**ISSN** 2409-2665
**Journal of Logistics, Informatics and Service Science**
Vol. 13 (2026) No. 6, pp. 129-150
**DOI:10.33168/JLISS.2026.0608**

# A Real-Time ICU Data Warehouse Architecture Integrating IoMT and OMOP-CDM for Interoperable Health Information Services

Manirath Wongsim[1], Nipotepat Muangkote[1*], Anupol Panitchote[2], Khaengkai Compapong[3], Chareamchai Ittharat[4], Ekkachai Naenudorn[1], Preecha sriprapakhan[5], Pisit Maneechot[6]

[1]Department of Business Computer, Faculty of Mahasarakham Business School, Mahasarakham University, Mahasarakham, Thailand

[2]Division of Critical Care Medicine, Department of Medicine, Faculty of Medicine, Khon Kaen University, Khon Kaen, Thailand

[3]Khon Kaen Rajanagarindra Psychiatric Hospital, Khon Kaen, Thailand

[4]North Eastern Institute of Child and Adolescent Mental Health, Khon Kaen, Thailand

[5]Applied physics, Mahasarakham University, Mahasarakham, Thailand

[6] School of Renewable Energy and Smart Grid Technology, Naresuan University, Phitsanulok, Thailand

*manirath.w@acc.msu.ac.th, nipotepat.m@msu.ac.th (Corresponding author), panupo@kku.ac.th, khaengkai@dmh.mail.go.th, i.chareamchai.itt@dmh.mail.go.th, ekkachai.n@acc.msu.ac.th, preecha.sr@msu.ac.th, pisitm@nu.ac.th*

**Abstract.** Intensive Care Units (ICUs) generate large volumes of heterogeneous data from electronic medical records, clinical research registries, and high-frequency Internet of Medical Things (IoMT) devices. However, conventional clinical data warehouses are primarily designed for episodic and structured data and are not well suited to integrating continuous physiological time-series with longitudinal clinical records, limiting advanced analytics and real-time clinical intelligence. This study presents the design and evaluation of an OMOP Common Data Model (OMOP-CDM)–based clinical data warehouse architecture for near-real-time integration of IoMT data in intensive care environments. The proposed architecture implements an end-to-end data pipeline that integrates hospital operational systems, a REDCap-based ICU research registry, and high-frequency IoMT device streams into a unified and semantically standardized clinical data warehouse. Data extraction, transformation, and loading processes are orchestrated using Apache Airflow, while semantic harmonization is achieved through international clinical terminologies, including SNOMED CT, LOINC, RxNorm, and ICD-10. Mechanisms supporting data governance—such as auditability, traceability, and controlled data processing—are incorporated within the architecture. The architecture was implemented and evaluated in a tertiary hospital ICU in Thailand, integrating data from 1,088 patients, 1,241 ICU encounters, and over 1.5 million high-frequency IoMT records. Data quality assessment using the OMOP Data Quality Dashboard achieved an overall pass rate of 98%, demonstrating high levels of conformance, completeness, and plausibility. System evaluation indicates stable workflow execution and consistent data ingestion under continuous operation, supporting near–real-time operational capability within the system design. This study demonstrates the feasibility of integrating IoMT-enabled critical care data with standardized clinical and research records within an OMOP-CDM–compliant data warehouse. The proposed architecture provides a scalable foundation for advanced ICU

analytics, predictive modeling, and future AI-driven clinical decision support, supporting interoperable and data-driven critical care systems.

**Keywords:** Clinical Data Warehouse, Internet of Medical Things, Intensive Care Unit OMOP-CDM, Real-Time Data Integration, Healthcare Interoperability

# 1. Introduction

Intensive Care Units (ICUs) represent one of the most data-intensive and clinically complex environments in contemporary healthcare systems, where critically ill patients require continuous physiological monitoring, rapid therapeutic interventions, and coordinated multidisciplinary decision-making under time-sensitive conditions. Consequently, ICUs generate massive volumes of heterogeneous data originating from multiple clinical and operational sources. These data include structured electronic medical records (EMRs), laboratory results, medication administrations, clinical documentation, research registries, and high-frequency physiological signals produced by Internet of Medical Things (IoMT) devices such as bedside monitors, mechanical ventilators, infusion pumps, and continuous renal replacement therapy (CRRT) machines (de Mul et al., 2012; Beunza et al., 2024). Effectively integrating these diverse data streams is critical for improving quality of care, enabling data-driven clinical decision-making, and supporting advanced analytics and artificial intelligence (AI) applications in critical care medicine.

Clinical Data Warehouses (CDWs) have emerged as foundational infrastructures for consolidating clinical and administrative data to support secondary uses, including quality improvement initiatives, operational management, population health surveillance, and clinical research. Prior studies demonstrate that CDWs enhance interoperability, improve data accessibility, and facilitate the development of learning health systems capable of continuously improving care through data feedback loops (Wang et al., 2024; de Mul et al., 2012). However, most existing CDWs were originally designed to manage episodic, low-frequency transactional data—such as admissions, diagnoses, procedures, and billing records—and are therefore poorly suited to ingest, store, and process the continuous, high-velocity data streams generated by IoMT devices in ICU environments (de Mul et al., 2012; Wang et al., 2024). This architectural mismatch limits the ability of conventional CDWs to fully support critical care analytics and real-time clinical intelligence.

The rapid expansion of the Internet of Medical Things has fundamentally transformed critical care practice by enabling continuous physiological monitoring, early detection of patient deterioration, and partial automation of routine clinical documentation. IoMT-enabled ICUs continuously generate granular time-series data at sub-minute or even second-level resolution, creating unprecedented opportunities for precision medicine, advanced predictive modeling, and proactive clinical intervention (Barroca Filho et al., 2021; Beunza et al., 2024). Despite this potential, current IoMT implementations in ICUs are predominantly focused on short-term operational dashboards, alarm systems, and device-specific visualization platforms (Shaik et al., 2024). These systems typically lack long-term data persistence, standardized semantic representation, and seamless integration with enterprise-level CDWs, thereby limiting their reuse for retrospective analysis, multi-center research, and AI model development.

Data standardization remains a major challenge in the integration of heterogeneous ICU data. The Observational Medical Outcomes Partnership Common Data Model (OMOP-CDM) has been widely adopted as an international standard for harmonizing clinical data across institutions and healthcare systems, enabling reproducible observational studies and cross-institutional interoperability (Paris & Parrot, 2020; Kallfelz et al., 2021). Although prior research has demonstrated the feasibility of mapping ICU datasets—such as the MIMIC database—to the OMOP-CDM, these efforts are largely batch-oriented and focused on retrospective data harmonization. They do not adequately address the challenges associated with real-time ingestion, transformation, and management of high-frequency

IoMT data streams within operational ICU environments (Kallfelz et al., 2021). As a result, the systematic integration of continuous IoMT-generated physiological data into OMOP-compliant CDWs remains an open and underexplored research problem.

Beyond technical integration challenges, security, privacy, and data governance concerns are significantly amplified in ICU data ecosystems. High-resolution physiological data are inherently sensitive and may increase the risk of patient re-identification if not managed appropriately. Existing CDW security frameworks often lack mechanisms specifically tailored to IoMT environments, such as device identity management, fine-grained access control, controlled data streaming pipelines, and comprehensive auditability (Thantilage et al., 2023). Addressing these issues is essential to support data governance and regulatory requirements, maintain clinician trust, and enable the sustainable adoption of IoMT-enabled CDWs in clinical practice. A formal evaluation of security mechanisms and governance frameworks is beyond the scope of this study.

Within Asian healthcare systems, including Thailand, digital health initiatives have expanded rapidly in recent years. Previous studies have reported successful CDW implementations for hospital reporting and management (Wongchinda, 2025), as well as the deployment of IoMT-based telemedicine and remote patient monitoring solutions (Islam et al., 2015; Botta et al., 2016; Emaliyawati et al., 2025). However, these developments have largely evolved in parallel, with limited integration between ICU IoMT infrastructures and standardized CDW environments. While several studies have explored the use of OMOP-CDM for ICU data standardization and retrospective database transformation, limited research has examined the integration of high-frequency IoMT physiological data with clinical registries and hospital operational systems within a unified OMOP-compliant warehouse architecture. In the context of Thai tertiary hospitals, the operationalization of near–real-time IoMT data integration within an OMOP-CDM–based clinical data warehouse remains largely unexplored. Therefore, this study focuses on designing and evaluating an integrated architecture that enables standardized ingestion and harmonization of heterogeneous ICU data sources in a real-world hospital environment. In addition to its technical contribution, the proposed ICU-CDW architecture can be interpreted as an informatics-enabled service system that facilitates real-time data services and information logistics in critical care environments. By enabling efficient data flow, integration, and standardization across heterogeneous sources, the architecture supports reliable analytical service delivery and enhances interoperability within healthcare information infrastructures. This perspective is consistent with recent research in informatics and service science, which highlights the importance of integrated data architectures for enabling scalable, data-driven service systems (Wongsim et al., 2025).

To address this gap, this study proposes and evaluates a comprehensive architecture for integrating medical IoT data into a clinical data warehouse for critical care patients. The proposed framework combines real-time data ingestion pipelines, staging and transformation processes, OMOP-CDM standardization, and analytical layers designed to support descriptive analytics, predictive modeling, and future Clinical Decision Support Systems (CDSS). The architecture is implemented and validated in a real-world ICU setting at a tertiary hospital in Thailand, with system performance evaluated in terms of latency, throughput, data completeness, and transformation fidelity. By providing a scalable, standards-based, and traceable blueprint for IoMT–CDW integration, this research advances the field of critical care informatics and establishes a foundation for AI-driven ICU applications and supports future interoperability initiatives.

To the best of our knowledge, limited prior studies have addressed the integrated design and real-world implementation of OMOP-CDM–based architectures for near–real-time IoMT data integration in ICU settings. The contributions of this study can be summarized as follows:

1. Architectural integration for ICU data warehousing

This study proposes an end-to-end architecture that integrates heterogeneous ICU data sources—including hospital information systems, REDCap research registries, and high-frequency IoMT device streams—into a unified OMOP-CDM–based clinical data warehouse.

2. OMOP-compliant near–real-time data standardization

The study develops and implements a multi-layer ETL pipeline that supports near–real-time ingestion and semantic harmonization of heterogeneous ICU data using international clinical terminologies such as SNOMED CT, LOINC, RxNorm, and ICD-10.

3. Real-world system validation in a tertiary ICU environment

The proposed architecture is implemented and evaluated in a real ICU setting, demonstrating successful data integration, high data quality, and stable operational performance for large-scale clinical analytics.

The remainder of this paper is organized as follows. Section 2 reviews the related literature. Section 3 presents the proposed CDW architecture in detail. Section 4 describes the experimental setup and reports the results, analysis, and discussion. Finally, Section 5 concludes the paper and outlines directions for future work.

# 2. Related work

## 2.1. Clinical data warehousing in healthcare and critical care

Clinical data warehouses (CDWs) are essential infrastructures for supporting secondary use of healthcare data, including analytics, quality improvement, and clinical research. By integrating heterogeneous data from Electronic Health Records (EHRs), laboratory systems, and administrative databases, CDWs improve data accessibility and enable longitudinal analysis, thereby supporting learning health systems (Wang et al., 2024; de Mul et al., 2012). Early CDW architectures primarily relied on batch-oriented ETL processes designed for structured, episodic data, which limits their applicability in data-intensive environments such as intensive care units (ICUs) (de Mul et al., 2012; Wang et al., 2024). Although ICU-focused CDWs support outcome and resource utilization analyses, most aggregate physiological data at coarse temporal resolutions, constraining real-time analytics and predictive modeling.

## 2.2. Standardization and common data models for ICU data

Data standardization is essential for achieving interoperability, reproducibility, and multi-center research collaboration (Wongsim et al., 2025). The observational medical outcomes partnership common data model (OMOP-CDM), developed under the observational health data sciences and informatics (OHDSI) initiative, has become one of the most widely adopted frameworks for harmonizing clinical data across institutions and countries. OMOP-CDM provides a standardized schema and controlled vocabularies that support large-scale observational research and federated analytics (Paris & Parrot, 2020).

Several studies have demonstrated the feasibility of mapping ICU datasets into OMOP-CDM. Notably, the MIMIC and MIMIC-IV databases have been transformed into OMOP-compliant formats, enabling cross-database analytics and methodological reproducibility (Paris & Parrot, 2020; Kallfelz et al., 2021). These efforts highlight the potential of OMOP-CDM for critical care research; however, they are largely based on retrospective, batch-processed datasets. Continuous physiological signals are typically pre-aggregated or partially derived into summary features before loading into the CDM, leaving raw or near-real-time IoT data outside the CDW ecosystem. Consequently, while OMOP-CDM provides a strong foundation for semantic interoperability, its operationalization for real-time ICU IoMT data ingestion and long-term warehousing remains an open research challenge

## 2.3. Internet of Medical Things (IoMT) in intensive care units

The Internet of Medical Things (IoMT) has transformed ICU environments by enabling continuous patient monitoring, automated data capture, and real-time clinical alerts. IoMT devices generate high-resolution physiological data streams that support early detection of patient deterioration and precision critical care (Barroca Filho et al., 2021; Beunza et al., 2024). Recent ICU platforms have demonstrated

the feasibility of IoMT-based visualization and alerting, including cloud-based monitoring systems and ICU command center architectures (Barroca Filho et al., 2021; Shaik et al., 2024). However, these solutions primarily address short-term operational needs and lack standardized, long-term data warehousing capabilities. Fog and edge computing improve real-time responsiveness but still require centralized clinical data warehouses to support longitudinal analytics, research, and AI model development.

### 2.4. Security, privacy, and governance in IoMT-enabled CDWs

Security and privacy are critical concerns in CDW architectures due to the sensitivity and longitudinal nature of clinical data. Prior research emphasizes the need for end-to-end security mechanisms encompassing data ingestion, storage, analytics, and visualization layers (Thantilage et al., 2023). Key requirements include encryption, access control, audit logging, and pseudonymization. High-frequency IoMT data introduce additional governance challenges. Physiological time-series data may increase the risk of patient re-identification, even when traditional anonymization techniques are applied. Thantilage et al. (2023) argue that CDWs integrating IoMT sources must incorporate device identity management, secure communication protocols, and fine-grained authorization models tailored to streaming data environments. Furthermore, regulatory and ethical considerations vary across regions. In Asian healthcare systems, where infrastructure maturity and regulatory frameworks differ, security-by-design approaches are particularly important to ensure trust and sustainability of IoMT-enabled CDWs (Emaliyawati et al., 2025).

### 2.5. IoMT, CDWs, and healthcare systems in Thailand and Asia

In recent years, Asian healthcare systems-including Thailand-have made significant progress in digital health adoption. Studies have reported successful implementations of cloud-based telemedicine platforms integrating IoMT device networks to support remote monitoring and consultations (Islam et al., 2015; Botta et al., 2016). Additionally, early CDW implementations in Thai hospitals demonstrate growing institutional capacity for data integration and analytics (Wongchinda, 2025). Despite these advances, existing efforts in Thailand have largely addressed CDWs and IoMT as separate initiatives. There is limited published research on their integration within ICU settings, particularly using international standards such as OMOP-CDM. This fragmentation constrains the potential for advanced analytics, AI-driven decision support, and national-level interoperability in critical care.

## 3. CDW architecture

Figure 1 illustrates the overall architecture of the proposed ICU-CDW system, comprising five integrated layers: data sources, loader, staging area, data warehouse, and analytics.

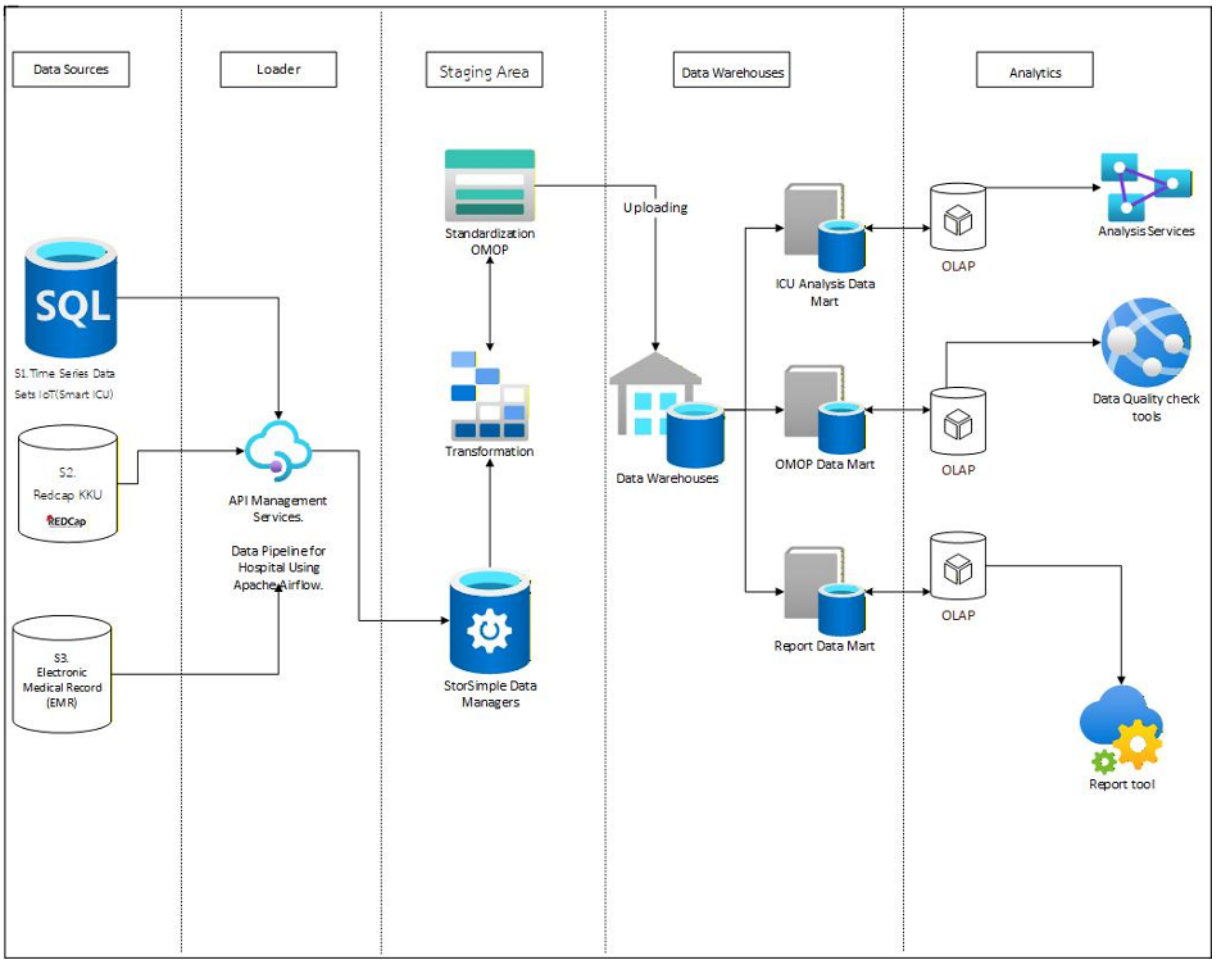


Fig. 1: Integrated ICU clinical data warehouse architecture combining IoMT streams, hospital systems, and OMOP-CDM standardization.

As illustrated in Fig. 1, the architecture delineates a multi-layered data pipeline in which heterogeneous clinical data sources are ingested through the loader layer, transformed and standardized into the OMOP common data model within the staging area, and subsequently stored in structured data marts in the data warehouse layer. The analytics layer further enables OLAP operations, data quality assessment, and advanced analytical services. The proposed system architecture comprises five integrated layers—data sources, loader, staging area, data warehouse, and analytics—forming an end-to-end pipeline for collecting, standardizing, and analyzing ICU and hospital data within an OMOP-compliant framework. The ICU-CDW integrates heterogeneous data from three primary sources: high-frequency IoMT time-series data from Smart ICU devices, research-oriented clinical observations collected via REDCap, and longitudinal EMR data. Each source contributes complementary clinical, operational, and research perspectives on patient care. Their integration enables comprehensive data harmonization and supports scalable interoperability, advanced analytics, machine learning applications, and the future development of clinical decision support systems in critical care environments.

### 3.1. Data Sources

This study integrates heterogeneous clinical and operational data from three primary source systems to construct a comprehensive Intensive Care Unit Clinical Data Warehouse (ICU-CDW). Each data source contributes distinct yet complementary perspectives on patient care, enabling both real-time monitoring and longitudinal clinical analysis. The integration of these sources is designed to support standardized data harmonization, advanced analytics, and the development of clinical decision support applications.

S1: Time-Series Internet of Medical Things (IoMT) Data (Smart ICU System)

The first data source consists of high-frequency time-series data generated by IoMT-enabled medical devices deployed in the ICU, collectively referred to as the Smart ICU system. These data are stored in a SQL-based operational database and include continuous physiological measurements and device operation logs captured from bedside multiparameter monitors, infusion pumps, ventilators, and other critical care equipment. Typical data elements include heart rate, blood pressure, oxygen saturation ($SpO_2$), respiratory parameters, infusion rates, and device status indicators, with sampling intervals ranging from seconds to minutes. This data source provides a granular, real-time view of patient physiological status and device utilization, forming the foundation for time-sensitive analytics, early warning systems, and predictive modeling.

S2: REDCap KKU Clinical Research and Registry Data

The second data source comprises clinical research and registry data maintained in REDCap (Research Electronic Data Capture) databases at Khon Kaen University. These datasets contain manually curated and research-specific information recorded by clinicians and researchers, including detailed treatment interventions, clinical assessments, protocol-driven observations, and outcome measures not routinely captured in hospital operational systems. REDCap data are structured according to predefined Case Record Forms (CRFs) and are collected under approved ethical protocols, ensuring high data validity and traceability. Integrating REDCap data into the ICU-CDW enriches routine clinical records with research-grade variables, enabling more comprehensive observational studies and hypothesis-driven analyses.

S3: Electronic Medical Record (EMR) Systems

The third data source includes structured clinical and administrative data extracted from the hospital's Electronic Medical Record (EMR) systems. These systems store core patient information such as demographics, admission and discharge details, diagnoses, procedures, medication orders and administrations, laboratory results, and billing records. EMR data provide the longitudinal clinical context necessary for reconstructing patient care trajectories, evaluating treatment outcomes, and performing population-level analyses. Within this study, EMR data serve as the backbone for mapping patient encounters and clinical events into standardized domains of the Observational Medical Outcomes Partnership Common Data Model (OMOP-CDM).

Together, these three data sources form a multi-dimensional representation of ICU care. The integration of high-frequency IoMT data (S1), research-oriented clinical observations (S2), and standardized operational records (S3) enables comprehensive analysis across temporal scales—from real-time physiological monitoring to retrospective outcome evaluation. This multi-source design supports the study's research objectives of developing a scalable, interoperable ICU data warehouse capable of powering advanced analytics, machine learning models, and future clinical decision support systems.

## 3.2. Loader layer and ETL pipeline

To ensure reliable, scalable, and auditable integration of heterogeneous ICU data sources, this study implements a dedicated Loader Layer that orchestrates the Extract–Transform–Load (ETL) pipeline for the ICU-CDW. This layer functions as the central coordination mechanism between operationals ource systems and downstream analytical infrastructures, ensuring that data ingestion processes are standardized, reproducible, and compliant with governance requirements.

The Loader Layer is designed to accommodate both batch-oriented clinical records and near–real-time IoMT data streams, addressing the distinct characteristics of each data modality. It encapsulates data extraction services, workflow orchestration, intermediate storage, and metadata management, thereby decoupling source systems from the core warehouse architecture and enhancing system robustness.

### *3.2.1. Data extraction mechanisms*

Data extraction from source systems is implemented using a combination of API-based access, scheduled queries, and automated data pipelines, depending on the nature of each data source.

For S1 (IoMT smart ICU data), physiological and device-level data are extracted through secure APIs and streaming connectors that retrieve time-stamped numerical measurements and device logs from the SQL-based Smart ICU database. These extraction processes are designed to support high-frequency data ingestion while preserving temporal integrity and minimizing latency.

For S2 (REDCap KKU data), clinical research and registry data are extracted using REDCap's secure RESTful API. This approach enables controlled access to research datasets in accordance with approved ethical protocols and supports incremental data retrieval based on study identifiers, timestamps, or admission records.

For S3 (EMR systems), structured clinical and administrative data are extracted via database connectors or service interfaces provided by the hospital information system. These extractions include patient demographics, encounters, diagnoses, procedures, medication orders, and laboratory results, forming the longitudinal clinical backbone of the ICU-CDW.

#### *3.2.2. Workflow orchestration and automation*

All extraction and loading tasks are orchestrated using Apache Airflow, which manages ETL workflows as directed acyclic graphs (DAGs). Airflow enables automated scheduling, dependency management, error handling, and logging across complex multi-stage pipelines. Each DAG represents a logically independent workflow-such as EMR extraction, REDCap synchronization, or IoMT ingestion-allowing fine-grained control over execution frequency and resource allocation.

The use of workflow orchestration ensures that ETL processes are repeatable and auditable, supporting research reproducibility and operational transparency. Failed tasks can be retried or isolated without disrupting the entire pipeline, which is particularly important in ICU environments where data availability and continuity are critical.

#### *3.2.3. Intermediate storage and data landing zone*

Extracted raw data are initially stored in an intermediate landing zone prior to transformation. This staging storage serves multiple purposes: buffering high-volume IoMT data, preserving raw data for traceability, and enabling quality checks before standardization. Temporary storage solutions, such as scalable file systems or managed storage services, are employed to support varying data velocities and volumes.

This design allows the system to handle bursty data patterns commonly observed in ICU IoMT streams while maintaining system stability and data integrity.

#### *3.2.4. Loader Layer, Workflow Orchestration, and Monitoring*

The Loader Layer plays a central role in ensuring data quality, governance, and methodological rigor within the proposed ICU-CDW architecture. By separating data extraction from transformation and analytical processes, the architecture enables independent validation of data ingestion and supports systematic evaluation of performance characteristics such as latency, throughput, and data completeness. Furthermore, this layer provides the foundation for standardized downstream processing, including consistent mapping to the Observational Medical Outcomes Partnership Common Data Model (OMOP-CDM), thereby supporting advanced analytics, machine learning applications, and Clinical Decision Support System (CDSS) development.

To operationalize and manage ETL processes within the Loader Layer, this study employs Apache Airflow as the primary workflow orchestration and monitoring platform. Airflow enables the definition, scheduling, execution, and monitoring of ETL workflows through Directed Acyclic Graphs (DAGs), providing a transparent and reproducible framework for managing complex ICU data integration pipelines. Each DAG consists of multiple interdependent tasks representing key stages of the ETL

process, including data extraction, validation, transformation, and loading into downstream storage layers.

The ETL pipeline is operationalized using scheduled workflow execution, in which data ingestion from source systems is triggered at regular intervals to support timely data availability. The workflow incorporates task-level retry mechanisms and logging to ensure robustness in the presence of transient failures. Data extraction and transformation processes are executed in batch-oriented stages within each workflow cycle, and task dependencies are managed through Directed Acyclic Graphs (DAGs). While specific parameter settings may vary depending on deployment conditions, the architecture is designed to support configurable scheduling, retry policies, and scalable data processing to accommodate varying workload characteristics. Detailed system performance observations are discussed in Section 4.5.

The Airflow monitoring interface provides real-time visibility into workflow execution through status indicators (e.g., success, failure, retry), facilitating efficient troubleshooting and system oversight. Workflow scheduling ensures temporal consistency in data ingestion, which is essential for aligning high-frequency IoMT data with episodic EMR and REDCap records. In addition, built-in logging and execution metadata enhance auditability and traceability, enabling verification of data provenance and reproducibility of experimental conditions.

From a system reliability perspective, Airflow supports robust pipeline execution through task-level failure handling and automatic retry mechanisms, ensuring continuity of data processing in dynamic ICU environments. The dashboard further provides aggregated views of DAG execution history and task outcomes, allowing continuous monitoring of pipeline health and early detection of anomalies or system bottlenecks.

This integration of the Loader Layer with Airflow-based orchestration and monitoring establishes a scalable, reliable, and reproducible ETL framework for supporting near–real-time clinical operations and research-oriented data analysis in ICU settings.

### *3.2.5. Role of airflow in supporting performance evaluation*

This section highlights the role of Apache Airflow in enhancing the reliability, transparency, and reproducibility of the proposed ICU Clinical Data Warehouse (ICU-CDW). By orchestrating ETL workflows through Directed Acyclic Graphs (DAGs), Airflow enables automated scheduling, centralized monitoring, and structured management of data integration across heterogeneous sources, including hospital systems, REDCap research databases, and IoMT streams. The monitoring dashboard and execution logs support continuous oversight, early detection of pipeline failures, and systematic performance evaluation in terms of latency, throughput, and stability. Moreover, the explicit representation of ETL logic as DAGs improves methodological transparency and facilitates replication across environments and institutions. The technical workflow for integrating REDCap-based clinical research data using this orchestration framework is illustrated in Fig. 2. Overall, Apache Airflow serves as a robust orchestration backbone that supports scalable ICU data integration while meeting the requirements of reliable clinical operation and reproducible research.

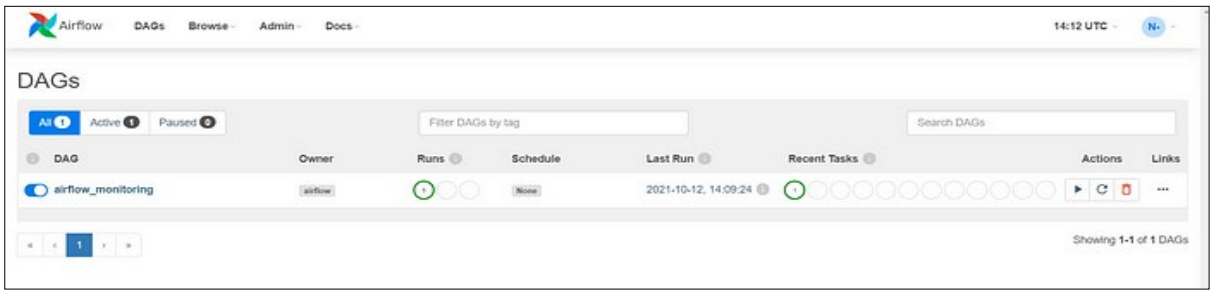


Fig. 2: REDCap data integration pipeline illustrating ETL processes for incorporating research registry data into the ICU-CDW.

This Fig. 2. illustrates the technical workflow for integrating REDCap clinical research data into the Intensive Care Unit Clinical Data Warehouse (ICU-CDW) using an automated ETL pipeline orchestrated by Apache Airflow, as shown in Fig. 3. The integration process is designed to support secure, periodic, and reproducible ingestion of research-grade clinical data while maintaining alignment with ICU operational and IoMT datasets.

On the left side of the figure, REDCap data are stored within a relational database environment, where research-specific tables—such as *redcap* and *redcap_admission*—contain structured clinical observations collected by clinicians and researchers. These datasets include protocol-driven variables, treatment interventions, and outcome-related information that are not routinely captured in the hospital operational system. REDCap records are typically updated on a daily basis, with new or modified entries reflecting ongoing clinical documentation and research activities.

The right side of the figure presents the Apache Airflow DAG (*02_redcap*) responsible for orchestrating the REDCap ETL workflow (Fig. 3). The DAG is executed according to a predefined schedule (e.g., hourly or daily), ensuring timely synchronization between the REDCap database and the ICU-CDW. The workflow consists of two primary tasks:

**get_data:** This task retrieves new or updated REDCap records from the source database using secure database connections or RESTful API calls. Data extraction is incremental, relying on timestamps or record identifiers to minimize redundancy and reduce system load. Extracted data are temporarily staged for validation and transformation.

**post_data:** Following extraction, the transformed REDCap data are loaded into intermediate or target tables within the ICU-CDW environment. During this step, records are linked to corresponding ICU admissions using unique patient identifiers (e.g., hospital number and admission ID), enabling seamless integration with EMR and IoMT datasets.

The use of Apache Airflow provides several methodological advantages for research-oriented data integration. First, DAG-based orchestration ensures process transparency and traceability, allowing each data ingestion step to be monitored, logged, and audited. Second, built-in task status indicators (e.g., running, success, failure) enable rapid detection of pipeline disruptions and support operational reliability. Third, the modular design of the DAG facilitates extensibility, allowing additional transformation or validation tasks—such as OMOP-CDM mapping or data quality checks—to be incorporated without redesigning the entire pipeline.

From a research perspective, this REDCap integration pipeline, as illustrated in Fig. 3, enriches the ICU-CDW by incorporating high-quality, manually curated clinical variables that complement routine EMR data and high-frequency IoMT streams. The resulting integrated dataset supports more comprehensive observational analyses, improves the validity of predictive modeling, and strengthens the foundation for Clinical Decision Support System (CDSS) development.

This workflow demonstrates a robust and scalable approach to integrating research registry data into a standardized ICU data warehouse, supporting consistency, reproducibility, and analytical readiness across heterogeneous clinical data sources.

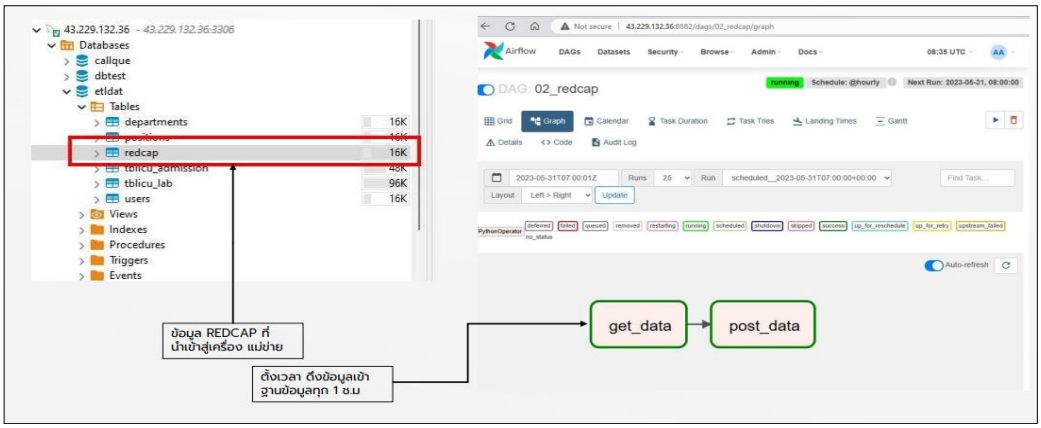


Fig. 3: Apache Airflow DAG illustrating workflow orchestration for REDCap data extraction, transformation, and loading.

### 3.3. Staging area and data transformation

The Staging area functions as an intermediate processing layer between the loader layer and the core CDW. Its primary role is to prepare raw data extracted from heterogeneous ICU data sources for standardized storage, integration, and downstream analytics. Given the diversity of data modalities-including high-frequency IoMT time-series data, structured EMR records, and research-oriented REDCap datasets-the staging process is essential for ensuring data quality, semantic consistency, and interoperability.

Within the Staging Area, all incoming data undergo a series of transformation procedures, including data cleansing, normalization, deduplication, and temporal alignment. Data cleansing addresses missing values, out-of-range measurements, and inconsistent coding practices across source systems. Temporal normalization ensures that timestamps from different systems are synchronized to a common reference, enabling accurate reconstruction of patient care timelines and alignment between continuous IoMT streams and episodic clinical events. During preprocessing, duplicate records originating from multiple data sources were identified and removed, and inconsistent timestamps from high-frequency IoMT devices were aligned to a unified temporal reference. Missing or incomplete values were handled through standardized preprocessing rules, and source-specific codes were normalized prior to mapping to OMOP standardized vocabularies. These preprocessing steps enhance data consistency and support reliable downstream transformation into standardized formats.

A key function of this layer is semantic standardization using the OMOP-CDM. Local data schemas and proprietary codes are mapped to standardized OMOP tables and vocabularies, including SNOMED CT for diagnoses and clinical findings, LOINC for laboratory and physiological measurements, RxNorm and ATC for medications, and standardized device concepts for IoMT equipment. This mapping process transforms heterogeneous source data into a unified, analysis-ready structure while preserving clinical meaning.

For high-frequency IoMT data, the staging process includes aggregation and feature preparation steps to balance analytical utility and storage efficiency. While raw waveform data are retained in a separate lakehouse environment for advanced signal analysis, numerical summaries and derived features-such as minute-level vital signs or device operation metrics-are transformed and prepared for loading into OMOP-CDM-compliant tables. This hybrid approach enables scalable data management while maintaining flexibility for future advanced analytics.

This process ensures that only validated, standardized, and high-quality data are propagated to the downstream warehouse, establishing a reliable foundation for clinical research, machine learning, and decision support applications.

### 3.4. Data Warehouse and Data Mart Architecture

Following transformation and standardization, processed data are loaded into the enterprise Clinical Data Warehouse, which serves as the central repository for longitudinal ICU patient data. The warehouse is designed according to OMOP-CDM specifications, enabling consistent representation of patient demographics, encounters, diagnoses, procedures, medications, observations, measurements, and device-related data across all integrated sources.

To support diverse analytical workloads and optimize performance, the architecture employs a multi–data mart strategy derived from the central warehouse. Each data mart is tailored to specific use cases while maintaining governance and consistency through the shared OMOP backbone.

The ICU Analysis Data Mart is optimized for detailed clinical and operational analysis within the ICU context. It supports queries related to patient flow, treatment patterns, resource utilization, and clinical outcomes, enabling both descriptive and exploratory analytics.

The OMOP Research Data Mart preserves the standardized OMOP-CDM structure to facilitate reproducible observational studies and interoperability with external research networks. This mart enables compatibility with established OHDSI analytical tools and supports multicenter and longitudinal research.

The Reporting Data Mart is denormalized and performance-tuned for rapid query execution and visualization. It supports dashboards, routine reports, and operational monitoring interfaces used by clinicians, administrators, and researchers.

This layered warehouse and data mart design balances scalability, performance, and governance. By separating analytical workloads while maintaining a single source of truth, the architecture supports both real-time operational needs and rigorous research applications in critical care.

## 3.5. Analytics Layer and Future Analytical Applications

The analytics layer represents the final component of the ICU Clinical Data Warehouse architecture, where standardized and integrated data are made available for analytical services and future clinical intelligence applications. In the current implementation, the analytics layer primarily supports descriptive and exploratory analysis through structured queries, reporting tools, and exploratory data analysis on the integrated OMOP-CDM dataset.

Using the standardized data marts derived from the ICU-CDW, clinicians and researchers can perform multidimensional analysis of ICU patient trajectories, including patterns of diagnoses, procedures, medication administration, laboratory results, and device-generated physiological measurements. These capabilities support retrospective clinical analysis, operational monitoring, and hypothesis generation for future research.

The analytics layer is designed as a future-ready infrastructure that supports descriptive analytics, dashboard development, predictive modeling, and clinical decision support applications. While the current study focuses on data integration and standardization, the proposed architecture provides a scalable and interoperable foundation for developing advanced analytics and AI-driven applications in future work. Although predictive models are not implemented within the scope of this study, the integrated dataset provides the necessary foundation for future machine learning applications, such as early warning systems, mortality risk prediction, and treatment optimization in ICU environments.

In summary, the analytics layer provides a standardized and interoperable foundation for future advanced analytics and AI-driven clinical applications.

To provide a structured evaluation of the proposed ICU-CDW architecture, the assessment framework is organized around four dimensions:

1. Integration success – the ability to integrate heterogeneous ICU data sources
2. OMOP-CDM compliance – structural and semantic conformity with the OMOP standard
3. Data quality – completeness, plausibility, and conformance assessed using the OHDSI Data Quality Dashboard
4. System performance – operational reliability and ingestion capability of the ETL pipeline

# 4. Experimental result and discussion

This section presents an experimental evaluation to verify the feasibility and correctness of standardizing and integrating heterogeneous ICU data into an OMOP-CDM–based Clinical Data Warehouse, rather than benchmarking device performance or specific algorithms. The evaluation focuses on the effectiveness of the ETL pipeline, compliance with OMOP-CDM structural and semantic standards, and overall data quality assessed using the OHDSI Data Quality Dashboard, thereby demonstrating the methodological soundness of the proposed CDW–OMOP architecture for integrated clinical and IoMT data analytics. The details are described as follows:

## 4.1. Data Acquisition

This study follows a system design-and-evaluation approach focusing on the implementation and validation of a clinical data warehouse architecture rather than a traditional clinical observational study design. Retrospective de-identified ICU data obtained from the Srinagarin Hospital ICU Registry were used to validate the proposed architecture and evaluate the effectiveness of data integration and standardization processes. The dataset comprised ICU admission records collected between April 8, 2023 and May 20, 2023. The evaluation focuses on verifying structural integration, OMOP-CDM compliance, and data quality of the integrated dataset rather than analyzing clinical outcomes. To ensure patient anonymity and confidentiality, all personal identifiers were removed prior to analysis. This de-identification process is consistent with established practices in similar clinical studies and supports the ethical use of sensitive medical data (Muangkote et al., 2025).

It is important to distinguish between the full integrated ICU-CDW dataset and the subset used for evaluation in this section. The overall architecture integrates data from 1,088 patients, 1,241 ICU encounters, and more than 1.5 million IoMT records, representing the complete data warehouse. In contrast, the dataset described in this section corresponds to a time-limited subset derived from the ICU registry (372 patients and 702 records), which is used for focused system validation and performance evaluation. This distinction clarifies the scope of each dataset and ensures transparency in the evaluation process.

## 4.2. ER-Diagram validation and structural integration

This section presents and discusses the experimental results obtained from implementing the proposed ICU-CDW architecture. The analysis focuses on data integration outcomes, structural validity of the OMOP-CDM–based schema, data quality, and the implications for advanced clinical analytics and decision support. The ER diagram shown in Fig. 4 illustrates the final integrated data model derived from heterogeneous ICU data sources.

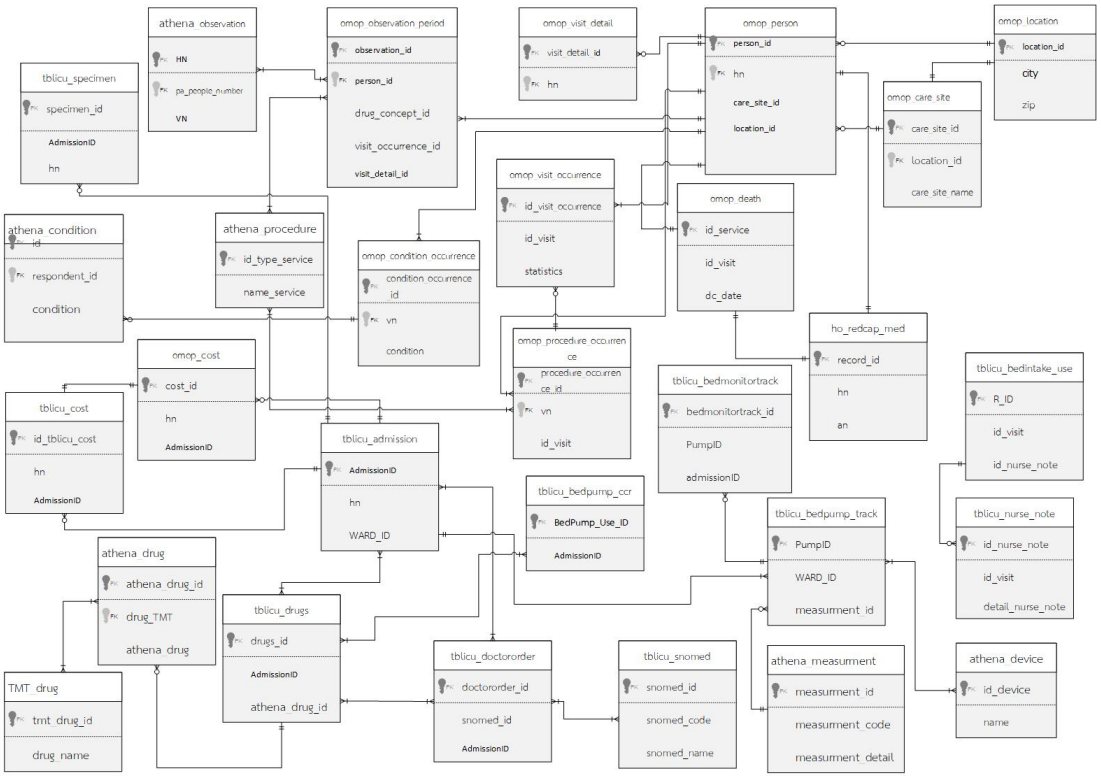


Fig. 4: ER diagram of the OMOP-CDM–based ICU clinical data warehouse showing key entities and relationships.

The experimental implementation resulted in a fully operational ER-Diagram that harmonizes data from hospital operational systems, REDCap clinical research databases, and IoMT-based ICU monitoring systems into a unified OMOP-CDM-compliant structure. As shown in Fig. 4, the *omop_person* table functions as the central anchor entity, linking patient demographics with clinical encounters, locations, care sites, and downstream clinical events.

Key OMOP domains-including *visit_occurrence, visit_detail, condition_occurrence, Procedure _occurrence, drug_exposure, measurement, observation,* and *death*-are structurally connected through consistent use of person identifiers and visit identifiers. This design enables longitudinal reconstruction of ICU patient trajectories from admission through treatment, monitoring, and outcomes. The experimental results confirm that the ER structure supports referential integrity across domains, with no orphan records observed during validation.

## 4.3. Integration of IoMT data with clinical and research Records

A major contribution of the proposed architecture is the successful integration of high-frequency IoMT data with traditional clinical and research datasets, as illustrated in Fig. 5 (Integration of IoMT Data with Clinical Records). IoMT-derived measurements—captured through bedside monitors and infusion pumps—were mapped to standardized *athena_measurement* and *athena_device* vocabularies before being linked to OMOP *measurement* records. Tables such as *tblicu_bedmonitortrack*, *tblicu_bedpump_track*, and *tblicu_bedpump_ccr* provide detailed temporal representations of physiological signals and device utilization, which are systematically associated with ICU admissions and visits.

The experimental results demonstrate that continuous physiological monitoring data can be coherently aligned with episodic EMR events and REDCap research observations. This alignment enables synchronized analysis across temporal scales, allowing researchers to examine how real-time physiological changes relate to diagnoses, interventions, medication administration, and clinical outcomes, as conceptually depicted in Fig. 5.

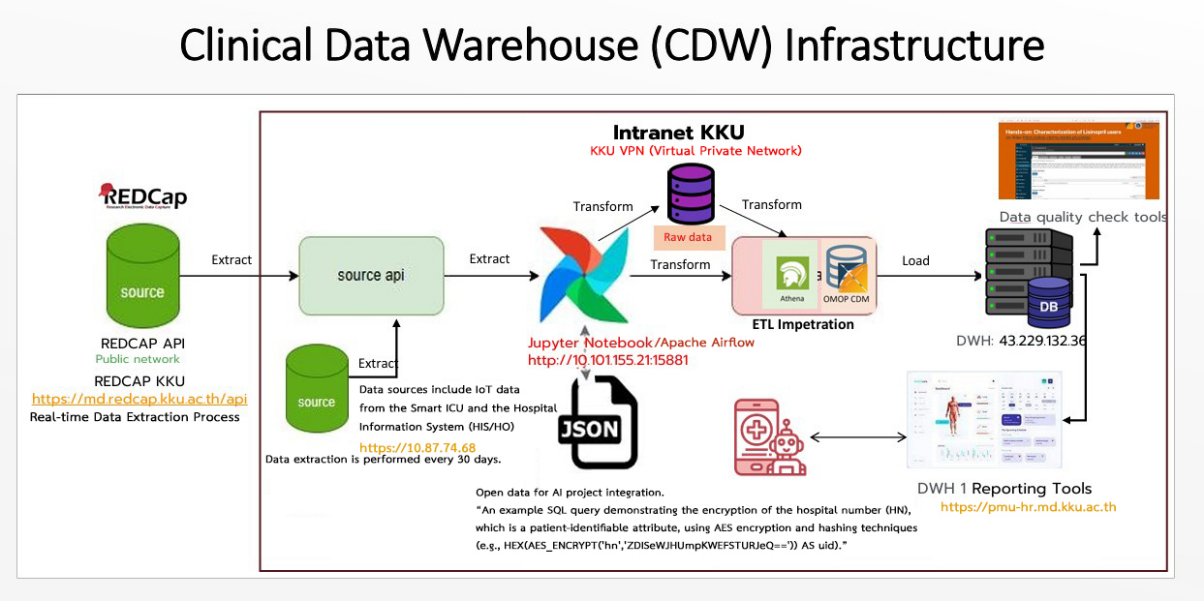


Fig. 5: Framework for integrating IoMT data with clinical and research records in the ICU-CDW.

Figure 5 illustrates the end-to-end data ingestion, transformation, and analytics architecture developed for integrating heterogeneous ICU data sources into the CDW at Khon Kaen University (KKU). The architecture operates within the KKU intranet environment, ensuring data security, controlled access, and compliance with institutional and regulatory requirements.

## 4.4. Data quality assessment results

The quality of the integrated ICU dataset was evaluated using the Data Quality Dashboard (DQD) developed by the Observational Health Data Sciences and Informatics (OHDSI) community. The DQD applies a standardized set of automated rules to OMOP-CDM–compliant databases, covering three key dimensions: conformance, completeness, and plausibility.

A total of 4,153 data quality checks were executed across both verification and validation phases to assess structural correctness, semantic consistency, and data coverage. Overall, 4,069 checks passed and 84 checks failed, resulting in a pass rate of 98%. A substantial portion of checks (2,752) were classified as Not Applicable due to unused tables or fields not relevant to the ICU context. When considering only applicable checks, the corrected pass rate was 94%.

The detailed results are summarized in Table 1. Across all dimensions, conformance and plausibility achieved consistently high pass rates (⩾99%), indicating strong adherence to OMOP-CDM structural and semantic requirements. Completeness showed slightly lower performance, particularly in the validation phase (75%), reflecting the presence of missing or partially populated fields in some data elements.

These findings suggest that the ETL pipeline effectively ensures structural integrity and semantic standardization, while highlighting areas for improvement in data completeness. Overall, the high pass rates demonstrate the robustness of the data integration process and support the reliability of the dataset for downstream clinical analytics and research applications.

Table 1. Data quality assessment results of the integrated ICU dataset using the OHDSI data quality Dashboard

| | Verification | | | | Validation | | | | Total | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Pass | Fail | Total | % Pass | Pass | Fail | Total | % Pass | Pass | Fail | Total | % Pass |
| Plausibility | 2179 | 36 | 2215 | 98% | 287 | 0 | 287 | 100% | 2466 | 36 | 2502 | 99% |
| Conformance | 996 | 11 | 1007 | 99% | 180 | 0 | 180 | 100% | 1176 | 11 | 1187 | 99% |
| Completeness | 415 | 33 | 448 | 93% | 12 | 4 | 16 | 75% | 427 | 37 | 464 | 92% |
| Total | 3590 | 80 | 3670 | 98% | 479 | 4 | 483 | 99% | 4069 | 84 | 4153 | **98%** |

The high data quality underscores the effectiveness of the ETL pipeline and establishes a solid foundation for advanced clinical analytics and AI-driven innovations in critical care.

## 4.5. System Performance Evaluation

To evaluate the operational performance of the proposed ICU-CDW architecture, system-level characteristics were examined, including data ingestion behavior, workflow reliability, and data processing continuity. The ETL pipeline orchestrated through Apache Airflow demonstrated stable performance under continuous operation. Data ingestion from IoMT streams was observed to occur in a timely manner within the scheduled workflow cycles, supporting near–real-time operational capability of the system. Physiological monitoring data were made available in the warehouse shortly after generation, enabling their use in downstream analytical processes. Workflow reliability was assessed using Apache Airflow DAG execution logs, which showed consistent task completion and automatic retry handling in the event of transient failures. During the evaluation period, the majority of ETL tasks completed successfully without manual intervention, indicating robust workflow orchestration and system stability. These observations indicate that the proposed architecture is capable of supporting continuous integration of high-frequency ICU data while maintaining stable and reliable operation under real-world conditions. However, a detailed quantitative performance evaluation—such as precise latency measurement, throughput benchmarking, and stress testing—is beyond the scope of this study and will be addressed in future work.

Table 2 presents the mapping of source clinical data to OMOP standardized vocabularies across major clinical domains. Local clinical codes and ICU device data were mapped to standardized vocabularies supported by the OMOP Common Data Model. Diagnoses were mapped to SNOMED CT concepts using ICD-10 and ICD-10-TM source codes. Laboratory measurements and physiological observations were harmonized with LOINC concepts, while medication data were standardized using RxNorm terminology. Device-related data generated from ICU monitoring systems were mapped to OMOP device domain concepts.

Table 2. Mapping of source clinical data to OMOP standardized vocabularies

| **Source System** | **Source Concept** | **Standard Vocabulary** | **Mapping Rate** |
|---|---|---|---|
| ICU diagnosis | ICD10-TM | SNOMED CT | 94% |
| Lab tests | local lab code | LOINC | 91% |
| Medication | hospital drug code | RxNorm | 89% |
| Device data | ICU device code | OMOP device | 87% |

Local clinical codes and ICU device data were mapped to standardized vocabularies supported by the OMOP Common Data Model. Diagnoses were mapped to SNOMED CT concepts using ICD-10 and ICD-10-TM source codes. Laboratory measurements and physiological observations were harmonized with LOINC concepts, while medication data were standardized using RxNorm terminology. Device-related data generated from ICU monitoring systems were mapped to OMOP device domain concepts.

Mapping procedures combined automated vocabulary lookup methods with manual validation by

domain experts when direct concept matches were unavailable. As shown in Table 3, mapping coverage exceeded 90% for major clinical domains, ensuring semantic interoperability while preserving the original clinical meaning of source data. The remaining unmapped concepts (approximately 6–13%) were primarily associated with device-specific or locally defined codes that lacked direct equivalents in standard vocabularies and were addressed through manual review or retained as non-standard entries where appropriate.

**Summary of Standardized Data tables in the ICU clinical data warehouse**

An overview of the resulting data schema is summarized in Tables 3-6, which outlines the core OMOP-CDM tables together with the ICU-specific operational and IoMT-related tables created in the proposed ICU-CDW architecture.

Table 3. Core OMOP-CDM clinical implemented in the ICU clinical data warehouse

| No. | Data Table | Description | Count |
|---|---|---|---|
| 1 | omop_person | Stores patient demographic information and unique identifiers. | 1,088 |
| 2 | observation_period | Defines time intervals during which clinical events occur. | 1,241 |
| 3 | visit_occurrence | Healthcare encounter records including ICU admissions. | 1,241 |
| 4 | visit_detail | Detailed departmental and sub-visit information. | 1,241 |
| 5 | condition_occurrence | Clinical diagnoses coded using ICD-10 or ICD-10-TM. | 10,167 |
| 6 | procedure_occurrence | Clinical procedures coded using ICD-9-CM. | 1,221 |
| 7 | death | Mortality information for ICU patients. | 24 |
| 8 | location | Geographic or residence information of patients. | 77 |
| 9 | care_site | Hospital department or clinical unit information. | 11 |

Table 4. Standardized OMOP vocabularies used in the ICU clinical data warehouse

| No. | Data Table | Description | Count |
|---|---|---|---|
| 10 | athena_drug | Standardized drug terminology (RxNorm). | 254,991 |
| 11 | athena_condition | Standardized terminology for clinical conditions (SNOMED CT). | 163,376 |
| 12 | athena_procedure | Standardized terminology for procedures. | 84,582 |
| 13 | athena_device | Standardized terminology for medical devices. | 218,097 |
| 14 | athena_observation | Standardized terminology for clinical observations. | 269,001 |
| 15 | athena_measurement | Standardized terminology for clinical measurements (LOINC). | 40,148 |

Table 5. REDCap research registry data in the ICU clinical data warehouse

| No. | Data Table | Description | Count |
|---|---|---|---|
| 16 | redcap_med_kku | Research registry data collected through REDCap for ICU clinical studies. | 37,308 |

Table 6. ICU operational and IoMT data in the clinical data warehouse

| No. | Data Table | Description | Count |
|---|---|---|---|
| 17 | kku_tblicu_admission | ICU admission records from the Smart ICU system. | 266 |
| 18 | tblicu_lab | Laboratory test results. | 2,753 |
| 19 | tblicu_bedmonitortrack | High-frequency bedside monitor vital signs (1-minute interval). | 36,155 |
| 20 | tblicu_bedintake_use | Nurse-recorded medication and fluid intake. | 22,435 |
| 21 | tblicu_bedpump_ccr | Hourly infusion pump output summary. | 42 |
| 22 | tblicu_bedpump_track | High-frequency infusion and syringe pump transmissions. | 1,537,272 |
| 23 | tblicu_doctororder | Physician diagnostic and treatment orders. | 18,251 |
| 24 | tblicu_snomed | SNOMED-CT coded diagnosis and procedures. | 3,604 |
| 25 | tblicu_nurse_note | Nursing clinical notes and documentation. | 82,311 |

Summarizes the standardized tables implemented in the ICU Clinical Data Warehouse (ICU-CDW). The schema includes four main categories: (1) OMOP clinical tables representing patient demographics, encounters, diagnoses, procedures, and outcomes; (2) standardized vocabulary tables from the OMOP Athena repository that ensure semantic interoperability; (3) research registry tables derived from REDCap clinical research databases; and (4) ICU operational and IoMT tables capturing real-time physiological monitoring data and ICU workflow records. Together, these tables integrate heterogeneous clinical, research, and device-generated data into a unified OMOP-CDM–compliant analytical environment.

At the core of the data model, the *omop_person* tablecontains unique demographic records for 1,088 patients, ensuring that each individual is represented by a single, unified record derived from multiple source systems. Temporal boundaries for clinical observation are defined in the *observation_period* table (1,241 records), which ensures that clinical events are analyzed within valid, non-overlapping time intervals.

Patient encounters are captured through the *visit_occurrence* and *visit_detail* tables, each comprising 1,241 records, enabling detailed tracking of ICU admissions, transfers, and departmental interactions. Clinical diagnoses are stored in the *condition_occurrence* table (10,167 records), coded using ICD-10 or ICD-10-TM standards, while interventional activities are represented in the *procedure_occurrence* table (1,221 records) using ICD-9-CM classifications. Mortality outcomes are documented in the *death* table (24 records), supporting outcome and survival analyses.

Contextual information related to patient residence and care delivery locations is provided by the *location* (77 records) and *care_site* (11 records) tables, which enable geographic and organizational analyses of ICU services.

To support semantic interoperability, the ICU-CDW incorporates extensive standardized vocabulary tables, including *athena_drug* (254,991 entries), *athena_condition* (163,376 entries), *athena_procedure* (84,582 entries), *athena_device* (218,097 entries), *athena_observation* (269,001 entries), and *athena_measurement* (40,148 entries). These vocabularies ensure consistent mapping of clinical concepts, procedures, devices, and measurements to international standards, facilitating reproducible research and cross-institutional data exchange.

Research-specific clinical variables are stored in the *redcap_med_kku* table (37,308 records), which integrates curated research data collected through REDCap with routine clinical and IoMT datasets. ICU operational data from the Smart ICU system are represented by tables such as *kku_tblicu_admission* (266 records) and *tblicu_lab* (2,753 records), providing admission-level and

laboratory information, respectively.

High-frequency physiological and device-generated data are captured in dedicated IoMT tables. These include *tblicu_bedmonitortrack* (36,155 records) for minute-level vital signs, *tblicu_bedintake_use* (22,435 records) for nurse-recorded medication and fluid intake, *tblicu_bedpump_ccr* (42 records) for summarized infusion pump outputs, and *tblicu_bedpump_track* (1,537,272 records) for high-resolution infusion and syringe pump transmissions. Clinical decision and care process information are further enriched by *tblicu_doctororder* (18,251 records), *tblicu_snomed* (3,604 records), and *tblicu_nurse_note* (82,311 records), capturing physician orders, SNOMED CT–coded clinical concepts, and narrative nursing documentation.

Among the 84 failed checks identified by the data quality dashboard, most were associated with completeness constraints rather than structural conformance errors. Many of these failures were related to ICU-specific data characteristics, such as optional clinical attributes that were not consistently recorded in the source systems. Some checks also failed due to missing mappings for certain device-specific or locally defined measurement codes. Importantly, these issues did not compromise the structural validity of the OMOP-CDM schema. Most failures were attributable to source-system limitations rather than errors in the ETL transformation pipeline. These findings indicate that the observed data quality issues are primarily driven by source-system limitations and ICU-specific data characteristics rather than deficiencies in the ETL design or OMOP-CDM implementation.

These results demonstrate the successful integration of heterogeneous clinical, research, and IoMT data into a unified, standardized ICU data warehouse (Fig. 6), providing a robust foundation for advanced analytics and CDSS development in critical care settings.

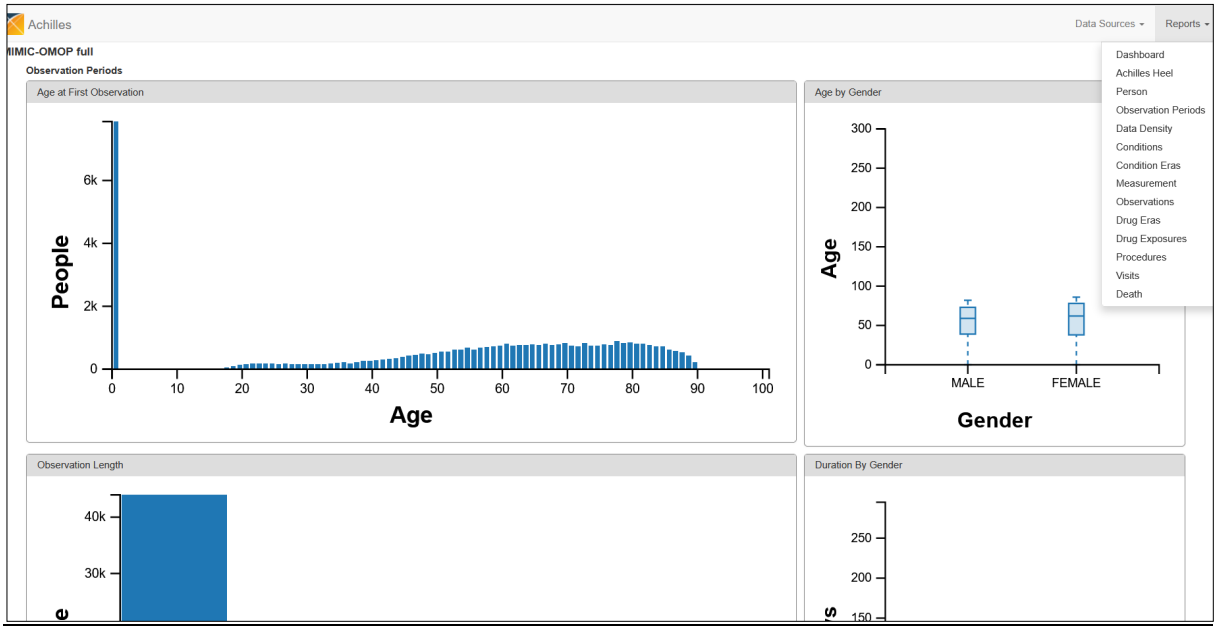


Fig. 6: Architecture for integrating heterogeneous clinical and IoMT data with standardized data harmonization in the ICU-CDW.

To further validate and characterize the standardized ICU dataset, exploratory analyses were conducted using ACHILLES (Automated Characterization of Health Information at Large-Scale Longitudinal Evidence Systems), a widely adopted analytical tool within the Observational Health Data Sciences and Informatics (OHDSI) ecosystem. ACHILLES provides automated descriptive statistics and visual summaries that enable rapid assessment of population characteristics, temporal coverage, and data plausibility in OMOP-CDM–compliant databases.

The ACHILLES analysis provided additional evidence supporting the validity and usability of the integrated dataset. Descriptive outputs confirmed consistent population characteristics, including patient demographics, visit distributions, and clinical event frequencies, aligned with expected ICU patterns. Temporal analyses demonstrated plausible distributions of observation periods and clinical events, indicating appropriate alignment of high-frequency IoMT data with episodic clinical records. Furthermore, schema-level summaries verified that key OMOP-CDM tables were populated with coherent and well-connected data structures. These findings complement the Data Quality Dashboard results and further support the successful implementation of a standardized and analytically ready ICU clinical data warehouse.

# 5. Conclusion

## 5.1. Conclusion and Discussion

This study presents the design and implementation of an OMOP-CDM–based Clinical Data Warehouse for integrating heterogeneous ICU data into a standardized and interoperable analytical environment. The experimental implementation demonstrates that diverse clinical and high-frequency IoMT data can be successfully ingested, transformed, and harmonized using a scalable ETL pipeline. Data quality assessment using the OHDSI Data Quality Dashboard showed high levels of conformance, completeness, and plausibility, achieving an overall pass rate of 98%, confirming the reliability of the integrated dataset for large-scale clinical analytics. From an informatics and service systems perspective, the proposed architecture functions as an interoperable data infrastructure that supports efficient data integration, standardized representation, and reliable analytical services. This enables improved data accessibility and supports the development of data-driven applications in critical care environments.

This study demonstrates the feasibility of integrating IoMT-enabled ICU data within an OMOP-CDM–compliant framework and highlights its potential as a foundation for advanced analytics and future clinical decision support applications.

## 5.2. Future work

Although the proposed IoMT–CDW architecture demonstrates strong feasibility and performance in a real-world ICU setting, several directions remain for future enhancement. First, subsequent work will focus on tighter integration of high-frequency waveform data (e.g., ECG, PPG, invasive blood pressure) by investigating efficient storage formats, compression techniques, and standardized feature extraction pipelines compatible with the OMOP-CDM. Second, the integrated multimodal dataset provides a robust foundation for developing predictive and prescriptive AI models, including early warning systems, risk stratification, and mortality prediction, with emphasis on interpretability and clinical validation. Third, future efforts will extend the architecture toward real-time clinical decision support through interactive dashboards and automated alerts integrated into ICU workflows. Finally, scalability will be evaluated through multicenter deployments, alongside strengthened data governance, security, and regulatory compliance to ensure ethical and secure use of high-resolution ICU data.

# Ethical approval

This study protocol was reviewed and approved by the Ethics Committee at the Center for Ethics in Human Research, Khon Kaen University (Institutional Review Board Number: IRB00013103; Federalwide Assurance Number: FWA00003418; Reference No: HE664004). The study utilized anonymized data obtained from the Srinagarind Hospital ICU registries, thereby eliminating the potential for identifying individual participants. In view of the retrospective nature of the study and the use of de-identified data, the requirement for informed consent was waived by the Khon Kaen University Ethics Committee for Human Research, in accordance with the Declaration of Helsinki and the international standard for clinical investigation of medical devices involving human subjects—Good Clinical Practice (ISO 14155:2011).

## Funding

This research has received funding support from the NSRF via the Program Management Unit for Human Resources & Institutional Development, Research and Innovation [grant number B04G650012].

## References

Barroca Filho, I. de M., Aquino, G., Malaquias, R. S., Girão, G., & Melo, S. R. M. (2021). An IoT-based healthcare platform for patients in ICU beds during the COVID-19 outbreak. *IEEE Access, 9*, 27262–27277. https://doi.org/10.1109/ACCESS.2021.3057276

Beunza, J.-J., Lafuente, J.-L., González, S., & Gómez-Tello, V. (2024). Artificial intelligence and the Internet of Medical Things in the ICU: Time for implementation. *Medicina Intensiva (English Edition), 48*(1), 56–58. https://doi.org/10.1016/j.medine.2023.06.002

Botta, A., de Donato, W., Persico, V., & Pescapé, A. (2016). Integration of Cloud computing and Internet of Things: A survey. Future Generation Computer Systems, 56, 684-700. https://doi.org/https://doi.org/10.1016/j.future.2015.09.021

de Mul, M., Alons, P. W. F., van der Velde, M. P., Konings, I., Bakker, J., & Hazelzet, J. A. (2012). Development of a clinical data warehouse from an intensive care clinical information system. *Computer Methods and Programs in Biomedicine, 105*(1), 22–30. https://doi.org/10.1016/j.cmpb.2010.07.003

Emaliyawati, E., Ibrahim, K., Kurniawan, T., Fitria, N., & Songwathana, P. (2025). A systematic review of internet-based remote patient monitoring systems for chronic disease management in Asian countries. *Patient Preference and Adherence, 19*, 2985–3000. https://doi.org/10.2147/PPA.S456789

Kallfelz, M., Tsvetkova, A., Pollard, T., Johnson, A., Mark, R., & Badawi, O. (2021). *MIMIC-IV demo data in the OMOP Common Data Model (Version 0.9)* [Data set]. PhysioNet. https://doi.org/10.13026/p1f5-7x35

Islam, S. M. R., Kwak, D., Kabir, M. H., Hossain, M., & Kwak, K. S. (2015). The Internet of Things for Health Care: A Comprehensive Survey. IEEE Access, 3, 678-708. https://doi.org/10.1109/ACCESS.2015.2437951

Muangkote, N., A.R. Wangkhamhan, and T. Wangkhamhan, Exploring social and environmental factors contributing to smoking initiation among Thai adolescents using advanced feature selection techniques. Tobacco Prevention & Cessation, 2025. 11(June): p. 1-15. https://doi.org/10.18332/tpc/205065

Paris, N., & Parrot, A. (2020). MIMIC in the OMOP Common Data Model. *medRxiv*. https://doi.org/10.1101/2020.08.14.20175141

Shaik, T., Tao, X., Li, L., Xie, H., & Velásquez, J. D. (2024). A survey of multimodal information fusion for smart healthcare: Mapping the journey from data to wisdom. Information Fusion, 102, 102040. https://doi.org/https://doi.org/10.1016/j.inffus.2023.102040

Thantilage, R. D., Le-Khac, N.-A., & Kechadi, M.-T. (2023). Healthcare data security and privacy in data warehouse architectures. *Informatics in Medicine Unlocked, 39*, 101270. https://doi.org/10.1016/j.imu.2023.101270

Wang, Z., Syed, M., Syed, S., Greer, M., Seker, E., Zozus, M. N., & Craven, C. K. (2024). Clinical data warehousing: A scoping review. *Journal of the Society for Clinical Data Management, 4*(1), 8. https://doi.org/10.47912/jscdm.8

Wongchinda, P. (2025). The implementation of data warehouse for result reporting system: The case of outpatient university clinic in Thailand. *Journal of the Thai Medical Informatics Association, 11*(1), 64–70.

Wongsim, M., Chanthinok, K., Satchawatee, N., & Udchachone, S. (2025). Unlocking data potential: A comprehensive framework for integrating open government data in Thailand's public sector. Journal of Logistics, Informatics and Service Science, 12(2), 26–46. https://doi.org/10.33168/JLISS.2025.0202