International Journal of Information Technology and Computer Science Applications (IJITCSA)
p-ISSN: 2964-3139 e-ISSN: 2985-5330
Vol. 04, No. 01, page 70 - 82
Submitted 11/2/2026; Accepted 1/4/2026; Published 14/4/2026

Healthcare Data Integration Through Enterprise Data Warehousing: Architecture, Conformance Pipeline, and Experimental Validation for Readmission Analytics

La Duy Ngôn
Information Technology, Can Tho University, Can Tho City, VIETNAM
e-mail: nayarunar11@gmail.com

Corresponding Autor: La Duy Ngôn

Publisher's Note: JPPM stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Copyright: © 2026 by the authors. Submitted for possible open access publication under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

Abstract
Healthcare organizations operate a fragmented digital landscape in which hospital information systems (HIS), electronic health records (EHR), laboratory systems, billing platforms, and departmental applications are optimized for transaction processing but not for integrated analysis. The resulting interoperability gaps, semantic inconsistency, duplicated records, and uneven data quality constrain enterprise reporting and limit higher-value analytics. This paper substantially proposes implementable enterprise data warehouse architecture, formalizing its data-quality and conformance mechanisms, and validating the design through experimental analytics use case. The proposed framework combines an integration layer for ETL/ELT, conformed dimensions, departmental marts, governance controls, and an analytics layer for OLAP and machine learning. To demonstrate practical value, the paper evaluates the framework on a de-identified inpatient diabetes dataset comprising 101,766 encounters and 50 raw attributes. The experimental pipeline performs profiling, conformance mapping, diagnosis grouping, missing-value treatment, and dimensional modeling before training benchmark readmission models. The best ranking performance is obtained by XGBoost with an AUROC of 0.688 and an AUPRC of 0.235, while threshold tuning improves recall-oriented operational utility. The results show that healthcare warehousing should not be framed merely as centralized storage; rather, it is an architectural mechanism for interoperability, data quality control, reproducible analytics, and decision support. The manuscript concludes with implementation guidance and limitations relevant to hospitals seeking a scalable, governance-aware warehousing program.

Keywords— Data integration, Healthcare data warehouse, HIS, EHR, Interoperability, Dimensional modeling, ETL, Readmission analytics.

1 Introduction
Healthcare institutions have become high-volume data producers. Clinical encounters, laboratory observations, medications, procedures, claims, device outputs, and administrative events are now captured continuously across multiple applications. Yet the analytical value of this abundance is rarely realized automatically, because most operational systems are built for speed, locality, and transaction integrity rather than enterprise integration. Hospitals therefore face a paradox. They are data-rich at the point of capture but insight-poor at the organizational level [1], [2].

The practical manifestation of this paradox is familiar. A physician may document care in the EHR, laboratory results may be stored in a separate subsystem, claims data may reside in revenue-cycle software, and managerial dashboards may be assembled manually from spreadsheets or extracts. Although each system performs its local task adequately, the organization struggles to answer cross-cutting questions such as which patient groups are associated with elevated readmission risk, how medication changes relate to length of stay, whether quality indicators differ across admission channels, and how operational bottlenecks influence outcomes [3], [4]. These questions require longitudinal, integrated, and quality-controlled data precisely the capabilities that operational silos do not provide.

©2026 La Duy Ngôn.

Most of the literature that we gathered positioned this challenge primarily as an issue of infrastructure fragmentation and security in HIS and EHR environments, and it proposed a healthcare data warehouse as a general solution concept. That framing is directionally correct, but contemporary healthcare warehousing literature now demands a more rigorous articulation of architecture, interoperability, data quality, governance, and empirical validation than a purely descriptive essay can deliver [5], [6], [7]. Research in this area must do more than restate that warehousing is useful. It must explain how the warehouse should be structured, how heterogeneous source data are conformed, what governance controls are required, and whether the resulting analytical layer measurably supports decision-making.

This study addresses that need. It expands the initial research into a full, evidence-based research by integrating three levels of contribution. First, it develops a practical enterprise architecture for healthcare data integration that unifies operational sources, staging, conformance, dimensional storage, governance, and analytics. Second, it formalizes the logic of conformance and data quality through mathematical formulations and an ETL algorithm suitable for implementation. Third, it validates the architecture with a real analytical workload such as 30-day hospital readmission prediction on a large de-identified inpatient dataset, designed to illustrate how a warehouse-backed analytical mart can support both descriptive and predictive intelligence.

The paper makes the following contributions:

1. It reframes healthcare warehousing from a storage-centric perspective to an integration-and-governance perspective grounded in current clinical data warehouse literature [5], [7].
2. It proposes a conformed healthcare warehouse architecture combining source adapters, quality gates, dimensional marts, and access controls for analytical reuse across departments.
3. It introduces a reproducible data processing workflow for transforming heterogeneous clinical-style records into warehouse-ready analytical structures.
4. It demonstrates the framework empirically using a readmission analytics scenario and reports benchmark machine learning results, thereby connecting architectural design to measurable analytical utility.

The rest of the paper is organized as follows. Section II reviews healthcare data integration challenges and motivates warehousing. Section III presents the proposed architecture, dimensional model, and formal processing framework. Section IV describes the experimental design. Section V reports the empirical results. Section VI discusses implementation implications, and Section VII concludes the paper.

2 Background and Problem Framing

2.1 Operational Systems, Interoperability, and Analytical Friction

Hospital information systems and EHR platforms are indispensable for day-to-day care delivery. They support registration, order entry, charting, medication administration, billing, and many other operational processes. However, their dominant design logic is OLTP-oriented, updating a transaction quickly, preserving local consistency, and serve a bounded workflow. This logic is appropriate for clinical operations but poorly aligned with enterprise analytics, which require integrated historical data, stable semantics, and efficient cross-subject querying [2], [8].

Interoperability remains a persistent difficulty even as standards mature. Recent reviews show that healthcare interoperability is not merely a transport problem but also a semantic, organizational, and governance problem. Systems may exchange messages while still disagreeing on code systems, data granularity, event timing, or update logic [9 – 11]. Consequently, many organizations continue to rely on brittle interfaces and custom extracts that satisfy immediate reporting needs but do not establish a durable analytical foundation.

This fragmentation creates three classes of analytical friction. The first is structural friction where records are distributed across heterogeneous schemas and applications. The second is semantic friction where identical concepts are represented differently across systems, for example through varying diagnosis code formats, category definitions, or missing-value placeholders. The third is temporal friction where data refresh cycles, event timestamps, and slowly changing reference data are not synchronized. Together these frictions undermine longitudinal analysis and reduce trust in analytics [12], [13].

2.2 Why Enterprise Data Warehousing Still Matters in Healthcare

Despite the rise of data lakes, cloud-native platforms, and interoperability APIs, the enterprise data warehouse remains central to healthcare analytics because it provides curated, governed, and query-efficient representations of organizational data [3], [4]. Current scoping reviews report that clinical data warehouses continue to serve as foundational infrastructure for cohort identification, quality measurement, translational research, operational reporting, and predictive modeling [5], [6].

A warehouse is especially useful when the institution needs repeatable metrics and decision support rather than raw data accumulation. In that context, dimensional modeling remains attractive because it exposes stable business concepts such as patients, encounters, admissions, diagnoses, medications, and time through conformed dimensions and measurable facts. Such models support explainable analytics and governance better than ad hoc extraction pipelines assembled independently for every dashboard or study [14], [15].

Warehousing also supports organizational scale. A carefully designed enterprise model can preserve common dimensions while allowing department-specific marts for radiology, laboratory performance, utilization review, chronic disease surveillance, or readmission management. This approach reduces duplication of analytical logic and creates a shared semantic layer for the institution [3], [7].

2.3 Data Quality, Governance, and Security as Architectural Requirements

Healthcare warehousing cannot be evaluated only by integration breadth. Data quality is a first-order requirement because poor completeness, inconsistency, plausibility, or provenance directly damages downstream analysis [16 – 18]. Reviews of the clinical data life cycle repeatedly show that quality failures emerge at multiple stages such as capture, coding, transmission, aggregation, transformation, and reuse [19], [20]. For this reason, quality should be embedded as a series of checkpoints in the integration of architecture rather than treated as an afterthought.

Governance and security are equally non-negotiable. Healthcare data contain sensitive personal information, and unauthorized access, weak lineage, or uncontrolled secondary use can undermine both compliance and trust. Modern governance frameworks emphasize role-based access, auditability, minimization, de-identification for secondary use, and policy alignment with regulatory and institutional obligations [21], [22]. Therefore, any proposed warehouse architecture must integrate governance into the design of staging, dimensional marts, metadata, and access pathways.

2.4 Research Gap

The literature strongly supports healthcare data warehousing, but a practical gap remains between high-level architectural advocacy and reproducible end-to-end demonstrations. Many articles discuss data warehouse value conceptually, while empirical modeling studies often jump directly to predictive performance without explaining the intermediate conformance and dimensional design that would make such analytics sustainable inside a hospital [5], [23], [24]. This paper bridges that gap by connecting architecture, ETL logic, dimensional modeling, and experimental analytics within a single manuscript.

3 Proposed Healthcare Data Integration Solution

3.1 Architecture Overview

Figure 1 presents the proposed enterprise architecture. The design contains four principal layers namely operational sources, integration, enterprise data warehouse, and analytics/access. A cross-cutting governance plane spans the entire architecture. This arrangement reflects the observation that warehousing success depends not only on storage design but also on the discipline of conformance, quality control, and security enforcement [3], [7].

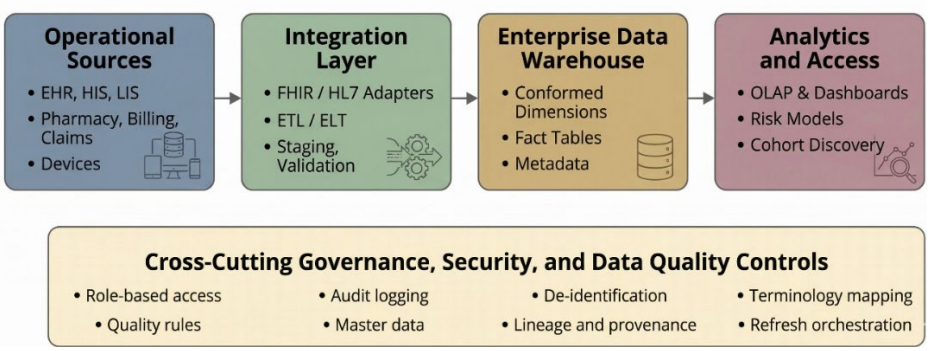

Figure 1: Proposed healthcare data integration architecture linking operational systems, integration services, enterprise warehousing, and governed analytics.

The *operational source layer* includes EHR, HIS, billing, pharmacy, laboratory, claims, and other departmental systems. These systems preserve their transactional role and are not replaced by the warehouse. The *integration layer* contains connectors, staging storage, validation services, terminology mapping, and ETL/ELT logic. Its objective is

to translate heterogeneous source data into stable analytical representations. The *enterprise warehouse layer* conformed to dimensions, fact tables, metadata, and subject-area marts. Finally, the *analytics/access layer* exposes OLAP, dashboards, cohort discovery workflows, and predictive models.

The governance plan is deliberately explicit. It includes role-based access control, audit logging, de-identification policies, provenance tracking, master data handling, and refresh orchestration. These controls are necessary because healthcare integration without governance may scale analytical reach while simultaneously scaling organizational risk [21], [22].

3.2 ETL and Conformance Pipeline

Figure 2 details the ETL and analytics pipeline. The central idea is that warehouse value emerges from conformance, not just ingestion. Raw extracts are profiled for missingness, duplication, drift, and cardinality anomalies. Attributes are then standardized, coded concepts are grouped into analytically meaningful categories, and only validated structures are loaded into dimensional storage.

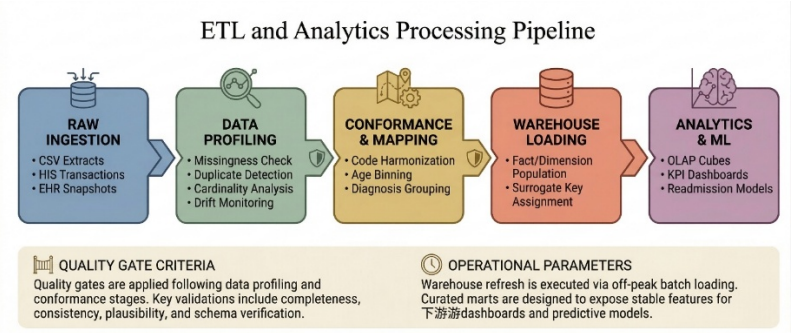

Figure 2: ETL and analytics pipeline from raw ingestion to conformed warehouse loading and machine-learning-ready marts.

The pipeline assumes off-peak batch refresh for the experimental scenario, which is appropriate for managerial analytics and risk stratification dashboards. More time-critical use cases can adopt near-real-time micro-batching, but the architectural principle remains identical: source capture must be followed by profiling, conformance, and governed loading before analytics are exposed [13], [25].

3.3 Dimensional Model for the Analytical Mart

The analytical mart used in the experiment is illustrated in Figure 3. The star schema centers on Fact Encounter, which stores encounter-level measures and foreign keys to conformed dimensions. The dimensions are Dim Patient, Dim Admission, Dim Diagnosis, Dim Medication, and Dim Time. This schema is intentionally compact: it demonstrates the dimensional logic needed for analytical reuse without reproducing the full complexity of a production hospital warehouse.

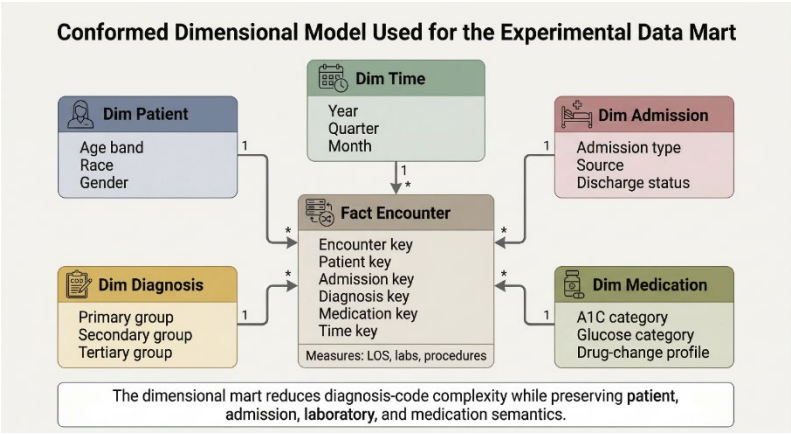

Figure 3: Conformed dimensional model for the experimental encounter mart.

The main design principle is semantic stabilization. For example, diagnosis codes are transformed into higher-level diagnostic groups so that analytics are not overwhelmed by extreme cardinality and coding sparsity. Medication-related variables are similarly organized into clinically interpretable features, while the time dimension supports consistent period-based slicing. This dimensional approach improves interpretability and facilitates both OLAP reporting and downstream predictive modeling [14], [15]. Table I summarizes the dimensional mart.

Table 1: Core entities in the experimental dimensional mart

Entity	Example attributes	Analytical purpose
Dim Patient	Age band, race, gender	Demographic stratification, equity and cohort analysis
Dim Admission	Admission type, source, discharge status	Utilization, care pathway, case-mix comparison
Dim Diagnosis	Primary/secondary/tertiary diagnosis groups	Clinical burden profiling and group-level analysis
Dim Medication	A1C category, glucose category, medication-change profile	Treatment pattern analysis and readmission risk characterization
Dim Time	Year, quarter, month	Trend analysis, seasonality, workload planning
Fact Encounter	Length of stay, procedure counts, laboratory counts, surrogate keys	Encounter-level measures and joins across dimensions

3.4 Mathematical Formulation of Conformance and Analytical Scoring

Let x_{ij} denote the value of attribute j for record i in a raw source extract. The conformance stage transforms raw values into warehouse-ready values through attribute-specific mapping

$$z_{ij} = \phi_j(x_{ij}) \tag{1}$$

Where $\phi_j(\cdot)$ may represent type conversion, code normalization, category grouping, or imputation logic. In practice, ϕ_j is not a single universal function but a family of domain-specific transformations coordinated through metadata and business rules.

For each attribute j, completeness after profiling can be quantified as

$$C_j = 1 - \frac{m_j}{n} \tag{2}$$

Where m_j is the number of missing or unusable values and n is the number of records. A broader warehouse quality score can then be defined as a weighted aggregation of quality dimensions such as completeness, consistency, plausibility, and uniqueness:

$$Q = \sum_{d=1}^{D} w_d \, q_d, \qquad \sum_{d=1}^{D} w_d = 1 \tag{3}$$

Where $q_d \in [0,1]$ is the score for dimension d and w_d is its relative importance. The practical implication is that warehouse loading should be conditional on quality gates rather than unconditional extraction.

For the predictive layer, the readmission probability is estimated with gradient-boosted trees. Given encounter feature vector $\mathbf{x}_i$, the model output is

$$\hat{p}(y_i = 1 \mid \mathbf{x}_i) = \sigma\left(\sum_{k=1}^{K} f_k(\mathbf{x}_i)\right) \tag{4}$$

Where f_k is the k-th decision tree and $\sigma(\cdot)$ is the logistic link. Ranking quality is evaluated with AUROC and AUPRC, while operational deployment can choose a threshold τ such that allowing the institution to trade precision against recall according to care-management capacity.

$$\hat{y}_i = \mathbb{I}[\hat{p}(y_i = 1 \mid \mathbf{x}_i) \geq \tau] \tag{5}$$

3.5 Warehouse Build Algorithm

Algorithm 1 summarizes the transformation from source extracts to an analytical mart. The algorithm is expressed generically so that it can be implemented with SQL-based ETL tools, Python pipelines, or enterprise data integration platforms.

Algorithm 1 Healthcare warehouse build and analytical mart preparation

Require: Source systems $S = \{s_1, s_2, \ldots, s_n\}$, metadata rules M, quality thresholds T
Ensure: Conformed warehouse tables and analytical mart

Initialize staging area and metadata log
for all source system $s \in S$ **do**
Extract records and schema descriptors from s
Profile missingness, duplicates, invalid codes, and type mismatches
Write profiling statistics to lineage metadata
if quality metrics violate T **then**
    quarantine offending batch and raise remediation task
else
    apply conformance mappings ϕ_j from M
    standardize keys, timestamps, and controlled vocabularies
    derive grouped diagnosis and medication features
    load conformed records into dimension and fact staging tables
end if
end for
Generate surrogate keys and populate warehouse fact/dimension tables
Refresh subject-area marts for reporting and predictive modeling
Train or score analytical models on curated mart features
Publish metrics, dashboards, and audit log entries

4 Experimental Design

4.1 Analytical Scenario and Dataset Description

To validate the proposed architecture, the paper uses a de-identified inpatient diabetes dataset widely employed in secondary health analytics. The dataset contains 101,766 encounters described by 50 raw attributes, including demographics, admission descriptors, utilization measures, laboratory indicators, diagnosis fields, and medication-related variables. The target variable indicates whether the patient was readmitted within 30 days of discharge. This use case is suitable for healthcare warehousing because readmission management requires cross-domain integration such as demographics, encounter history, diagnoses, medication patterns, and utilization variables all contribute to risk stratification [23 – 26].

The dataset is analytically attractive for three reasons. First, it is sufficiently large to expose realistic quality and conformance challenges. Second, it mixes categorical and numerical fields typical of hospital systems. Third, the target class is imbalanced, which makes the evaluation of ranking metrics and threshold selection directly relevant to operational deployment.

Table II summarizes the analytical dataset after the main population definition step. Consistent with common readmission modeling practice, encounters associated with discharge categories not meaningful for standard

readmission follow-up were excluded from the modeling population. This exclusion is not a cosmetic preprocessing step. It represents a governance decision about what analytical question the mart is designed to answer.

Table 2: Dataset summary for the experimental use case

Characteristic	Value
Raw encounters	101,766
Raw attributes	50
Modeling encounters after discharge filtering	99,343
Target event rate (readmitted within 30 days)	11.39%
Numerical predictors used in mart	11
Categorical predictors used in mart	33
Train/test split	80% / 20%
Evaluation setting	Stratified hold-out with fixed random seed

4.2 Preprocessing and Feature Engineering

The preprocessing logic mirrors the warehouse conformance pipeline rather than a purely machine-learning-oriented cleanup. First, administrative identifiers not suitable for predictive learning were removed from the feature set. Second, placeholder symbols used to denote missingness were converted into proper null values. Third, highly incomplete attributes were dropped when their information value did not justify operational maintenance cost. Fourth, primary, secondary, and tertiary diagnosis codes were grouped into broader diagnostic categories to reduce sparsity and improve business interpretability. Fifth, indicators were derived for the presence of glycated hemoglobin and serum glucose results. Finally, numerical variables were median-imputed and standardized for linear modeling, while categorical variables were mode-imputed and encoded for baseline models.

This sequence is important from a warehousing perspective. The goal is not to maximize model accuracy by arbitrary feature manipulation, but to create stable, repeatable, and semantically interpretable features that could plausibly be maintained inside a production mart. Conformance should therefore reduce complexity while preserving decision-relevant signal [17], [20].

4.3 Benchmark Models and Evaluation Metrics

Four benchmark models were assessed, such as a dummy baseline, logistic regression, random forest, and XGBoost. The dummy model serves as a no-skill reference. Logistic regression represents a transparent linear baseline. Random forest captures nonlinearities and interaction effects while remaining widely understood in health analytics. XGBoost was selected as a strong tree-ensemble benchmark commonly used in structured clinical prediction problems [23], [24].

Given class imbalance, the evaluation prioritizes AUROC and AUPRC for discrimination and ranking quality, while accuracy, precision, recall, and F1-score are reported for operational context. Accuracy alone is insufficient because a majority-class classifier can achieve deceptively high accuracy in imbalanced readmission tasks. To reflect practical deployment, the best model is additionally evaluated under threshold tuning to improve recall, which is often more important when the objective is to flag patients for transitional-care intervention.

4.4 Reproducibility Considerations

The experiment was designed as a reproducible analytical demonstration rather than a one-off benchmark. The package includes the LaTeX manuscript, bibliography, figures, and a Python script for experiment reproduction. Fixed random seeds were used for data splitting and model fitting. The analytical workflow was constrained to features that could be derived from the warehouse mart without privileged future knowledge. This design choice guards against leakage and keeps the experiment aligned with realistic deployment assumptions.

5 Results

5.1 Data Quality and Conformance Outcomes

Figure 4 summarizes the main data quality profile. Before conformance, the raw data contained 374,017 missing or placeholder cells. After population selection and attribute rationalization, 185,243 unresolved missing values remained for imputation-aware modeling, while five highly incomplete attributes were removed from the mart design. Diagnosis-code complexity was reduced from hundreds of granular categories to a smaller set of conformed diagnosis groups suitable for dimensional storage and reporting.

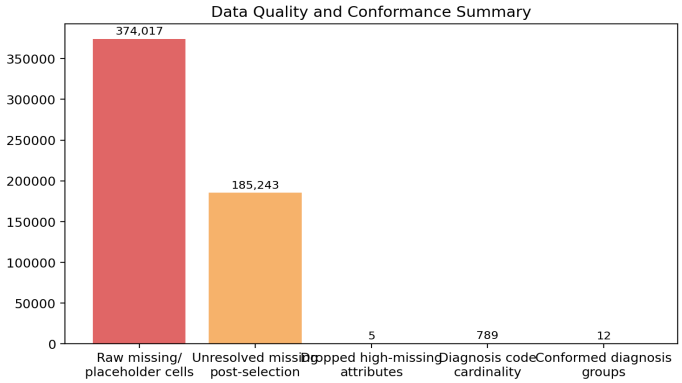

Figure 4: Data quality and conformance summary for the experimental pipeline.

Table 3 interprets these outcomes from an architectural perspective. The point is not only that missingness exists, but that the warehouse process makes missingness explicit, measurable, and governable. This is one of the major distinctions between enterprise warehousing and ad hoc reporting: warehouse design exposes data quality as metadata and policy rather than hiding it inside isolated analyst scripts [16], [18].

Table 3: Interpretation of data quality and conformance outputs

Indicator	Value	Interpretation
Raw missing/placeholder cells	374,017	High pre-integration incompleteness typical of heterogeneous operational capture
Unresolved missing after selection	185,243	Remaining nulls handled through model-aware imputation and explicit profiling
Dropped high-missing attributes	5	Warehouse rationalization removes low-value, maintenance-heavy attributes
Diagnosis-code cardinality	789	Raw coding space too sparse for robust reporting without conformance
Conformed diagnosis groups	12	Reduced complexity improves interpretability and dimensional stability

5.2 Benchmark Model Performance

The predictive results are presented in Table 4 and Figure 5. XGBoost achieved the best ranking performance with AUROC =0.688 and AUPRC =0.235. Random forest followed closely, while logistic regression provided lower discrimination but much higher recall at the default decision threshold because class balancing shifted the operating point toward sensitivity. The dummy classifier, as expected, exhibited no useful ranking ability.

Table 4: Benchmark performance for 30-day readmission prediction

Model	Accuracy	Precision	Recall	F1	AUROC	AUPRC
Dummy	0.8884	0	0	0	0.5	0.1116
Logistic Regression	0.6646	0.1704	0.5187	0.257	0.6458	0.1996

Random Forest	0.889	0.5524	0.0255	0.049	0.6587	0.217
XGBoost	0.8889	0.6136	0.0119	0.023	**0.6875**	**0.235**

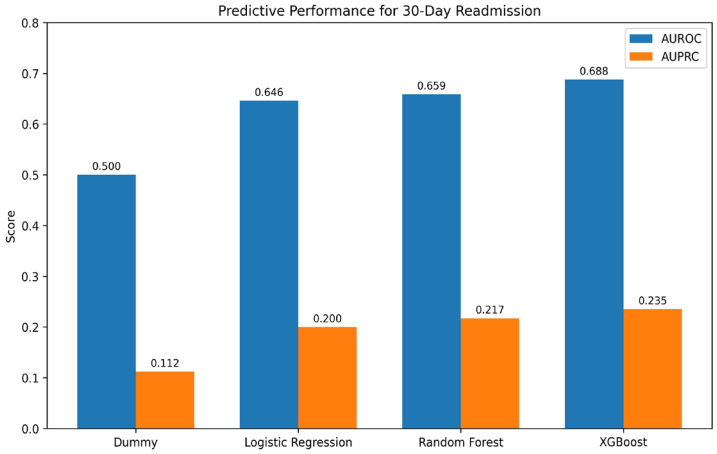

Figure 5: AUROC and AUPRC comparison for benchmark models

These results should be interpreted carefully. The three ensembles achieve better ranking ability, but their default 0.5 threshold is poorly calibrated for an imbalanced readmission task. In operational deployment, the institution is unlikely to retain such a threshold if the purpose is early intervention. Threshold selection must therefore be treated as part of the analytical governance layer, not as an afterthought buried inside the model code.

5.3 Threshold Tuning for Operational Use

To illustrate this point, the best model was re-evaluated using an F1-optimized threshold of $\tau = 0.1394$. The tuned operating point reduced accuracy to 0.7531 but increased recall to 0.4527 and improved F1-score to 0.2904. Table 5 shows the result.

Table 5: Threshold-tuned operational performance for the best model

Model	**Threshold**	**Accuracy**	**Precision**	**Recall**	**F1**	**AUROC**	**AUPRC**
XGBoost (threshold tuned)	0.1394	0.7531	0.2137	0.4527	0.29	0.6875	0.235

From a clinical operations perspective, this trade-off is often more useful than the default threshold. A transitional-care team may accept more false positives if the intervention cost is moderate and the cost of missing a likely readmission is high. The warehouse architecture supports this by keeping both the stable mart features and the scored outputs accessible to dashboards, cohort queues, and audit logs.

5.4 Feature Importance and Interpretability

Figure 6 shows the most influential warehouse features for the XGBoost model. The most informative variables include laboratory procedure volume, A1C result category, time in hospital, glucose-result availability, procedure count, and diagnosis grouping. These variables are clinically plausible: they proxy acuity, diagnostic workload, chronic disease control, and complexity of care. Their importance also validates the dimensional design, since several top predictors arise directly from the conformed mart rather than raw identifier fields.

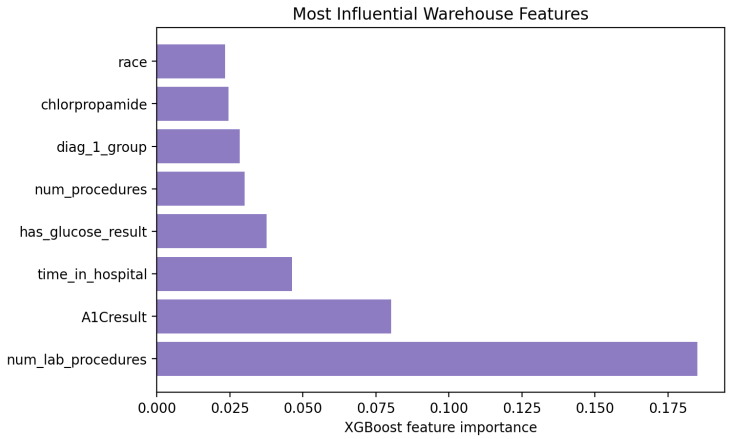

Figure 6: Top warehouse-derived features contributing to XGBoost predictions.

This finding matters for architecture. A warehouse is valuable not only because it stores data centrally, but because it exposes meaningful and reusable feature semantics. If a model's most useful signals align with stable warehouse dimensions and facts, then the model can be refreshed and governed more reliably than if it depended on fragile, analyst-specific transformations outside the warehouse.

6 Discussion

6.1 From Fragmented Systems to Decision Infrastructure

The results support a broader claim that healthcare data warehousing should be understood as decision infrastructure. The warehouse resolves structural fragmentation, codifies conformance rules, measures quality, and produces reusable analytical entities. The experimental readmission scenario demonstrates that these architectural decisions have direct consequences for downstream analytics. Without diagnosis grouping, stable medication descriptors, missingness handling, and encounter-level measures, model development would be less interpretable and less reproducible.

This interpretation aligns with recent literature emphasizing that data warehouses remain essential not despite modern interoperability standards, but because standards alone do not solve the full problem of analysis-ready integration [3], [7], [9]. APIs can improve exchange, but warehouses create institutional memory, historical harmonization, and governed access patterns.

6.2 Why the Performance Levels Are Still Useful

The predictive performance reported here is moderate rather than extraordinary, and that is precisely why it is informative. Readmission prediction on routine hospital data is a difficult problem with complex social and clinical determinants. A modest AUROC in a reproducible, warehouse-aligned workflow is more valuable than an apparently superior number produced by an opaque, irreproducible feature pipeline. In healthcare deployment, governance, interpretability, and refreshability often matter as much as marginal gains in headline accuracy [23], [24].

Moreover, the threshold-tuning result shows that warehouse-backed models can be adapted to operational priorities. If the hospital wants a high-recall surveillance queue, it can lower the threshold. If resources are scarce and only the highest-risk patients should be escalated, it can raise the threshold. This reinforces the argument that the warehouse should serve both descriptive and prescriptive analytics.

6.3 Implementation Implications for Hospitals

Several practical implications emerge. First, hospitals should treat data profiling as a mandatory stage in every refresh cycle. Missingness, duplicates, invalid values, and vocabulary drift must be surfaced continuously. Second, conformed dimensions should be designed around recurring business questions rather than around source-system convenience. Third, governance metadata should be captured alongside transformed data so that analysts can trace provenance and quality status. Fourth, predictive use cases should be implemented through warehouse marts rather than bespoke extracts whenever possible, because marts provide stable semantics and access control. Finally, deployment teams should separate ranking quality from decision threshold selection; the latter should be aligned with clinical operations, not assumed by default.

7 Limitations and Threats to Validity

This study has several limitations. First, the experiment uses a single de-identified inpatient dataset rather than a live multi-system hospital deployment. Although the dataset is realistic and sufficiently rich for demonstration, it cannot capture all interface, latency, governance, and terminology challenges present in production environments. Second, the analytical mart is intentionally compact and does not model every possible hospital subject area, such as procedures at finer clinical granularity, imaging workflows, or cost accounting. Third, the predictive evaluation relies on a hold-out split rather than external-site validation. Therefore, the results should be interpreted as proof of architectural and analytical feasibility, not as a universal benchmark for readmission performance.

A further limitation is that the dimensional conformance rules necessarily simplify raw clinical coding. Such simplification is useful for organization-level analytics but may obscure nuances needed for specialized clinical research. Future work should therefore study how semantic layers can support both coarse managerial marts and more granular research views without duplicating governance effort.

8 Conclusion

This paper introduces the contribution that unifies architecture, conformance logic, governance, and experimental analytics. The proposed enterprise data warehouse architecture addresses the structural weaknesses of fragmented HIS and EHR ecosystems by introducing an integration layer, quality gates, conformed dimensions, subject-area marts, and governed analytics access. The experimental validation on a large de-identified inpatient dataset showed that the resulting mart can support meaningful predictive modeling for 30-day readmission risk, with XGBoost providing the best ranking performance and threshold tuning improving operational recall.

The central conclusion is straightforward. A healthcare data warehouse is not merely a repository for historical records. Properly designed, it becomes the semantic and governance backbone of institutional analytics. It enables interoperability beyond message exchange, data quality beyond one-time cleaning, and decision support beyond isolated reports. For hospitals aiming to strengthen analytics maturity, the critical investment is therefore not only in storage technology but in conformance design, metadata discipline, access governance, and reproducible analytical marts.

BIBLIOGRAPHY

[1]. Stylianou, A., & Talias, M. A. Big data in healthcare: a discussion on the big challenges. Health and Technology, 7(1), 97–107, 2017. DOI: 10.1007/s12553-016-0152-4. URL: https://doi.org/10.1007/s12553-016-0152-4

[2]. Shen, Y., Yu, J., Zhou, J., & Hu, G. Twenty-Five Years of Evolution and Hurdles in Electronic Health Records and Interoperability in Medical Research: Comprehensive Review. Journal of Medical Internet Research, 27, e59024, 2025. DOI: 10.2196/59024. URL: https://www.jmir.org/2025/1/e59024/

[3]. Campion, T. R., Jr., Craven, C. K., Dorr, D. A., Bernstam, E. V., & Knosp, B. M. Understanding enterprise data warehouses to support clinical and translational research: impact, sustainability, demand management, and accessibility. Journal of the American Medical Informatics Association, 31(7), 1522–1528, 2024. DOI: 10.1093/jamia/ocae111. URL: https://doi.org/10.1093/jamia/ocae111

[4]. Knosp, B. M., Craven, C. K., Dorr, D. A., Bernstam, E. V., & Campion, T. R., Jr. Understanding enterprise data warehouses to support clinical and translational research: enterprise information technology relationships, data governance, workforce, and cloud computing. Journal of the American Medical Informatics Association, 29(4), 671–676, 2022. DOI: 10.1093/jamia/ocab256. URL: https://doi.org/10.1093/jamia/ocab256

[5]. Wang, Z., Craven, C., Syed, M., Greer, M., Seker, E., Syed, S., & Zozus, M. N. Clinical Data Warehousing: A Scoping Review. Journal of the Society for Clinical Data Management, 4(1), Article 8, 1–19, 2024. DOI: 10.47912/jscdm.320. URL: https://doi.org/10.47912/jscdm.320

[6]. Lyu, S., Craig, S., O'Reilly, G., Taniar, D., et al. The development and use of data warehousing in clinical settings: a scoping review. Frontiers in Digital Health, 7, 1599514, 2025. DOI: 10.3389/fdgth.2025.1599514. URL: https://doi.org/10.3389/fdgth.2025.1599514

[7]. Knezevic Ivanovski, T., Honap, S., Matic, R., Markovic, S., & Peyrin-Biroulet, L. Building a healthcare data warehouse: considerations, opportunities, and challenges. Frontiers in Digital Health, 7, 1691142, 2025. DOI: 10.3389/fdgth.2025.1691142. URL: https://doi.org/10.3389/fdgth.2025.1691142

[8]. Sabooniha, N., Toohey, D. P., & Lee, K. An evaluation of hospital information systems integration

approaches. In Proceedings of the International Conference on Advances in Computing, Communications and Informatics (ICACCI 2012), pp. 498–504, 2012. DOI: 10.1145/2345396.2345479. URL: https://doi.org/10.1145/2345396.2345479

[9]. Tabari, P., Costagliola, G., De Rosa, M., & Boeker, M. State-of-the-Art Fast Healthcare Interoperability Resources (FHIR)–Based Data Model and Structure Implementations: Systematic Scoping Review. JMIR Medical Informatics, 12, e58445, 2024. DOI: 10.2196/58445. URL: https://medinform.jmir.org/2024/1/e58445

[10]. El-Yafouri, R., & Klieb, L. A scoping review of electronic health records interoperability levels, expectations, approaches, and problems. Health Informatics Journal, 31(4), 2025. DOI: 10.1177/14604582251385986. URL: https://doi.org/10.1177/14604582251385986

[11]. Adegoke, K., Adegoke, A., Dawodu, D., Adekoya, A., Bayowa, A., Kayode, T., & Singh, M. Interoperability as a Catalyst for Digital Health and Therapeutics: A Scoping Review of Emerging Technologies and Standards (2015–2025). International Journal of Environmental Research and Public Health, 22(10), 1535, 2025. DOI: 10.3390/ijerph22101535. URL: https://doi.org/10.3390/ijerph22101535

[12]. Zhang, H., Lyu, T., Yin, P., Bost, S., He, X., Guo, Y., Prosperi, M., Hogan, W. R., & Bian, J. A scoping review of semantic integration of health data and information. International Journal of Medical Informatics, 165, 104834, 2022. DOI: 10.1016/j.ijmedinf.2022.104834. URL: https://doi.org/10.1016/j.ijmedinf.2022.104834

[13]. Marfoglia, A., Nardini, F., Arcobelli, V. A., Moscato, S., Mellone, S., & Carbonaro, A. Towards real-world clinical data standardization: A modular FHIR-driven transformation pipeline to enhance semantic interoperability in healthcare. Computers in Biology and Medicine, 187, 109745, 2025. DOI: 10.1016/j.compbiomed.2025.109745. URL: https://doi.org/10.1016/j.compbiomed.2025.109745

[14]. Sahama, T. R., & Croll, P. R. A data warehouse architecture for clinical data warehousing. In J. F. Roddick & J. R. Warren (Eds.), Proceedings of the First Australasian Workshop on Health Knowledge Management and Discovery (HKMD 2007), CRPIT, Vol. 68, pp. 227–232, 2007. Persistent URL: https://dl.acm.org/doi/10.5555/1274531.1274560

[15]. Berndt, D. J., Fisher, J. W., Hevner, A. R., & Studnicki, J. Healthcare data warehousing and quality assurance. Computer, 34(12), 56–65, 2001. DOI: 10.1109/2.970578. URL: https://doi.org/10.1109/2.970578

[16]. Lighterness, A., Adcock, M., Scanlon, L. A., & Price, G. Data Quality–Driven Improvement in Health Care: Systematic Literature Review. Journal of Medical Internet Research, 26, e57615, 2024. DOI: 10.2196/57615. URL: https://www.jmir.org/2024/1/e57615/

[17]. Penev, Y. P., Buchanan, T. R., Ruppert, M. M., Liu, M., Shekouhi, R., Guan, Z., Balch, J., Ozrazgat-Baslanti, T., Shickel, B., Loftus, T. J., & Bihorac, A. Electronic Health Record Data Quality and Performance Assessments: Scoping Review. JMIR Medical Informatics, 12, e58130, 2024. DOI: 10.2196/58130. URL: https://medinform.jmir.org/2024/1/e58130/

[18]. Hosseinzadeh, E., Afkanpour, M., Momeni, M., et al. Data quality assessment in healthcare, dimensions, methods and tools: a systematic review. BMC Medical Informatics and Decision Making, 25, 296, 2025. DOI: 10.1186/s12911-025-03136-y. URL: https://doi.org/10.1186/s12911-025-03136-y

[19]. An, D., Lim, M., Lee, S., et al. Challenges for Data Quality in the Clinical Data Life Cycle: Systematic Review. Journal of Medical Internet Research, 27, e60709, 2025. DOI: 10.2196/60709. URL: https://www.jmir.org/2025/1/e60709/

[20]. Declerck, J., Kiliç, Ö. D., Erol, E. E., et al. Assessing Data Quality in Heterogeneous Health Care Integration: Simulation Study of the AIDAVA Framework. JMIR Medical Informatics, 13, e75275, 2025. DOI: 10.2196/75275. URL: https://medinform.jmir.org/2025/1/e75275/

[21]. Faridoon, A., & Kechadi, M. T. Healthcare Data Governance, Privacy, and Security. In Body Area Networks: Smart IoT and Big Data for Intelligent Health Management, pp. 261–271. Springer, Cham, 2024. DOI: 10.1007/978-3-031-72524-1_19. URL: https://doi.org/10.1007/978-3-031-72524-1_19

[22]. Ahmed, A., Shahzad, A., Naseem, A., Ali, S., & Ahmad, I. Evaluating the effectiveness of data governance frameworks in ensuring security and privacy of healthcare data: A quantitative analysis of ISO standards, GDPR, and HIPAA in blockchain technology. PLOS ONE, 20(5), e0324285, 2025. DOI: 10.1371/journal.pone.0324285. URL: https://doi.org/10.1371/journal.pone.0324285

[23]. Emi-Johnson, O. G., & Nkrumah, K. J. Predicting 30-Day Hospital Readmission in Patients With Diabetes Using Machine Learning on Electronic Health Record Data. Cureus, 17(4), e82437, 2025. DOI: 10.7759/cureus.82437. URL: https://doi.org/10.7759/cureus.82437

[24]. Mishra, V., Tanniru, M. R., & Sreedharan, J. Prediction of 30-day readmission in diabetes management

using machine learning. Computers in Biology and Medicine, 195, 110616, 2025. DOI: 10.1016/j.compbiomed.2025.110616. URL: https://doi.org/10.1016/j.compbiomed.2025.110616

[25]. Chen, E. T. Implementation Issues of Enterprise Data Warehousing and Business Intelligence in the Healthcare Industry. Communications of the IIMA, 12(2), Article 3, 2012. DOI: 10.58729/1941-6687.1186. URL: https://scholarworks.lib.csusb.edu/ciima/vol12/iss2/3/

[26]. Strack, B., DeShazo, J. P., Gennings, C., Olmo, J. L., Ventura, S., Cios, K. J., & Clore, J. N. Impact of HbA1c Measurement on Hospital Readmission Rates: Analysis of 70,000 Clinical Database Patient Records. BioMed Research International, 2014, 781670, 2014. DOI: 10.1155/2014/781670. URL: https://doi.org/10.1155/2014/781670