

AI-Driven Real-Time Data Quality Validation in Healthcare ETL Pipelines

Sudhakar Guduri

Jawaharlal Nehru Technological University, India

ARTICLE INFO

Received: 11 March 2026

Accepted: 15 March 2026

ABSTRACT

Pressure on healthcare data pipelines is growing to provide accurate, consistent, and regulatory-compliant data in real-time; traditional extract-transform-load validation models are still rooted in the concept of ensuring that data is correct, consistent, and regulatory-compliant but fail structurally to meet the velocity and complexity of current healthcare data environments. Smart, self-evolving validation built into streaming ETL processes is an architectural breakthrough that enables data quality testing to become more than a reactionary activity after the data has loaded; instead, it's a capacity that executes while the system is operational. With machine learning-based anomaly detection, such as Isolation Forest, autoencoder neural networks, and statistical modeling and schema drift monitoring and threshold adaptation reinforced by reinforcement learning, ETL pipelines can have the ability to detect and intervene in data integrity failures before they escalate into downstream clinical, financial, and regulatory systems. Explainable AI systems make sure that each automated quality decision is supported by a mode of interpretation, meeting the traceability and auditability standards that the healthcare regulatory frameworks have established regarding the protected health information. Unalterable audit logging transforms compliance records from a periodical manual process into a pipeline property that is run on an automatic basis. Automated correction, quarantine, and intelligent reprocessing of anomalous records are all possible with self-healing remediation capabilities without interrupting the pipeline and the manual intervention burden inherent to the traditional quality assurance models. These intelligent validation capabilities can remain sustained at enterprise volumes of healthcare data without placing throughput pressure on distributed computing architectures that have the required horizontal scaling. The combination of all these capabilities creates an ETL infrastructure that proactively protects the integrity of data instead of just passively accepting records, which creates a reliable basis of data to make clinical decisions, model population health, operate the revenue cycle, and provide regulatory reporting.

Keywords: Healthcare ETL Pipelines, Real-Time Data Quality Validation, Anomaly Detection, Explainable Artificial Intelligence, Self-Healing Remediation

1. Introduction

Healthcare data ecosystems create high-velocity (continuous) streams of structured and semi-structured data based on electronic health records, payer claims systems, laboratory information platforms, and interoperability interfaces that are regulated by standards like HL7 and FHIR. This heterogeneity and size add compounding data integrity risks, which traditional Extract, Transform, Load validation architectures cannot handle. Conditions of predictable, low-velocity data environments, which are the subject of design of static rule engines, batch-oriented null checks, and post-load schema matching, are no longer relevant to the operational reality of the contemporary healthcare data infrastructure.

Clinical and financial effects of unmanaged data quality failures are massive. Mistakes within patient records, claims forms, and diagnostic data sets have the potential to misrepresent clinical decision support reports, undermine population health models, and create regulatory liability that spans across various compliance systems. Deep learning applications on electronic health record data have shown that when it is ensured systematically that the quality of inputs to the model is of high quality, then scalable and accurate modeling of healthcare data is possible, which supports the argumentative premise that intelligent validation is a precondition to trustworthy downstream analytics [1]. To this end, studies of EHR data quality measurement have determined that the reusability of clinical data to be used in research and operational activities directly relies on the rigor of the quality dimensions of completeness, correctness, concordance, plausibility, and currency in data ingestion and transformation [2].

The distance between what statistical validation provides and what healthcare needs has resulted in an ever-growing need to prioritize AI-based and real-time validation as a research and implementation urgency. In this paper, the authors suggest a layered model that directly integrates machine learning-driven anomaly detection, schema drift detection, explainable AI, and automated remediation into streaming ETL pipelines that can transform data quality assurance into an active, adaptable, and regulation-grade operation through a continuous, adaptive, and regulation-grade operating mode.

2. Limitations of Conventional ETL Validation

The historical ETL validation models were designed based on a pre-underlying assumption that has become unsustainable in healthcare settings, namely, the assumption that one can adequately measure data quality once the data have been introduced into target systems. The standard toolkit row count reconciliation, null field checks, schema conformance matching, referential integrity constraints, and basic duplicate detection toolkit deal with structural properties of data with reasonable effectiveness but are fundamentally blind to semantic inconsistencies, which are some of the most serious failure modes in healthcare data pipelines. A syntactically correct procedure code that is clinically non-codable with the diagnosis code it is intended to pair corresponds to all structural validation rules and silently compromises the integrity of claims analytics, clinical decision support systems, and population health reporting reliant on correct code pair relationships.

The dimension-based EHR data quality assessment model has long known that structural correctness is not the sole one among a variety of important quality properties that the healthcare data should meet. The completeness, correctness, concordance, plausibility, and currency are all different failure vectors against which the post-load batch validation is ill-placed to provide any countermeasures [2]. Missing demographic attributes, procedure modifiers, or incomplete encounter records. Completeness failures: Completeness check failures may not be a constraint of the schema, but they have profoundly negative analytic value. Plausibility failures, in which the recorded values are valid in each case, but implausible on a larger scale in the clinical context, demand reasoning in many fields and historical trends beyond the capability of the SQL rules. Currency failures, in which data is received out of tolerable timeliness intervals without being noticed, are completely opaque to batch ingestion that does not consider temporal characteristics.

These limitations on detection are combined with a structural latency issue in the operational model of batch validation. In the cases where validation logic is performed only after data has been persisted, there are no errors that exist and spread. The imperfect records are processed at downstream systems, and errors become more entrenched in data assets with each processing step. Investigations into the application of big data in health informatics have continued to highlight data quality as the core limitation to the potential value that can be derived from large-scale healthcare data streams, with the scale of health data streams increasing the inapplicability of manual or batch-based data quality

control strategies to data infrastructure size [3]. The phase cost of remediation of errors found late in the pipeline is not linear; it increases with the number of downstream systems impacted and how many derived datasets, aggregations, and analytical results have been corrupted with bad data.

Also, the graphical rule engines are unable to support the ongoing development of healthcare data standards. The structure and semantics of incoming data are changed by incremental changes to the ICD coding, CPT updates, LOINC version updates, and EHR schema changes initiated by vendors on a rolling basis. A validation framework based upon fixed rules gradually becomes increasingly out of sync with the real data that it is assessing, creating false negatives that enable real anomalies to escape and false positives that get in the way of legitimate records. Studies that consider the data mining techniques used in health informatics have pointed out that the complex and dynamic environment of healthcare data requires actively changing and learning quality mechanisms as opposed to fixed constraint checking [3]. The failure of traditional structures to evolve unless systematically serviced with rules makes persistent and accumulating quality assurance debt, which grows with each standard release, each vendor release, and each new source of data added to the pipeline.

Validation Limitation	Root Cause	Healthcare Impact
Semantic blindness	Rule-based structural checks only	Clinically incompatible code pairs pass undetected
Post-load detection	Batch execution after persistence	Errors propagate into downstream systems before discovery
Static rule rigidity	No adaptive learning capability	Validation logic becomes misaligned with updated standards
No temporal evaluation	The timeliness dimension not assessed	Late-arriving records accepted without currency checks
Manual rule maintenance	No self-updating mechanism	Growing quality assurance debt across standard update cycles

Table 1 - Limitations of Conventional ETL Validation [3, 4]

3. Architecture of an AI-Integrated Validation Framework

The suggested framework is built in five functional layers that will result in an active and intelligent data quality system out of a passive ETL pipeline. The ingestion layer is concerned with the source reception and routing of nonhomogeneous healthcare data streams. The streaming validation layer applies structural and semantic quality constraints to incoming records in real time and then persists the records. The AI quality engine uses machine learning models to identify anomalies, monitor schema drift, and tune validation thresholds. The compliance and audit layer creates regulatory-quality records of all validation events. The remediation layer is a self-healing layer where automated corrective measures are taken against reported errors. The whole architecture is based on distributed computing platforms that can support near real-time validation levels at healthcare enterprise scale, where record volumes regularly reach into the millions per hour in parallel data streams.

3.1 Streaming Validation Layer

The streaming validation layer is the most primordial architectural element of the conventional ETL design. Instead of performing quality checks on data that has already been written to target storage, records are intercepted by the streaming layer itself and checked against a set of quality rules that are

constantly maintained before a record can be allowed to continue on its way towards persistence. It is an in-stream positioning that removes the latency window, which characterizes batch validation and which is the structural requirement of all downstream AI-based quality capabilities. The five fundamental data quality dimensions implemented at this layer, completeness, accuracy, consistency, conformity, and timeliness, are considered simultaneously, and therefore, a record should meet the quality standard of each dimension before proceeding through the pipeline. In-stream validation frameworks based on real-time detection of anomalies in streaming healthcare data have shown that with in-stream validation architectures based on distributed processing, detection performance and throughput levels fit the enterprise healthcare workload, making the practicability of the streaming validation approach a practical demonstration [4].

3.2 AI Quality Engine

The analytical heart of the framework is the AI quality engine, which applies various machine learning approaches to the live data stream and detects the anomalies that structural validation rules are incapable of detecting. The main task of the engine is anomaly detection, and there are three analogous models that cover various anomaly signatures. Isolation Forest models are provided on high-dimensional claims data, in which the capacity of the algorithm to isolate multivariate outliers by random partitioning renders them especially useful in detecting abnormal combinations of procedure codes, diagnosis codes, and billed amounts that are not learned in historical trends. Autoencoder neural networks are reconstruction-error-based detectors of complex non-linear anomaly patterns in laboratory and diagnostic data and learn to compress normal data behavior and indicate records where the reconstruction error is larger than the learned thresholds. Z-score statistical modeling also detects univariate outliers of numeric fields whose distributions can be parametrically approached. Theoretical underpinnings and performance features in comparison to each other of these anomaly detection methods have been methodically investigated in the literature of anomaly detection, which identifies isolation-based and reconstruction-based techniques as especially appropriate to the high-dimensional, heterogeneous format of healthcare data [5].

Schema drift schemes take the engine abilities from examining records separately to surveying the structural patterns of the data in successive versions. Since vendor updates, standard modifications, and integration developments change the metadata signatures of the incoming data streams; the drift detection component recognizes the difference between the pre-established baseline schemas and sends warning signals before structurally varied records corrupt downstream datasets. Reinforcement learning gives the engine its adaptive aspect, which continuously retunes the validation thresholds according to the error distributions observed and seasonal patterns in healthcare data and feedback on remediation progress, such that the quality engine advances with experience of operation but not diminishes with changing data landscapes.

Architectural Layer	Primary Function	Key Technology
Ingestion Layer	Reception and routing of heterogeneous data streams	HL7, FHIR, X12 interfaces
Streaming Validation Layer	In-stream structural and semantic quality enforcement	Apache Spark Streaming
AI Quality Engine	Anomaly detection, drift monitoring, threshold adaptation	Isolation Forest, Autoencoders, Reinforcement Learning
Compliance and Audit Layer	Immutable event logging and regulatory documentation	Audit log pipelines, XAI output
Self-Healing Remediation Layer	Automated correction, quarantine, and reprocessing	Reference data sync, intelligent reprocessing

Table 2 - Architecture of an AI-Integrated Validation Framework [5, 6]

4. Regulatory Compliance and Explainability

Healthcare data systems have one of the most rigorous regulatory environments in any industry sector, with the responsibilities of federal privacy and security laws and developing state-level health data requirements offering a compliance environment where validation frameworks should actively facilitate but not passively address. When AI-based decision-making is incorporated into the data pipeline activities, some new regulatory considerations emerge: the automated systems that would identify, quarantine, or alter health records should be able to generate interpretable, auditable explanations for their conclusions. A validation framework that enhances the performance of detection and decreases the transparency of audits generates a liability on compliance that compensates for the operational advantages. This tension is resolved in the architecture determined in this paper because the two concepts of explainability and audit logging are not seen as add-ons to the design but as first-class architectural constraints implemented at the design level.

4.1 Explainable AI for Audit Transparency

The explainability layer of the framework creates detailed, human-explicable explanations of all the anomaly determinations of the AI quality engine. Every flagged record is provided with a root cause classification that defines what particular driver led to the anomaly score, a feature contribution breakdown estimate of the relative contribution of the individual data fields to the detection outcome, and a historical comparison contextualizing the flagged record against baseline behavioral patterns. This is a systematic output that is utilized in many operational aspects at the same time. To compliance officers involved in regulatory audits, it offers the traceability documentation needed to prove that automated decisions impacting protected health information have been made by accountable, interpretable processes. It offers diagnostic data that speeds up the process of root cause identification and remediation to data engineers who seek to solve problems with pipelines. This is handled by the framework by making sure that the explainability layer generates outputs that are organized based on the particular documentation categories upon which compliance frameworks draw: decision basis, confidence level, comparison against established norms, and action taken.

The issue of explainability in medical AI systems has received significant scrutiny in the scholarly literature, with discussions on what must be in place to create explainable AI in medical diagnosis consistently determining interpretability as a condition to clinical and regulatory approval of automated decision systems [6]. The interpretability requirements that apply to clinical AI systems are equally applicable to the data quality validation systems that dictate the information received by those clinical AI systems. A validation framework that is an opaque black box, irrespective of its accuracy in detection, cannot meet the level of transparency that is required by healthcare regulators and institutional review procedures. This is handled by the framework by making sure that the explainability layer generates outputs that are organized based on the particular documentation categories upon which compliance frameworks draw: decision basis, confidence level, comparison against established norms, and action taken.

4.2 Immutable Audit Logging

In addition to the explainability layer, the audit logging component produces an exhaustive, non-modifiable log of all validation events flowing through the pipeline. Each log record logs the time of occurrence of an event, with enough accuracy to allow reconstructing the activity in the pipeline chronologically, the identity of the user or automated process that submitted the data, the validation rule or machine learning model that triggered that event, the confidence score of the AI engine that made that determination, and the remedial action taken. The cumulative audit log is an ongoing, methodical document of the quality assurance exercise of the pipeline throughout its entire history of operation.

The importance of this capability goes beyond the normal compliance documentation. Studies on the need to develop computational technology in the context of effective healthcare information systems

have stressed the fact that dependable health data infrastructure must not only involve proper processing of health data but must also have verifiable accountability over all transformation and quality determination of health records [7]. Unalterable audit trails can meet this accountability criterion by making sure that the history of quality decisions cannot be changed afterwards, which provides a secure evidentiary basis to regulatory reviews, litigation support, or internal governance audits. The framework has the benefit of removing a significant amount of manual documentation overhead that traditional workflows of QA activities would otherwise place on data engineering and compliance teams, and the benefit is also improving the coverage of the audit process since the validation event will never go undocumented.

Compliance Requirement	Framework Capability	Regulatory Relevance
Audit traceability	Structured XAI output per validation event	HIPAA enforcement documentation
Decision interpretability	Feature contribution breakdown per anomaly flag	Automated decision accountability
Tamper-resistant records	Immutable audit log generation	Regulatory examination evidentiary support
Confidence documentation	AI confidence score captured per event	Quality assurance defensibility
Continuous coverage	Every validation event is logged automatically	Eliminates manual documentation gaps

Table 3 - Regulatory Compliance and Explainability [7, 8]

5. Self-Healing Remediation and Scalability

The most operationally transformative shift of the traditional ETL design thought is the self-healing nature of the proposed framework. Conventional pipeline designs view error detection as becoming a terminal event: an error in a validation check causes the pipeline to stop, and an alert is issued and requires human intervention to troubleshoot the failure, introduce a fix, and restart the damaged data. The interrupt-and-wait model can be taken into consideration in low-volume and low-velocity data environments where manual QA cycles can be performed within operational cycles [11]. The interrupt-and-wait model produces unacceptable latency and introduces data readiness delays that persist across the chain of dependencies and is engineering resource-unscaled to the scale of increased data volumes in healthcare enterprise environments where data volumes scale to millions of records (at least) in a single processing cycle and the data made available by downstream systems is relied upon by subsequent processing chains.

5.1 Automated Remediation Workflows

The self-healing layer overcomes this limitation by proposing a progressive sequence of automated remedial actions that have been tuned to the level of severity, type, and confidence level of each identified anomaly group [12]. Format errors, misformed date values, and nonconformant identifier values are automatically fixed against continuously synchronized reference data, and records with incorrectly encoded categorical values are automatically patched without human intervention, and the records are allowed to continue flowing down the pipeline. The records with suspicious billing patterns, statistically implausible sets of clinical coding, or anomaly scores beyond configurable confidence limits are automatically quarantined in separate holding partitions pending organized review, which ensures that they do not persist in operational systems, but the records are retained to

be reviewed clinically or financially. The reference data synchronization is ongoing to enable coding validation to capture current versions of the ICD, CPT, and LOINC standards without a manual update cycle, which creates maintenance windows and version delays. Records that have been quarantined can be intelligently reprocessed after the underlying problems that caused the quarantine have been addressed, to allow continuity in the pipeline and avoid the permanent loss of data.

This remediation capability (automated) has a high operational and financial value. The revolution of big data in healthcare has already determined that the capability to produce value out of large-scale health data assets are directly limited by the practicality and cost-efficiency of the data quality management procedures, and the decrease in the burden of manual intervention of data remediation is a prerequisite to achieving the full analytical potential of healthcare data infrastructure [8]. Direct solutions to this limitation include automated self-healing to replace the cost-structuring of quality assurance as reactive, labor-intensive remediation with proactive, automated correction to lower the downstream remediation expenses and speed the timelines of data readiness at the same time[9].

5.2 Distributed Scalability

The scalability design of the framework means the introduction of an extensive AI-based validation will not add throughput bottlenecks that can weaken the performance of the pipeline. Split data validation ensures the quality checks are allocated to parallel compute nodes, resulting in validation having a horizontal scale of performance with the volume of data, as opposed to a bottleneck. A hash-based row comparison can be used to perform efficient duplicate detection on very large datasets without the computational cost of a record-by-record comparison. Predicate pushdown filtering reduces the amount of data that needs to be transferred over the pipeline before being evaluated by the quality predicate by bringing the filter conditions as close to the source of data as they can. Frequently accessed validation artifact reference code sets are stored in in-memory caching, which removes repeated retrieval latency, and schema definitions are stored in distributed memory. Parallel anomaly scoring spreads model inference throughout the compute cluster, where the machine learning aspects of the quality engine do not limit it but instead run it at pipeline speed [13].

This scalability architecture is based on the distributed systems foundations that are well-established. Studies of resilient distributed datasets and fault-tolerant abstractions of in-memory cluster computing have shown that in-memory cluster computing distributed processing frameworks can scale to the throughput and fault-tolerant demands of large-scale data processing workloads and that the low-latency performance properties of real-time validation can be achieved [14]. The computational efficiency plans that the framework uses take advantage of these distributed systems' capabilities so that validation intelligence is balanced by execution efficiency, providing the detection accuracy of state-of-the-art machine learning within the performance band that healthcare enterprise data operations must operate in[15].

Capability	Mechanism	Operational Benefit
Format auto-correction	Reference data matching and normalization	Records corrected without pipeline interruption
Anomalous record quarantine	Confidence-threshold-based isolation	Prevents persistence of high-risk records
Reference data synchronization	Continuous coding standard version alignment	Eliminates manual update cycles and version lag
Intelligent reprocessing	Condition-cleared record re-evaluation	Prevents permanent data loss from temporary issues
Distributed anomaly scoring	Parallel model inference across compute nodes	Validation throughput scales with data volume

Table 4 - Self-Healing Remediation and Scalability [9, 10]

Conclusion

The sufficiency of rule-based validation schemes has been continuously undermined by the increasing speed, volume, and heterogeneity of healthcare data facilities to the extent that batch-based quality assurance can no longer be used as a core integrity scheme to orchestrate enterprise data pipelines. Brainy, intelligent, adaptive validation built into streaming ETL infrastructure deals directly with this erosion and puts quality assurance back to being an ongoing, proactive, operating activity, as opposed to an infrequent, responsive gateway. The anomaly detection proposed by machine learning, relying on isolation-based, reconstruction-based, and statistical modeling methods, has detection features far beyond the structural consistency of traditional validation, revealing semantic inconsistencies, clinically implausible code sets, and behavior anomalies that can be categorically hidden by a set of static rules. Schema drift detection maintains validation values throughout the ongoing development of healthcare coding standards and vendor data schemas, whereas reinforcement learning permits threshold adjustment that is enhanced with operational experience. Explainable AI is designed to make automated decisions on quality have certain interpretability and auditability, which meet the transparency requirements that healthcare regulatory frameworks place on systems dealing with protected health information. Unchangeable audit logging creates accountability that can be verified in all quality events in the entire pipeline history. Self-healing remediation changes the pipeline interruption signal into a signal to begin a resolution to minimize the need for manual intervention and speed up the preparation of data. The intelligent capabilities are supported by distributed computing foundations with the scalability that will enable the enterprise data volumes to be handled without performance loss. The combination of these architectural components creates data pipelines, which are able to provide the integrity, consistency, and regulatory reliability that clinical decision-making, financial functions, and population health programs seriously demand.

References

- [1] Alvin Rajkomar et al., "Scalable and accurate deep learning with electronic health records," *npj Digital Medicine*, 2018. [Online]. Available: <https://doi.org/10.1038/s41746-018-0029-1>
- [2] Nicole Gray Weiskopf and Chunhua Weng, "Methods and dimensions of electronic health record data quality assessment: enabling reuse for clinical research," *Journal of the American Medical Informatics Association*, 2013. [Online]. Available: <https://doi.org/10.1136/amiajnl-2011-000681>
- [3] Matthew Herland, Taghi M Khoshgoftaar, and Randall Wald, "A review of data mining using big data in health informatics," *Journal of Big Data*, 2014. [Online]. Available: <https://doi.org/10.1186/2196-1115-1-2>
- [4] Jieru Ding et al., "TDM-MIMO Automotive Radar Point-Cloud Detection Based on the 2-D Hybrid Sparse Antenna Array," *IEEE Transactions on Geoscience and Remote Sensing*, 2022. [Online]. Available: <https://ieeexplore.ieee.org/document/9756021>
- [5] Varun Chandola et al., "Anomaly detection: A survey," *ACM Computing Surveys (CSUR)*, 2009. [Online]. Available: <https://doi.org/10.1145/1541880.1541882>
- [6] Vishakh Hegde and Reza Zadeh, "FusionNet: 3D Object Classification Using Multiple Data Representations," *arXiv preprint, arXiv:1607.05695*, 2016. [Online]. Available: <https://arxiv.org/abs/1607.05695>
- [7] William W. Stead and Herbert S. Lin, "Computational Technology for Effective Health Care: Immediate Steps and Strategic Directions," *National Research Council of the National Academies*, 2009. [Online]. Available: https://www.nlm.nih.gov/pubs/reports/comptech_prepub.pdf
- [8] Peter Groves et al., "The 'big data' revolution in healthcare: Accelerating value and innovation," *McKinsey Center for US Health System Reform business technology office*, 2013. [Online].

Available:

https://www.mckinsey.com/~/media/mckinsey/industries/healthcare%20systems%20and%20services/our%20insights/the%20big%20data%20revolution%20in%20our%20health%20care/the_big_data_revolution_in_healthcare.pdf

- [9] Matei Zaharia et al., "Resilient Distributed Datasets: A Fault-Tolerant Abstraction for In-Memory Cluster Computing," 2012. [Online]. Available: <https://www.usenix.org/system/files/conference/nsdi12/nsdi12-final138.pdf>
- [10] Shan Yin et al., "Dependency-aware task cooperative offloading on edge servers interconnected by metro optical networks," *Journal of Optical Communications and Networking*, 2022. [Online]. Available: <https://ieeexplore.ieee.org/document/9762345>
- [11] G. A. Ascanio, "Bathrooms as spaces of recovery: Wellness-oriented design strategies in domestic architecture," *Evolutionary Studies in Imaginative Culture*, pp. 117–124, 2023.
- [12] R. Chhibber, "Channel-based sales strategy for sustainable enterprise revenue streams," *Journal of International Crisis and Risk Communication Research*, vol. 5, no. S12, pp. 123–132, 2022.
- [13] A. Kejriwal, "Decision-making dynamics in multi-stakeholder project negotiations," *Journal of Information Systems Engineering and Management*, vol. 7, no. 2, pp. 1–9, 2022.
- [14] D. Puthiya, "AI-enabled growth architectures for digitally mature organizations," *Journal of Computational Analysis and Applications*, vol. 30, no. 2, pp. 1113–1129, 2022.
- [15] P. A. Diaz Munoz, "Bridging architecture and urban systems: An interdisciplinary approach to built environments," *Evolutionary Studies in Imaginative Culture*, vol. 7, no. 2, suppl. 1, pp. 109–116, 2023.