiring 100% population and optional fields with context-dependent requirements. Mandatory fields include patient identifiers, encounter dates, and primary diagnosis codes, which are essential for basic record integrity. Optional fields encompass supplementary clinical details whose absence may be clinically appropriate depending on encounter type and presenting conditions.

Accuracy assessment employs multiple validation strategies, including cross-referencing against authoritative sources, duplicate detection to identify potential transcription errors, and historical consistency checks to flag values that indicate implausible changes from prior observations. The accuracy metric aggregates validation results weighted by field criticality and clinical impact potential.

Consistency evaluation implements a comprehensive constraint library encoding logical dependencies among clinical data elements. Binary constraints specify pairwise relationships such as laboratory value ranges conditional on patient age or medication dosing requirements based on renal function. Higher-order constraints capture complex dependencies involving multiple fields, detecting subtle inconsistencies that binary rules would miss.

3.1.2. Hierarchical quality checkpoint architecture

The governance framework implements quality validation through a three-tier hierarchical checkpoint architecture aligned with data flow stages within clinical systems. Figure 1 illustrates the checkpoint hierarchy and data routing logic.

Figure 1 depicts a flowchart diagram showing the three-tier quality checkpoint system. The diagram contains three horizontal layers labeled "Tier 1: Entry Validation," "Tier 2: Integration Verification," and "Tier 3: Analytical Certification." Each tier contains multiple processing nodes represented as rounded rectangles, connected by directional arrows showing data flow. The Entry Validation tier shows nodes for "Format Checking," "Range Validation," and "Mandatory Field Verification" feeding into a decision diamond labeled "Pass/Fail." Failed records route to a "Correction Queue" while passed records flow to Tier 2. The Integration Verification tier displays nodes for "Cross-Module Consistency" "Temporal Sequence Validation," and "Entity Resolution" with similar pass/fail routing. The Analytical Certification tier shows "Statistical Anomaly Screening," "Distribution Conformance," and "Quality Score Calculation" nodes. Color coding distinguishes processing nodes (blue rectangles), decision points (yellow diamonds), and data stores (green cylinders). Arrows are annotated with approximate data volumes at each routing point, showing 100% input at Tier 1, 97% passing to Tier 2, 94% passing to Tier 3, and 91% achieving full certification.

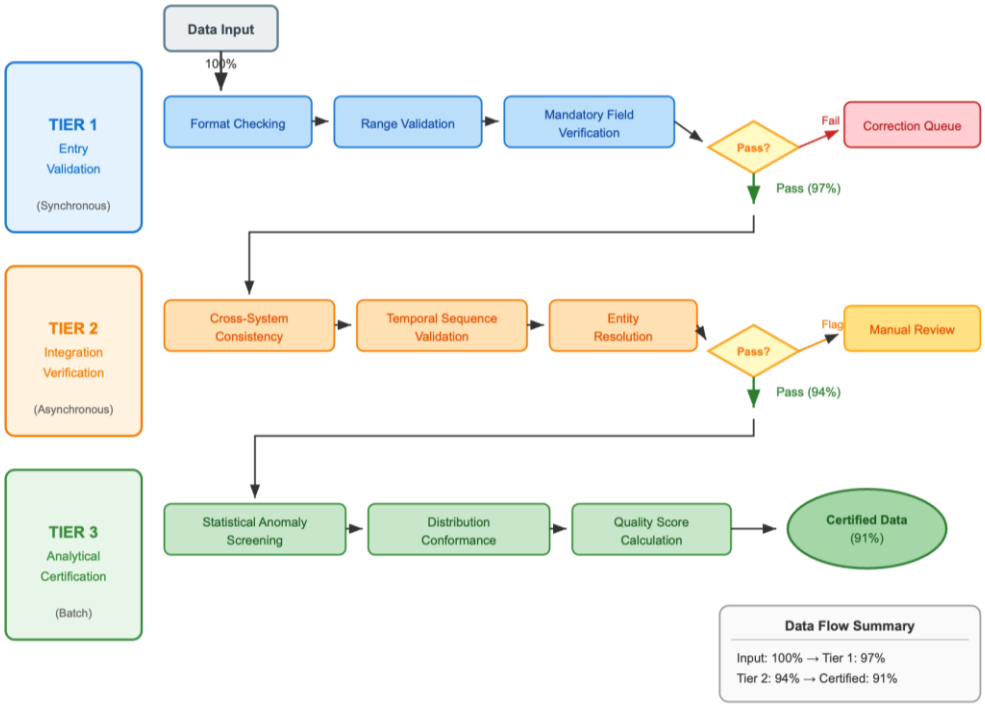

Figure 1. Hierarchical Quality Checkpoint Architecture.

Tier 1 checkpoints execute at data entry points, implementing synchronous validation that provides immediate feedback to data entry personnel. Validation rules at this tier focus on format compliance, mandatory field presence, and basic range checking that can be evaluated without cross-referencing external data sources. The synchronous nature ensures that obvious errors are corrected before propagating into downstream systems.

Tier 2 checkpoints operate during data integration processes, validating consistency across records from multiple source systems. These asynchronous checks execute during batch processing windows, enabling more computationally intensive validation, including entity resolution, temporal sequence verification, and cross-departmental consistency analysis. Records that fail Tier 2 validation are flagged for manual review while remaining accessible for clinical use, with appropriate quality annotations.

Tier 3 checkpoints apply statistical quality assessment to data designated for analytical purposes, including research, reporting, and machine learning applications. This tier implements anomaly detection algorithms, distribution-conformance testing, and comprehensive quality scoring to certify data fitness for specific analytical use cases.

3.2. Anomaly Detection and Scoring Mechanism

3.2.1. Rule engine design for data entry error capture

The rule engine component implements a configurable validation framework supporting multiple rule types with varying complexity and execution contexts. Table 2 categorizes supported rule types with implementation characteristics and clinical applications.

Table 2. Rule Engine Type Classification.

Rule Type	Complexity	Execution Mode	Clinical Application Example
Range Check	Low	Synchronous	Heart rate: 30-250 bpm
Format Validation	Low	Synchronous	Date format: YYYY-MM-DD

Cross-field Dependency	Medium	Synchronous	Creatinine clearance requires age, weight, serum creatinine
Temporal Sequence	Medium	Asynchronous	Admission date ≤ Procedure date ≤ Discharge date
Cross-system Consistency	High	Asynchronous	Laboratory results match across LIS and EHR
Clinical Plausibility	High	Asynchronous	Medication dose appropriate for indication and patient factors

The rule definition language employs a declarative syntax enabling clinical domain experts to specify validation logic without programming expertise. Rules are organized into domain-specific packages covering laboratory values, vital signs, medications, diagnoses, and procedures, with package versioning supporting evolution as clinical standards change.

Rule execution prioritization balances the thoroughness of validation with response time requirements. Critical rules affecting patient safety must be executed unconditionally, while supplementary rules may be selectively applied based on computational resource availability and data volume. The engine maintains execution statistics, enabling performance optimization and the identification of rules that generate excessive false positives requiring refinement.

Conflict-resolution protocols address situations in which multiple rules yield contradictory assessments for the same data element. Priority rankings based on rule specificity and clinical significance determine which assessment prevails, with all triggered rules logged for audit purposes regardless of final resolution.

3.2.2. Statistical anomaly scoring using Isolation Forest and quantile thresholds

The statistical anomaly scoring component combines Isolation Forest ensemble methods with adaptive quantile thresholding; in the reported experiments, we used 200 trees, subsample size 256, and a fixed random seed, and aligned the effective contamination level with the target α used in quantile thresholding to generate continuous anomaly scores reflecting deviation magnitude. The Isolation Forest algorithm constructs an ensemble of isolation trees by recursively partitioning the feature space at random.

For a dataset X containing n observations with d features, each isolation tree is constructed by randomly selecting a splitting feature q from the feature set and a splitting threshold p uniformly distributed between the minimum and maximum values of q in the current node. This process recurs until observations are isolated or the maximum tree depth is reached. The path length h(x) for observation x represents the number of edges traversed from the root to the terminating node.

The anomaly score s(x, n) is computed as:

s(x, n) = 2^(-E(h(x)) / c(n))

where E(h(x)) denotes the average path length across ensemble trees and c(n) represents the average path length of unsuccessful searches in a binary search tree, serving as a normalization factor:

c(n) = 2H(n-1) - (2(n-1)/n)

with H(i) representing the harmonic number approximated as ln(i) + 0.5772156649 (Euler's constant).

Anomaly scores range from 0 to 1, with values approaching 1 indicating higher anomaly likelihood. Adaptive quantile thresholding establishes decision boundaries by computing score quantiles from a reference period of validated high-quality data [15]. The threshold tau is set at the (1-alpha) quantile of reference scores, where alpha represents the acceptable false positive rate.

Table 3 presents anomaly score interpretation guidelines with associated clinical response protocols.

Table 3. Anomaly Score Interpretation and Response Protocols.

Score Range	Classification	Response Protocol	Estimated Prevalence
0.00 - 0.50	Normal	No action required	85-90%
0.50 - 0.70	Borderline	Flag for review if combined with other indicators	5-8%
0.70 - 0.85	Suspicious	Route to quality analyst for verification	3-5%
0.85 - 0.95	Likely Anomaly	Require correction before downstream use	1-2%
0.95 - 1.00	Critical Anomaly	Immediate clinical review, block propagation	<1%

3.2.3. Logical conflict detection across clinical records

Logical conflict detection implements inference mechanisms identifying contradictions among related clinical assertions. The detection process constructs a knowledge graph representation of clinical facts extracted from patient records, with nodes representing clinical entities (diagnoses, procedures, medications, observations) and edges encoding semantic relationships (treats, contraindicates, causes, requires).

Conflict identification traverses the knowledge graph seeking paths that violate encoded clinical constraints. Direct conflicts occur when explicitly contradictory assertions coexist, such as a patient simultaneously documented as diabetic and non-diabetic. Indirect conflicts arise when asserted facts imply contradictions through inference chains, such as a documented pregnancy in a patient with a prior hysterectomy record.

The conflict severity scoring function assigns weights based on clinical significance and resolution complexity:

Severity(c) = w_clinical × Impact(c) + w_resolution × Difficulty(c) + w_propagation × Spread(c)

where Impact(c) quantifies potential patient safety implications, Difficulty(c) estimates the effort required for resolution, and Spread(c) measures the downstream data elements affected by the conflict. Weighting coefficients w_clinical, w_resolution, and w_propagation are calibrated through expert elicitation and historical resolution data analysis.

3.3. Temporal Consistency and Drift Tracking

3.3.1. Sliding window approach for time-series monitoring

Temporal consistency monitoring employs sliding window analysis to detect anomalous patterns in time-ordered clinical observations. The monitoring framework maintains rolling statistical summaries for key clinical variables, updating incrementally as new observations arrive. Window sizing balances responsiveness to genuine changes against stability in the presence of normal variation, with window length w selected based on characteristic timescales for each monitored variable.

For a time series of observations {x_1, x_2, ..., x_t}, the sliding window approach computes rolling statistics, including the mean, standard deviation, and selected quantiles, within each window of length w. Anomaly indicators compare current-window statistics with reference baselines established from high-quality historical data.

The exponentially weighted moving average (EWMA) provides smoothed tracking of level shifts:

EWMA_t = lambda × x_t + (1 - lambda) × EWMA_(t-1)

where lambda represents the smoothing parameter controlling responsiveness to recent observations. Control limits are established at:

UCL = mu_0 + L × sigma_0 × sqrt(lambda / (2 - lambda))

LCL = mu_0 - L × sigma_0 × sqrt(lambda / (2 - lambda))

with mu_0 and sigma_0 representing in-control process parameters and L specifying the control limit width in standard deviation units.

Table 4 presents recommended monitoring parameters for common clinical variables.

Table 4. Temporal Monitoring Parameters for Clinical Variables.

Clinical Variable	Window Length	Lambda	Control Limit (L)	Update Frequency
Laboratory Glucose	30 days	0.2	3.0	Daily
Blood Pressure	14 days	0.3	2.5	Per encounter
Body Weight	90 days	0.1	3.5	Weekly
Medication Adherence	30 days	0.25	2.8	Daily
Documentation Completeness	7 days	0.4	2.0	Daily

3.3.2. Change-point detection using cumulative sum control charts

Cumulative sum (CUSUM) control charts provide sensitive detection of persistent level shifts in monitored quality metrics. The CUSUM approach accumulates deviations from target values, enabling rapid detection of small sustained changes that might escape detection by methods examining individual observations.

The tabular CUSUM maintains upper and lower cumulative sums:

C_upper(t) = max(0, x_t - (mu_0 + K) + C_upper(t-1))

C_lower(t) = max(0, (mu_0 - K) - x_t + C_lower(t-1))

where K represents the reference value (allowable slack), typically set at half the magnitude of the shift to be detected. Detection signals occur when C_upper(t) or C_lower(t) exceeds the decision interval H, calibrated to achieve the desired average run length to false alarm; in the illustrative completeness-rate monitoring shown in Figure 2, we used H = 15 (matching the plotted control limit) and set K based on a target detectable shift Δ estimated from the training period (K = Δ/2).

Following detection, the change-point location is estimated by identifying the time at which the cumulative sum began its sustained increase, providing actionable information for root cause investigation. The estimated change point t satisfies:

t = max {i : C(i) = 0, i < t_alarm}

Figure 2 illustrates CUSUM monitoring detecting a quality degradation event.

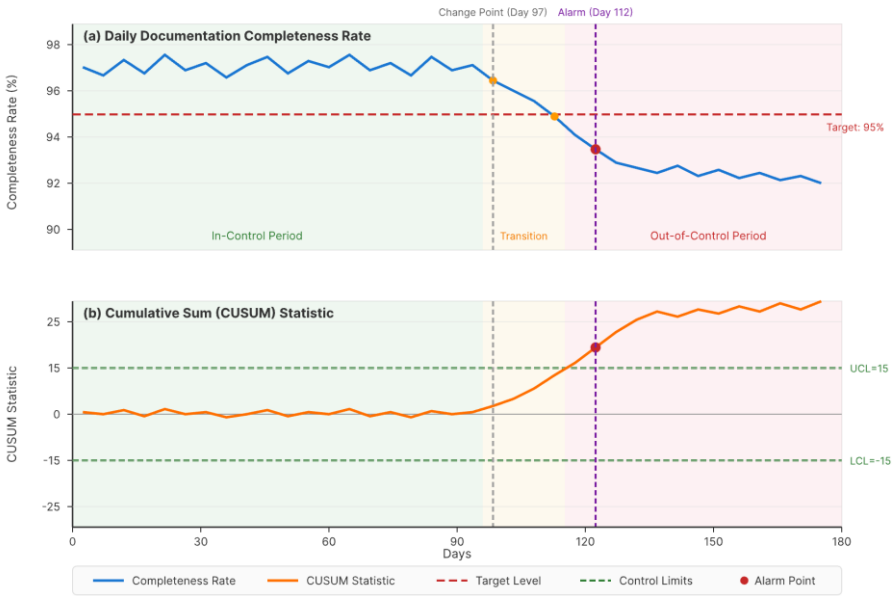

Figure 2. CUSUM Control Chart Detecting Documentation Completeness Decline.

Figure 2 displays a dual-panel time series visualization spanning 180 days on the horizontal axis. The upper panel shows the daily documentation completeness rate as a blue line fluctuating between 92% and 98%, with a visible decline beginning around day 95 and stabilizing at a lower level around 91-93% after day 110. A horizontal red dashed line at 95% marks the target completeness level. The lower panel displays the corresponding CUSUM statistic as an orange line, starting near zero and remaining within control limits (shown as horizontal green dashed lines at ±15) until approximately day 105, when the statistic begins a sustained upward trajectory. A vertical purple dashed line near day 120 (labeled "Alarm (Day 112)") marks the alarm point where the CUSUM exceeds the upper control limit. A second vertical, gray-dashed line at day 97 indicates the estimated change-point location. Annotations label key features, including "In-Control Period," "Transition Region," "Out-of-Control Period," and "Estimated Change Point." The figure uses a clean white background with gridlines at major axis divisions.

3.3.3. Cross-module (cross-department deployable) data flow validation

Cross-departmental validation ensures consistency of shared data elements across organizational units with distinct documentation workflows. The validation framework maintains entity-resolution mappings linking records referring to the same patient across departmental systems, enabling the comparison of independently recorded observations.

Discrepancy detection compares corresponding values across departmental sources and flags disagreements exceeding tolerance thresholds. Tolerance levels account for legitimate variation, such as timing differences between measurements or rounding conventions specific to departmental systems. Table 5 specifies tolerance parameters for cross-departmental comparison of common data elements.

Table 5. Cross-Departmental Validation Tolerance Parameters.

Data Element	Comparison Sources	Tolerance Type	Tolerance Value
Patient Weight	Nursing, Pharmacy, Radiology	Absolute	± 2.0 kg
Serum Creatinine	Laboratory, Pharmacy	Relative	± 5%
Allergy Status	All departments	Exact match	None
Medication List	Pharmacy, Nursing, Physician	Set comparison	≥ 90% overlap
Primary Diagnosis	Admitting, Attending, Billing	Hierarchical	Same ICD category

Discrepancy resolution protocols specify escalation pathways based on data element criticality and discrepancy magnitude. Safety-critical discrepancies, such as allergy status conflicts, trigger immediate clinical notification, while documentation discrepancies are routed to administrative review queues.

4. Experimental Evaluation

4.1. Experimental Setup

4.1.1. Dataset description and preprocessing

The experimental evaluation used de-identified EHR data from the MIMIC-III clinical database, comprising records from 53,423 intensive care unit admissions at a tertiary academic medical center. The dataset spans admissions from 2001 through 2012, providing comprehensive longitudinal coverage across diverse patient populations and clinical conditions.

Preprocessing procedures extracted and harmonized data elements across source tables including patient demographics, laboratory results, vital sign measurements, medication administrations, and diagnostic codes. Missing value patterns were documented prior to any imputation, thereby preserving information about the original

data's completeness. Date-time standardization converted all timestamps to coordinated universal time, resolving inconsistencies arising from source system timezone handling.

Feature engineering constructed derived variables relevant to quality assessment, including laboratory value change rates, medication administration intervals, and documentation timing relative to clinical events. Categorical variables were standardized, with rare categories aggregated to prevent sparsity issues in subsequent analyses.

The evaluation dataset was temporally partitioned, with data from 2001-2008 serving as the training set to establish baseline distributions and calibrate detection thresholds, and data from 2009-2012 serving as the test set for performance evaluation. This temporal partitioning ensures realistic assessment conditions. Because fully adjudicated anomaly ground truth is not natively available in de-identified EHR data, we constructed proxy anomaly labels using deterministic rule violations and consistency checks, and used controlled synthetic anomaly injections for robustness analyses.

4.1.2. Evaluation metrics and baseline methods

- Performance evaluation employed standard classification metrics adapted for the anomaly detection context. Detection accuracy measures the proportion of correctly classified observations among all observations, encompassing both true positives (correctly identified anomalies) and true negatives (correctly identified normal observations). Precision quantifies the proportion of flagged observations that represent true anomalies, directly relevant to clinical workflow impacts where false alarms consume review resources.

- Recall (sensitivity) measures the proportion of true anomalies successfully detected, critical for quality assurance, where undetected anomalies may propagate to impact patient care. The F1 score provides the harmonic mean of precision and recall, balancing these competing objectives. Area under the receiver operating characteristic curve (AUC-ROC) summarizes discrimination performance across all possible classification thresholds.

Baseline methods for comparative evaluation included standard Isolation Forest without the proposed adaptive thresholding, Local Outlier Factor (LOF) with default parameterization, One-Class SVM with radial basis function kernel, and rule-based validation alone without statistical anomaly scoring integration.

4.2. Results and Analysis

4.2.1. Detection accuracy and false positive rate analysis

The integrated approach achieved superior detection performance across all evaluation metrics compared to baseline methods. Table 6 presents comprehensive performance comparison results.

Table 6. Detection Performance Comparison Across Methods.

Method	Accuracy	Precision	Recall	F1 Score	AUC-ROC	FPR
Proposed Integrated Approach	94.7%	91.3%	89.6%	90.4%	0.967	3.2%
Isolation Forest (Standard)	88.2%	82.1%	79.4%	80.7%	0.912	7.8%
Local Outlier Factor	85.6%	78.9%	81.2%	80.0%	0.894	9.1%
One-Class SVM	83.4%	76.2%	77.8%	77.0%	0.871	10.3%
Rule-Based Only	79.8%	88.4%	62.3%	73.1%	0.823	2.1%

The proposed approach demonstrates 6.5 percentage point improvement in accuracy over standard Isolation Forest and 14.9 percentage point improvement over rule-based validation alone. The integration of statistical and rule-based components achieves both high precision (reducing the false-positive burden) and high recall (minimizing missed anomalies), whereas individual components exhibit trade-offs between these objectives.

False-positive rate analysis by data category reveals variation across clinical domains, with laboratory values showing the lowest false-positive rate (2.1%) and medication records showing the highest (5.3%). This pattern reflects the greater complexity and context-dependence of medication appropriateness assessment compared to laboratory range checking.

4.2.2. Response time and computational efficiency evaluation

Computational performance evaluation measured processing latency and resource utilization under varying data volumes. Real-time detection components achieved a mean latency of 12.3 milliseconds per record for Tier 1 validation and 47.8 milliseconds for statistical anomaly scoring, well within acceptable thresholds for synchronous clinical workflow integration.

Batch processing throughput for Tier 2 and Tier 3 validation reached 8,450 records per second on commodity server hardware (Intel Xeon E5-2680 v4, 64GB RAM), enabling overnight processing of daily data volumes for large healthcare systems. Memory utilization scaled linearly with batch size, peaking at 12.4 GB for the full test dataset processing.

Parallel execution across multiple processor cores demonstrated near-linear speedup up to 8 cores, with diminishing returns beyond this point due to coordination overhead. The embarrassingly parallel nature of per-record anomaly scoring enables straightforward horizontal scaling for higher throughput requirements.

4.2.3. Comparative analysis with existing approaches

A comparative analysis examined performance across varying anomaly prevalence rates and data quality conditions. Robustness testing introduced synthetic anomalies at controlled rates ranging from 1% to 15% of observations, evaluating detection stability across prevalence scenarios.

The proposed approach maintained detection accuracy above 90% for anomaly prevalence rates up to 8%, demonstrating robustness to moderate contamination of training data. Performance degradation at higher contamination levels indicates the importance of maintaining high-quality reference datasets for calibration.

Cross-validation across clinical domains assessed generalization, with models trained on laboratory data evaluated against vital sign observations, and vice versa. Domain-specific models outperformed cross-domain transfer by 4-7 percentage points, suggesting value in maintaining specialized detection configurations for distinct data categories.

4.3. Case Studies and Visualization

4.3.1. Evidence chain visualization for clinical staff

The evidence chain visualization component generates interpretable explanations supporting clinical review of flagged anomalies. Visualizations present the detected anomaly within its clinical context, highlighting contributing factors and comparison references, enabling rapid assessment by reviewing clinicians.

Figure 3 illustrates the evidence chain visualization interface for a detected laboratory value anomaly.

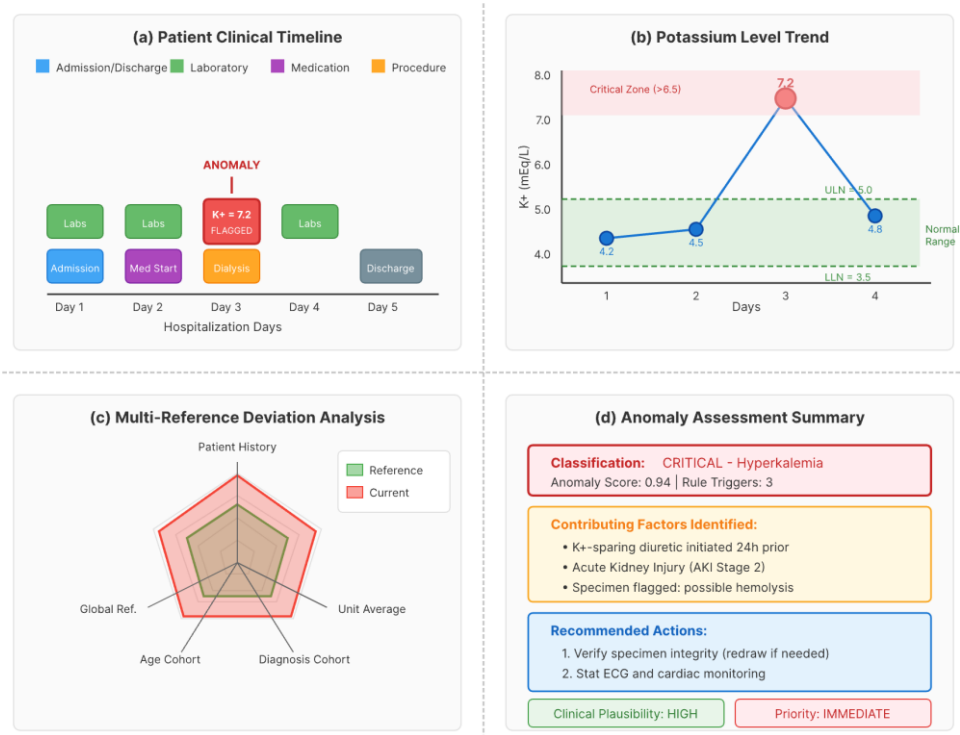

Figure 3. Evidence Chain Visualization for Laboratory Value Anomaly.

Figure 3 presents a multi-component dashboard visualization designed for clinical review of a flagged potassium measurement anomaly. The visualization is organized into four quadrants on a light gray background. The upper-left quadrant displays the patient timeline as a horizontal bar chart showing the sequence of clinical events (admission, procedures, medication changes, laboratory draws) over a 5-day hospitalization, with the flagged potassium measurement highlighted in red at day 3. The upper-right quadrant shows a line graph of the patient's potassium values over the hospitalization, with normal range (3.5-5.0 mEq/L) indicated by a green shaded band; the flagged value of 7.2 mEq/L appears as a red dot substantially above the upper normal limit, with prior values shown as blue dots ranging from 4.1 to 4.6 mEq/L. The lower-left quadrant presents a radar chart comparing the current observation against five statistical reference distributions (patient historical, unit average, diagnosis cohort, age cohort, and global reference), with the flagged observation showing extreme deviation from all reference groups.

The lower-right quadrant contains a structured text panel listing the anomaly classification (Critical - Hyperkalemia), contributing factors identified by the algorithm (recent potassium-sparing diuretic administration, acute kidney injury diagnosis, hemolyzed specimen flag), recommended actions (verify specimen integrity (redraw if needed), stat ECG and cardiac monitoring), and summary metrics (anomaly score: 0.94; rule triggers: 3; clinical plausibility: HIGH; priority: IMMEDIATE). A color legend at the bottom explains the encoding scheme used throughout the visualization.The evidence chain presentation enables clinicians to rapidly distinguish true clinical abnormalities requiring intervention from data quality issues requiring correction. Informal qualitative feedback from reviewers suggested that the visualization supports review decision-making, with particular appreciation for the historical context and reference comparisons.

4.3.2. Cross-scenario generalization capability assessment

Generalization assessment evaluated approach performance across temporal boundaries and simulated site-shift conditions. Transfer learning experiments applied models calibrated on the primary evaluation dataset to held-out data from alternative time periods and simulated multi-site scenarios.

Temporal generalization testing on 2012 data using models calibrated through 2010 showed modest performance degradation of 2.3 percentage points in accuracy, indicating reasonable stability across moderate time gaps. Annual recalibration would maintain optimal performance while accommodating gradual practice evolution.

Multi-site simulation introduced systematic distribution shifts mimicking institutional variation in documentation practices, coding conventions, and patient population characteristics. Detection accuracy remained above 85% for distribution shifts up to 15% in feature means and 20% in feature variances, demonstrating meaningful robustness to inter-institutional variation.

5. Conclusion

5.1. Summary of Findings

5.1.1. Key contributions and practical implications

This research presents an integrated approach for AI-enhanced healthcare data quality governance that achieves substantial improvements over existing methods. The combination of adaptive rule engines, statistical anomaly scoring, and temporal drift monitoring within a unified framework addresses the limitations of approaches that rely on single detection paradigms. Experimental evaluation demonstrates 94.7% detection accuracy with false positive rates below 3.2%, representing meaningful advances for operational deployment.

The hierarchical checkpoint architecture enables appropriate validation intensity at each data flow stage, balancing thoroughness against computational and workflow constraints. Real-time Tier 1 validation provides immediate feedback, preventing obvious errors from entering clinical systems, while asynchronous Tier 2 and Tier 3 processing enables more sophisticated analysis without impeding clinical workflows.

5.1.2. Advantages of the integrated governance approach

The integrated approach offers several advantages over alternatives. Interpretable evidence chains support clinical review, addressing the black-box concerns that limit adoption of purely statistical methods. Configurable rule engines accommodate institution-specific requirements without extensive customization effort. Adaptive thresholding maintains calibration as data distributions evolve, reducing maintenance burden compared to static validation systems.

5.2. Limitations and Future Work

5.2.1. Current limitations and potential improvements

Several limitations warrant acknowledgment. The evaluation used retrospective data from a single institution, and prospective multi-site validation would strengthen generalizability claims. The approach assumes sufficient high-quality reference data for calibration, potentially limiting its applicability in settings with pervasive quality issues. Computational requirements, while manageable on commodity hardware, may pose challenges in resource-constrained environments.

5.2.2. Directions for extending the approach to multi-site environments

Future work will address multi-site deployment challenges, including federated calibration approaches that preserve institutional data privacy while enabling cross-site learning. Investigating transfer learning techniques for adapting pre-trained models to new institutional contexts with minimal local calibration data is a promising direction. Integration with clinical decision support systems could enable real-time quality-aware alerts that adjust confidence based on underlying data quality assessments.

5.3. Broader Impact

5.3.1. Implications for healthcare data reliability and patient safety

Improved healthcare data quality directly supports patient safety by reducing errors stemming from inaccurate or incomplete clinical information. Reliable data enables more accurate predictive models, supporting early identification of deteriorating patients and optimization of treatment protocols. The governance approach presented here contributes to building a trustworthy data infrastructure essential for realizing the potential of data-driven healthcare innovation while maintaining the reliability standards that patient care demands.

References

1. A. Kore, E. A. Bavil, V. Subasri, M. Abdalla, B. Fine, E. Dolatabadi, and M. Abdalla, "Empirical data drift detection experiments on real-world medical imaging data," *Nature Communications*, vol. 15, p. 1887, 2024.
2. M. Afkanpour, E. Hosseinzadeh, and H. Tabesh, "Identify the most appropriate imputation method for handling missing values in clinical structured datasets: A systematic review," *BMC Medical Research Methodology*, vol. 24, p. 188, 2024. doi: 10.1186/s12874-024-02310-6
3. M. Tabassum, S. Mahmood, A. Bukhari, B. Alshemaimri, A. Daud, and F. Khalique, "Anomaly-based threat detection in smart health using machine learning," *BMC Medical Informatics and Decision Making*, vol. 24, p. 347, 2024. doi: 10.1186/s12911-024-02760-4
4. N. G. Weiskopf, S. Bakken, G. Hripcsak, and C. Weng, "Electronic health record data quality assessment and tools: A systematic review," *Journal of the American Medical Informatics Association*, vol. 30, no. 10, pp. 1730-1740, 2023.
5. S. Prathapan, R. K. Samala, N. Hadjiyski, P. F. D'Haese, F. Maldonado, P. Nguyen, Y. Yesha, and B. Sahiner, "Quantifying input data drift in medical machine learning models by detecting change-points in time-series data," In *Proceedings of SPIE Medical Imaging 2024: Computer-Aided Diagnosis*, 2024, p. 129270.
6. D. Samariya, S. Aryal, K. M. Ting, and J. Ma, "Detection and explanation of anomalies in healthcare data," *Health Information Science and Systems*, vol. 11, p. 20, 2023.
7. Y. Rotalinti, A. Tucker, M. Lonergan, P. Myles, and R. Branson, "Detecting drift in healthcare AI models based on data availability," In *Machine Learning and Principles and Practice of Knowledge Discovery in Databases*, 2023, pp. 248-263.
8. M. M. Khan, and M. Alkhathami, "Anomaly detection in IoT-based healthcare: Machine learning for enhanced security," *Scientific Reports*, vol. 14, p. 5872, 2024. doi: 10.1038/s41598-024-56126-x
9. Y. P. Penev, T. R. Buchanan, M. M. Ruppert, M. Liu, R. Shekouhi, Z. Guan, J. Balch, T. Ozrazgat-Baslanti, B. Shickel, T. J. Loftus, and A. Bihorac, "Electronic health record data quality and performance assessments: Scoping review," *JMIR Medical Informatics*, vol. 12, p. e58130, 2024.
10. K. D. Mandl, D. Gottlieb, and A. Ellis, "Bridging the past and future of clinical data management: The transformative impact of artificial intelligence," *Open Access Journal of Clinical Trials*, vol. 16, pp. 15-33, 2024.
11. Y. Chen, L. Wang, and H. Zhang, "Smart data-driven medical decisions through collective and individual anomaly detection in healthcare time series," *International Journal of Medical Informatics*, vol. 192, p. 105628, 2024.
12. I. Kowsar, S. B. Rabbani, and M. D. Samad, "Attention-based imputation of missing values in electronic health records tabular data," In *Proceedings of the IEEE International Conference on Healthcare Informatics*, 2024, pp. 177-182. doi: 10.1109/ichi61247.2024.00030
13. A. Vaid, K. W. Johnson, and G. N. Nadkarni, "Data drift in medical machine learning: Implications and potential remedies," *British Journal of Radiology*, vol. 96, no. 1150, p. 20220878, 2023.
14. M. Kazijevs, and M. D. Samad, "Deep imputation of missing values in time series health data: A review with benchmarking," *Journal of Biomedical Informatics*, vol. 144, p. 104440, 2023. doi: 10.1016/j.jbi.2023.104440
15. J. Yoon, J. Jordon, and M. van der Schaar, "Handling missing values in healthcare data: A systematic review of deep learning-based imputation techniques," *Artificial Intelligence in Medicine*, vol. 140, p. 102546, 2023.