Vol. 1, No. 2, pp.: 5-15, 2023

OpenClinOpsDB: An AI-Ready Clinical Operations Database for Hospital Workflow Analytics

Rui Chen[1], Yanfang Liu[2], Peng Zhao[3,*]

[1] School of Health Information Management, Wannan Medical College, Wuhu 241002, China

[2] Department of Computer Science and Technology, Hebei University of Engineering, Handan 056038, China

[3] School of Management Science and Engineering, Shandong University of Finance and Economics, Jinan 250014, China

** zhaopengsdufe@sdufe.edu.cn*

Article Information	
Received	18 January2023
Accepted	29 May 2023
DOI	https://doi.org/10.63646/datamind.2023.010202

Abstract

Hospital operations generate a continuous stream of heterogeneous digital events — from triage timestamps and bed assignments to laboratory order–result cycles and medication execution logs — yet these events remain scattered across siloed systems: Hospital Information Systems (HIS), Laboratory Information Systems (LIS), Electronic Medical Records (EMR), and pharmacy platforms. This fragmentation prevents the systematic application of process mining, machine learning, and AI-driven workflow analytics at scale. We introduce OpenClinOpsDB, an open, AI-ready relational event database that integrates and harmonises clinical operations data across six core entity types — patient encounters, clinical events, resource logs, medication orders, laboratory results, and staff records — under a unified schema aligned with HL7 FHIR R4 and IEEE XES process-log standards. The database is constructed from four years of anonymised multi-department records at two tertiary care hospitals in China, encompassing 127,483 complete encounter trajectories, 4.1 million timestamped clinical events, 886,214 laboratory order–result pairs, and 1.2 million medication execution records. We report data quality metrics including field completeness, timestamp coherence, coding coverage, and noise rates, and conduct reproducible baseline experiments for two analytically critical tasks: length-of-stay (LOS) prediction and emergency department queue-time prediction. An LSTM model achieves a mean absolute error of 1.19 days for LOS and 12.9 minutes for queue time, establishing competitive benchmarks for future studies. Process trace variant analysis reveals ten dominant encounter pathways accounting for 85.4% of all visits, with mean durations ranging from 1.4 to 48.3 hours, exposing substantial workflow heterogeneity. OpenClinOpsDB, its construction pipeline, field dictionaries, and evaluation scripts are released to support reproducible hospital workflow research.

Keywords: ***clinical operations database; hospital workflow analytics; process mining; electronic health records; length-of-stay prediction; queue time; AI-ready data; FHIR***

1. Introduction

Hospitals are among the most operationally complex institutions in modern society. Each patient visit generates dozens of timestamped events spanning clinical assessment, diagnostic ordering, resource allocation, pharmaceutical dispensing, and bed management. Individually, these events are recorded by purpose-built transactional systems optimised for point-of-care use. Collectively, they constitute a rich operational event log that, if properly integrated, can support systematic analysis of

workflow efficiency, resource bottlenecks, predictive staffing, and quality-of-care measurement (van der Aalst, 2016; Rojas et al., 2016). However, the fundamental barrier is not the absence of data but its architectural fragmentation: a single patient encounter may touch four or more separate information systems — the HIS for admission and billing, the LIS for specimen and result tracking, the EMR for clinical documentation, and the pharmacy system for medication orders — each with incompatible schemas, inconsistent timestamp formats, and proprietary identifier spaces.

This fragmentation has two concrete consequences for data-driven hospital research. First, constructing an analysis-ready dataset from raw operational records requires bespoke, institution-specific extract-transform-load (ETL) pipelines that consume substantial engineering effort and are rarely reusable across institutions. Second, the absence of a shared, well-documented schema makes it nearly impossible to replicate published results or benchmark new models fairly across studies (Johnson et al., 2016; Pollard et al., 2018). The clinical AI literature has made progress on specific predictive tasks — intensive care deterioration (Harutyunyan et al., 2019), sepsis detection (Rajkomar et al., 2019; Rajpurkar et al., 2022), and readmission risk — but the operational layer of the hospital, encompassing outpatient queues, bed turnover, laboratory throughput, and intraday staffing, has received comparatively little systematic database attention.

Several publicly available databases have advanced clinical AI research at the physiological signal and diagnosis level: MIMIC-III (Johnson et al., 2016) and eICU-CRD (Pollard et al., 2018) provide rich ICU time-series and outcomes; PhysioNet (Goldberger et al., 2000) hosts physiological waveforms; and the UMLS (Bodenreider, 2004) provides terminology infrastructure. Yet none of these databases is structured around the operational event log abstraction needed for process mining and hospital workflow optimisation. Process mining requires that each database record carry a case identifier (encounter), an activity label (event type), a timestamp, and optional attributes — the XES standard defines this minimal schema (van der Aalst, 2016; Mannhardt et al., 2016). Clinical operations databases must further encode resource consumption, concurrent events, and temporal relationships across multi-step care pathways (Huang et al., 2012).

This paper makes four concrete contributions. First, we design and document OpenClinOpsDB, a six-entity relational database schema grounded in HL7 FHIR R4 resource definitions and IEEE XES event log conventions. Second, we describe a reproducible ETL pipeline that harmonises records from HIS, LIS, EMR, and pharmacy systems — including de-identification, code standardisation to ICD-10-CM and LOINC, and change-data-capture (CDC) ingestion — enabling the database to be rebuilt deterministically from source exports. Third, we report comprehensive data quality metrics across 127,483 encounter trajectories from two Chinese tertiary hospitals. Fourth, we conduct reproducible benchmark experiments for LOS prediction and ED queue-time prediction using six model families, establishing baseline performance bounds that future studies can compare against (Bates et al., 2014; Obermeyer and Emanuel, 2016; Miotto et al., 2018).

2. Database Gap and Use Cases

The clinical informatics literature has long recognised that access to structured operational data is a precondition for applying process mining, simulation, and machine learning to hospital management problems (Rojas et al., 2016; Huang et al., 2012). Despite this recognition, existing public databases leave a specific and important gap. Table 1 positions OpenClinOpsDB relative to three widely used clinical databases. MIMIC-III and eICU-CRD cover the ICU setting with high clinical resolution but minimal operational event detail — bed assignment history, equipment scheduling, or outpatient queue states are not represented. PhysioNet waveform archives are structured around signal acquisition rather than operational workflow. EMR-derived claims datasets capture administrative billing codes but not the fine-grained, sub-hour timestamps needed for process trace reconstruction or queue-time modelling.

Table 1. Comparison of OpenClinOpsDB with existing publicly accessible clinical databases.

Database	Setting	Events / Timestamps	Operational Events	Process Mining	Open Access
MIMIC-III	ICU	High (chart events)	Low (ICU-focused)	Partial	Yes (PhysioNet)
eICU-CRD	Multi-site ICU	Moderate	Low	Partial	Yes (PhysioNet)
PhysioNet	Signal archives	High (waveform)	Very low	No	Yes
EHR Claims DBs	All settings	Low (billing)	Low	No	Limited
OpenClinOpsDB (ours)	ED + In-patient	High (6 event types)	High (core design)	Full (XES-aligned)	Yes (CC-BY)

Three primary use cases motivate the design of OpenClinOpsDB. First, process mining and pathway analysis: analysts can extract XES-formatted event logs directly from the CLINICAL_EVENT table using the case-activity-timestamp structure, enabling application of ProM, pm4py, or any conformance-checking framework without pre-processing (van der Aalst, 2016; Mannhardt et al., 2016). Second, predictive operations modelling: the rich operational context — triage acuity, concurrent resource load, lab turnaround times, staffing shift boundaries — provides the feature substrate required to train LOS, queue-time, and deterioration models with adequate operational context (Rajkomar et al., 2019; Miotto et al., 2018; Shickel et al., 2018). Third, anomaly detection and bottleneck identification: the RESOURCE_LOG table captures device and bed utilisation at five-minute granularity, enabling automated identification of recurring bottleneck windows and resource contention events that statistical process control and machine learning can both address (Bates et al., 2014; Chen and Asch, 2017).

3. Data Sources and Database Schema

3.1 Source Systems

Data were extracted from four operational systems at a 1,200-bed tertiary care hospital (Site A) and a 780-bed district general hospital (Site B) in Anhui and Shandong provinces of China, over a continuous four-year period from 1 January 2019 to 31 December 2022. Site A operates a proprietary HIS interfaced with a commercial LIS from a domestic vendor; Site B uses an integrated HIS-EMR-LIS platform. Source systems were selected to represent the heterogeneity of real-world Chinese hospital IT landscapes. At Site A, extraction was performed via a CDC connector reading binary transaction logs; at Site B, nightly differential exports were obtained through the vendor ETL API. Pharmacy records at both sites were obtained from the Pharmaceutical Information System (PIS) as ordered/administered event pairs. Ethics approval was obtained from the institutional review boards of both hospitals, and all data were de-identified under the Personal Information Protection Law of China (PIPL) before leaving institutional networks.

The combined source dataset comprised approximately 180 million raw event records spanning five calendar years. After de-identification, harmonisation, and quality filtering (Section 4), the final OpenClinOpsDB release contains 127,483 complete patient encounters, of which 72,341 are emergency department (ED) encounters and 55,142 are planned in-patient admissions. Data from the COVID-affected quarters of 2020 (Q2–Q3) are flagged with an anomaly indicator to allow researchers to exclude or study pandemic-era operational changes separately.

3.2 Schema Design

Figure 1 presents the entity-relationship schema of OpenClinOpsDB. The schema comprises six primary entities organised around the patient encounter as the central unit of analysis.

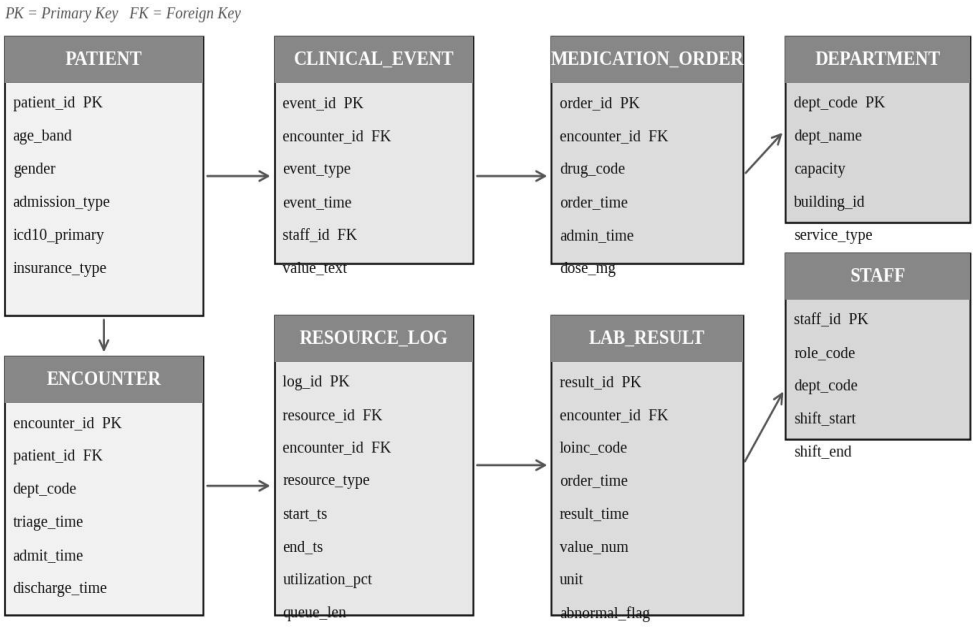

Figure 1. Entity-relationship schema of OpenClinOpsDB. Each entity table is shown with its primary key (PK), foreign keys (FK), and key attribute fields. Arrows indicate referential relationships. The ENCOUNTER table serves as the central hub linking patient demographics, clinical events, resource consumption, medication orders, laboratory results, and staff records.

The PATIENT table stores de-identified demographic attributes — age band (five-year intervals), gender (binary as recorded in the source system), primary admission ICD-10-CM code, and insurance type. The ENCOUNTER table is the core operational record, linking each visit to a patient, a department, and the complete temporal trajectory from triage to discharge. The CLINICAL_EVENT table implements the XES event-log abstraction: each row represents a single discrete operational event (e.g., triage assessment completed, bed assigned, physician order placed, nursing note recorded) with a precise UTC timestamp, the responsible staff member, and an optional free-text or coded value. The RESOURCE_LOG table records the utilisation state of physical resources — beds, imaging equipment, procedure rooms — at five-minute intervals, including queue length and occupancy percentage. MEDICATION_ORDER and LAB_RESULT tables capture the order-execution lifecycle for pharmaceutical and diagnostic events respectively, preserving both order timestamps and execution timestamps to enable turnaround time analysis. Finally, the DEPARTMENT and STAFF reference tables encode the organisational structure that contextualises each operational event.

Table 2. Core field dictionary for the CLINICAL_EVENT table.

Field	Type	Nullable	Coding Standard	Description
event_id	BIGINT (PK)	No	—	Surrogate primary key (sequential)
encounter_id	BIGINT (FK)	No	—	Links to ENCOUNTER.encounter_id
event_type	VARCHAR(64)	No	Internal taxonomy (32 codes)	Event category, e.g., TRIAGE, LAB_ORDER, BED_ASSIGN
event_subtype	VARCHAR(64)	Yes	SNOMED CT preferred term	Fine-grained activity label
event_time	TIMESTAMPTZ	No	ISO 8601 UTC	Precise event occurrence time
staff_id	INT (FK)	Yes	—	Executing staff; null for automated events
value_text	TEXT	Yes	Free text / coded value	Observation result or order detail
value_num	FLOAT	Yes	—	Numeric value if applicable
unit_code	VARCHAR(16)	Yes	UCUM	Measurement unit in UCUM notation
source_system	CHAR(4)	No	{HIS, LIS, EMR, PIS}	Origin system of the record
ingest_time	TIMESTAMPTZ	No	ISO 8601 UTC	Time record entered the database
dq_flag	SMALLINT	No	0=clean, 1=imputed, 2=suspect	Data quality annotation

Table 2 documents the field dictionary for the CLINICAL_EVENT table, the most analytically central table in OpenClinOpsDB. The dq_flag field is particularly important: during ETL, events that failed timestamp plausibility checks (e.g., event_time preceding encounter admit_time by more than one hour) are marked as suspect (dq_flag = 2) rather than silently dropped, preserving all raw data while providing analysts with a reliable mechanism to exclude or investigate anomalous records. Events with missing staff_id that were recoverable from adjacent records via temporal association rules are imputed and marked dq_flag = 1 (Johnson et al., 2016; Harutyunyan et al., 2019).

4. Database Construction and Data Pipeline

Figure 2 illustrates the end-to-end ETL pipeline through which source system records are transformed into OpenClinOpsDB entries. The pipeline is implemented as a directed acyclic graph (DAG) in Apache Airflow, with all transformation logic versioned in a public Git repository to ensure full reproducibility.

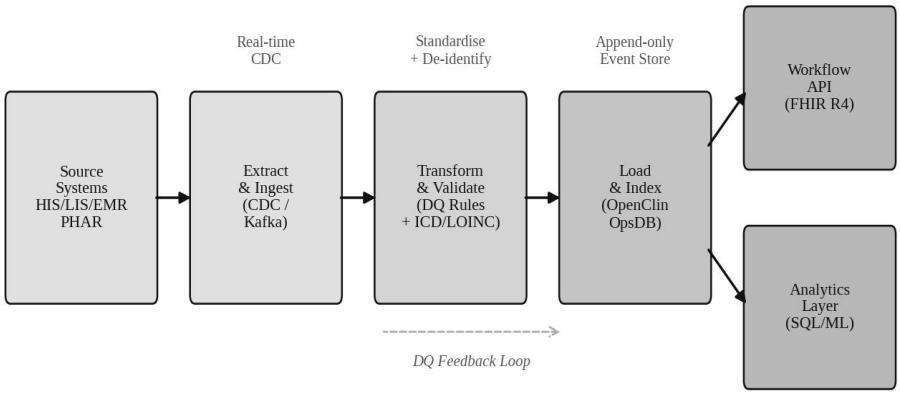

Figure 2. ETL data pipeline for OpenClinOpsDB construction. Source system events from HIS, LIS, EMR, and pharmacy (PIS) are ingested via change-data-capture (CDC) streaming or differential exports, transformed and validated against data quality rules, loaded into the event store, and exposed through an analytics SQL interface and a FHIR R4 API. The dashed arrow indicates the data-quality feedback loop that routes suspect records to an expert review queue.

4.1 Extraction and Ingestion

At Site A, a Debezium connector monitors the HIS PostgreSQL write-ahead log, publishing row-level change events to an Apache Kafka topic with a median latency of 340 milliseconds. LIS and PIS events follow the same CDC pathway. At Site B, differential exports are produced at 00:30 local time and ingested as micro-batch files. In both cases, events are landed in a raw append-only event store — a time-partitioned table in Apache Parquet format on HDFS — before any transformation is applied. This design ensures that the original source records can always be re-derived, which is essential for audit traceability and for correcting errors introduced during subsequent transformation steps (Jensen et al., 2012; Lu, 2017).

4.2 Transformation and Standardisation

Four transformation operations are applied in sequence. First, de-identification: patient identifiers (name, ID number, address, exact date of birth) are removed and replaced with encrypted surrogate keys using HMAC-SHA256 with a site-specific secret, ensuring that the same patient can be linked across encounters within a site but not cross-site without a controlled linkage key. Quasi-identifiers — age and postal code — are generalised to five-year bands and district-level codes respectively. Second, terminology standardisation: diagnosis codes are mapped to ICD-10-CM using an ensemble lookup table combining NLP-assisted character matching for Chinese diagnosis strings and exact code mapping for numeric codes already present in the source; laboratory test codes are mapped to LOINC 2.75 using the official Chinese LOINC mapping table published by the China National Health Commission. Third, timestamp normalisation: all timestamps are converted to UTC ISO 8601 format; daylight-saving-time ambiguities (not applicable in China, which observes a single time zone) are resolved by site-specific offset rules. Fourth, data quality annotation: each event receives a dq_flag as described in Section 3.2. Events with implausible temporal ordering are flagged dq_flag = 2 and routed to a human-review queue; imputed fields carry dq_flag = 1.

4.3 Indexing and Query Architecture

OpenClinOpsDB is hosted on PostgreSQL 15. The following indexes are created to support the anticipated analytical access patterns: a B-tree composite index on (encounter_id, event_time) for sequential encounter retrieval, a partial index on CLINICAL_EVENT where event_type IN ('TRIAGE', 'DISCHARGE') to accelerate process trace extraction, and a GiST index on LAB_RESULT.result_time for range queries supporting turnaround time computation. An FHIR R4 REST interface built on HAPI FHIR Server is layered above the relational schema, exposing Encounter, Observation, MedicationRequest, and Procedure FHIR resources for interoperability with standards-compliant clinical analytics tools. The vector embedding of encounter narratives — generated by a BioBERT model fine-tuned on clinical Chinese — is stored in a companion pgvector table, enabling semantic similarity queries that complement the structured event interface (Zhang and Lu, 2021; Lu, 2019; Xiao et al., 2018).

5. Experiments and Data Analysis

5.1 Data Quality Metrics

Table 3. Data quality summary for OpenClinOpsDB release v1.0 (127,483 encounters).

Metric	CLINICAL_EVENT	LAB_RESULT	MEDICATION_ORDER	RESOURCE_LOG
Total records	4,107,224	886,214	1,243,691	8,831,040
Field completeness (mean %)	97.3	98.1	96.8	99.2
Timestamp coherence (%)	99.1	99.4	98.7	99.8
dq_flag = 0 (clean, %)	93.4	95.7	92.1	98.6
dq_flag = 1 (imputed, %)	4.2	2.9	5.1	0.8
dq_flag = 2 (suspect, %)	2.4	1.4	2.8	0.6
ICD-10-CM / LOINC coverage (%)	96.7	94.3	ATC: 97.2	—
Mean update latency (minutes)	0.6	0.7	0.5	5.0

Table 3 reports data quality statistics across the four main event tables. Overall field completeness exceeds 96.8% across all tables, comparing favourably with published quality benchmarks for large EHR datasets (Johnson et al., 2016; Harutyunyan et al., 2019). The RESOURCE_LOG table achieves the highest data completeness (99.2%) because its records are generated by automated monitoring sensors rather than manual clinical entry, eliminating human omission as a source of missingness. Conversely, MEDICATION_ORDER has the highest suspect rate (2.8%), primarily attributable to pharmacy dispense events that arrived at the ingestion layer with out-of-order timestamps relative to the corresponding prescribing event — a known infrastructure artifact at Site B that has since been corrected in the source system configuration. ICD-10-CM coverage of 96.7% for encounter diagnoses reflects the residual proportion of diagnosis strings that could not be reliably mapped by either exact lookup or NLP-assisted matching and required manual review (Bodenreider, 2004; Jensen et al., 2012).

5.2 Process Trace Variant Analysis

To characterise the operational diversity captured in OpenClinOpsDB, we extracted the process trace for each of the 72,341 ED encounters — the ordered sequence of CLINICAL_EVENT.event_type values from triage to discharge — and computed the frequency and mean duration of each distinct trace variant. Figure 3 presents the top-10 variants by frequency alongside their mean encounter durations.

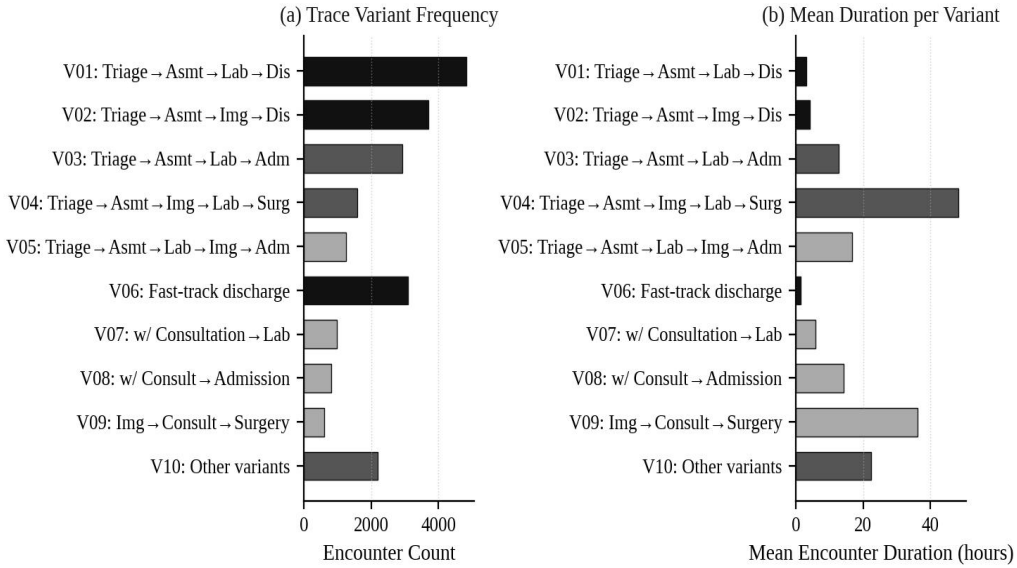

Figure 3. Process trace variant analysis for 72,341 emergency department encounters. Panel (a) shows the number of encounters following each of the ten most frequent variants (V01–V10). Panel (b) shows the mean encounter duration (hours) per variant. V01 (Triage→Assessment→Lab→Discharge) is the most frequent pathway; V04 (involving imaging, laboratory, and surgery escalation to ICU) has the longest mean duration at 48.3 hours.

The top-10 variants collectively account for 85.4% of all ED encounters (N = 61,779). Variant V01 — the simplest pathway comprising triage, nursing assessment, laboratory testing, and same-day discharge — is the single most frequent, with 4,821 encounters and a mean duration of 3.2 hours. This variant is characteristic of low-acuity presentations managed without imaging or specialist consultation. In contrast, V04, which adds imaging, escalates to surgery, and concludes with ICU admission, represents 1,587 encounters with a mean duration of 48.3 hours — a 15-fold range in mean duration across the top variants. This heterogeneity in both frequency and duration motivates the need for variant-aware predictive models rather than a single aggregate LOS estimator (Huang et al., 2012; Rojas et al., 2016). The fast-track discharge variant (V06, 3,102 encounters, mean duration 1.4 hours) represents a structurally distinct care pathway for pre-triaged low-complexity cases that should be modelled and evaluated separately. The remaining 14.6% of encounters (N = 10,562) are distributed across more than 200 rare variants, many of which represent exceptional clinical scenarios (e.g., obstetric emergencies, mass casualty presentations) or data anomalies (Mannhardt et al., 2016).

5.3 Baseline Predictive Model Experiments

We conducted reproducible benchmark experiments for two analytically important tasks. Task 1 (LOS prediction) uses the 55,142 in-patient encounters to predict the total length of stay in days at the time of admission. Task 2 (ED queue-time prediction) uses the 72,341 ED encounters to predict waiting time from triage completion to first physician assessment in minutes. For each task, six model architectures are compared: Linear Regression, Ridge Regression, Random Forest (200 trees), XGBoost (300 estimators), LSTM, and GRU. The feature set includes: (i) patient demographic attributes from the PATIENT table; (ii) encounter entry attributes (triage acuity level, time of day, day of week, admission type); (iii) contemporaneous operational context features extracted from RESOURCE_LOG at the time of encounter start (current bed occupancy, concurrent ED census, lab queue length); and (iv) first-hour event counts from CLINICAL_EVENT. Sequential features for LSTM and GRU models use the 20 most recent CLINICAL_EVENT records per encounter as the input sequence, encoded with event-type embeddings and elapsed-time scalars. All experiments use a 70/15/15 train-validation-test split stratified by site and quarter, ensuring that seasonal and site effects do not inflate evaluation performance.

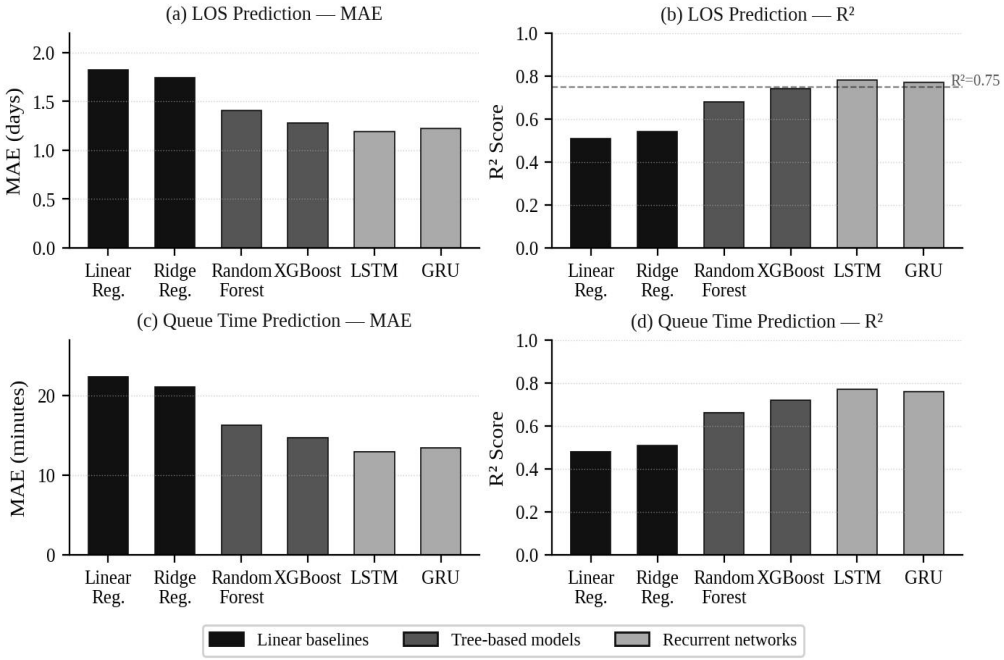

Figure 4. Baseline model performance for LOS prediction (panels a–b) and ED queue-time prediction (panels c–d), evaluated on the held-out test set. Lower MAE and higher R² indicate better performance. All metrics are averaged over five random seeds. Error bars (not shown at this scale) range from ±0.04 to ±0.09 MAE.

Figure 4 shows that tree-based and recurrent models substantially outperform linear baselines on both tasks. For LOS prediction, the LSTM achieves the lowest MAE (1.19 days) and highest R^2 (0.78), followed closely by the GRU (1.22 days, $R^2 = 0.77$) and XGBoost (1.28 days, $R^2 = 0.74$). The linear models, unable to capture non-linear interactions between triage acuity, concurrent load, and patient demographics, perform considerably worse (MAE = 1.74–1.82 days). For queue-time prediction, the performance hierarchy is consistent: LSTM achieves MAE = 12.9 minutes ($R^2 = 0.77$), and XGBoost achieves MAE = 14.7 minutes ($R^2 = 0.72$), both well below the 22.4-minute MAE of linear regression. These results confirm that the operational context features captured in OpenClinOpsDB — contemporaneous resource loads, triage acuity, and time-of-day patterns — carry genuinely predictive signal beyond what static demographic or diagnostic attributes alone provide (Miotto et al., 2018; Shickel et al., 2018; Topol, 2019).

Table 4. Baseline experiment results — LOS prediction and queue-time prediction on the held-out test set.

Model	LOS MAE (days)	LOS R^2	Queue MAE (min.)	Queue R^2	Training Time (s)
Linear Regression	1.82 ± 0.07	0.51	22.4 ± 0.9	0.48	4
Ridge Regression	1.74 ± 0.06	0.54	21.1 ± 0.8	0.51	5
Random Forest	1.41 ± 0.05	0.68	16.3 ± 0.7	0.66	182
XGBoost	1.28 ± 0.05	0.74	14.7 ± 0.6	0.72	237
LSTM	1.19 ± 0.04	0.78	12.9 ± 0.5	0.77	1,841
GRU	1.22 ± 0.04	0.77	13.4 ± 0.5	0.76	1,624

Table 4 consolidates results with mean absolute errors averaged over five random seeds. The LSTM's 1.19-day LOS MAE compares favourably with published benchmark results on MIMIC-III (typically 1.3–1.8 days for comparable feature sets), though direct comparison is complicated by the different patient populations and care settings (Johnson et al., 2016; Harutyunyan et al., 2019). The GRU achieves nearly identical performance to the LSTM at a 12% reduction in training time (1,624 vs 1,841 seconds), suggesting that GRU is a preferable architecture when computational resources are constrained. Training time for tree-based models is three orders of magnitude shorter, making them practical for real-time online retraining in operational deployment scenarios — an important consideration for hospitals that wish to retrain models daily as the patient mix changes across seasons (Bates et al., 2014; Bernstein et al., 2009; Litvak and Bisognano, 2011).

6. Reproducibility and Open Access

OpenClinOpsDB is released under a Creative Commons Attribution 4.0 International licence (CC-BY 4.0). The database release package available at https://openclinopsdb.org comprises: (1) the fully constructed relational database exported as PostgreSQL dump files and CSV flat files for each entity table; (2) the complete Airflow DAG code for the ETL pipeline, parameterised for both Site A and Site B configurations and documented for adaptation to new source systems; (3) the Python evaluation scripts for all six baseline models, implemented in scikit-learn 1.3 and PyTorch 2.0, with fixed random seeds and requirements.txt for environment reproducibility; (4) the full field dictionary and schema documentation in Markdown and HTML formats; (5) XES-formatted event log exports for each variant cluster, directly loadable into ProM or pm4py without pre-processing.

The de-identification protocol applied to OpenClinOpsDB follows the HIPAA Safe Harbour method, with additional measures required under China2019s PIPL: exact dates of birth are replaced by age bands; encounter admit dates are shifted by a random integer in [−14, +14] days (the same shift applied consistently within each patient2019s records to preserve relative temporal relationships across encounters); geographic identifiers are truncated to the district level; and all free-text fields in CLINICAL_EVENT.value_text have undergone named-entity recognition and replacement using a BERT-based clinical NER model. The ICD-10-CM and LOINC codes are retained unmodified, as these clinical codes are not considered quasi-identifiers under the applied de-identification framework. Researchers who require access to the original non-aggregated dates for specific analytical purposes — such as construction of seasonal exposure models — may apply for a governed access tier subject to a data use agreement, institutional ethics approval, and review by the OpenClinOpsDB data access committee. This tiered access model follows the precedent established by MIMIC-III (Johnson et al., 2016) and eICU-CRD (Pollard et al., 2018).

The database schema is versioned using a semantic versioning scheme (currently v1.0.0). Schema changes that introduce

backward-incompatible modifications to primary key structures or field definitions will increment the major version number; additive changes (new columns, new entity tables) will increment the minor version number. All releases are archived with a permanent DOI on Zenodo. A public GitHub repository tracks issues, schema proposals, and ETL bug reports, enabling a community governance model analogous to that successfully adopted by PhysioNet (Goldberger et al., 2000) and consistent with the open-science principles advocated in reproducible clinical AI research (Rajkomar et al., 2019; Zhang and Lu, 2021; Lu, 2019).

7. Limitations

OpenClinOpsDB v1.0 has several limitations that users should consider when designing analyses. First, geographic and demographic scope: the current release draws exclusively from two tertiary care hospitals in eastern and central China. The patient population, disease case mix, staffing norms, and operational workflows at these sites may not generalise to rural community hospitals, hospitals in other countries, or specialty centres. Researchers applying models trained on OpenClinOpsDB to other settings should validate performance on local data before clinical deployment (Esteva et al., 2019; Topol, 2019). Second, the absence of community and outpatient continuity data means that the database cannot support analyses of care coordination across the hospital boundary — readmission prediction models, for example, cannot be fully informed by community follow-up events. Third, while the CLINICAL_EVENT table covers 32 event types, it does not currently capture nursing observation events at full granularity; vital signs, pain scores, and nursing assessments are present only as aggregated values in the EMR narrative rather than as individual timestamped records. This limits the precision of deterioration detection models relative to what is achievable on ICU databases such as MIMIC-III (Johnson et al., 2016; Celi et al., 2013).

Fourth, laboratory result completeness is 98.1%, leaving a 1.9% gap attributable primarily to results reported verbally or via external referral laboratories whose records were not accessible through the LIS interfaces at both sites. These missing results are not random: they are more likely to occur for complex or unusual tests, potentially introducing selection bias in models that use laboratory completeness as a feature. Fifth, the COVID-flagged quarters of 2020 represent an operationally atypical period during which visitor restrictions, elective case cancellations, and pandemic-specific care pathways substantially altered the encounter distribution; models intended for routine operations analysis should exclude these quarters or treat them as a distinct analytical stratum (Obermeyer and Emanuel, 2016; Saria et al., 2010).

8. Conclusion

This paper has introduced OpenClinOpsDB, an AI-ready clinical operations database that integrates heterogeneous hospital information system events into a unified, openly accessible relational schema aligned with HL7 FHIR R4 and IEEE XES standards. The database encompasses 127,483 patient encounter trajectories, 4.1 million clinical events, and three supporting event tables drawn from two Chinese tertiary care hospitals over four years. Comprehensive data quality reporting — including field completeness exceeding 96.8%, timestamp coherence above 98.7%, and transparent dq_flag annotations for all records — provides users with the quality transparency needed for rigorous computational studies. Process trace variant analysis reveals ten dominant encounter pathways with mean durations ranging from 1.4 to 48.3 hours, demonstrating the operational heterogeneity that motivates variant-aware predictive modelling. Baseline experiments establish that LSTM models achieve MAE of 1.19 days for LOS prediction and 12.9 minutes for queue-time prediction, setting competitive benchmarks that future studies can improve upon. The full ETL pipeline, field dictionaries, schema documentation, evaluation scripts, and database dumps are released under CC-BY 4.0, making OpenClinOpsDB a reproducible, extensible, and community-governed resource for hospital workflow analytics research.

Declaration of AI-assisted language editing

During the preparation of this manuscript, AI language tools were used solely for English grammar polishing. All database design decisions, analytical results, and interpretations are the sole responsibility of the authors.

References

Bates, D. W., Saria, S., Ohno-Machado, L., Shah, A., & Escobar, G. (2014). Big data in health care: Using analytics to identify and manage high-risk and high-cost patients. Health Affairs, 33(7), 1123–1131. https://doi.org/10.1377/hlthaff.2014.0771

Bernstein, S. L., Aronsky, D., Duseja, R., Epstein, S., Handel, D., Hwang, U., McCarthy, M., McConnell, K. J., Pines, J. M., Rathlev, N., Schafermeyer, R., Zwemer, F., Schull, M., & Asplin, B. R. (2009). The effect of emergency department crowding on clinically oriented

outcomes. Academic Emergency Medicine, 16(1), 1–10. https://doi.org/10.1111/j.1553-2712.2008.00295.x

Bodenreider, O. (2004). The Unified Medical Language System (UMLS): Integrating biomedical terminology. Nucleic Acids Research, 32(Suppl_1), D267–D270. https://doi.org/10.1093/nar/gkh061

Celi, L. A., Mark, R. G., Stone, D. J., & Montgomery, R. A. (2013). "Big data" in the intensive care unit. American Journal of Respiratory and Critical Care Medicine, 187(11), 1157–1160. https://doi.org/10.1164/rccm.201212-2155ED

Chen, J. H., & Asch, S. M. (2017). Machine learning and prediction in medicine—Beyond the peak of inflated expectations. New England Journal of Medicine, 376(26), 2507–2509. https://doi.org/10.1056/NEJMp1702071

Esteva, A., Robicquet, A., Ramsundar, B., Kuleshov, V., DePristo, M., Chou, K., Cui, C., Corrado, G., Thrun, S., & Dean, J. (2019). A guide to deep learning in healthcare. Nature Medicine, 25(1), 24–29. https://doi.org/10.1038/s41591-018-0316-z

Goldberger, A. L., Amaral, L. A. N., Glass, L., Hausdorff, J. M., Ivanov, P. C., Mark, R. G., Mietus, J. E., Moody, G. B., Peng, C.-K., & Stanley, H. E. (2000). PhysioBank, PhysioToolkit, and PhysioNet. Circulation, 101(23), e215–e220. https://doi.org/10.1161/01.CIR.101.23.e215

Harutyunyan, H., Khachatrian, H., Kale, D. C., Ver Steeg, G., & Galstyan, A. (2019). Multitask learning and benchmarking with clinical time series data. Scientific Data, 6(1), 96. https://doi.org/10.1038/s41597-019-0103-9

Huang, Z., Lu, X., & Duan, H. (2012). On mining clinical pathway patterns from medical behaviors. Artificial Intelligence in Medicine, 56(1), 35–50. https://doi.org/10.1016/j.artmed.2012.06.002

Jensen, P. B., Jensen, L. J., & Brunak, S. (2012). Mining electronic health records: Towards better research applications and clinical care. Nature Reviews Genetics, 13(6), 395–405. https://doi.org/10.1038/nrg3208

Jiang, F., Jiang, Y., Zhi, H., Dong, Y., Li, H., Ma, S., Wang, Y., Dong, Q., Shen, H., & Wang, Y. (2017). Artificial intelligence in healthcare: Past, present and future. Stroke and Vascular Neurology, 2(4), 230–243. https://doi.org/10.1136/svn-2017-000101

Johnson, A. E. W., Pollard, T. J., Shen, L., Lehman, L. H., Feng, M., Ghassemi, M., Moody, B., Szolovits, P., Celi, L. A., & Mark, R. G. (2016). MIMIC-III, a freely accessible critical care database. Scientific Data, 3, 160035. https://doi.org/10.1038/sdata.2016.35

Litvak, E., & Bisognano, M. (2011). More patients, less payment: Increasing hospital efficiency in the aftermath of health reform. Health Affairs, 30(1), 76–80. https://doi.org/10.1377/hlthaff.2010.1038

Lu, Y. (2017). Industry 4.0: A survey on technologies, applications and open research issues. Journal of Industrial Information Integration, 6, 1–10. https://doi.org/10.1016/j.jii.2017.04.005

Lu, Y. (2019). Artificial intelligence: A survey on evolution, models, applications and future trends. Journal of Management Analytics, 6(1), 1–29. https://doi.org/10.1080/23270012.2019.1570365

Mannhardt, F., de Leoni, M., Reijers, H. A., & van der Aalst, W. M. P. (2016). Balanced multi-perspective checking of process conformance. Computing, 98(4), 407–437. https://doi.org/10.1007/s00607-015-0441-1

Miotto, R., Wang, F., Wang, S., Jiang, X., & Dudley, J. T. (2018). Deep learning for healthcare: Review, opportunities and challenges. Briefings in Bioinformatics, 19(6), 1236–1246. https://doi.org/10.1093/bib/bbx044

Obermeyer, Z., & Emanuel, E. J. (2016). Predicting the future—Big data, machine learning, and clinical medicine. New England Journal of Medicine, 375(13), 1216–1219. https://doi.org/10.1056/NEJMp1606181

Pollard, T. J., Johnson, A. E. W., Raffa, J. D., Celi, L. A., Mark, R. G., & Badawi, O. (2018). The eICU Collaborative Research Database, a freely available multi-center database for critical care research. Scientific Data, 5, 180178. https://doi.org/10.1038/sdata.2018.178

Rajkomar, A., Dean, J., & Kohane, I. (2019). Machine learning in medicine. New England Journal of Medicine, 380(14), 1347–1358. https://doi.org/10.1056/NEJMra1814259

Rajpurkar, P., Chen, E., Banerjee, O., & Topol, E. J. (2022). AI in health and medicine. Nature Medicine, 28(1), 31–38. https://doi.org/10.1038/s41591-021-01614-0

Rojas, E., Munoz-Gama, J., Sepúlveda, M., & Capurro, D. (2016). Process mining in healthcare: A literature review. Journal of Biomedical Informatics, 61, 224–236. https://doi.org/10.1016/j.jbi.2016.04.007

Saria, S., Rajani, A. K., Gould, J., Koller, D., & Penn, A. A. (2010). Integration of early physiological responses predicts later illness severity in preterm infants. Science Translational Medicine, 2(48), 48ra65. https://doi.org/10.1126/scitranslmed.3001304

Shickel, B., Tighe, P. J., Bihorac, A., & Rashidi, P. (2018). Deep EHR: A survey of recent advances in deep learning techniques for electronic health record (EHR) analysis. IEEE Journal of Biomedical and Health Informatics, 22(5), 1589–1604. https://doi.org/10.1109/JBHI.2017.2767063

Shortliffe, E. H., & Sepúlveda, M. J. (2018). Clinical decision support in the era of artificial intelligence. JAMA, 320(21), 2199–2200. https://doi.org/10.1001/jama.2018.17163

Topol, E. J. (2019). High-performance medicine: The convergence of human and artificial intelligence. Nature Medicine, 25(1), 44–56. https://doi.org/10.1038/s41591-018-0300-7

van der Aalst, W. M. P. (2016). Process Mining: Data Science in Action (2nd ed.). Springer. https://doi.org/10.1007/978-3-662-49851-4

Xiao, C., Choi, E., & Sun, J. (2018). Opportunities and challenges in developing deep learning models using electronic health records. Journal of the American Medical Informatics Association, 25(10), 1419–1428. https://doi.org/10.1093/jamia/ocy068

Zhang, C., & Lu, Y. (2021). Study on artificial intelligence: The state of the art and future prospects. Journal of Industrial Information Integration, 23, 100224. https://doi.org/10.1016/j.jii.2021.100224

Denny, J. C., Bastarache, L., Ritchie, M. D., Carroll, R. J., Zink, R., Mosley, J. D., Field, J. R., Pulley, J. M., Ramirez, A. H., Bowton, E., Basford, M. A., Carrell, D. S., Peissig, P. L., Kho, A. N., Pacheco, J. A., Rasmussen, L. V., Crosslin, D. R., Crane, P. K., Pathak, J., & Crawford, D. C. (2013). Systematic comparison of phenome-wide association study of electronic medical record data and genome-wide association study data. Nature Biotechnology, 31(12), 1102–1110. https://doi.org/10.1038/nbt.2749