International Journal of Computational and Experimental Science and ENgineering (IJCESEN)

Vol. 12-No.1 (2026) pp. 1389-1397
http://www.ijcesen.com

ISSN: 2149-9144

Research Article

Data Aggregation of HEDIS Measures using ETL Technologies

Avinash Dulam*

Osmania University, Hyderabad, India
* **Corresponding Author Email:** reachme.avinashdulam@gmail.com - **ORCID:** 0000-0002-0047-4477

Article Info:

DOI: 10.22399/ijcesen.5034
Received : 05 December 2025
Revised : 06 February 2026
Accepted : 08 February 2026

Keywords

HEDIS Measures,
ETL (Extract, Transform, Load),
Healthcare Data Integration,
Quality Performance Reporting,
Data Warehousing

Abstract:

The Healthcare Effectiveness Data and Information Set (HEDIS) has standardized performance measures. These measures aim to confirm that the health care services provided and utilized are of high quality. For reporting HEDIS measures, the data must be collected and aggregated from a variety of sources, including electronic health records (EHRs), claims data, pharmacy data, and lab information systems. This article uses Extract, Transform, Load (ETL) technologies to orchestrate data integration processes and provide trustable measure calculations. In the extract process, heterogeneous data sources are connected to the ETL pipelines by database interfaces, application programming interfaces (APIs), and streaming ingestion protocols, with security and compliance requirements. This phase includes cleansing and validation checks, as well as standardizing and recoding all of the different coding systems and heterogeneous constructs into a common structure that is consistent with the HEDIS specifications. The loading phase involves complex calculations and performance reporting. Business benefits include improved data quality via data validation at the source, process efficiency via workflow automation and scalability, and better decision support with improved visibility for all stakeholders in real-time operational performance. Implementation considerations include enterprise platforms, open-source and cloud-native applications, interoperability via Fast Healthcare Interoperability Resources (FHIR) standards, and available software development kits (SDKs). As new technologies like real-time data processing, cloud computing, and AI-based validation develop, healthcare organizations will choose their reporting technology based on their size, technical skills, and available resources, while also improving data management and adapting to new standards and regulations in a data-heavy environment.

1. Introduction

1.1 Overview of HEDIS Measures and Performance Metrics

The Healthcare Effectiveness Data and Information Set, also known as HEDIS, is a set of standardized performance measures that supports health plans, integrated delivery systems and managed care organizations to measure and report the quality of the health care provided to their enrolled population. Many different health care settings use HEDIS measures to indicate clinical performance, preventive care, and chronic disease care. A third purpose of HEDIS measures is to assess the degree to which clinical interventions improve a member's health status, the degree to which services are being delivered to members in need of care for them, and the degree to which members are satisfied with the care they receive. HEDIS measures are standardized. Regulators, accreditors, purchasers, and consumers can compare the quality of health plans across healthcare and geographic areas. This allows the comparison. HEDIS reports measure health plan performance for accreditation, for regulatory compliance, for public reporting, and for an increasing number of value-based reimbursement contracts. In order for a health plan to accurately report on its HEDIS measures and thereby its overall organizational performance, a health plan needs to be able to accurately collect, store, aggregate, and analyze its clinical and administrative data.

1.2 The Role of Data Aggregation in Healthcare Quality Reporting

The data used to calculate HEDIS measures come from different source information elements. The medical data are often stored in different source systems with different structures and semantics. The data are typically stored in the range of electronic health record systems, claims processing systems, pharmacy management systems, laboratory information systems, and many other administrative information systems that comprise a healthcare organization's technology ecosystem. Each typically has its own independent data model, optimized for its particular use case. The heterogeneity of the data in terms of coding systems, granularity, temporal resolution, and identifier schemas complicates record linkages and measures computation. Data aggregation must address representing the data in a way that maintains data lineage, referential integrity, and integrity of transformed data and represents an accurate view of the clinical and administrative reality captured in source systems. ETL technologies deliver the architecture with which to automate the workflow processes that extract data from source systems, transform the different data representations into a common structure that adheres to the HEDIS technical specifications, and load these data into analytical data stores to calculate and report healthcare quality measures. Healthcare organizations apply ETL principles. They can regularly develop repeatable, auditable, and scalable processes for timely, accurate, and complete HEDIS reporting. This minimizes the operational burden linked to manual data reconciliation and quality assurance (QA) activities.

1.3 Article Scope and Objectives

This article looks at the technical architecture and implementation issues that arise when using ETL-based methods to combine HEDIS data. It focuses on the design principles, processing patterns, and technology selection criteria that affect how well a system works and how well an organization does. The subsequent sections provide detailed analysis of the data source landscape relevant to HEDIS reporting, the core components and workflow orchestration patterns within ETL architectures, the strategic benefits realized through systematic data aggregation, and the technology platforms that enable implementation at scale. The audience for the book is healthcare IT staff responsible for designing, building, and maintaining the organization's data integration architecture; data engineers responsible for ETL jobs and data transformation logic; and quality reporting staff responsible for understanding the technical underpinnings of the measures they are calculating and reporting against. The book discusses architecture, integration patterns, and other design trade-offs, rather than implementation procedures. The topics discussed are generalizable to many different organizations and technology stacks.

2. Data Source Landscape for HEDIS Reporting

2.1 Electronic Health Records (EHR) Systems

Electronic Health Record systems constitute the primary repository of clinical data elements essential for calculating quality measures that assess the processes and outcomes of direct patient care delivery. These systems capture comprehensive patient demographics, including age, gender, race, ethnicity, and contact information, that enable population stratification and risk adjustment methodologies required by HEDIS specifications. Clinical documentation within EHR platforms encompasses structured data elements such as problem lists containing diagnosis codes, procedure documentation with associated temporal and provider attribution, medication orders and administration records, vital signs and anthropometric measurements, laboratory test orders and results, immunization records, and clinical observations recorded during patient encounters [3]. The role of EHRs in capturing point-of-care documentation is particularly critical for measures requiring evidence of specific clinical interventions, counseling activities, screening procedures, or diagnostic evaluations that may not generate billable events and therefore remain absent from claims data. The extraction of data from EHR systems presents significant technical challenges stemming from the diversity of platform architectures, proprietary data models, and the coexistence of structured coded data alongside unstructured narrative documentation in clinical notes [3]. This process requires complex natural language processing (NLP) to extract clinically relevant concepts for measure calculations. Important heterogeneity exists between and within healthcare systems for EHR system implementation, such as workflow, documentation practice, terminology system, and data validation rules. These factors create challenges to developing a standard approach for extraction logic and require flexible ETL architectures.

2.2 Claims and Administrative Systems

Claims and administrative data sets contain thorough records of health service usage based on billing transactions that identify the diagnoses that lead to care, the services and procedures provided,

the providers and facilities involved in the provision of services, and the chronology of the clinical events in the continuum of care. This data is standardized and employs several coding schemes, including the International Classification of Diseases for diagnosis coding, Current Procedural Terminology and Healthcare Common Procedure Coding System (CPT and HCPCS, respectively) for procedure documentation, and National Drug Codes for pharmaceutical products, to ensure consistent interpretation of the data through ETL pathways for HEDIS measure computation. Because claims data reflect longitudinal care across multiple providers, practice settings, and healthcare organizations, they may be used to assess measures of care coordination, follow-up after acute care, and high-value chronic care delivery. However, completeness of claims data may be compromised by provider billing, claims submission, and claims adjudication cycles that lead to lag time between service delivery and availability of claims data for analysis. ETL processes may also handle timeliness issues by capturing claim run-out dates in measure definitions and allowing for the reprocessing of late-arriving claim data to correct previously reported measurement results. Timeliness issues could also apply to the interim reporting requirement and potentially real-time quality monitoring initiatives, which intrinsically process claims data after a lag. This scenario creates a trade-off between having a timely view of performance and ensuring that measures are complete to avoid systematically undercounting the services provided.

2.3 Pharmacy Systems and Ancillary Data Repositories

Pharmacy systems track medication dispensing events through records of prescription fills that document the specific pharmaceutical products dispensed, quantities supplied, days of supply provided, prescribing providers, and fill dates essential for calculating medication adherence measures that assess patient persistence with evidence-based pharmacological therapies for chronic disease management. These systems enable the assessment of prescription fill patterns over time, the identification of gaps in medication availability, and the calculation of the proportion of days covered metrics that serve as proxy indicators for patient adherence to prescribed therapeutic regimens across multiple HEDIS measure domains. Laboratory information systems contribute discrete test results, reference ranges, and specimen collection dates necessary for measures requiring evidence of appropriate monitoring, diagnosis confirmation, or disease control assessment through specific biomarkers and diagnostic values [4]. Health information exchanges facilitate the aggregation of clinical data across organizational boundaries, providing visibility into care delivered by external providers and enabling more complete measure calculation in scenarios where patients receive services from multiple healthcare organizations not accessible through internal data repositories. Specialty data aggregators compile information from focused clinical domains such as immunization registries, cancer registries, and disease-specific databases that contribute targeted data elements for measures addressing specific clinical conditions or preventive services. The contribution of each ancillary source type varies across HEDIS measure domains, with pharmacy data being essential for medication adherence measures, laboratory data supporting diabetes and cardiovascular disease management measures, and supplemental clinical registries providing critical information for cancer screening and immunization measures that may be incompletely documented in primary organizational systems.

3. ETL Architecture and Core Processing Framework

3.1 Extract Phase: Data Acquisition from Heterogeneous Sources

The extraction phase establishes connectivity pathways to source systems and orchestrates systematic data retrieval required for HEDIS measure calculation through diverse technical approaches tailored to each repository's architectural characteristics. Database extraction uses structured query language interfaces through direct connections or database-specific drivers that speed up query execution while putting less strain on the operating system [5]. Application programming interfaces provide standardized access mechanisms for modern healthcare platforms implementing service-oriented architectures or microservices patterns that expose discrete data operations through representational state transfer endpoints. Flat file extraction processes structured text files and delimited formats used for data exchange between systems lacking real-time integration capabilities. Streaming sources enable continuous event ingestion as activities occur within source systems. Temporal extraction strategies present fundamental trade-offs between batch processing that executes on predetermined schedules to retrieve accumulated changes and real-time capture mechanisms that reduce latency but

increase system complexity and resource utilization. Batch scheduling needs to take into account the need to limit impact on source systems during business hours, scheduling for internal and external reporting requirements, and the need for sufficient throughput for downstream consumers. Data security and privacy requirements include encrypting data in transit, authentically ensuring the identity of the system performing the operation, audit logging of data access, and complying with legislation enforced on PHI [6]. The extraction architecture should accommodate heterogeneous source system schemas, various data availability patterns, and data access constraints while also ensuring process stability and data quality assurance.

3.2 Transform Phase: Data Cleansing, Validation, and Standardization

The transformation phase converts extracted data into standardized representations suitable for analytical processing, addressing quality deficiencies and structural inconsistencies that compromise HEDIS reporting validity. Data cleansing operations remediate anomalies, including duplicate records from overlapping sources, erroneous values violating domain constraints, and missing values requiring imputation or exclusion logic [5]. Deduplication implements matching algorithms accounting for identifier format variations, typographical errors, and temporal factors to consolidate multiple entity representations into canonical records. Data mapping translates source representations into target schemas aligned with measure requirements, resolving semantic terminology differences, harmonizing disparate coding practices, and establishing cross-system linkages. Transformation logic converts data types for analytical platform compatibility, standardizes temporal representations across systems with different granularities, normalizes textual data for consistent pattern matching, and applies business rules encoding clinical knowledge relevant to measure specifications. HEDIS-specific transformations include tests to see if a value set member is clinically significant, the calculation of derived attributes like member age or enrollment status, and the use of exclusion criteria to take members out of measure denominators based on documented conditions [6]. Transformation complexity necessitates careful design, maintaining traceability between source data and outputs, enabling validation of correctness, and supporting iterative refinement as specifications evolve or quality issues emerge.

3.3 Load Phase: Data Persistence and Target System Integration

The loading phase persists, transforming data into target repositories optimized for analytical queries, measure calculation and implementing strategies balancing data freshness, system performance, and historical retention. Data warehousing organizes data into dimensional models supporting complex aggregations and temporal analyses required for HEDIS calculation. Data lakes provide flexible storage for diverse data types, accommodating structured clinical data and semi-structured content informing future analytics. The choice between incremental loading (appending only changed records) and full refresh (replacing entire datasets) involves trade-offs between processing efficiency, consistency guarantees, and change detection complexity [5]. Incremental loads reduce processing time by limiting data movement to affected records but require sophisticated mechanisms for tracking lineage, managing slowly changing dimensions, and ensuring referential integrity. Full refresh simplifies processing logic but imposes higher computational costs and longer windows, constraining reporting timeliness. Metadata management captures technical, business, and operational metadata describing structures, transformation logic, quality metrics, and dependencies, enabling impact analysis and compliance documentation. Data lineage tracking establishes transparent relationships between source elements, transformations, and outputs, supporting auditability requirements and facilitating root cause analysis when quality issues affect calculations [6]. The loading architecture must be scalable to accommodate increasingly complex analytical requirements, as well as changing volumes, operational constraints (such as maintenance windows), and availability targets.

4. Strategic Benefits of ETL-Enabled HEDIS Reporting

4.1 Data Quality and Consistency Improvements

An important part of the ETL is the standardization, where the ETL process systematically converts heterogeneous source data to the same format to comply with the HEDIS use cases' technical specifications and analytic requirements. Data mapping is the process of harmonizing heterogeneous coding systems to resolve semantic discrepancies and structural variations among data formats in the integrated data [7]. Data quality validation rules can be incorporated into the ETL processes to assess and ensure data quality, rules

that can examine completeness, enforce domain constraints, identify temporal and referential integrity issues, and flag data as outliers or implausible with respect to known clinical or administrative patterns. Automated data quality tests detect problems as early as data ingestion to prevent poor-quality performance data from reaching the publishing stage. ETL workflow transformations can include error correction for established data quality problems by using commonly accepted imputation formulas, code mapping tables, or business rules that encode domain knowledge of valid data states and ranges for specific variables when common issues arise. ETL-enabled data quality improvement systems lead to more accurate reporting of HEDIS measures. This minimizes the risk that measurement error and exclusion and inclusion bias will considerably distort the reported measure result and that the reported measure result reflects the quality of care provided rather than artifacts of data quality deficiencies [8]. It further supports more reliable performance comparison across health plans, compliance with regulatory This affects reporting requirements and ultimately the reliability of quality measures used for accreditation decisions, public reporting, and value-based payment models.

4.2 Operational Efficiency and Scalability

Without ETL automation, data extraction, transformation and loading processes must be performed manually, which can be time-consuming for data analysts and has the potential for human error. ETL tools help automate the data workflows, enabling complex integration logic to be executed reliably, calculations of measures to be performed quickly by minimizing data preparation time, and reporting cycles to be reduced by using scheduled, repeatable extraction, transformation, and loading operations [7]. Other benefits of automation include auditability of data lineage and transformation processes through detailed data and transformation logging capabilities and better reproducibility of analytics results by reducing the reliance on domain-specific technical expertise for routine data processing tasks. Another important benefit of automation for health care organizations is having scalable systems to manage growing member populations and additional data sources to meet changing measurement specifications and growing clinical and administrative transaction volumes. ETL architectures designed for scalability employ distributed processing frameworks that partition workloads across computational resources, implement parallel execution patterns that process independent data segments concurrently, and utilize incremental processing strategies that limit computational scope to changed data rather than reprocessing entire datasets. Performance optimization techniques address the computational demands of large-scale healthcare datasets through query optimization strategies that minimize data movement and maximize processing efficiency, indexing approaches that accelerate data retrieval operations, and caching mechanisms that store frequently accessed data in high-performance storage tiers [8]. The architectural flexibility of modern ETL platforms enables organizations to scale processing capacity dynamically in response to workload demands, accommodate seasonal variations in data volumes associated with reporting cycles, and integrate emerging data sources without requiring fundamental redesign of integration infrastructure.

4.3 Enhanced Reporting Capabilities and Decision Support

Real-time or near real-time ETL configurations transform HEDIS reporting from retrospective performance assessment to dynamic quality monitoring by reducing latency between clinical service delivery and data availability for measure calculation. The acceleration of data integration cycles enables healthcare organizations to access current performance indicators that reflect recent care delivery activities, supporting timely identification of quality gaps, performance trends, and emerging issues requiring operational intervention. This temporal proximity between clinical activities and performance visibility facilitates proactive quality improvement initiatives that can address deficiencies while care episodes remain active rather than discovering gaps retrospectively when intervention opportunities have passed [7]. Trend analysis capabilities enabled by consolidated historical data support longitudinal assessment of performance trajectories, identification of seasonal patterns in quality metric achievement, and evaluation of improvement initiative effectiveness over time. Gap closure initiatives benefit from enhanced reporting capabilities through targeted identification of members requiring specific preventive services or disease management interventions, enabling outreach programs and care coordination activities informed by current data. ETL outputs are used to create decision support functions in business intelligence systems that convert raw performance data into interactive dashboards, scheduled reports, and visual analytics that convey complex dimensions and quality metrics to stakeholders,

including clinical leadership, quality improvement teams, and executive management [8]. Decision support capabilities extend to external public reporting, benchmarking with peer organizations, and creation of Business intelligence capabilities identify performance-based, data-driven strategies for investments in quality improvement and operational priorities.

5. Implementation Technologies and Platform Ecosystems

5.1 ETL Software Solutions and Platforms

Enterprise ETL tools ease the creation, orchestration, governance, and management of complex data integration processes. Commercial products like IBM DataStage, Informatica PowerCenter, and Talend Data Integration offer user-friendly tools that help technical staff create and manage data changes, connect to healthcare data sources, and handle important data information for analyzing impacts and managing changes. Enterprise offerings may also include additional capabilities such as data quality profiling, automated code generation, workflow scheduling, error handling, performance tuning, etc. Open-source options like Apache NiFi provide flexible and scalable data flow management, leveraging a distributed architecture to implement real-time data ingestion, routing, transforming and system mediation, as well as providing visual programming interfaces and extensibility for implementing custom data integration logic. Open-source ETL platforms might cost less for licenses, benefit from community development, be more transparent, and be easier to customize for specific organizational needs compared to enterprise software that comes with vendor support, high-quality features, and the ability to handle internal setup Cloud-native ETL platforms such as Databricks may be built on top of distributed computing frameworks and elastic cloud infrastructure to support modern data processing architectures, such as scalable workloads, batch and streaming processing patterns, and integration of machine learning capabilities for advanced analytics. Factors influencing ETL tool choice include the technical expertise of the organization, existing technology stack, data volume to process, data velocity, integration complexity, budget, and preference between using managed cloud services or on-premises infrastructure.

5.2 Data Warehousing and Storage Infrastructure

Centralized data warehouses serve as target repositories for transformed HEDIS data, providing optimized storage structures and query processing capabilities that support the complex analytical operations required for measure calculation, performance reporting, and quality improvement analytics. Data warehouse architectures for storing healthcare data may employ dimensional modeling techniques, which organize data according to business processes and analytical perspectives. Star schema and snowflake schema are popular dimensional modeling techniques that can be used for building data integration models that support ad hoc queries and aggregations [9]. HEDIS reporting dimensional models often include fact tables for measuring events such as clinical visits, laboratory tests, and medication dispensing events, as well as dimension tables for organizing descriptive attributes such as patients, providers, time, and clinical categories. These models efficiently answer common queries and can be used flexibly to define additional report queries and measures. Physical data warehouses exist as specialized database systems, storing data in column-oriented data structures for performance and in compressed formats to save space. A data warehousing architecture performs processing in parallel using distributed computing. Cloud data warehouses, such as Amazon Redshift, Google BigQuery, and Azure Synapse Analytics, provide elastic computing and storage resources to accommodate changing demand and allow organizations to move away from on-premises data warehouse systems. Additionally, there is no need for expensive hardware upfront, and users can scale storage up and down as needed, benefit from pay-as-you-go pricing, receive automatic software upgrades, and integrate with other cloud services such as data loading, data governance, data security, and business intelligence services. Data warehouse architecture decisions concern the use of on-premises, cloud, or hybrid data warehouses. Influencing this choice are factors such as data sovereignty, latency, existing infrastructure investment, cloud usage policies, total cost of ownership (TCO) of infrastructure, software licensing, and operations

5.3 Interoperability Standards and Integration Frameworks

Fast Healthcare Interoperability Resources (FHIR) is a standard describing data formats, data elements, and application programming interfaces (APIs) for exchanging electronic health information for different kinds of health data system interoperability. FHIR's modular resource-based

architecture provides granular representations of clinical and administrative entities such as patients, encounters, observations, medications, and diagnostic reports, enabling fine-grained data access patterns that align with modern API-driven integration approaches [9]. The adoption of FHIR as a common data representation framework simplifies ETL implementation by reducing the semantic mapping complexity between source systems and target analytical repositories, as systems implementing FHIR interfaces expose data in standardized formats that require less transformation logic than proprietary data models. Standardized data models beyond FHIR, including HL7 Version 2 messaging standards, Clinical Document Architecture specifications, and domain-specific terminologies such as SNOMED CT and LOINC, improve ETL mapping efficiency by establishing common vocabularies and structural conventions that reduce the variability across source systems. APIs are used as the integration mechanism for cloud-based healthcare systems and modern application architecture use cases. Real-time data access from multiple sources and systems is possible through RESTful APIs that expose parameterized queries, fine-grained data, and event notifications instead of reloading files in batches. This improves the total data freshness of the end-to-end process between updates in the source systems of record and the analytics systems. It also allows one to deconstruct monolithic integration logic into domain-centric microservice components for consumption and deployment independently. Standards-based integration frameworks may also need to verify identities and permissions, keep data private, manage errors, handle problems with specific strategies, and control versions to ensure that integration remains stable even when the underlying standards or system setups change.

***Table 1**: HEDIS Data Source Classification [3, 4]*

Data Source	Key Data Elements	Primary HEDIS Application
Electronic Health Records	Demographics, diagnoses, procedures, vital signs, lab orders, immunizations	Clinical interventions, preventive screening, non-billable services
Claims Systems	Billing transactions, diagnosis/procedure codes, provider identifiers	Care coordination, longitudinal tracking, service utilization
Pharmacy Systems	Prescription fills, medication quantities, days supply, fill dates	Medication adherence, therapeutic persistence
Laboratory Systems	Test results, biomarker values, specimen dates	Disease monitoring, diagnosis confirmation
Health Information Exchanges	Cross-organizational clinical data, external encounters	Multi-organization care visibility
Specialty Registries	Immunization records, cancer registries, disease databases	Preventive services, supplemental documentation

***Table 2:** ETL Processing Phases for HEDIS Data Integration [5, 6]*

ETL Phase	Technical Approaches	Key Operations	Strategic Considerations
Extract	Database connections (SQL), APIs (REST), flat files, streaming ingestion	Source system connectivity, batch/real-time scheduling, data retrieval	Minimize operational load, ensure security/encryption, manage access patterns
Transform	Data cleansing, deduplication, mapping, validation	Standardization, error correction, value set testing, business rule application	Maintain traceability, enable iterative refinement, ensure HEDIS compliance
Load	Data warehouse loading, dimensional modeling, incremental/full refresh	Data persistence, metadata management, lineage tracking	Balance freshness and performance, ensure auditability, support scalability

***Table 3:** Strategic Benefits of ETL-Enabled HEDIS Reporting [7, 8]*

Data Quality & Consistency	Automated validation, standardization, error correction, code mapping	Reduced measurement error, regulatory compliance, stakeholder confidence	Accurate performance metrics, reliable comparisons, valid quality assessment
Operational Efficiency	Workflow automation, reduced manual handling, repeatable processes,	Resource optimization, faster reporting cycles, improved reproducibility	Timely measure calculation, scalable processing, reduced human error

Decision Support	Real-time monitoring, trend analysis, interactive dashboards, gap	Proactive interventions, data-driven planning, performance visibility	Timely care gap closure, longitudinal tracking, evidence-based improvements

Table 4: *ETL Implementation Technology Comparison [9, 10]*

Technology Category	Solutions	Key Features	Implementation Considerations
Enterprise ETL Platforms	IBM DataStage, Informatica PowerCenter, Talend Data Integration	Graphical development environments, pre-built connectors, metadata management, automated code generation	Vendor support, enterprise features, licensing costs, technical expertise requirements
Open-Source ETL Tools	Apache NiFi	Distributed architecture, real-time ingestion, visual programming, custom extensibility	Lower licensing costs, community development, transparency, implementation complexity
Cloud-Native Platforms	Databricks	Distributed computing, elastic infrastructure, batch/streaming processing, ML integration	Scalability, workload flexibility, cloud adoption strategy, consumption-based pricing
Data Warehousing	Amazon Redshift, Google BigQuery, Azure Synapse Analytics	Dimensional modeling, columnar storage, parallel processing, elastic resources	Data sovereignty, network latency, existing infrastructure, total cost of ownership
Interoperability Standards	FHIR, HL7 V2, CDA, SNOMED CT, LOINC	Standardized data models, API-driven integration, reduced mapping complexity	Authentication/authorization, data privacy, version management, integration stability

6. Conclusions

ETL technologies help ensure accurate reporting of HEDIS measures. ETL processes use electronic health record (EHR) data, as well as medical and pharmacy claims data, from central data warehouses, claims systems, and pharmacy claims databases to assist with measure reporting. ETL extraction, transformation, and loading processes empower workflows that lessen variation in data processing, better measure quality, and make steady versions of HEDIS technical specifications to receive performance metrics for accreditation, regulatory compliance, or value-based reimbursement. Operationally, a strong ETL process enables automation. A strong ETL process accelerates reporting cycles. A strong ETL process scales with increasing data volumes. Strategically, it improves decision-making, enables focused quality improvement activities, and provides an integrated view of performance. With the evolution of real-time ETL processes, healthcare has shifted from retrospective identification of quality of care gaps to real-time monitoring of quality to identify and close quality gaps in clinical episodes underway. The rise of cloud-native platforms is impacting HEDIS reporting infrastructure with elastic computing and advanced analytics tooling. Emerging technologies, such as artificial intelligence and machine learning, may improve data quality with clever outlier detection, automate validation logic, and offer predictive analytic capabilities for measurement calculation. Healthcare organizations building ETL must assess the scale of their operations, technical capabilities, and financial resources. We must weigh the trade-offs between enterprise, open-source, cloud-native, and customary on-premise ETL tools. Thorough data governance frameworks, security controls for regulations and security policy compliance, and continuous optimization to meet HEDIS specification and interoperability framework (e.g., FHIR) changes are also important. Thorough planning for ETL capabilities is needed to harness the planned value of high-quality integrated data in support of clinical excellence, operational efficiencies, and long-term data-driven competitive advantage and differentiation in the healthcare marketplace.

Author Statements:

- **Ethical approval:** The conducted research is not related to either human or animal use.
- **Conflict of interest:** The authors declare that they have no known competing financial interests or personal relationships that could

have appeared to influence the work reported in this paper

- **Acknowledgement:** The authors declare that they have nobody or no-company to acknowledge.
- **Author contributions:** The authors declare that they have equal right on this paper.
- **Funding information:** The authors declare that there is no funding to be acknowledged.

- **Data availability statement:** The data that support the findings of this study are available on request from the corresponding author. The data are not publicly available due to privacy or ethical restrictions.
- **Use of AI Tools:** The author(s) declare that no generative AI or AI-assisted technologies were used in the writing process of this manuscript.

References

[1] Julia Adler-Milstein et al., "Electronic Health Record Adoption In US Hospitals: Progress Continues, But Challenges Persist," Health Affairs Vol. 34, No. 12: Affordability, Access, Models Of Care, & More, 2015. Available: https://www.healthaffairs.org/doi/10.1377/hlthaff.2015.0992

[2] Sharon Silow-Carroll et al., "Using Electronic Health Records to Improve Quality and Efficiency: The Experiences of Leading Hospitals," Commonwealth Fund Issue Briefs 17:1-40 (2012). Available: https://www.researchgate.net/publication/230570249

[3] Kristiina Häyrinen et al., "Definition, structure, content, use and impacts of electronic health records: A review of the research literature," International Journal of Medical Informatics, 2008. Available: https://www.sciencedirect.com/science/article/abs/pii/S1386505607001682

[4] Clemens Scott Kruse et al., "Challenges and Opportunities of Big Data in Health Care: A Systematic Review," JMIR Publications Advancing Digital Health & Open Sciences, 2016. Available: https://medinform.jmir.org/2016/4/e38/

[5] Panos Vassiliadis et al., "A Survey of Extract-Transform-Load Technology," International Journal of Data Warehousing and Mining 5:1-27, 2009. Available: https://www.researchgate.net/publication/220613761_A_Survey_of_Extract-Transform-Load_Technology

[6] Moh'd Alsqour et al., "A survey of data warehouse architectures—Preliminary results," IEEE 2012 Federated Conference on Computer Science and Information Systems (FedCSIS), 2012. Available: https://ieeexplore.ieee.org/document/6354451

[7] Erhard Rahm and Hong Hai Do, "Data Cleaning: Problems and Current Approaches," Research Gate, 2000. Available: https://www.researchgate.net/publication/220282831

[8] Carlo Batini et al., "Methodologies for data quality assessment and improvement," AACM Computing Surveys (CSUR), Volume 41, Issue 3, 2009. Available: https://dl.acm.org/doi/10.1145/1541880.1541883

[9] I.R. Mansuri and S. Sarawagi, "Integrating Unstructured Data into Relational Databases," IEEE : 22nd International Conference on Data Engineering (ICDE'06), 2006. Available: https://ieeexplore.ieee.org/document/1617397

[10] Michael Armbrust et al., "A View of Cloud Computing," Communications of the ACM, Volume 53, Issue 4, 2010. Available: https://dl.acm.org/doi/10.1145/1721654.1721672