International Journal of Artificial Intelligence, Data Science, and Machine Learning
Grace Horizon Publication | Volume 6, Issue 4, 218-226, 2025
ISSN: 3050-9262 | https://doi.org/10.63282/3050-9262.IJAIDSML-V6I4P130

Original Article

Adaptive Data Quality Management for Multi-Cloud Healthcare Warehouses: FHIR-Aware Semantics and Unsupervised Thresholding

Sai Kiran Yadav Battula
Independent Researcher, Pittsburgh, Pennsylvania, United States.

Received On: 28/10/2025 ***Revised On: 02/12/2025*** ***Accepted On: 10/12/2025*** ***Published on: 27/12/2025***

Abstract - *The rapid proliferation of multi-cloud architectures in healthcare promises elastic scalability and regional redundancy, but it also introduces acute challenges in data consistency, latency, and governance. Traditional, centrally orchestrated, rule-based Data Quality Management (DQM) tools are ill-equipped to handle the volume, heterogeneity, and semantic complexity of distributed electronic health records (EHRs) and claims data. As schemas drift and new data sources are onboarded, static checks generate escalating false positives, incur avoidable data movement costs, and contribute to "data swamps" that compromise clinical decision-making. This paper presents an Adaptive Data Quality Management framework for multi-cloud healthcare warehouses that combines unsupervised anomaly detection with FHIR-aware semantic validation. The framework deploys lightweight quality components alongside analytic workloads to profile and score data streams, while a cloud-agnostic control layer dynamically adjusts quality thresholds using rolling statistics over anomaly scores. A FHIR-based semantic distance metric decomposes deviations into structural, vocabulary, and cardinality components, enabling graded policies rather than binary pass/fail checks.*

Using a synthetic but structurally realistic workload of approximately 500,000 patients generated by a Synthea-style engine and partitioned across AWS, Azure, and GCP, we evaluate the framework under controlled "chaos engineering" scenarios including schema drift, value-set drift, and volume anomalies. Compared with a centralized, rule-based DQM baseline, the adaptive framework reduces false-positive quality alerts by roughly 40% while increasing precision from about 0.62 to 0.74 at comparable recall. These results demonstrate that combining FHIR-aware semantics with unsupervised, adaptively thresholded quality scoring can substantially reduce noise in quality monitoring while preserving anomaly detection performance in multi-cloud healthcare analytics and decision-support systems.

Keywords *- Multi-Cloud Healthcare, Data Quality Management, FHIR, Semantic Validation, Schema Drift, Unsupervised Anomaly Detection, Isolation Forest, Autoencoder, Adaptive Thresholding, Synthetic Health Data.*

1. Introduction

The digitization of healthcare has progressed from siloed, on-premise transactional systems to highly distributed, cloud-native platforms that span multiple regions and providers. Health systems, payers, and digital health vendors increasingly rely on multi-cloud architectures, combining services from AWS, Microsoft Azure, and Google Cloud Platform (GCP), to integrate electronic health records (EHRs), claims, Internet of Medical Things (IoMT) telemetry, and social determinants of health (SDOH) into unified analytics environments.

While this multi-cloud strategy delivers elastic scalability, geographic redundancy, and vendor diversification, it simultaneously amplifies the complexity of Data Quality Management (DQM). Data now flows through heterogeneous storage engines, schema conventions, security policies, and regional regulations, making it harder to maintain consistency, trust, and governance across the analytic surface. Under regulatory frameworks such as the ONC's HTI-1 final rule, where transparency and safe use of predictive decision-support interventions are emphasized, data quality incidents that silently propagate into risk scores, risk-adjusted payment models, or population health dashboards pose both clinical and compliance risk.

1.1. The Limitations of Traditional DQM in Multi-Cloud Settings

Traditional DQM practices in healthcare evolved around centralized data warehouses. Data was periodically extracted from source systems, loaded into a single staging area, and subjected to static rule sets, "age must be between 0 and 120," "discharge date must be after admission date," "diagnosis code must belong to a known value set", before being committed to curated marts. These approaches assume that:

1. Data can be copied into a single inspection zone with negligible cost.

2. Schemas and coding practices are relatively stable over time.
3. Quality can be expressed as a fixed, hand-tuned collection of rules.

In modern multi-cloud healthcare environments, each of these assumptions breaks down:

1.1.1. Schema Drift and Rule Fragility.

Hospitals upgrade EHR platforms, payers refine claims adjudication logic, and IoMT vendors revise firmware. These changes can introduce new columns, modify datatypes, or alter coding distributions without prior notice. Static rule sets quickly fall out of sync with reality, generating floods of false-positive alerts or, worse, failing silently when rules no longer apply.

1.1.2. Semantic Heterogeneity.

Diagnoses, lab results, and procedures may be encoded using different vocabularies (ICD-10-CM, SNOMED CT, LOINC, RxNorm) or local codes across systems. Traditional DQM tools concentrate on syntactic validity (non-null, datatype, simple ranges) rather than semantic relationships (e.g., whether a code actually belongs to the expected value set for a measure).

1.1.3. The Observability Gap.

Data observability tooling has improved visibility into freshness, volume, and pipeline failures, but it remains largely system-centric. These tools can indicate that "something changed" in a particular dataset but rarely explain whether the change is clinically meaningful, benign, or catastrophic for downstream decision-support. In practice, this gap can cause subtle but serious failures, for example, a schema change in an Epic or Cerner upgrade that silently drops fields used to populate care-gap registries, or a shift in diagnosis coding that causes readmission risk models to under-estimate risk for specific subpopulations. Without semantics-aware quality signals, such issues may go undetected until they surface as anomalous quality scores, denial patterns, or adverse events months later.

The result is a growing disconnect between the sophistication of multi-cloud infrastructure and the brittleness of quality controls. Organizations oscillate between over-sensitive rules that produce alert fatigue and blind spots where consequential drifts escape detection.

1.2. Adaptive, Semantics-Aware Data Quality

To address this gap, there is a need for DQM approaches that:

1. Adapt Automatically to changing data distributions and schemas rather than relying solely on manual rule curation.
2. Reason About Semantics, not just structure, using standards such as HL7 FHIR to validate that codes, relationships, and temporal patterns align with clinical expectations.
3. Scale Across Clouds, supporting heterogeneous analytic stacks and storage systems without enforcing a single data platform.

This paper focuses on the methodological core of such a framework: the combination of unsupervised anomaly detection with FHIR-aware semantic scoring and adaptive thresholding. We assume a multi-cloud setting but treat deployment and infrastructure choices (e.g., edge vs. serverless) as secondary to the underlying quality logic; those aspects are explored in a companion paper.

1.3. Contributions

1.3.1. This work makes three main contributions:

1.3.1.1. FHIR-Aware Semantic Distance Metric.

We formalize a semantic distance metric for FHIR resources that decomposes deviations into structural, vocabulary, and cardinality components. This allows the quality layer to distinguish "technically invalid but still useful" records from genuinely broken ones and to apply graded policies rather than binary pass/fail outcomes.

1.3.1.2. Unsupervised Anomaly Detection and Adaptive Thresholding for Healthcare DQM.

We integrate Isolation Forest (iForest) for efficient point-anomaly detection and deep autoencoders for distributional drift detection, coupled with exponentially weighted moving average (EWMA) thresholds over anomaly scores. This combination reduces rule fragility and alert fatigue while preserving sensitivity to clinically relevant shifts.

1.3.1.3. Empirical Evaluation on Synthetic, FHIR-Compatible Multi-Cloud Workloads.

Using a Synthea-style generator, we construct a synthetic yet structurally realistic FHIR dataset (~500k patients over five years) partitioned across AWS, Azure, and GCP. We then inject controlled "chaos" scenarios, schema drift, vocabulary drift, and volume anomalies, and show that the adaptive, semantics-aware approach yields approximately 40% fewer false positives than a centralized, static rule baseline at comparable or higher recall.

The remainder of this paper is organized as follows. Section 2 reviews related work in healthcare data quality, FHIR-based validation, synthetic data, and unsupervised anomaly detection. Section 3 introduces the semantic distance model. Section 4 details the unsupervised detection and adaptive thresholding methodology. Section 5 describes the experimental setup and results. Section 6 discusses implications and threats to validity, and Section 7 concludes.

2. Related Work

2.1. Data Quality in Healthcare

Data quality has long been recognized as a critical bottleneck in clinical research, quality measurement, and operational analytics. Foundational work by Wang and Strong framed data quality as multi-dimensional, encompassing accuracy, completeness, timeliness, consistency, and relevance from the consumer's perspective rather than as a single scalar "score." Subsequent healthcare-focused reviews have cataloged recurring quality issues in EHR and claims data, linking them to misclassification of

conditions, biased effect estimates, and unreliable performance indicators.

Frameworks for healthcare DQM often combine rule engines with governance processes and manual review. A harmonized terminology for assessing the secondary use of EHR data emphasizes that quality must be judged relative to context: data may be "good enough" for population-level surveillance but insufficient for patient-level decision support. However, most published frameworks assume centralized warehouses and batch-oriented checks. They provide limited guidance on adapting quality controls to multi-cloud, streaming, or rapidly evolving data landscapes.

2.2. FHIR-Based Validation and Semantic Interoperability

The HL7 FHIR standard has emerged as the dominant approach to representing and exchanging healthcare information as web-friendly resources. FHIR profiles specify structural constraints (elements, datatypes, cardinalities) and bind elements to terminology value sets (e.g., ICD-10, SNOMED CT, LOINC). Official validators and open-source libraries such as HAPI FHIR can enforce these constraints at ingest time.

Recent work has extended FHIR validation beyond syntax, leveraging FHIR-encoded data for quality measurement and exploring how resource-level constraints affect downstream analytics and reporting. Initiatives such as NCQA's Bulk FHIR quality efforts and eCQM-on-FHIR implementation guides highlight both the potential and fragility of FHIR-based pipelines: conformant resources can still be semantically inconsistent (e.g., clinically impossible combinations, inconsistent cross-resource references) or incomplete for specific quality measures.

Most FHIR-focused tools remain point-in-time validators; they are applied at the perimeter of a system or exchange. Less attention has been given to continuous FHIR-aware quality scoring within analytic warehouses, or to integrating FHIR semantics with machine-learned anomaly detection.

2.3. Unsupervised Anomaly Detection and Observability

Unsupervised anomaly detection methods, such as Isolation Forest and autoencoders, are widely used to detect outliers and distribution shifts in settings where labeled "bad data" is scarce. Isolation Forest isolates anomalies by randomly partitioning the feature space; anomalous points tend to require fewer splits. Autoencoders, by contrast, learn a compressed representation of "normal" data and flag records with high reconstruction error as potential anomalies.

In parallel, data observability platforms have emerged to monitor pipeline health via metrics on volume, freshness, latency, and schema changes. These systems typically apply generic anomaly detection over metadata (e.g., row counts, null rates, schema signatures). While they can flag unexpected behaviors, they often lack domain semantics and generate noisy alerts when legitimate workflow changes occur, for example, a new documentation policy that alters diagnosis distributions.

Existing work therefore provides building blocks, unsupervised models and observability patterns, but does not directly address how to combine them with healthcare semantics in a way that is both adaptive and clinically meaningful.

2.4. Synthetic Health Data and Evaluation

Synthetic health data has become a practical tool for evaluating algorithms without exposing real protected health information (PHI). Synthea, an open-source synthetic patient generator, simulates disease progression and care pathways to produce realistic-but-not-real patient records in formats such as FHIR, C-CDA, and CSV. Synthetic populations such as SyntheticMass have been used to benchmark algorithms and pipelines under realistic yet privacy-safe conditions.

Best-practice guidance stresses documenting generator assumptions, target distributions, and validation procedures to ensure transparency about how synthetic data differs from real-world PHI. These recommendations are especially important for evaluating DQM and observability systems, which must cope not only with realistic baseline data but also with the "messiness" induced by schema changes, coding drift, and pipeline failures.

Our work follows this guidance by using a Synthea-style generator and explicitly injecting data quality perturbations to stress-test the proposed semantics-aware, unsupervised DQM framework.

3. FHIR-Aware Semantic Distance Metric

A central contribution of this paper is a semantic distance function that translates FHIR validation results into a continuous quality signal. Rather than treating FHIR validation as binary (pass vs. fail), we define a graded measure of how far a given resource deviates from its intended profile.

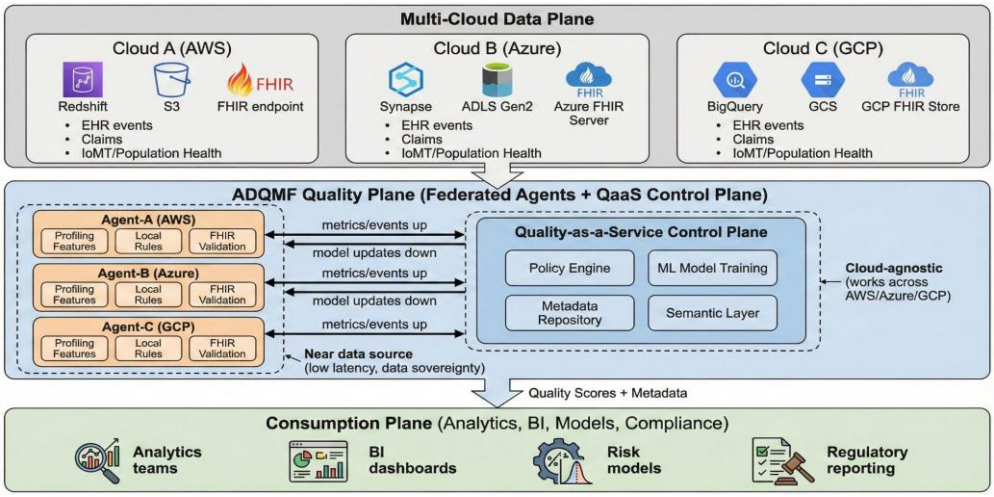

Fig 1: ADQMF Multi-Cloud Architecture (3-Layer Conceptual Diagram)

3.1. Design Objectives

The semantic distance metric, denoted $Dsem(R, P)$ for a FHIR resource R and profile P, is designed to:

- Capture different types of deviations (missing elements vs. incorrect codes vs. cardinality violations).
- Support soft policies (e.g., allow minor vocabulary mismatches under certain workflows) rather than enforcing a single hard boundary.
- Integrate naturally with anomaly scores from unsupervised models, enabling combined decision logic.

3.2. Decomposition into Structural, Vocabulary, and Cardinality Components

We decompose $Dsem$ into three components:

$$D_{\text{sem}}(R,P) = w_s \cdot \delta_{struct}(R,P) + w_v \cdot \delta_{vocab}(R,P) + w_c \cdot \delta_{card}(R,P)$$

Where:

- $\delta struct$ (Structural Distance): The proportion of required elements that are missing or malformed relative to the profile's snapshot definition, plus a penalty for unexpected elements when the profile disallows extensions.

$\delta vocab$ (Vocabulary Distance):
A terminology-aware penalty. For each coded element:

- Distance 0 if the code is in the required value set.
- Small penalty (e.g., 0.2) if the code is in a closely related hierarchy (e.g., via SNOMED CT "is-a" relationships).
- Larger penalty (e.g., 0.7) for codes in an allowed but loosely related set.
- Maximum penalty (1.0) for local or unmapped codes or free text.

The element-level penalties are aggregated (e.g., via mean or max) to yield $ab \in [0,1]$.

- $\delta card$(Cardinality Distance): A measure of how often cardinality constraints are violated, for example, missing required repeats (min > 0) or exceeding maximum occurrences.

Weight Selection and Interpretation

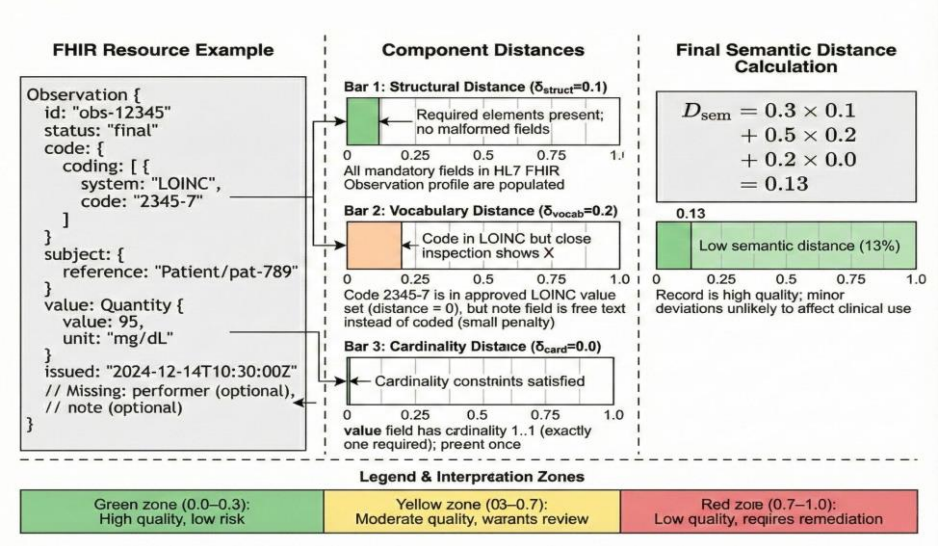

Fig 2: Semantic Distance Metric Decomposition (Component Example)

The weights w_s, w_v, w_c express the relative importance of each dimension for a given workflow and quality context. In this work, we select weights to reflect typical analytic and quality-measurement workloads:

$$w_s = 0.3, w_v = 0.5 \; and \; w_c = 0.2$$

These values emphasize the importance of correct clinical vocabularies and value sets (e.g., LOINC, ICD-10-CM, SNOMED CT), which are critical for clinical decision-making and quality measure computation, while still penalizing missing mandatory fields and cardinality violations. The weights were chosen in consultation with domain experts and exploratory analysis of incident patterns in healthcare data pipelines.

Different workflows may benefit from alternative weights. For example:

- Quality measurement workflows might emphasize vocabulary accuracy ($w_v = 0.6$) to ensure codes align with measure definitions.
- Real-time clinical decision support might weight structure more heavily ($w_s = 0.4$) to ensure mandatory fields for care algorithms are always present.
- Research data warehouses might prioritize completeness/cardinality ($w_c = 0.3$) to ensure sufficient data for statistical analysis.

Learning weights directly from steward feedback, incident outcomes, and downstream measure performance is identified as important future work; this would enable context-aware and outcome-driven quality policies.

3.3. Semantic Distance as a Quality Signal

- For each resource (e.g., Patient, Observation, Condition), a semantic distance score is computed. The scores can be:
- Aggregated per table or domain (e.g., median $Dsem$ for all Observations in a day).
- Used directly in alerting (e.g., raise a warning when the share of resources with $Dsem$ > 0.7 exceeds a threshold).
- Combined with anomaly scores from unsupervised models (Section 4) to prioritize incidents where both semantic distance and anomaly scores are high.

This provides a semantics-aware complement to metadata-based observability metrics, closing part of the "observability gap" described in the introduction.

4. Unsupervised Detection and Adaptive Thresholding

We now describe the unsupervised learning components used to detect anomalies and drift, and how their scores are converted into adaptive thresholds.

4.1. Feature Extraction and Profiling

For each dataset or partition (e.g., per facility per day), we compute lightweight profiling features, including:

- Basic statistics (mean, variance, min, max) for numerical fields (age, lab values).
- Frequency distributions and entropy for categorical fields (diagnosis codes, procedure codes).
- Missingness rates and uniqueness ratios (e.g., IDs).
- Aggregated semantic features (e.g., summary statistics of $Dsem$ across resources).

These features form an input vector for anomaly detection models. They are computed either on micro-batches or over rolling windows, depending on latency requirements.

4.2. Point Anomalies with Isolation Forest

At the level of individual records or small batches, we use Isolation Forest due to its efficiency and suitability for high-volume streams.

Principle. Anomalies are "few and different." Randomly partitioning the feature space isolates such points with fewer splits than typical points.

Algorithm. An ensemble of t random trees is built over the data. For a point x, the average path length $[h(x)]$ across trees is used to define an anomaly score:

$$s(x, n) = 2^{-\frac{E(h(x))}{c(n)}}$$

Where n is the sample size and (n) is the average path length in a binary search tree of size n. Scores closer to 1 indicate stronger anomalies.

Records whose scores exceed a dynamic threshold (Section 4.4) are flagged for further inspection; their semantic distance scores are also examined to prioritize cases where both syntactic and semantic anomalies co-occur.

4.3. Drift Detection with Autoencoders

To detect subtler distributional shifts that may not manifest as obvious outliers, such as gradual changes in diagnosis coding patterns or lab ranges, we employ deep autoencoders at an aggregate level.

An autoencoder learns to reconstruct input vectors x (e.g., profiling features over a time window) via a lower-dimensional latent representation z:

$$x \rightarrow z = f_{enc}(x) \rightarrow \hat{x} = f_{dec}(z)$$

The reconstruction error is defined as:

$$RE(x) = \| x - \hat{x} \|^2$$

The model is trained on a sliding window of historical data assumed to be representative of "normal" behavior. When the underlying data distribution drifts, reconstruction errors increase, providing a signal of concept drift or schema drift. Rather than directly treating every high-error window as an incident, we again rely on adaptive thresholding over the RE time series.

4.4. Adaptive Thresholding via EWMA Bands

Static thresholds on anomaly scores are brittle in the face of changing workloads, e.g., backfills of legacy data, seasonal spikes in claims, or incremental rollout of new documentation practices. To mitigate this, we maintain rolling statistics over anomaly scores and compute dynamic thresholds using an exponentially weighted moving average (EWMA):

Let at be a sequence of anomaly scores (from iForest or autoencoders) aggregated for a domain and time t

We compute:
μt: EWMA of scores.
σt: EWMA-based estimate of variability.
We define a time-varying threshold as:

$$\theta t = \mu t + \alpha \cdot \sigma t$$

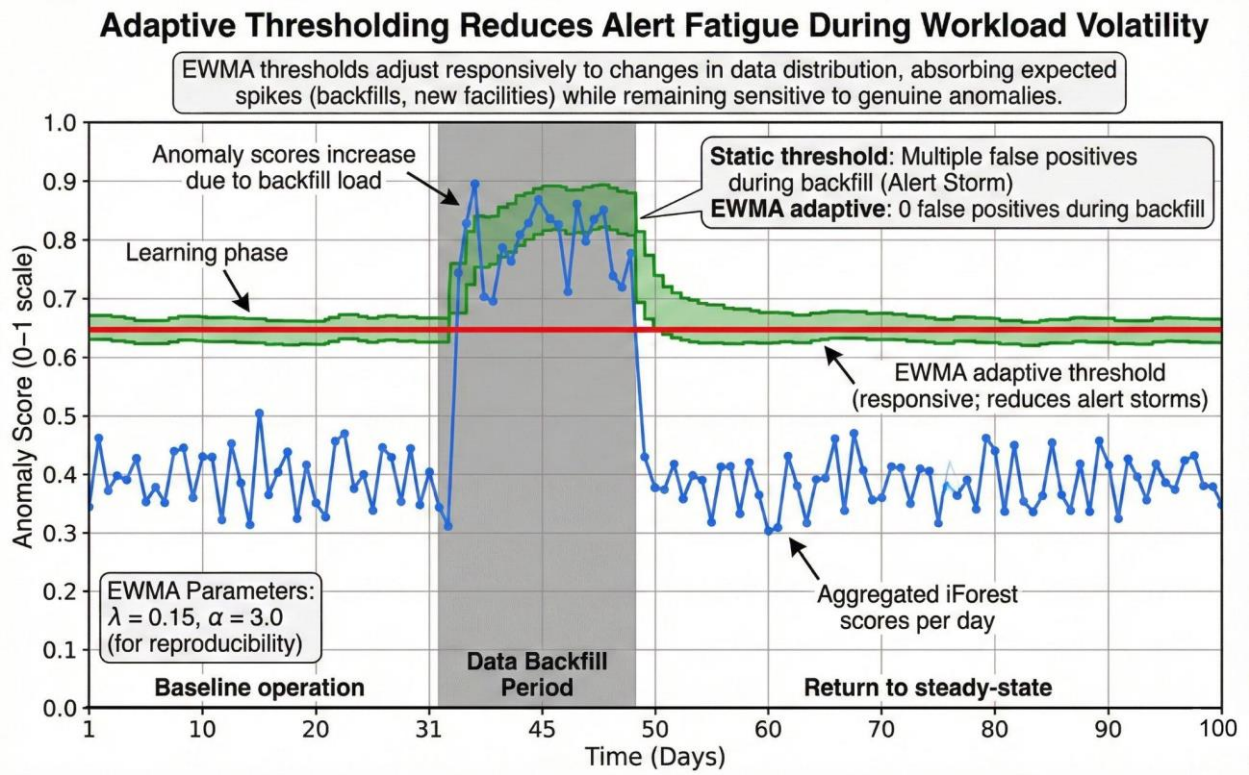

Fig 3: EWMA Adaptive Thresholding over Time (Threshold Adaptation during Backfill)

Where μt and σt are computed using exponentially weighted moving averages and α\alphaα controls sensitivity. In our experiments, we use smoothing factors (EWMA decay parameters) in the range λ∈[0.1,0.2], which provide a reasonable trade-off between responsiveness and stability for daily and hourly aggregates, and set α∈[2.5,3.0], analogous to a 2.5–3σ control band. Lower α values increase sensitivity at the cost of more alerts, whereas higher values produce more conservative thresholds and fewer alerts. For each domain (e.g., laboratory results vs. claims), λ and α can be tuned to match operational tolerance for volatility.

4.5. Cold-Start and Human-in-the-Loop Review

During initialization, unsupervised models lack sufficient history. To avoid erratic behavior in this "cold start" period, we:

- Apply conservative rule-based checks (e.g., FHIR cardinality, simple ranges) while models establish baselines over a configurable window (e.g., 7–14 days).
- Gradually ramp up reliance on learned thresholds as variance estimates stabilize.

For high-severity issues, alerts are routed to data stewards. To preserve data sovereignty, the central quality layer exposes pointers (e.g., time-bound pre-signed URLs or dataset identifiers) rather than copying PHI into a separate system. Stewards review the flagged records directly in their source environments, update local rules if needed, and feed back incident labels that can be used to refine weights or model parameters.

5. Experimental Evaluation

5.1. Dataset and Workload

We generate a synthetic, FHIR-compatible dataset of approximately 500,000 patients spanning five years using a Synthea-style synthetic patient generator. The generator simulates longitudinal clinical trajectories and care encounters, producing "synthetic, realistic but not real" FHIR bundles including Patient, Encounter, Condition, Procedure, and Observation resources.

To emulate a multi-cloud warehouse environment:

- EHR-style events are partitioned across three logical clouds (AWS, Azure, GCP).

- Each cloud exposes both FHIR endpoints and flattened analytic tables derived from FHIR resources.
- Per-cloud schemas vary slightly to mimic real-world heterogeneity (e.g., column naming conventions, datatype choices).

5.2. Chaos Scenarios

To stress-test the framework, we inject several controlled perturbations:

Schema Drift:

- Type changes (e.g., integer to string).
- Column addition/removal (e.g., new lab field; deprecation of a legacy field).

Vocabulary Drift:

- Gradual replacement of standard codes with local codes.
- Shifts in underlying code distributions (e.g., new COVID-19 diagnosis codes entering the mix).

Volume and Freshness Anomalies:

- Sudden drops in event counts for specific facilities or regions.
- Delayed batches simulating pipeline slowdowns.

Ground-truth labels record which partitions and time windows are affected, enabling computation of anomaly detection metrics (precision, recall, false-positive rate).

5.3. Baseline: Centralized Static Rules

As a baseline, we implement a conventional centralized DQM pipeline in which data from all clouds is replicated into a single warehouse and checked via static SQL rules. Representative rules include:

Range checks:

- $0 \leq \text{age} \leq 120$ years,
- $60 \leq$ systolic_blood_pressure ≤ 260 mmHg,
- $25 \leq$ body_mass_index ≤ 70 for adult obesity registries.

Null and uniqueness constraints:

- encounter identifiers and patient identifiers must be non-null and unique within a facility-day,
- admission and discharge timestamps must not be null for inpatient encounters.

Basic value-set membership:

- Diagnosis codes must belong to known ICD-10-CM chapters for cardiovascular registries,
- Laboratory test codes must appear in a curated LOINC value set.

Thresholds for allowable null rates or out-of-range proportions are fixed per field based on an initial profiling window (e.g., "no more than 1% of systolic blood pressure values may be outside [60, 260]"). These thresholds are not updated as distributions change, reflecting the static nature of many current production DQM configurations.

5.4. Evaluation Metrics

We evaluate:

- RQ1 – Detection Performance: How does the adaptive, semantics-aware framework (ADQMF) compare to a centralized static-rule baseline in terms of detecting injected data quality anomalies (precision, recall, and false-positive rate)?
- RQ2 – Semantic Signal Quality: How informative is the semantic distance metric D_{sem} in distinguishing clinically meaningful deviations (e.g., vocabulary drift into local codes) from benign structural changes or expected distribution shifts?
- RQ3 – Alert Volume and Stability: Does adaptive thresholding reduce alert volume and volatility over time relative to static thresholds, particularly during periods of planned volatility such as backfills or onboarding of new facilities?

5.5. Results Overview

Across all scenarios and clouds, the adaptive semantics-aware framework (ADQMF) achieves:

- Comparable or higher recall than the static baseline for detecting injected anomalies.

Approximately 40% relative reduction in false-positive rate, with representative aggregate values:

- Baseline: FPR ≈ 0.25, Recall ≈ 0.87, Precision ≈ 0.62
- ADQMF: FPR ≈ 0.15, Recall ≈ 0.89, Precision ≈ 0.74

The FPR reduction is computed as:

$$\frac{\mathbf{0.25 - 0.15}}{\mathbf{0.25}} = \mathbf{0.40}$$

Table 1: Performance Comparison of Static Rule-Based and Adaptive Semantics-Aware Models for Detection Accuracy

Model / Approach	Precision	Recall	False-Positive Rate	Relative Δ Precision	Relative Δ FPR
Centralized static rules	0.62	0.87	0.25	–	–
ADQMF (adaptive + semantics)	0.74	0.89	0.15	+19%	−40%

Relative improvements are computed as:

- $\Delta precision = (0.74 - 0.62)/0.62 \approx 0.19$ (19% relative improvement)
- $\Delta FPR = (0.25 - 0.15)/0.25 = 0.40$ (40% relative reduction).

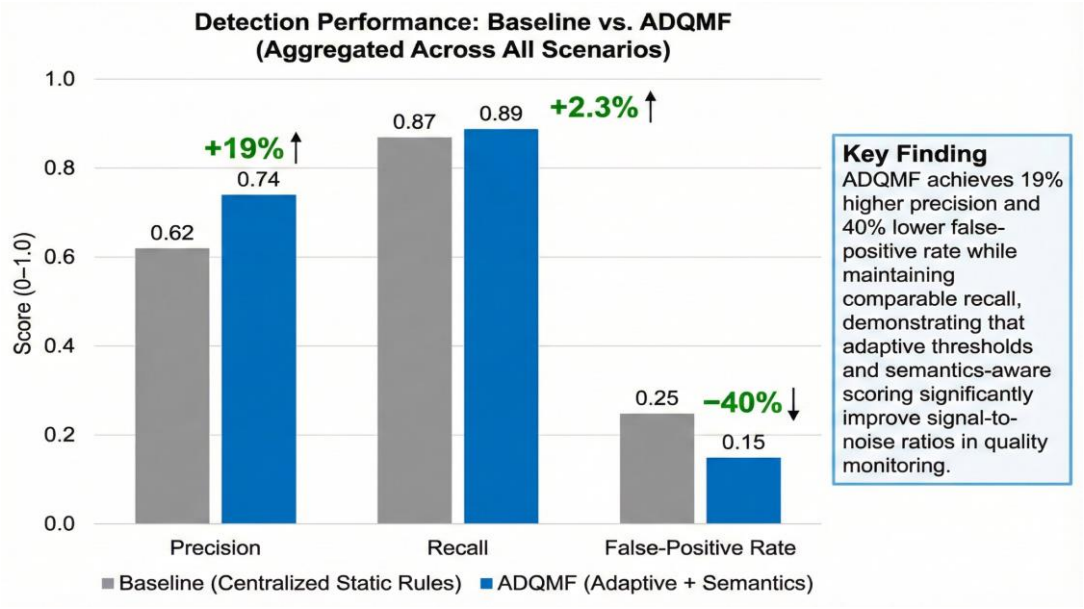

Fig 4: Results Comparison (Precision, Recall, False-Positive Rate)

As summarized in Table 1, ADQMF achieves a precision of 0.74 vs. 0.62 for the static-rule baseline (a 19% relative improvement) at slightly higher recall (0.89 vs. 0.87). The false-positive rate drops from 0.25 to 0.15, a 40% relative reduction. These gains directly address RQ1 by demonstrating that adaptive thresholds and semantics-aware scoring can significantly improve signal-to-noise ratios in quality monitoring without sacrificing anomaly detection sensitivity.

Qualitatively, this manifests as fewer "noise" alerts following benign distribution changes and more targeted alerts on genuinely problematic drift.

5.6. Role of Semantic Distance

In vocabulary drift scenarios, relying solely on structural checks and generic anomaly scores leads to frequent false positives when coding practices change in predictable ways (e.g., adopting new standard codes). By incorporating D_{sem}:

- Codes that remain within the intended value sets yield low vocabulary distance and do not trigger alerts even if volume patterns shift.
- Codes that move to related but broader hierarchies incur moderate distance, prompting warnings rather than hard failures.
- Local or unmapped codes generate high distance, triggering high-priority incidents.

This graded behavior improves precision for semantic issues: the system more accurately distinguishes benign code churn from truly incompatible or low-quality coding practices.

5.7. Impact of Adaptive Thresholding

Static thresholds applied to anomaly scores are sensitive to workload changes. For example, backloading historical data or onboarding a new facility can trigger sustained periods of "anomalous" scores that are actually expected. Under EWMA-based adaptive thresholds:

- Thresholds widen during periods of volatility, absorbing predictable shifts.
- Thresholds tighten in steady-state, maintaining sensitivity to unusual spikes.

Empirically, we observe smoother alert volume over time and reduced "alert storms" during planned data migrations, without masking the injected errors.

6. Discussion and Threats to Validity

6.1. Practical Implications

For healthcare organizations increasingly operating across multiple clouds and analytic platforms, the proposed framework offers several practical benefits:

- Reduced Alert Fatigue: Fewer false positives mean data engineers and stewards can focus on high-value issues.
- Semantics-Centric Quality: FHIR-aware scoring ensures that quality assessments align more closely with clinical realities and quality measurement requirements.
- Portability across Platforms: Because the approach relies on profiling features, anomaly scores, and FHIR profiles, it can be implemented on top of different warehouses and orchestration tools.

6.2. Threats to Validity

Several limitations should be acknowledged:

- Synthetic vs. Real PHI: Real multi-cloud PHI datasets cannot be shared or instrumented as freely as synthetic workloads, making synthetic data a necessary starting point. Although Synthea produces realistic patterns, real data may exhibit more complex and messy behaviors; these privacy and regulatory constraints are also a central

motivation for exploring federated deployment patterns in related work.

- Manual Weighting of Semantic Components: We manually choose ws_i. Different choices could shift the balance of what is considered "severe." Learning these weights from incident logs and downstream outcome impact is an important direction for future work.
- Model and Parameter Choices: Isolation Forest and autoencoders are representative unsupervised methods, but other models (e.g., random cut forests, variational autoencoders) might yield different trade-offs. Similarly, EWMA parameters affect sensitivity and response time.
- Operational Constraints: This paper focuses on methodology rather than on detailed deployment patterns, resource management, or regulatory approvals for running such systems in live PHI environments.

7. Conclusion and Future Work

This paper introduced an adaptive data quality management approach for multi-cloud healthcare warehouses that combines FHIR-aware semantic validation with unsupervised anomaly detection and adaptive thresholding. By moving beyond static rule sets and binary validation outcomes, the framework reduces false-positive alerts while maintaining sensitivity to clinically relevant data issues.

Evaluation on a synthetic, FHIR-compatible multi-cloud workload suggests that integrating semantic distance metrics and dynamic thresholds into DQM pipelines can significantly improve signal-to-noise ratios in quality monitoring. These gains are especially relevant as healthcare organizations navigate increasingly complex, distributed analytics environments under tightening regulatory and transparency expectations.

Future work will focus on:

- Deploying and evaluating the framework in real health systems under PHI-aware governance.
- Learning semantic weights and alerting policies from steward feedback and downstream measure performance.
- Integrating the quality signals with digital quality measure ecosystems (e.g., DEQM, eCQM-on-FHIR) and model monitoring platforms.
- Exploring richer combinations of topology-aware deployment and FinOps-oriented optimization, which will be the focus of a companion paper on edge-centric architectures for healthcare DQM.

References

[1] N. G. Weiskopf and C. Weng, "Methods and dimensions of electronic health record data quality assessment: Enabling reuse for clinical research," J. Am. Med. Inform. Assoc., vol. 20, no. 1, pp. 144–151, Jan. 2013.

[2] M. G. Kahn et al., "A harmonized data quality assessment terminology and framework for the secondary use of electronic health record data," EGEMS (Wash. DC), vol. 4, no. 1, p. 1244, 2016.

[3] C. Batini and M. Scannapieco, Data Quality: Concepts, Methodologies and Techniques. Berlin, Germany: Springer, 2006.

[4] A. E. Lewis et al., "Electronic health record data quality assessment and tools: A systematic review," J. Am. Med. Inform. Assoc., vol. 30, no. 10, pp. 1730–1742, Oct. 2023.

[5] Z. Wang, J. R. Talburt, N. Wu, S. Dagtas, and M. N. Zozus, "A rule-based data quality assessment system for electronic health record data," Appl. Clin. Inform., vol. 11, no. 4, pp. 622–634, Aug. 2020.

[6] R. Y. Wang and D. M. Strong, "Beyond accuracy: What data quality means to data consumers," J. Manage. Inf. Syst., vol. 12, no. 4, pp. 5–33, Spring 1996.

[7] J. A. Walonoski et al., "Synthea: An approach, method, and software mechanism for generating synthetic electronic health records," J. Am. Med. Inform. Assoc., vol. 25, no. 3, pp. 230–238, Mar. 2018.

[8] National Academies of Sciences, Engineering, and Medicine, "Synthetic health data generation engine to accelerate patient-centered outcomes research," in Understanding the Impacts of OS-PCORTF Projects on Data Infrastructure and Methods. Washington, DC, USA: National Academies Press, 2023.

[9] Health Level Seven International, "FHIR Release 4 (R4): Fast Healthcare Interoperability Resources," 2019. [Online]. Available: http://hl7.org/fhir/R4/

[10] C. N. Vorisek et al., "Fast Healthcare Interoperability Resources (FHIR) for interoperability in health research: A systematic review," JMIR Med. Inform., vol. 10, no. 7, p. e35724, Jul. 2022.

[11] B. Moses, "The rise of data observability: Architecting the future of data trust," Monte Carlo Data, 2021. [Online]. Available: https://www.montecarlodata.com/blog/the-rise-of-data-observability/

[12] F. T. Liu, K. M. Ting, and Z.-H. Zhou, "Isolation forest," in Proc. 8th IEEE Int. Conf. Data Mining (ICDM), 2008, pp. 413–422.

[13] U.S. Dept. of Health and Human Services, Office of the National Coordinator for Health IT, "Health Data, Technology, and Interoperability: Certification Program Updates, Algorithm Transparency, and Information Sharing (HTI-1) Final Rule," Fed. Regist., vol. 89, no. 7, pp. 1192–1279, Jan. 9, 2024.

[14] N. Yaraghi, A. A. Seixas, and F. Zizi, "How ONC can strengthen its HTI-1 rule to ensure transparency, fairness, and equity in AI," Health Affairs Forefront, Jun. 26, 2024.