The CRITICAL Records Integrated Standardization Pipeline (CRISP): End-to-End Processing of Large-scale Multi-institutional OMOP CDM Data

Xiaolong Luo XIAOLONGLUO@FAS.HARVARD.EDU
School of Engineering and Applied Sciences, Harvard University, Cambridge, MA 02138, USA

Michael Lingzhi Li MILI@HBS.EDU
Harvard Business School, Harvard University, Boston, MA 02163, USA

Abstract

Large-scale critical care datasets have driven major progress in clinical AI, yet most remain limited to single institutions. The newly released CRITICAL dataset expands this scope, linking 1.95 billion records from 371,365 patients across four CTSA sites and capturing longitudinal patient journeys from pre-ICU to post-ICU care. Its scale and diversity enable more generalizable modeling but introduce significant challenges in data cleaning, vocabulary harmonization, and computational efficiency.

We introduce CRISP (CRITICAL Records Integrated Standardization Pipeline), a scalable framework that transforms the raw CRITICAL resource into machine-learning–ready form. CRISP performs systematic data validation, cross-vocabulary mapping, and unit standardization while maintaining full auditability. Through parallelized optimization, it processes the entire dataset in under a day on standard computing hardware. The pipeline also provides reproducible baselines across multiple clinical prediction tasks, substantially reducing data preparation time and enabling consistent, multi-institutional evaluation. All code, documentation, and benchmarks are publicly available to support transparent and scalable clinical AI research.

Keywords: Electronic Health Records, CRITICAL Dataset, Intensive Care Unit, Multi-institutional Data, Critical Care, Data Processing Pipeline, Healthcare AI, Real-world Data

Data and Code Availability The CRITICAL dataset is available under a data use agreement at https://critical.fsm.northwestern.edu. CRISP source code, documentation, and processing scripts are publicly available at https://github.com/AaronLuo00/CRISP-Pipeline.

Institutional Review Board (IRB) This research was reviewed and determined by the Harvard University Institutional Review Board to be Not Human Subjects Research (IRB Protocol #IRB25-0399 and #IRB25-0870).

1. Introduction

The rapid advancement of artificial intelligence (AI) has revolutionized healthcare through its integration with electronic health records (EHRs), enabling unprecedented capabilities in clinical prediction, diagnosis, and decision support. Recent breakthroughs have demonstrated the ability to accurately predict multiple medical events from EHR data (Rajkomar et al., 2018; Jiang et al., 2023; Grout et al., 2024; Hegselmann et al., 2025), with models achieving performance comparable to clinical experts in various domains including mortality prediction, disease diagnosis, and treatment recommendation. These successes have increased interest in developing AI systems that can assist clinicians in real-time decision-making (Tomasev et al., 2019), predict patient deterioration hours before clinical manifestation (Lauritsen et al., 2020), and optimize resource allocation in increasingly strained critical care settings where timely intervention can significantly impact patient outcomes (Komorowski et al., 2018; Gutierrez, 2020).

However, the promise of AI in healthcare critically depends on access to large-scale, diverse, and well-structured clinical data, a fundamental challenge that limit the development of truly generalizable AI models (Futoma et al., 2020; Kelly et al., 2019). Most existing models are trained on single-institution datasets, raising concerns about their transferability across different healthcare systems, patient populations, and clinical practice patterns (Zech et al., 2018; Nestor et al., 2019). Furthermore, the hetero-

geneity in data collection practices, vocabulary usage, and documentation standards across institutions creates a substantial barrier to developing robust, multi-institutional AI systems that can benefit diverse patient populations (Gianfrancesco et al., 2018; Obermeyer et al., 2019). This challenge is particularly acute in critical care settings, where the complexity of patient conditions, the diversity of monitoring equipment, and the urgency of clinical decisions demand models that can generalize across varying institutional protocols and patient demographics (Sendak et al., 2020; Shah et al., 2019).

The Observational Medical Outcomes Partnership (OMOP) Common Data Model (CDM) responds to this fragmentation as an open community standard that harmonizes the structure and content of observational data while pairing the schema with Observational Health Data Sciences and Informatics (OHDSI) standardized vocabularies to enable reliable, large-scale analytics (Hripcsak et al., 2015). By transforming disparate EHR and claims databases into a common representation, OMOP allows researchers to run characterization, population-level effect estimation, and patient-level prediction studies with the same analytic routines across institutions. Adoption has accelerated globally, with surveys reporting more than 450 databases—spanning hundreds of millions of patient records across North America, Europe, and Asia—now operating in the CDM (Hallinan et al., 2024; Reinecke et al., 2021; Ahmadi et al., 2022). Within this ecosystem, CRITICAL is the first publicly accessible longitudinal, multi-institutional EHR dataset released entirely in OMOP CDM format, motivating the CRISP framework introduced next.

1.1. The Data Challenge in Healthcare AI

Pioneering datasets like MIMIC-III and MIMIC-IV have established the foundation for critical care AI research, with MIMIC-IV comprising over 65,000 intensive care unit (ICU) patients and more than 200,000 emergency department (ED) patients, enabling groundbreaking advances in mortality prediction, treatment optimization, and clinical decision support (Johnson et al., 2016, 2023). The eICU Collaborative Research Database further expanded the field by demonstrating the value of multi-center data (∼200,000 ICU admissions across 208 hospitals), pioneering cross-institutional research approaches (Pollard et al., 2018). These foundational resources have trained a generation of researchers and established methodological standards that continue to guide the field.

Building upon these essential contributions, the CRITICAL dataset extends the research landscape by providing 1.95 billion records from 371,365 patients across four Clinical and Translational Science Awards (CTSA) sites, offering complementary strengths including full-spectrum patient journeys (pre-ICU, ICU, and post-ICU), extended longitudinal tracking, and diverse geographic representation (The CRITICAL Consortium, 2025). While this scale and diversity enable more generalizable modeling, they also introduce cross-site semantic heterogeneity—differences in vocabulary usage, coding practices, units, and temporal granularity—demanding transparent, reproducible data preprocessing for cleaning, standardization, and harmonization. CRISP addresses these challenges by providing a modular, reusable pipeline that not only accelerates research but also ensures consistency with the methodological standards established by the MIMIC and eICU communities.

1.2. Our Contributions

To address these critical needs for standardization and unified data formats, we present CRISP (CRITICAL Records Integrated Standardization Pipeline), a comprehensive solution that transforms CRITICAL's 1.95 billion raw records into ML-ready formats. Our contributions include:

(1) Five-stage preprocessing pipeline that systematically transforms raw OMOP data through exploratory analysis, data cleaning, vocabulary mapping, standardization, and ICU cohort extraction with comprehensive audit trails ensuring reproducibility.

(2) Scalable parallel architecture that processes the entire 278.97 GB dataset in under 24 hours using 12 CPU cores and 64GB RAM through optimized chunked processing and parallel optimization, making large-scale multi-institutional data processing accessible to resource-constrained research teams.

(3) Comprehensive benchmarks across four critical prediction tasks using multiple model architectures, establishing reproducible baselines for the research community.

(4) Open-source implementation with complete code, documentation, and processed datasets, saving researchers months of preprocessing effort.

2. Related Work

2.1. Clinical Data Processing Pipelines

As clinical datasets grow in scale and complexity, the need for standardized, reusable pipelines becomes increasingly critical. In particular, there has been many processing pipelines designed for MIMIC-III, MIMIC-IV, eICU and other EHR datasets. Notable contributions include MIMIC-Extract (Wang et al., 2020) for cohort extraction, multitask benchmarks (Harutyunyan et al., 2019) establishing standard prediction tasks, COP-E-CAT (Mandyam et al., 2021) for modular preprocessing, an extensive MIMIC-IV pipeline (Gupta et al., 2022), METRE (Liao and Voldman, 2023) for cross-database validation, and reproducibility MIMIC benchmark (Purushotham et al., 2018).

These pipelines have established a solid foundation for processing single-institution EHR datasets and have achieved remarkable success within their respective domains. However, extending them to multi-institutional environments poses several challenges. Existing pipelines are typically optimized for dataset-specific schemas (e.g., MIMIC's custom structure or eICU's format). Additionally, they often assume a single-vocabulary system, whereas multi-institutional environments incorporate multiple overlapping vocabularies requiring sophisticated cross-vocabulary harmonization. Finally, many of these pipelines utilize single-threaded architectures that are sufficient for moderate-scale datasets, but inadequate for multi-site, billion-row CDM tables.

CRISP builds upon these prior efforts by introducing a parallelized pipeline tailored specifically for a multi-institutional setting. It incorporates systematic data cleaning, schema standardization, and multi-vocabulary harmonization to enable large-scale, multi-institutional processing. Our implementation achieves a 4–6× speedup compared to serial execution and completes full-dataset processing in approximately 20 hours on commodity hardware. We also release reproducible benchmarks covering both traditional machine learning models and deep learning architectures.

2.2. Broader Context and Challenges

Multi-institutional Harmonization and Standardization: The heterogeneity of medical vocabularies across institutions creates fundamental challenges for multi-site data integration. Henke et al. (2024) proposed systematic harmonization approaches to address schema heterogeneity across OMOP implementations, identifying multiple sources of incompatibility requiring comprehensive harmonization processes. Wang et al. (2025) examined OMOP CDM adoption challenges in specialized domains like oncology, revealing significant gaps in cancer-specific concept coverage. This vocabulary heterogeneity leads to severe feature matrix sparsity: when identical clinical concepts are encoded differently across sites, each feature is populated only by a subset of institutions. The resulting matrices are dominated by missing values, introducing noise and degrading model performance (Che et al., 2018). As demonstrated in Figure 2, this vocabulary heterogeneity is particularly pronounced in multi-institutional datasets. CRISP addresses this challenge by systematically analyzing vocabulary distributions in the CRITICAL dataset, constructing cross-vocabulary mappings to unified Systematized Nomenclature of Medicine Clinical Terms (SNOMED-CT[1]) standards, and standardizing different units to Unified Code for Units of Measure (UCUM) specifications (Schadow and McDonald, 2009), thereby consolidating fragmented features into dense, semantically consistent representations essential for robust multi-institutional EHR model training.

Clinical Benchmarks: Foundational work established standard prediction tasks (e.g., mortality, length-of-stay) and demonstrated that preprocessing strongly affects performance (Johnson et al., 2017; Rocheteau et al., 2021). Building on this, we release reproducible benchmarks over CRITICAL using CRISP-processed OMOP data, enabling fair comparisons across models.

3. The CRITICAL Dataset

The CRITICAL dataset represents the first cross-CTSA initiative to create a multi-site, multi-modal, de-identified clinical dataset combining both deep longitudinal coverage and broad institutional diversity. Developed collaboratively across four CTSA sites (Northwestern, Tufts, Washington University in St. Louis, and University of Alabama at Birmingham), CRITICAL encompasses 1.95 billion records from 371,365 patients (The CRITICAL Consortium, 2025), establishing it as the largest publicly shared, disease-independent benchmarking dataset for criti-

1. Hereafter refer as SNOMED

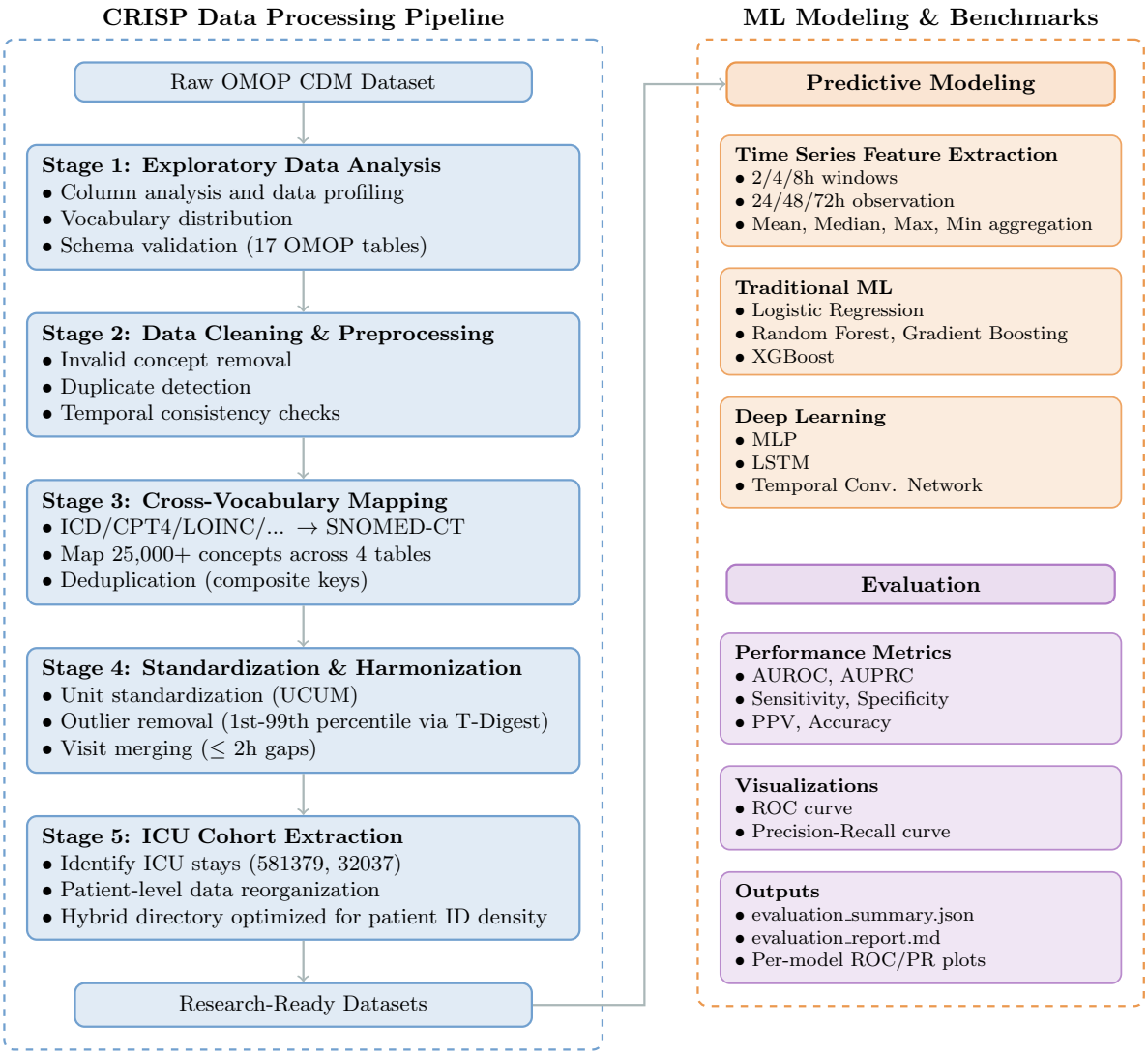

Figure 1: CRISP Pipeline Architecture: Five-stage data processing pipeline for the CRITICAL dataset.

cal care research. Built on OMOP CDM v5.3, the repository spans 17 tables totaling 278.97 GB (Table 4), with MEASUREMENT alone containing 1.4 billion rows. The dataset includes 38 million visits and 28 million unit-level records, averaging 5,242 rows per patient across all tables.

CRITICAL provides comprehensive patient care journeys with a median observation period of 3.11 years and maximum spanning 31.8 years (Table 5). This extensive temporal coverage captures pre-ICU, ICU, and post-ICU encounters across both inpatient and outpatient settings, with patients averaging 102.3 visits throughout their observation periods. This multi-institutional, longitudinal perspective introduces substantial vocabulary heterogeneity—150,671 unique source concepts across 30 vocabularies[2], with SNOMED alone accounting for 58.0% (87,453 concepts). Figure 2 illustrates this vocabulary heterogeneity across major tables within the dataset (see Appendix B for detailed distribution analysis), requiring systematic harmonization to unified standards.

2. After deduplication across related tables, the dataset contains over 110,000 unique clinical concepts.

4. Data Pipeline Overview

4.1. Pipeline Architecture

To harness the CRITICAL dataset's scale and multi-institutional diversity described in Section 3, CRISP employs a five-stage processing framework that transforms the raw OMOP CDM tables into ML-ready dataset. The pipeline architecture consists of two primary components: (1) a core data processing module executing sequential stages of exploratory analysis, data cleaning, cross-vocabulary mapping, standardization, and patient data extraction with label generation; and (2) an optional predictive modeling module offering baseline implementations and evaluation benchmarks. This modular design allows researchers to utilize individual stages independently, customize processing parameters, or extend the pipeline with task-specific modification. (Figure 1).

The implementation leverages parallel processing strategies across all computationally intensive operations. Through chunked data loading and concurrent table processing, the pipeline handles billion-row tables within memory constraints while maintaining processing efficiency. Every transformation generates detailed audit trails—tracking removed records, vocabulary mappings, unit conversions, and outlier statistics—enabling complete reproducibility. This architecture processes the entire 278.97 GB CRITICAL dataset in under 24 hours using standard computational resources (12 CPU cores, 64GB RAM).

4.2. Five-Stage Processing Pipeline

Stage 1: Exploratory Data Analysis. This stage generates comprehensive dataset statistics that guide subsequent processing modules and provide researchers with detailed data understanding. The pipeline analyzes all 17 OMOP tables, producing: (1) column-level missingness analysis, automatically flagging columns with >95% missing values for removal to simplify downstream processing; (2) table-level summaries including row counts (1.95 billion total), unique patient counts ,memory usage, and temporal coverage (date ranges for each table); and (3) population-level statistics such as ICU admission rates, mortality rates , gender distributions, and age ranges. All statistics are exported as structured JSON files that subsequent modules consume to parameterize their operations.

Stage 2: Data Cleaning and Preprocessing. Building on Stage 1's column analysis, this stage systematically cleans 14 tables[3] by addressing data quality issues. The pipeline performs three parallel cleaning operations: (1) *invalid concept removal* filters out records where the primary concept ID field (e.g., measurement_concept_id, procedure_concept_id) is null, empty, or zero—these represent unmapped or invalid clinical codes that cannot be interpreted, eliminating approximately 2-5% of records; (2) *duplicate elimination* identifies and removes redundant records using composite keys (person_id + concept_id + datetime), preventing the same clinical event from being counted multiple times; and (3) *temporal validation* ensures chronological consistency by verifying start times precede end times, removing records with future dates or impossible sequences. All removed records are archived in structured directories with detailed logs, enabling quality assessment and potential recovery.

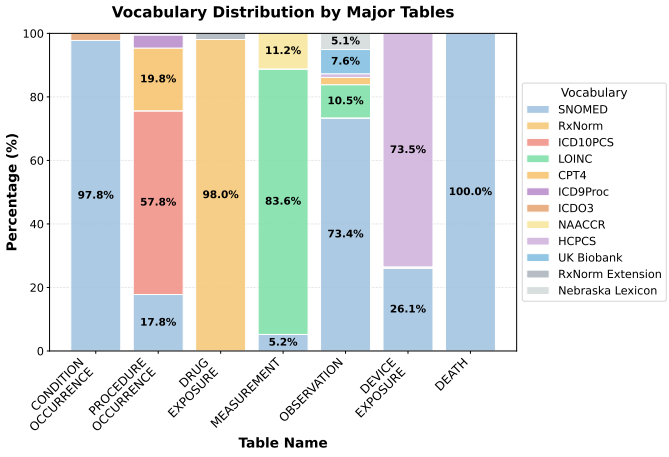

Figure 2: Vocabulary distribution across major OMOP tables demonstrates significant vocabulary heterogeneity.

Stage 3: Cross-Vocabulary Concept Mapping. This stage addresses the fundamental challenge of vocabulary heterogeneity in multi-institutional datasets, where different sites may use entirely different coding systems for the same clinical concepts. SNOMED was selected as the target vocabulary because it already represents the majority of concepts both globally (58.0% of all concepts) and within most individual tables (Figure 2), mini-

3. The three tables not processed are LOCATION, CARE_SITE, and PROVIDER, as they either lack meaningful information or are independent of patient-level data.

mizing the required mapping effort while maximizing coverage. The pipeline targets four key tables[4] that exhibit the most complex vocabulary heterogeneity—for instance, PROCEDURE_OCCURRENCE contains concepts from ICD10PCS (57.8%), CPT4 (19.8%), and SNOMED (15.4%) as shown in Figure 2. These tables are critical for clinical prediction yet suffer from severe fragmentation without harmonization. The pipeline applies pre-computed crosswalks to map diverse source vocabularies (including ICD9CM, ICD10CM, ICD10PCS, CPT4, LOINC, HCPCS, RxNorm, and others shown in Figure 3) to unified SNOMED codes, harmonizing over 25,000 unique source concepts. Post-mapping deduplication removes redundancies from many-to-one mappings using composite keys (person_id + SNOMED_id + datetime), consolidating multiple source representations of the same clinical concept.

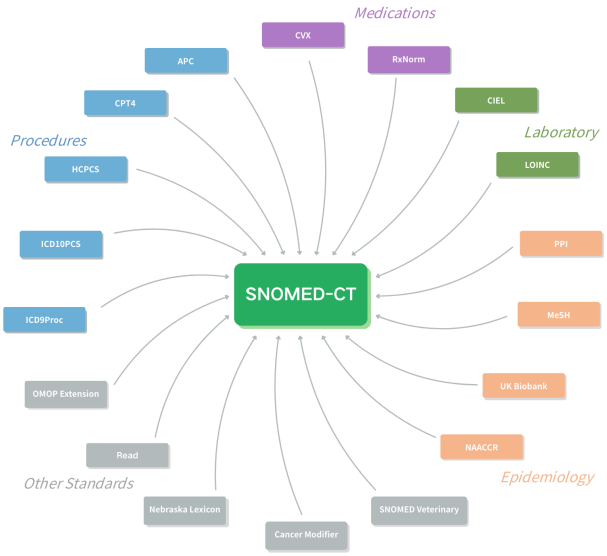

Figure 3: Vocabulary harmonization: mapping diverse medical terminologies to unified SNOMED standards.

Stage 4: Data Standardization and Unit Harmonization. This stage ensures measurement consistency and temporal coherence across the harmonized dataset. Processing tables from Stage 3's output, the pipeline performs four key standardization operations. First, *outlier removal* applies T-Digest algorithms (Dunning and Ertl, 2019) specifically to MEASUREMENT tables, computing memory-efficient percentiles across 1.4 billion records and filtering out outliers beyond the 1st and 99th percentiles[5]. Second, *unit standardization* converts heterogeneous measurement units to UCUM standards—for example, temperature from Fahrenheit to Celsius, weight from pounds to kilograms, and height from inches to centimeters—while removing physiologically implausible values (illustrated in Figure 4). Third, *visit consolidation* merges fragmented VISIT_DETAIL and VISIT_OCCURRENCE records within 2-hour windows, reconstructing continuous care episodes from 66 million visit records. Fourth, *data type standardization* ensures consistent representation—NaN for missing values, integers for IDs, floats for measurements, and ISO format for datetimes. After Stage 4 completion, all critical tables have unified concepts, with missing, erroneous, and implausible values removed, and data formats standardized for machine learning readiness.

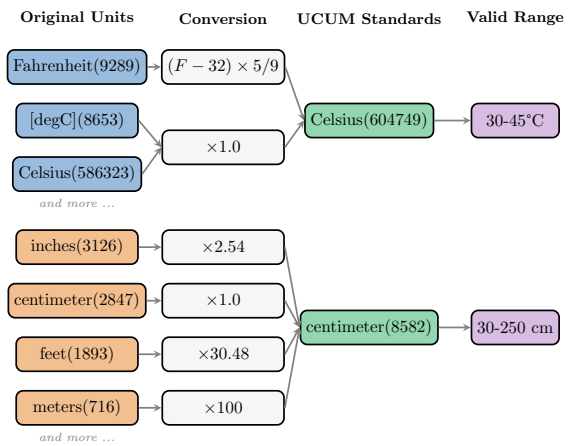

Figure 4: Unit standardization pipeline: converting diverse measurement units to UCUM standards with outlier filtering.

Stage 5: Patient Data Extraction and Label Generation. This final stage transforms table-centric OMOP data into patient-centric structures required for machine learning. This stage performs three key operations. First, *patient-level aggregation*

4. MEASUREMENT, OBSERVATION, PROCEDURE_OCCURRENCE, DEVICE_EXPOSURE

5. Configurable parameters, following the approach of Gupta et al. (2022).

reorganizes 1.95 billion records into 371,365 individual patient directories, consolidating each patient's complete medical history. Second, *ICU cohort identification* scans VISIT_DETAIL tables for ICU-specific concept IDs[6], and generating temporal marks for different patient stages (pre-ICU, during-ICU, post-ICU). The extraction employs parallel chunked processing to efficiently handle billion-scale data while maintaining patient-level integrity.

The resulting patient-indexed structure is organized using a hybrid directory system (Figure 5) that adapts to the uneven distribution of 371,365 patient IDs. Low-density prefixes (<30,000 patients) use direct folder organization, while high-density prefixes like 600000071 (>30,000 patients) employ subdirectory layering (e.g., 002000-002999) to prevent file system degradation. This adaptive structure ensures efficient I/O performance while enabling direct patient queries for diverse ML tasks.

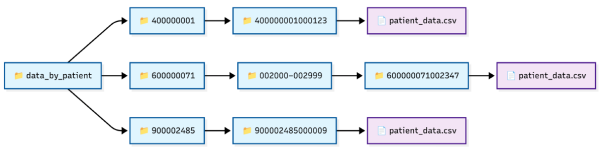

Figure 5: Hybrid directory structure adapting to patient ID density for optimal file system performance.

4.3. Computational Performance Analysis

CRISP leverages parallel processing across all five pipeline stages to handle large-scale clinical data efficiently. Each stage employs configurable worker pools (default 8 workers) for concurrent processing. This comprehensive parallelization achieves 4-6× speedup compared to sequential processing, reducing total pipeline execution time from approximately 4 days to ∼20 hours on standard hardware (12-core CPU, 64GB RAM). The substantial acceleration, combined with memory-efficient strategies like T-Digest for percentile computation, chunked I/O operations, and the hybrid directory structure design, makes large-scale multi-institutional data processing feasible for resource-constrained research teams.

6. (Using concept ID: 581379 for ICU stays, 32037 for critical care)

5. Benchmark Tasks and Models

To validate CRISP's effectiveness and establish reproducible baselines, we conduct comprehensive benchmark experiments on the harmonized CRITICAL dataset. Our evaluation framework tests both traditional and deep learning models across multiple clinical prediction tasks, revealing current performance limitations and opportunities for methodological advancement.

5.1. Experimental Setup and Methodology

Following MIMIC-Extract (Wang et al., 2020), we select the 800 most frequent clinical concepts from the harmonized dataset, extracting features from five key tables (MEASUREMENT, OBSERVATION, DRUG_EXPOSURE, CONDITION_OCCURRENCE, PROCEDURE_OCCURRENCE). The observation window spans the first 24 hours of ICU admission, discretized into 4-hour bins to capture temporal dynamics. We evaluate eight model architectures spanning traditional ML (Logistic Regression, Random Forest, Gradient Boosting, XGBoost) and deep learning approaches (Multi-Layer Perceptron (MLP), Long Short-Term Memory networks (LSTM), Temporal Convolutional Networks (TCN), and a Transformer encoder (TRF)). Detailed LSTM, TCN, and Transformer configurations are provided in Appendix C. All models employ 5-fold cross-validation, with the final results reported using 80% of data for training and 20% for testing, and we additionally run leave-one-site-out evaluations to quantify cross-site robustness (Section D).

5.2. Clinical Prediction Tasks

Inspired by Gupta et al. (2022), we define four binary classification tasks following standard ICU outcome prediction benchmarks. All tasks use features extracted from the first 24 hours of ICU admission, and patients with ICU stays shorter than 24 hours are excluded to ensure complete temporal coverage.:

Mortality Prediction: Predicting in-hospital mortality within 7 and 30 days after ICU admission using the first 24 hours of data.

Length of Stay: Predicting prolonged ICU stays exceeding 3 and 7 days based on the first 24 hours of ICU measurements.

Readmission Risk: Predicting 7-day, 30-day, and 90-day hospital readmissions using the same 24-

Table 1: Clinical Prediction Performance (AUROC)

Task Category	Prediction Target	LR	RF	GB	XGB	MLP	LSTM	TCN	TRF
Mortality	7-day	0.684	0.676	0.763	0.781	0.814	0.697	0.705	0.721
	30-day	0.708	0.732	0.802	0.804	0.835	0.764	0.778	0.746
Length of Stay	LOS $>$ 3 days	0.651	0.702	0.732	0.735	0.756	0.689	0.711	0.693
	LOS $>$ 7 days	0.666	0.702	0.746	0.748	0.767	0.687	0.719	0.735
Readmission	7-day	0.641	0.703	0.739	0.755	0.748	0.663	0.681	0.668
	30-day	0.619	0.699	0.726	0.735	0.743	0.652	0.698	0.683
	90-day	0.635	0.695	0.741	0.746	0.737	0.663	0.702	0.633
In ICU Sepsis	After ICU	0.879	0.895	0.897	0.898	0.883	0.884	0.871	0.887
	Within 48h	0.836	0.870	0.883	0.882	0.912	0.799	0.845	0.835
	Within 7 days	0.904	0.899	0.904	0.908	0.902	0.852	0.876	0.901

Table 2: Clinical Prediction Performance (LOSO Site 4; AUROC)

Task Category	Prediction Target	LR	RF	GB	XGB	MLP	LSTM	TCN	TRF
Mortality	7-day	0.609	0.594	0.719	0.733	0.518	0.654	0.575	0.701
	30-day	0.611	0.626	0.729	0.762	0.563	0.659	0.581	0.743
Length of Stay	LOS $>$ 3 days	0.587	0.683	0.704	0.705	0.594	0.581	0.579	0.651
	LOS $>$ 7 days	0.605	0.644	0.747	0.750	0.595	0.610	0.592	0.689
Readmission	7-day	0.507	0.598	0.596	0.610	0.494	0.510	0.540	0.520
	30-day	0.514	0.608	0.628	0.627	0.502	0.511	0.489	0.487
	90-day	0.505	0.626	0.650	0.658	0.509	0.531	0.486	0.540
In ICU Sepsis	After ICU	0.881	0.877	0.868	0.863	0.760	0.844	0.826	0.752
	Within 48h	0.573	0.875	0.884	0.804	0.537	0.451	0.545	0.602
	Within 7 days	0.569	0.900	0.879	0.847	0.546	0.638	0.504	0.600

hour admission window. Readmission outcomes are defined after discharge.

Sepsis Onset: Detecting sepsis development after ICU admission (post-ICU, within 48 hours, within 7 days) using only pre-ICU static covariates to avoid temporal overlap.

All non-sepsis tasks therefore share an identical 24-hour observation period, enabling consistent benchmarking across outcomes. The combined 800-feature vector (400 MEASUREMENT, 200 OBSERVATION, 100 DRUG_EXPOSURE, 50 CONDITION_OCCURRENCE, 50 PROCEDURE_OCCURRENCE) balances clinical coverage.

The comprehensive results on the full CRITICAL dataset (Table 1) reveal significant room for improvement in multi-institutional clinical prediction. Even with CRISP's harmonization and sparsity reduction, the best models achieve only 0.619–0.755 AUROC for readmission tasks, highlighting the inherent complexity of predicting patient trajectories across heterogeneous institutions. Leave-one-site-out experiments (Table 2 and Appendix D) reinforce this difficulty: AUROCs drop whenever Sites 4, 6, or 9 are held out because the training pool shrinks while residual differences in documentation practices and temporal granularity persist despite CRISP's harmonization. This performance gap motivates further research into advanced feature selection and utilization strategies, novel model architectures, as well as leveraging the extensive pre-ICU information uniquely available in CRITICAL.

6. Discussion and Conclusion

Core Contribution. CRISP is the first end-to-end processing pipeline specifically designed for the large-scale multi-institutional CRITICAL dataset. The pipeline systematically addresses the CRITICAL dataset's complexity—150,671 unique concepts across 30 vocabularies and 1.95 billion records—through five integrated stages: (1) systematic exploratory data analysis (2) comprehensive cleaning that removes invalid concepts, duplicates, and temporal inconsistencies across 14 tables; (3) cross-vocabulary mapping that harmonizes four key tables to unified SNOMED standards; (4) data stan-

dardization with outlier removal, unit conversion, and visit merging; and (5) patient-centric extraction that generates ML-ready features. Through optimized parallel processing and chunking strategies, CRISP achieves 4-6× speedup over sequential approaches, processing the entire 278.97 GB dataset in approximately 20 hours on standard hardware. The pipeline also provides comprehensive ML benchmarks across seven model architectures and four clinical prediction tasks, establishing reproducible baselines for future research.

Broader Impact. By making the groundbreaking CRITICAL dataset immediately accessible, CRISP democratizes multi-institutional healthcare AI research. The pipeline transforms months of manual data curation into a ready-to-deploy solution, lowering the barrier from requiring specialized data engineering expertise to simply executing pre-configured scripts. This accessibility significantly lowers the barrier to entry, enabling researchers to focus on developing and testing innovative models rather than wrestling with data preprocessing challenges. CRISP's comprehensive infrastructure enables the research community to quickly leverage CRITICAL's unprecedented scale and diversity, accelerating progress toward robust, generalizable clinical AI systems. The combination of systematic data processing, concept harmonization, and reproducible benchmarks establishes a foundation for collaborative advancement in cross-institutional healthcare ML.

Limitations and Future Work. While current vocabulary mapping prioritizes four critical tables that exhibit the most complex heterogeneity and are essential for clinical prediction, future releases will progressively extend mapping coverage to all tables requiring harmonization. The pipeline currently employs empirically-derived parameters—such as 99th percentile outlier thresholds and 2-hour visit merging windows—that provide robust general-purpose processing but may not be optimal for specific clinical tasks. Future work may develop task-adaptive parameter selection, leverage CRITICAL's unique longitudinal coverage spanning pre-ICU, ICU, and post-ICU periods to explore novel prediction tasks with high clinical value, and investigate advanced feature utilization strategies for the dataset's extensive concepts. We are actively extending CRISP to support more OMOP datasets and will update the public release whenever CRITICAL or other OMOP resources include unstructured text or imaging modalities.

References

Najia Ahmadi, Yuan Peng, Markus Wolfien, Michéle Zoch, and Martin Sedlmayr. Omop cdm can facilitate data-driven studies for cancer prediction: a systematic review. *International journal of molecular sciences*, 23(19):11834, 2022.

Zhengping Che, Sanjay Purushotham, Kyunghyun Cho, David Sontag, and Yan Liu. Recurrent neural networks for multivariate time series with missing values. *Scientific reports*, 8(1):6085, 2018.

Ted Dunning and Otmar Ertl. Computing extremely accurate quantiles using t-digests. *arXiv preprint arXiv:1902.04023*, 2019.

Joseph Futoma, Morgan Simons, Trishan Panch, Finale Doshi-Velez, and Leo Anthony Celi. The myth of generalisability in clinical research and machine learning in health care. *The Lancet Digital Health*, 2(9):e489–e492, 2020. doi: 10.1016/S2589-7500(20)30186-2.

Milena A Gianfrancesco, Suzanne Tamang, Jinoos Yazdany, and Gabriela Schmajuk. Potential biases in machine learning algorithms using electronic health record data. *JAMA Internal Medicine*, 178(11):1544–1547, 2018. doi: 10.1001/jamainternmed.2018.3763.

Robert Grout, Rishab Gupta, Ruby Bryant, Mawada A Elmahgoub, Yijie Li, Khushbakht Irfanullah, Rahul F Patel, Jake Fawkes, and Catherine Inness. Predicting disease onset from electronic health records for population health management: a scalable and explainable deep learning approach. *Frontiers in Artificial Intelligence*, 6:1287541, 2024.

Mehak Gupta, Brennan Gallamoza, Nicolas Cutrona, Pranjal Dhakal, Raphael Poulain, and Rahmatollah Beheshti. An extensive data processing pipeline for mimic-iv. In *Machine learning for health*, pages 311–325. PMLR, 2022.

Gabriel Gutierrez. Artificial intelligence in the intensive care unit. *Critical Care*, 24(1):101, 2020. doi: 10.1186/s13054-020-2785-y.

Christine Mary Hallinan, Roger Ward, Graeme K Hart, Clair Sullivan, Nicole Pratt, Ashley P Ng, Daniel Capurro, Anton Van Der Vegt, Siaw-Teng Liaw, Oliver Daly, et al. Seamless emr data access: Integrated governance, digital health and the

omop-cdm. *BMJ health & care informatics*, 31(1): e100953, 2024.

Hrayr Harutyunyan, Hrant Khachatrian, David C Kale, Greg Ver Steeg, and Aram Galstyan. Multitask learning and benchmarking with clinical time series data. *Scientific data*, 6(1):96, 2019.

Stefan Hegselmann, Georg von Arnim, Tillmann Rheude, Noel Kronenberg, David Sontag, Gerhard Hindricks, Roland Eils, and Benjamin Wild. Large language models are powerful electronic health record encoders. *arXiv preprint arXiv:2502.17403*, 2025.

E Henke, M Zoch, Y Peng, et al. Conceptual design of a generic data harmonization process for omop common data model. *BMC Medical Informatics and Decision Making*, 24:58, 2024. doi: 10.1186/s12911-024-02458-7.

George Hripcsak, Jon D Duke, Nigam H Shah, Christian G Reich, Vojtech Huser, Martijn J Schuemie, Marc A Suchard, Rae Woong Park, Ian Chi Kei Wong, Peter R Rijnbeek, et al. Observational health data sciences and informatics (ohdsi): opportunities for observational researchers. *Studies in health technology and informatics*, 216:574, 2015.

Lavender Y Jiang, Shengyi Liu, Dan Chen, Will Ning, David Zhang, Joyce Kim, Roxana Daneshjou, Junyang Duan, Peyton Chen, Dmitriy Lituiev, et al. Health system-scale language models are all-purpose prediction engines. *Nature*, 619(7969): 357–362, 2023.

Alistair E. W. Johnson, Lucas Bulgarelli, Lu Shen, Alvin Gayles, Ayad Shammout, Steven Horng, Tom J. Pollard, Sicheng Hao, Benjamin Moody, Brian Gow, Li-wei H. Lehman, Leo A. Celi, and Roger G. Mark. Mimic-iv, a freely accessible electronic health record dataset. *Scientific Data*, 10 (1):1, 2023. doi: 10.1038/s41597-022-01899-x.

Alistair EW Johnson, Tom J Pollard, Lu Shen, Li-wei H Lehman, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G Mark. Mimic-iii, a freely accessible critical care database. *Scientific data*, 3 (1):1–9, 2016.

Alistair EW Johnson, Tom J Pollard, and Roger G Mark. Reproducibility in critical care: a mortality prediction case study. In *Machine Learning for Healthcare Conference*, pages 361–376. PMLR, 2017.

Christopher J Kelly, Alan Karthikesalingam, Mustafa Suleyman, Greg Corrado, and Dominic King. Key challenges for delivering clinical impact with artificial intelligence. *BMC Medicine*, 17(1):195, 2019. doi: 10.1186/s12916-019-1426-2.

Matthieu Komorowski, Leo A Celi, Omar Badawi, Anthony C Gordon, and A Aldo Faisal. The artificial intelligence clinician learns optimal treatment strategies for sepsis in intensive care. *Nature Medicine*, 24(11):1716–1720, 2018. doi: 10.1038/s41591-018-0213-5.

Simon Meyer Lauritsen, Mads Kristensen, Mathias Vassard Olsen, Mads Stenhuus Larsen, Katrine Meyer Lauritsen, Marianne Johansson Jørgensen, Jeppe Lange, and Bo Thiesson. Explainable artificial intelligence model to predict acute critical illness from electronic health records. *Nature Communications*, 11(1):3852, 2020. doi: 10.1038/s41467-020-17431-x.

Wei Liao and Joel Voldman. A multidatabase extraction pipeline (metre) for facile cross validation in critical care research. *Journal of Biomedical Informatics*, 141:104356, 2023.

Aishwarya Mandyam, Elizabeth C Yoo, Jeff Soules, Krzysztof Laudanski, and Barbara E Engelhardt. Cop-e-cat: cleaning and organization pipeline for ehr computational and analytic tasks. In *Proceedings of the 12th ACM International Conference on Bioinformatics, Computational Biology, and Health Informatics*, pages 1–9, 2021.

Bret Nestor, Matthew BA McDermott, Willie Boag, Gabriela Berner, Tristan Naumann, Michael C Hughes, Anna Goldenberg, and Marzyeh Ghassemi. Feature robustness in non-stationary health records: caveats to deployable model performance in common clinical machine learning tasks. In *Machine Learning for Healthcare Conference*, pages 381–405. PMLR, 2019.

Ziad Obermeyer, Brian Powers, Christine Vogeli, and Sendhil Mullainathan. Dissecting racial bias in an algorithm used to manage the health of populations. *Science*, 366(6464):447–453, 2019. doi: 10.1126/science.aax2342.

Tom J Pollard, Alistair EW Johnson, Jesse D Raffa, Leo A Celi, Roger G Mark, and Omar Badawi. The eicu collaborative research database, a freely available multi-center database for critical care research. *Scientific data*, 5(1):1–13, 2018.

Sanjay Purushotham, Chuizheng Meng, Zhengping Che, and Yan Liu. Benchmarking deep learning models on large healthcare datasets. *Journal of biomedical informatics*, 83:112–134, 2018.

Alvin Rajkomar, Eyal Oren, Kai Chen, Andrew M Dai, Nissan Hajaj, Michaela Hardt, Peter J Liu, Xiaobing Liu, Jake Marcus, Mimi Sun, et al. Scalable and accurate deep learning with electronic health records. *NPJ Digital Medicine*, 1(1):18, 2018.

Ines Reinecke, Michéle Zoch, Christian Reich, Martin Sedlmayr, and Franziska Bathelt. The usage of ohdsi omop–a scoping review. *German Medical Data Sciences 2021: Digital Medicine: Recognize–Understand–Heal*, pages 95–103, 2021.

Emma Rocheteau, Pietro Liò, and Stephanie Hyland. Temporal pointwise convolutional networks for length of stay prediction in the intensive care unit. In *Proceedings of the Conference on Health, Inference, and Learning*, pages 58–68. PMLR, 2021.

Gunther Schadow and Clement J McDonald. The unified code for units of measure. *Regenstrief Institute and UCUM Organization: Indianapolis, IN, USA*, page 99, 2009.

Mark P Sendak, Michael Gao, Nathan Brajer, and Suresh Balu. A path for translation of machine learning products into healthcare delivery. *EMJ Innovations*, 10:19–00172, 2020.

Pratik Shah, Francis Kendall, Sean Khozin, Ryan Goosen, Jianying Hu, Jason Laramie, Michael Ringel, and Nicholas Schork. Artificial intelligence and machine learning in clinical development: a translational perspective. *NPJ Digital Medicine*, 2 (1):69, 2019. doi: 10.1038/s41746-019-0148-3.

The CRITICAL Consortium. CRITICAL dataset: A large-scale, multi-site dataset for critical care research. https://critical.fsm.northwestern.edu, 2025. Accessed: 2025-08-28. Additional information available at https://amia.org/webinar-library/critical-consortium-and-dataset.

Nenad Tomasev, Xavier Glorot, Jack W Rae, Michal Zielinski, Harry Askham, Andre Saraiva, Anne Mottram, Clemens Meyer, Suman Ravuri, Ivan Protsyuk, et al. A clinically applicable approach to continuous prediction of future acute kidney injury. *Nature*, 572(7767):116–119, 2019. doi: 10.1038/s41586-019-1390-1.

L Wang, A Wen, S Fu, et al. A scoping review of omop cdm adoption for cancer research using real world data. *npj Digital Medicine*, 8:189, 2025. doi: 10.1038/s41746-025-01581-7.

Shirly Wang, Matthew BA McDermott, Geeticka Chauhan, Marzyeh Ghassemi, Michael C Hughes, and Tristan Naumann. Mimic-extract: a data extraction, preprocessing, and representation pipeline for mimic-iii. *Proceedings of the ACM Conference on Health, Inference, and Learning*, pages 222–235, 2020.

John R Zech, Marcus A Badgeley, Manway Liu, Anthony B Costa, Joseph J Titano, and Eric Karl Oermann. Variable generalization performance of a deep learning model to detect pneumonia in chest radiographs: a cross-sectional study. *PLoS Medicine*, 15(11):e1002683, 2018. doi: 10.1371/journal.pmed.1002683.

Appendix A. Dataset Demographics and Volume Statistics

This appendix provides detailed demographic and volume statistics that demonstrate the multi-institutional heterogeneity CRISP was designed to address. Table 3 presents the comprehensive demographic breakdown of 371,365 patients across four CTSA sites, showcasing the substantial diversity that validates CRISP's generalizability across different patient populations. The demographic representation across race, ethnicity, and age groups exemplifies the health equity research opportunities enabled by CRISP's systematic vocabulary harmonization, which ensures consistent feature representation across diverse institutional practices and coding standards.

Table 4 illustrates the scale of vocabulary harmonization challenges addressed by CRISP—processing 1.95 billion records across 17 OMOP CDM tables with systematic standardization. The MEASUREMENT table alone contains 1.4 billion records, representing the largest single source of clinical data and exemplifying the computational challenges that CRISP efficiently handles through parallel processing and intelligent chunking. These volumes, combined with the extensive temporal coverage shown in Table 5 (median observation period of 3.11 years, with 65.6% of patients having multi-year longitudinal data), demonstrate CRISP's ability to transform massive, heterogeneous multi-institutional data into ML-ready datasets within approximately 20 hours on standard hardware, democratizing access to large-scale critical care data for the broader research community.

Appendix B. Vocabulary Distribution Analysis

This section visualizes the vocabulary heterogeneity challenges that necessitate CRISP's cross-vocabulary mapping (Stage 3). The fragmentation of clinical concepts across multiple vocabularies creates sparse feature matrices that impede effective machine learning, making systematic harmonization to unified SNOMED standards essential.

Figure 6 shows the distribution of unique concepts across vocabularies, with SNOMED representing the majority while substantial portions use RxNorm, ICD10PCS, and other specialized terminologies. Figure 7 reveals the absolute concept counts within major OMOP tables, demonstrating how vocabulary usage varies significantly across clinical domains—from CONDITION_OCCURRENCE's 39,544 concepts to DRUG_ERA's focused 2,534 concepts. These visualizations underscore the complexity of harmonizing diverse medical terminologies across multi-institutional datasets.

Appendix C. Deep Learning Model Architectures

This section presents the detailed architectures of the deep learning models used in our benchmark experiments. The LSTM (Figure 8) and TCN (Figure 9) models employ hybrid architectures that integrate static patient features with temporal clinical measurements, enabling comprehensive representation learning from the multi-modal CRITICAL dataset. Our Transformer baseline uses three encoder layers with hidden size 256, eight attention heads, feed-forward width 1,024, and learnable positional encodings.

Appendix D. Cross-site Evaluation Tables

This appendix details the leave-one-site-out (LOSO) and site-held-out evaluations used to assess cross-institutional generalization. Across most prediction tasks, Transformer-based models achieve performance comparable to or slightly exceeding the best supervised baselines (Gradient Boosting, XGBoost), demonstrating that the CRISP-processed data support modern deep architectures. LOSO performance drops relative to the random-split setting for two primary reasons: (1) the training pool shrinks substantially while each held-out site remains large (Site 4: 54,603 patients; Site 6: 134,148; Site 9: 127,015), and (2) strong cross-site heterogeneity in documentation practices and temporal granularity makes transfer learning challenging even after concept and unit harmonization. Site 7 (˜55k patients) is excluded because timestamps are currently available only at the day level; we are coordinating with the CRITICAL Consortium to obtain full-resolution data and will incorporate those results once available.

Table 3: Demographic characteristics of 371,365 patients in the CRITICAL dataset across four CTSA sites, demonstrating the multi-institutional diversity in race, ethnicity, gender, and age distributions

Variable		Gender		Total	
		Male	Female	N	%
Race					
	Asian[a]	6,750 (55.17%)	5,483 (44.83%)	12,233	3.29%
	Black/African American	39,142 (52.62%)	35,237 (47.38%)	74,382	20.03%
	Native American	338 (55.23%)	274 (44.77%)	612	0.16%
	Pacific Islander[b]	426 (54.27%)	359 (45.73%)	785	0.21%
	White	141,993 (56.72%)	108,327 (43.28%)	250,328	67.41%
	Multiple Race	1,612 (59.01%)	1,120 (41.00%)	2,732	0.74%
	Unknown	9,217 (56.43%)	7,116 (43.57%)	16,337	4.40%
	Other/Refused	8,121 (58.19%)	5,830 (41.81%)	13,956	3.76%
Ethnicity					
	Hispanic/Latino	12,065 (56.76%)	9,195 (43.24%)	21,260	5.72%
	Not Hispanic/Latino	185,417 (55.78%)	146,949 (44.22%)	332,376	89.50%
	Unknown	10,117 (57.07%)	7,602 (42.93%)	17,729	4.77%
Age at First Visit[c]					
	<18	40,591 (55.34%)	32,753 (44.66%)	73,352	19.75%
	18-30	13,775 (53.60%)	11,926 (46.40%)	25,703	6.92%
	31-50	39,876 (56.62%)	30,545 (43.38%)	70,421	18.96%
	51-70	79,063 (58.41%)	56,306 (41.59%)	135,372	36.45%
	>70	34,294 (51.56%)	32,216 (48.44%)	66,517	17.91%
Visit Type[d]					
	Outpatient	7,542,966 (51.50%)	7,105,372 (48.50%)	14,648,554	52.37%
	Inpatient (non-ICU)	3,335,192 (53.02%)	2,954,536 (46.98%)	6,289,940	22.49%
	ICU	506,178 (57.08%)	380,685 (42.92%)	886,896	3.17%
	Emergency	385,437 (51.04%)	369,800 (48.96%)	755,243	2.70%
	Other	2,868,848 (53.21%)	2,522,983 (46.79%)	5,391,831	19.28%
Mortality[e]					
	Alive	162,062 (56.02%)	127,243 (43.98%)	289,321	77.91%
	Deceased	45,537 (55.50%)	36,503 (44.50%)	82,044	22.09%
Total[f]		207,599 (55.91%)	163,746 (44.09%)	371,365	100.00%

Notes:

[a] Asian includes: Asian (11,227), Asian Indian (506), Korean (206), Chinese (100), Japanese (83), Vietnamese (49), Filipino (25), Thai (24), Cambodian (13)

[b] Pacific Islander includes: Native Hawaiian (557), Native Hawaiian or Other Pacific Islander (228)

[c] Age calculated as (Earliest Visit Date - Birth Date) / 365.25

[d] Visit Type based on 27,972,464 visit details from VISIT_DETAIL table. ICU identified by concept IDs 32037, 581379

[e] Mortality status reflects patient vital status at last recorded encounter in the dataset

[f] Total includes 20 additional patients with unknown or missing gender concept IDs

Table 4: Data volume distribution across 17 OMOP CDM tables in the CRITICAL dataset, totaling 1.95 billion records, with average records per patient demonstrating the comprehensive clinical coverage

Table Name	Row Count	Size (GB)	% of Total	Per Patient
MEASUREMENT	1,403,627,644	194.00	72.08%	3,779.1
OBSERVATION	174,355,400	21.06	8.95%	469.4
DRUG_EXPOSURE	160,361,417	27.18	8.24%	431.7
CONDITION_OCCURRENCE	138,749,128	17.63	7.13%	373.6
VISIT_OCCURRENCE	38,000,960	5.29	1.95%	102.3
CONDITION_ERA	35,921,008	2.73	1.85%	96.7
PROCEDURE_OCCURRENCE	31,905,907	3.71	1.64%	85.9
VISIT_DETAIL	27,972,464	4.44	1.44%	75.3
DRUG_ERA	24,322,578	1.88	1.25%	65.5
DEVICE_EXPOSURE	5,212,843	0.66	0.27%	14.0
LOCATION	4,875,096	0.10	0.25%	13.1
SPECIMEN	2,123,886	0.17	0.11%	5.7
PROVIDER	623,239	0.02	0.03%	1.7
PERSON	371,365	0.04	0.02%	1.0
OBSERVATION_PERIOD	244,350	0.01	0.01%	0.7
DEATH	82,064	0.004	0.004%	0.2
CARE_SITE	5,966	0.0002	0.0003%	/
TOTAL	**1,947,180,421**	**278.97**	**100.00%**	**5,242.2**

Table 5: Temporal characteristics of the CRITICAL dataset showing extensive longitudinal coverage with median observation period of 3.11 years and 65.6% of patients having multi-year data

Temporal Characteristic	Value
Time Span Statistics	
Mean observation period	1,983.6 days (5.43 years)
Median observation period	1,137.0 days (3.11 years)
Standard deviation	2,168.9 days (5.94 years)
Minimum time span	0 days
Maximum time span	11,631 days (31.8 years)
25th percentile	94.0 days (0.26 years)
75th percentile	3,463.0 days (9.49 years)
Visit Frequency	
Mean visits per patient	102.3
Median visits per patient	24
Patient Records Time Span Distribution[a]	
Single visit (0 days)	2,823 (0.8%)
Under 1 year	127,692 (34.4%)
1-5 years	100,125 (27.0%)
5-10 years	56,139 (15.1%)
10-15 years	46,182 (12.4%)
15-20 years	40,356 (10.9%)
Over 20 years	871 (0.2%)
Total Patients	**371,365 (100.0%)**

Notes:
[a] Time span calculated as the difference between the last and first visit times in the VISIT_OCCURRENCE table for each patient.

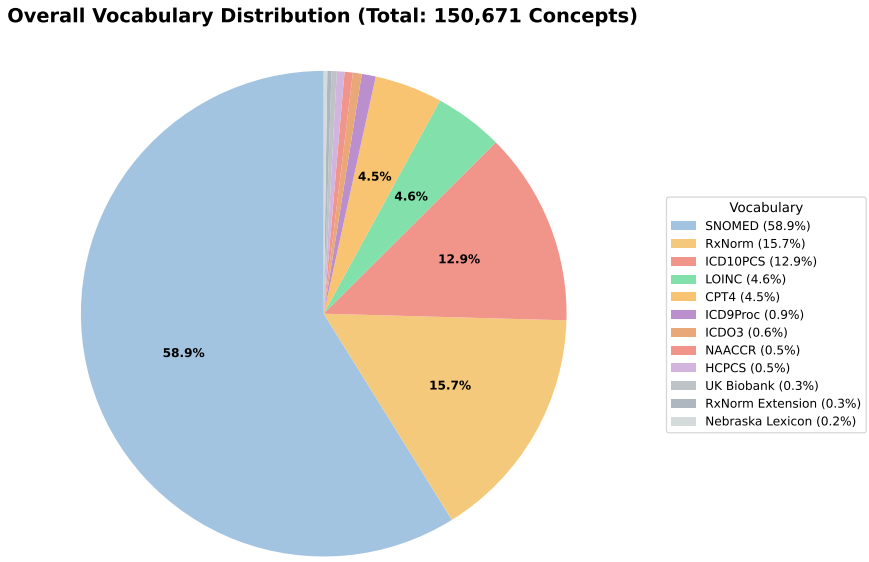

Figure 6: Overall vocabulary distribution across 150,671 unique concepts in the CRITICAL dataset. SNOMED represents 58.0% (87,453 concepts), followed by RxNorm (15.7%), ICD10PCS (12.9%), illustrating the heterogeneity challenge addressed by CRISP's mapping stage.

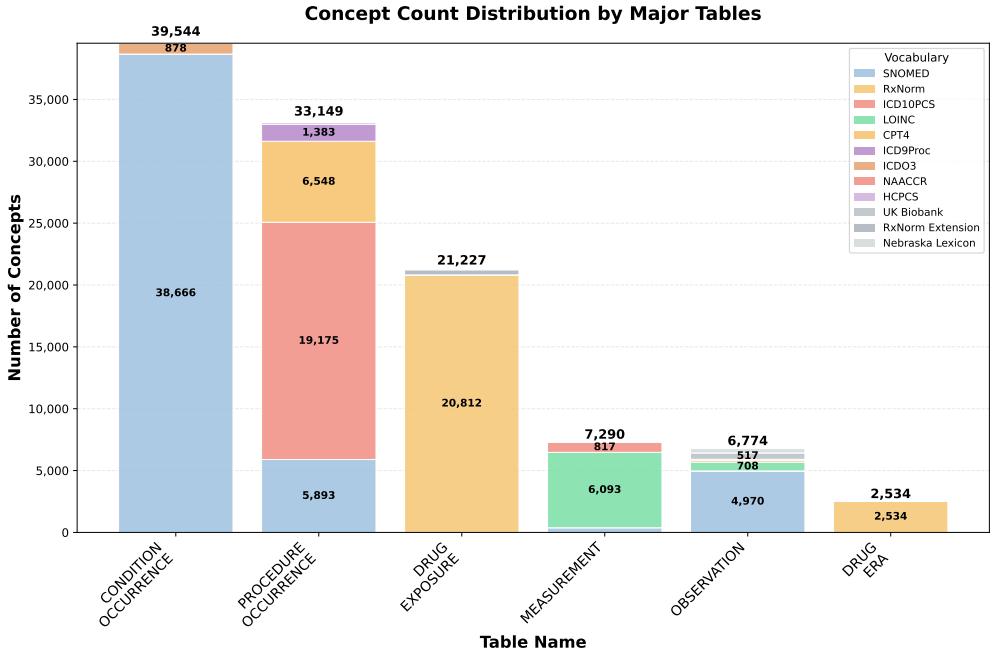

Figure 7: Absolute concept count distribution across major OMOP tables. CONDITION_OCCURRENCE exhibits the highest vocabulary diversity (39,544 unique concepts), followed by PROCEDURE_OCCURRENCE (33,149 concepts), while specialized tables like DRUG_ERA show focused vocabularies (2,534 concepts), demonstrating domain-specific vocabulary patterns.

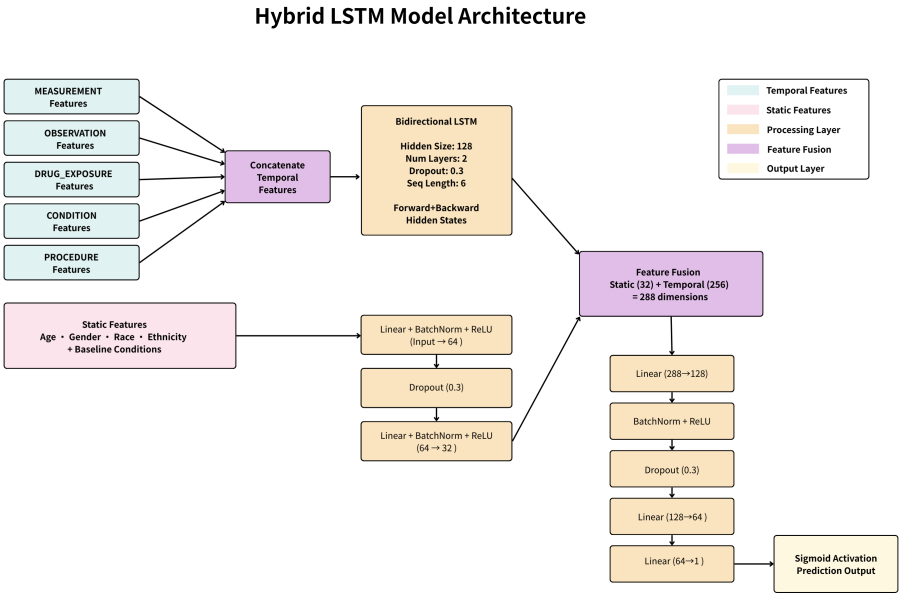

Figure 8: LSTM hybrid architecture with bidirectional LSTM layers for temporal features and separate encoding for static patient features.

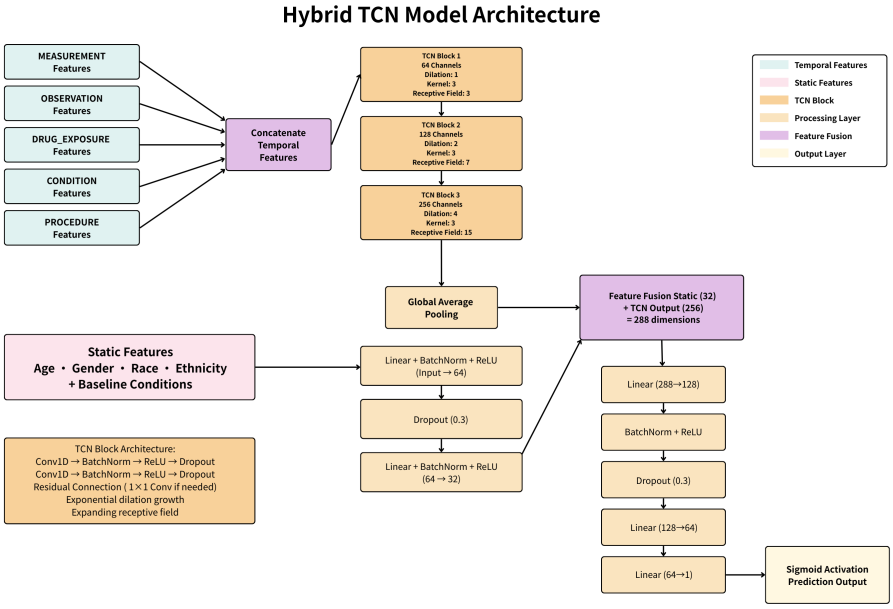

Figure 9: TCN hybrid architecture with dilated causal convolutions (dilation factors: 1, 2, 4) for temporal features and parallel processing for static features. The temporal features composition is identical to the LSTM architecture.

Table 6: Clinical Prediction Performance (LOSO Site 6; AUROC)

Task Category	Prediction Target	LR	RF	GB	XGB	MLP	LSTM	TCN	TRF
Mortality	7-day	0.586	0.672	0.778	0.747	0.595	0.617	0.590	0.688
	30-day	0.575	0.750	0.752	0.777	0.612	0.615	0.600	0.721
Length of Stay	LOS > 3 days	0.608	0.670	0.702	0.698	0.621	0.603	0.608	0.619
	LOS > 7 days	0.640	0.723	0.743	0.745	0.665	0.670	0.648	0.687
Readmission	7-day	0.530	0.637	0.626	0.653	0.521	0.524	0.509	0.539
	30-day	0.526	0.676	0.662	0.680	0.505	0.527	0.520	0.551
	90-day	0.540	0.677	0.670	0.693	0.531	0.533	0.517	0.580
In ICU Sepsis	After ICU	0.803	0.880	0.867	0.851	0.506	0.567	0.571	0.612
	Within 48h	0.532	0.896	0.879	0.837	0.435	0.534	0.496	0.619
	Within 7 days	0.530	0.901	0.899	0.873	0.508	0.474	0.541	0.523

Table 7: Clinical Prediction Performance (LOSO Site 9; AUROC)

Task Category	Prediction Target	LR	RF	GB	XGB	MLP	LSTM	TCN	TRF
Mortality	7-day	0.572	0.600	0.698	0.707	0.557	0.535	0.534	0.586
	30-day	0.585	0.590	0.696	0.698	0.563	0.538	0.542	0.580
Length of Stay	LOS > 3 days	0.572	0.576	0.597	0.606	0.577	0.597	0.563	0.592
	LOS > 7 days	0.563	0.562	0.576	0.569	0.554	0.553	0.546	0.549
Readmission	7-day	0.573	0.602	0.617	0.658	0.554	0.520	0.576	0.599
	30-day	0.568	0.616	0.645	0.668	0.595	0.536	0.593	0.573
	90-day	0.559	0.626	0.615	0.657	0.568	0.559	0.567	0.584
In ICU Sepsis	After ICU	0.850	0.872	0.847	0.864	0.830	0.844	0.832	0.851
	Within 48h	0.829	0.867	0.833	0.845	0.806	0.783	0.833	0.830
	Within 7 days	0.848	0.866	0.857	0.851	0.826	0.851	0.861	0.858