Original Paper

Scalable and Privacy-Conscious End-to-End Processing of Large-Scale Clinical Data for Precision Medicine: Empirical Evaluation Study

Jungwoo Lee[1], PhD; Sangwon Hwang[2], PhD; Kyu Hee Lee[2], PhD

[1]Artificial Intelligence Big Data Medical Center, Yonsei University Wonju College of Medicine, Wonju, Republic of Korea

[2]Department of Precision Medicine, Yonsei University Wonju College of Medicine, Wonju, Republic of Korea

Corresponding Author:
Kyu Hee Lee, PhD
Department of Precision Medicine
Yonsei University Wonju College of Medicine
20 Ilsan-ro, Wonju-si, Gangwon-do
Wonju, 26417
Republic of Korea
Phone: 82 82 741 0466
Email: powerpc@yonsei.ac.kr

Abstract

Background: In large-scale clinical data analysis, CSV and traditional relational database management system–based approaches are widely used but impose substantial storage and processing constraints that delay research preparation and hinder multicenter collaboration. Although column-oriented storage formats such as Apache Parquet have gained attention in data science, systematic end-to-end evaluations in clinical environments remain limited, particularly regarding efficiency and scalability.

Objective: This study aimed to empirically evaluate whether a Parquet-based end-to-end pipeline could improve computational efficiency and scalability in large-scale clinical data analysis while preserving predictive performance and protecting privacy.

Methods: Electronic health record data comprising 13.76 million rows from a large academic medical center in Korea were analyzed using Parquet, CSV, PostgreSQL, and DuckDB environments. Standardized SQL workloads and multilabel classification models—implemented using graphics processing unit–accelerated Extreme Gradient Boosting and classifier chain (CC) ensembles to address class imbalance—were applied to evaluate storage efficiency, time to analysis, and predictive performance. Statistical equivalence testing with prespecified clinical margins and bootstrap resampling ensured rigorous comparison, while privacy risks were assessed through advanced membership inference attacks (MIA), including shadow MIA and likelihood ratio attacks.

Results: Compared with CSV, Parquet demonstrated enhanced computational efficiency by lowering disk access from 940.2 to 44.2 seconds (95.3% reduction). End-to-end processing latency was substantially reduced across feature transformation (15.0 vs 9.3 s) and model training (8.1 vs 6.7 s). To address complex clinical correlations, we implemented CC and one-vs-rest architectures, which effectively captured interdependencies between disease labels. Classification performance remained statistically equivalent across area under the receiver operating characteristic curve, area under the precision-recall curve, accuracy, and F_1-score, with all differences falling within prespecified clinical equivalence margins (P<.001). Notably, the CC ensemble demonstrated high technical rigor, minimizing Hamming loss (2.2×10^{-4}) and ensuring robustness even in imbalanced cohorts. MIA performed at chance level (area under the curve=0.500), suggesting no measurable increase in privacy risk.

Conclusions: By significantly mitigating data processing bottlenecks, a Parquet-based pipeline enabled high-throughput, large-scale clinical evidence generation without compromising model integrity or patient privacy. This framework provides a scalable and robust infrastructure for precision medicine, facilitating agile multicenter collaborations and real-world data analysis in resource-constrained clinical environments.

(JMIR Med Inform 2026;14:e83487) doi: 10.2196/83487

KEYWORDS

electronic health records; medical informatics; predictive performance; data privacy; precision medicine

Introduction

Clinical data, including diagnostic reports, prescriptions, laboratory results, and imaging data, hold substantial potential for precision medicine, predictive modeling, and collaborative research. The volume of electronic health records (EHRs) generated in clinical practice has grown exponentially with the expansion of digital health systems [1-4]. However, as datasets expand, bottlenecks, such as storage overhead, transmission delays, and slow query execution, become increasingly common. These challenges cannot be resolved simply by expanding storage capacity; instead, they necessitate more efficient frameworks for data storage and processing.

In medical informatics, data are most commonly stored in row-oriented formats, such as CSV files, spreadsheets, and relational databases. Although valued for their simplicity and interoperability, these approaches exhibit clear limitations in storage efficiency and analytic scalability. CSV files are uncompressed and large in size, restricting input and output throughput, while relational databases often underperform in column-level operations. In contrast, column-oriented formats offer high compression and selective input and output, making them well suited for high-volume and cloud-based environments. Among these, Apache Parquet is widely used as a standard solution owing to its scalability and interoperability [5,6]. However, few studies have systematically evaluated its end-to-end performance in clinical pipelines that incorporate deidentification procedures.

A further challenge in clinical data use is balancing privacy protection with data utility. Deidentification is mandatory in multicenter research and machine learning, but it may affect statistical distributions or compromise predictive performance. Understanding how storage architectures interact with deidentification to influence analytic outcomes is therefore of both academic and practical importance [7-13].

Unlike prior studies that focused solely on benchmarking computational efficiency, this work integrates privacy risk assessment and formal equivalence testing into a real-world end-to-end pipeline, thereby providing a clinically meaningful evaluation of Parquet as a foundation for precision medicine and collaborative research. Against this background, this study developed a scalable Parquet-based pipeline for high-volume clinical data and empirically assessed its storage efficiency, computational throughput, and preservation of predictive performance. This study also sought to establish a reliable infrastructure that enables cohort construction, secure data sharing, and clinical decision support in multicenter and precision medicine research [14-23].

Methods

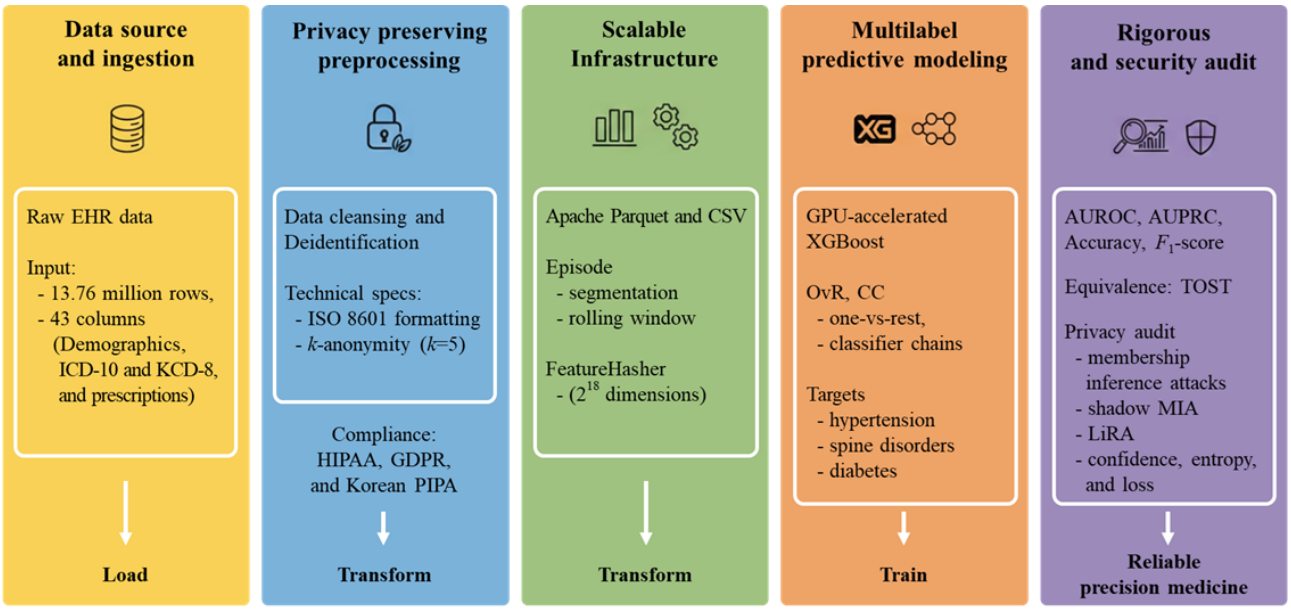
Ethical Considerations

This retrospective observational study was conducted using real-world EHR data. The study protocol was approved by the institutional review board of Wonju Severance Christian Hospital (CR325065). The requirement for informed consent was waived by the institutional review board due to the retrospective nature of the study and the use of deidentified data. All data processing was performed within an independent honest broker system, and investigators had access only to deidentified data. No compensation was provided to the participants as there was no direct contact or intervention involved. The entire process complied with the Personal Information Protection Act of Korea, the Health Insurance Portability and Accountability Act Safe Harbor standard, and the General Data Protection Regulation [24-28].

Integrated End-to-End Clinical Data Processing Workflow

We developed an integrated pipeline connecting 5 core stages: raw EHR ingestion, privacy-preserving preprocessing (ISO 8601 and *k*-anonymity), scalable infrastructure using Apache Parquet, multilabel modeling via graphics processing unit (GPU)–accelerated Extreme Gradient Boosting (XGBoost), and rigorous statistical and privacy audits (Figure 1). To optimize analytic performance, we adopted a column-oriented storage format (Apache Parquet) over the traditional row-oriented structure (eg, CSV) [6]. For comparative evaluation, the same dataset was also maintained in CSV format. While row-oriented structures excel in transactional tasks, columnar storage significantly reduces input and output overhead and enhances compression efficiency for complex analytic queries. This architectural choice is specifically tailored for the high dimensionality and sparse distribution of clinical datasets, ensuring both system scalability and robust privacy protection across the entire workflow.

Figure 1. Integrated workflow for scalable and privacy-preserving clinical data processing. AUROC: area under the receiver operating characteristic curve; AUPRC: area under the precision-recall curve; CC: classifier chains; EHR: electronic health record; GDPR: General Data Protection Regulation; GPU: graphics processing unit; HIPAA: Health Insurance Portability and Accountability Act; ICD-10: International Classification of Diseases, Tenth Revision; KCD: Korean Classification of Diseases, Eighth Revision; LiRA: Likelihood Ratio Attack; MIA: membership inference attack; OvR: one-vs-rest; PIPA: Personal Information Protection Act; TOST: two 1-sided tests; XGBoost: extreme gradient boosting.



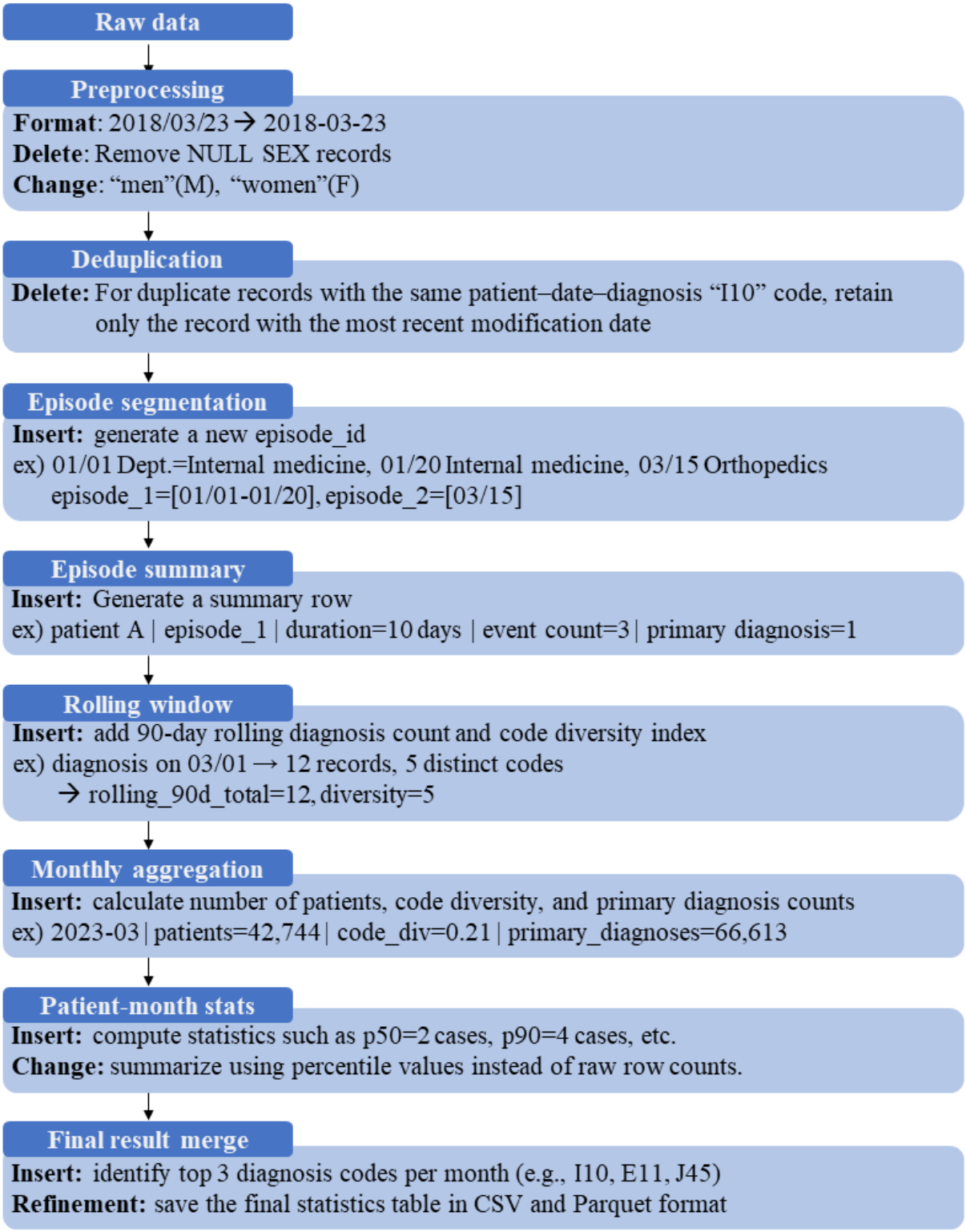
Data Preparation and Transformation

The dataset included patient demographics, diagnostic codes based on the International Classification of Diseases, Tenth Revision (ICD-10) [29-35] and the Korean Classification of Diseases, Eighth Revision (KCD-8), visit type, and prescription records. In total, the dataset comprised approximately 13.76 million rows (43 columns, approximately 3.2 GB), as detailed in Multimedia Appendix 1. After replacing patient identifiers with randomly generated surrogate IDs, quasi-identifiers were generalized to satisfy *k*-anonymity (*k*=5). Data cleansing further ensured quality control by standardizing date fields into the ISO-compliant format (YYYY-MM-DD).

Clinical Data Processing Pipeline

Clinical data processing followed a stepwise pipeline designed to reflect clinical context. In preprocessing, we standardized date formats, removed records with missing sex, and normalized sex codes (eg, “men” to M and “women” to F). Duplicate entries were resolved by retaining only the most recent record for identical combinations of patient, date, and diagnosis code. Visits were segmented into episodes by department changes or temporal gaps, and each episode was assigned an episode identifier. Episode-level summaries included episode duration, event count, and primary diagnosis status. To capture temporal continuity, a 90-day rolling window was applied to compute diagnosis counts and code diversity per patient. Monthly aggregates were then derived, including patient counts, code diversity, and primary diagnosis counts, with patient-month statistics summarized using quantiles (eg, p50 and p90). The diagnostic records for the targeted conditions—including I10.9, E10-E14, and M48.06—were identified and extracted to construct the final summary tables, which were then stored in both CSV and Parquet formats for comparative analysis. The stepwise process is illustrated in Figure 2.

Figure 2. Stepwise scenario of the clinical data processing pipeline.



Database Workload Benchmarking

To evaluate the efficiency of storage formats, scenario-based workloads were executed in DuckDB (DuckDB Labs) [36] and PostgreSQL (PostgreSQL Global Development Group) [37] environments. Synthetic example data are provided in Multimedia Appendix 2. Identical datasets were loaded into Parquet-based DuckDB and row-store PostgreSQL, and complex SQL queries—adapted from TPC-DS patterns [38,39] (eg, window functions, multidimensional aggregation, and ROLLUP)—were applied to compare performance. Resource use was further examined by varying chunk sizes (10,000-50,000 records) with detailed results in Multimedia Appendix 3.

Predictive Modeling and Multilabel Strategy

Predictive modeling assessed the impact of storage formats on classification performance for hypertension (I10.9), spinal disorders (M48.06), and diabetes (E10-E14). Patient-level

diagnostic histories were transformed into 2^{18} dimensional sparse vectors (approximately 262,000 features) via FeatureHasher (scikit-learn; Python Software Foundation) [40-42]. The dataset was randomly partitioned into training (65%), validation (15%), and test (20%) subsets with a fixed seed of 42. Classification models were implemented using GPU-accelerated XGBoost with positive class weighting to address imbalance [43-45]. Hyperparameter details are provided in Multimedia Appendix 4.

For multilabel classification, we applied one-vs-rest (OvR) [46] and classifier chains (CC) [47]. To mitigate label-order dependence and evaluate interlabel dependencies, we constructed ensembles of 5 strategically ordered chains encompassing a prevalence gradient—from high-prevalence anchors (hypertension and diabetes) to the sparse target, spinal stenosis (M48.06, 0.50%; Multimedia Appendix 5). Prioritizing M48.06, a domain-independent and sparse condition, served to demonstrate model robustness even without the clustering benefits of related disease families (eg, I-series codes). To ensure the rigorous prevention of label leakage and validate the integrity of the performance gains, all defining ICD-10 codes were strictly excluded from the feature space. This setup ensured that the model's predictive capability relied solely on statistical associations across hundreds of thousands of nondiagnostic features. Model performance was evaluated using area under the receiver operating characteristic curve (AUROC), area under the precision-recall curve (AUPRC), accuracy, and F_1-score, calculated for each label as well as macroaverages and microaverages.

Statistical Evaluation and Privacy Assessment

Model performance was evaluated using AUROC, AUPRC, accuracy, and F_1-score. To assess model reliability, calibration was further evaluated using the Brier score and expected calibration error (ECE), with detailed results provided in Multimedia Appendix 6. CSV and Parquet results were compared using paired runs with shared identifiers. Normality was assessed via the Shapiro-Wilk test, followed by paired 2-tailed *t* tests or Wilcoxon signed-rank tests with Bonferroni correction (P<.05) [48-50].

Equivalence was assessed via two 1-sided tests (TOST) [51] and 95% bootstrap CIs. Equivalence margins (δ) were prespecified at 0.02 for AUROC or AUPRC and 0.01 for accuracy or F_1-score. These thresholds were selected to be narrower than the 0.05 margin commonly used in clinical equivalence trials [52] and are more conservative than performance variances typically observed in the external validation of clinical artificial intelligence (AI) systems [53]. Such strict margins were intended to constrain observed differences to ranges generally regarded as clinically negligible, particularly for high-performing models operating within an elevated accuracy range (0.98-1.00).

To mitigate stochastic variation, a fixed random seed was applied across all experiments, ensuring that observed differences reflected data handling effects rather than algorithmic randomness. Privacy was further audited using membership inference attacks (MIA)—including confidence, entropy, loss-based (Yeom), shadow, and Likelihood Ratio Attack (LiRA) strategies—under patient-level stratified splits (train: 240,596; test: 240,597) to verify model robustness.

Experimental Setup and Reproducibility

All experiments were conducted on a Linux Ubuntu 22.04 server with a 24-core central processing unit (CPU), 128-GB memory, and an NVIDIA RTX 5090 GPU (CUDA 12.2, cuDNN 8.9, driver 550.54). Data processing used Python (version 3.10; Python Software Foundation) with Pandas 1.5.3, PyArrow 12.0.1, scikit-learn 1.3.2, and XGBoost 2.0.3 (GPU-hist). Multiprocessing was implemented using the standard Python multiprocessing module.

As specified in the modeling strategy, XGBoost was configured with a learning rate of 0.08, maximum depth of 6, and 1000 estimators. To handle class imbalance, task-specific *scale_pos_weight* and an ensemble of 5 CCs were used. All random operations used a fixed seed of 42 for reproducibility. For robust performance evaluation, we conducted 20 independent experimental runs and used 5000-iteration bootstrap resampling to compute 95% CIs. Detailed hyperparameters and the statistical framework for equivalence testing are provided in Multimedia Appendix 4.

Results

Storage and Pipeline Efficiency

Conversion to Parquet substantially improved storage efficiency, reducing file size by nearly 6.9-fold relative to CSV. As summarized in Table 1, Parquet also consistently outperformed CSV across the end-to-end pipeline. File loading (input and output) time decreased by more than 95%, and both feature transformation (FeatureHasher) and model training (XGBoost) were faster under Parquet. Although peak CPU memory was about 1.3 GB higher, GPU memory consumption decreased by 326 MB, resulting in stable overall resource use. Resource consumption remained stable at 9 to 10 GB across chunk sizes, reflecting consistent scalability, with efficiency varying by chunk size. The values in Table 1 represent peak use in end-to-end runs, with minor variations arising from measurement conditions, whereas CPU, memory, and disk input and output patterns remained consistent across chunk sizes (Multimedia Appendix 3).

Table 1. System performance comparison between CSV and Parquet formats.

Metric	CSV, mean (SD)	Parquet, mean (SD)	Difference[a] (95% CI)
Time (s)			
Input and output duration	940.2 (53.3)	44.2 (1.6)	−896.0 (−930.0 to −860.0)
Feature transformation	15.0 (0.9)	9.3 (0.7)	−5.7 (−7.0 to −4.5)
Model training	8.1 (1.2)	6.7 (0.9)	−1.4 (−2.5 to −0.3)
Resource use			
Memory maximum (GB)	21.2 (0.8)	22.5 (0.6)	1.3 (0.1 to 2.0)
GPU[b] memory maximum (MB)	4873.0 (582.6)	4547.0 (596.9)	−326.0 (−450.0 to −200.0)

[a]Calculated as the Parquet mean minus CSV mean.

[b]GPU: graphics processing unit.

Classification Performance and Calibration

Single-label results are presented in Table 2. All 3 tasks achieved AUROC values above 0.920, with AUPRC ranging from 0.488 to 0.580. Hypertension (I10.9) yielded the highest F_1-score (0.599), followed by spinal disorders (M48.06, 0.536) and diabetes (E10-E14, 0.521). Macroaveraged performance reached an AUROC of 0.920, an AUPRC of 0.545, an accuracy of 0.934, and an F_1-score of 0.552. Calibration analyses [54,55], reported in Multimedia Appendix 6, demonstrated excellent reliability for hypertension (Brier score 2.1×10^{-5}; ECE 6.0×10^{-6}), while spinal disorders (ECE=0.016) and diabetes (ECE=0.018) showed mild miscalibration but remained within recommended thresholds (ECE<0.020).

Table 2. Per-label classification performance using extreme gradient boosting.

Labels	AUROC[a]	AUPRC[b]	Accuracy	F_1-score
Hypertension	0.920	0.580	0.906	0.599
Spine disorder	0.941	0.488	0.982	0.536
Diabetes[c]	0.899	0.567	0.915	0.521
Macroaverage	0.920	0.545	0.934	0.552

[a]AUROC: area under the receiver operating characteristic curve.

[b]AUPRC: area under the precision-recall curve.

[c]Diabetes mellitus (International Classification of Diseases, Tenth Revision codes: E10-E14).

Table 2 evaluates label-specific discrimination under substantial class imbalance, whereas Table 3 summarizes overall multilabel prediction performance using microaveraged aggregation across all label–instance pairs. Specifically, Table 2 focuses on the predictive performance of individual sparse labels, while Table 3 reflects aggregated performance across all prediction events. Multilabel results in Table 3 confirmed high-performance consistency for both OvR and CC, with AUROC >0.999 and AUPRC >0.998. Subset accuracy was identical (0.999), but CC showed slight improvements in Hamming loss and Jaccard index [56], suggesting modest benefits from capturing label dependencies. A numerical gap was observed between per-label AUPRC (approximately 0.5; Table 2) and multilabel AUPRC (near 1.0; Table 3). This reflects the technical difference in aggregation: multilabel metrics were calculated using microaveraging to evaluate the overall model efficacy.

Table 3. Multilabel classification performance.

Models	AUROC[a]	AUPRC[b]	Hamming loss	Subset accuracy	Jaccard index
One-vs-rest					
Microaverage	0.999	0.998	2.4×10^{-4}	0.999	0.997
Macroaverage	0.999	0.998	—[c]	—	—
Classifier chains					
Microaverage	0.999	0.998	2.2×10^{-4}	0.999	0.997
Macroaverage	0.999	0.998	—	—	—

[a]AUROC: area under the receiver operating characteristic curve.

[b]AUPRC: area under the precision-recall curve.

[c]Not applicable.

Although discrimination remained high across all tasks, calibration analyses revealed modest overconfidence for diabetes and spinal disorder models. These deviations, while within accepted thresholds, emphasize the importance of calibration-aware adjustments when deploying predictive models in clinical workflows.

Database Benchmarking

As summarized in Table 4, DuckDB executed analytic workloads with an approximately 28-fold reduction in wall-clock time compared with PostgreSQL, which required 2546.4 seconds. DuckDB also exhibited higher CPU use (72.1% vs 39.6%), while memory consumption was comparable across systems.

Table 4. Performance comparison of DuckDB and PostgreSQL in scenario-based SQL workloads.

Metrics	DuckDB[a]	PostgreSQL[b]
Wall-clock time (s)	90.2	2546.4
CPU[c] use (%)	72.1	39.6
Memory use (GB)	13.7	14.8
Read throughput (MB/s)	554.7	365.2
Write throughput (MB/s)	1116.0	998.0

[a]DuckDB: column-oriented database optimized for analytical workloads.

[b]PostgreSQL: row-oriented relational database primarily designed for transactional workloads, with support for analytical queries.

[c]CPU: central processing unit.

Equivalence Analysis

Paired experiments (20 runs per condition) were conducted using shared run identifiers, with median and IQR values provided in Table S1 in Multimedia Appendix 7. Several metrics converged across all repetitions, producing zero IQR values and demonstrating high algorithmic stability. Overall, both formats achieved robust numerical consistency across all experimental iterations; for instance, mean AUROC was 1.000 for CSV and 0.984 for Parquet ($\Delta=-0.015$). While other metrics (AUPRC, accuracy, and F_1-score) were slightly higher under the CSV format, paired *t* tests and Wilcoxon signed-rank tests confirmed statistical significance after Bonferroni correction ($P<.001$; Table S2 in Multimedia Appendix 7). However, these numerical variations remained minimal.

Notably, all observed differences fell within the prespecified equivalence margins ($\delta=0.02$ for AUROC and AUPRC and $\delta=0.01$ for accuracy and F_1-score). TOST confirmed statistical equivalence ($P<.05$) across all evaluated metrics, and bootstrap 95% CIs for AUROC Δ (−0.016 to −0.015) further reflected near-zero variability consistent with asymptotic performance stability (Table S3 in Multimedia Appendix 7).

Privacy Assessment

The MIA evaluations, performed based on established methodologies [57-59] and summarized in Table 5, produced area under the curve (AUC) values of 0.501, 0.502, and 0.499 for confidence, entropy, and loss-based strategies, respectively. The maximum advantage did not exceed 0.007, and the gain over random guessing was ≤0.003. Attack performance was indistinguishable from random guessing, indicating no additional privacy risk. Additional shadow and LiRA [60,61] results are available in Multimedia Appendix 8, corroborating these findings.

Table 5. Membership inference attack results across different attack strategies.

Attack strategy	Attack AUC[a] (95% CI)	Advantage	Best accuracy	Gain vs baseline
Confidence	0.501 (0.501-0.503)	0.006	0.503	0.003
Entropy	0.502 (0.502-0.503)	0.007	0.503	0.003
Loss (Yeom)	0.499 (0.498-0.499)	0.002	0.500	0.001

[a]AUC: area under the curve.

Discussion

Principal Findings

This study successfully implemented an integrated end-to-end pipeline (Figure 1) that ensures a seamless transition from raw EHR data ingestion to predictive modeling. By leveraging Apache Parquet, the framework effectively addressed storage and processing bottlenecks—specifically, input and output latency—inherent in traditional row-oriented workflows, substantially enhancing data throughput and accelerating the analytic preparatory phase while maintaining high predictive fidelity. Furthermore, our findings indicate that the minor performance discrepancies between formats are attributable to systematic data handling effects under a strictly controlled environment. Beyond these computational efficiencies, privacy evaluations demonstrated that the optimized workflow maintains rigorous data protection standards, providing a secure foundation for multicenter research operations without introducing additional vulnerabilities.

Although paired statistical tests yielded highly significant *P* values ($P<.001$), this result reflects the sensitivity of paired testing under large-scale data and repeated bootstrap evaluation rather than clinically meaningful divergence. From an equivalence perspective, all performance differences remained well within the prespecified margins, confirming that the Parquet-based pipeline achieves practical performance equivalence to the traditional CSV workflow.

The substantial improvement in multilabel AUPRC compared with per-label performance highlights the system-level predictive advantage gained by explicitly modeling interlabel dependencies. While individual disease prediction remains challenging under substantial class imbalance, the multilabel formulation enables the model to capture clinically meaningful comorbidity structures, thereby reframing the observed performance gap as a reflection of holistic predictive efficacy rather than a metric artifact. This gap was also driven by the extreme class imbalance inherent in clinical datasets (eg, 0.50% prevalence for spinal disorders), where high AUROCs can coexist with lower F_1-scores. By integrating these dependencies within the CC architecture, the model leveraged frequent clinical co-occurrence patterns—such as hypertension and diabetes—to stabilize predictions for sparse conditions, including spinal stenosis. Importantly, the consistent performance observed despite the strict exclusion of all defining ICD-10 codes provided strong evidence against label leakage and underscored the model's reliance on nondiagnostic clinical associations.

Clinical Implications and Research Operations

The clinical implications were considerable. By substantially shortening preparatory phases and enabling analyses to be completed within hours rather than days, Parquet-based pipelines allowed investigators to progress more rapidly from raw data to actionable evidence. Such improvements not only accelerated study initiation but also facilitated the integration of predictive models into research and clinical workflows. In multicenter settings, scalability supported the efficient integration of heterogeneous patient populations without excessive computational cost, thereby enhancing reproducibility and advancing precision medicine initiatives. These efficiencies were directly relevant to clinical workflows and regulatory-grade analyses. They enabled interim analyses and timely regulatory submissions, supporting more responsive and efficient research operations.

Privacy and Security Safeguards

Privacy evaluation reinforced the suitability of this approach. Deidentification procedures satisfied k-anonymity, and MIA yielded chance-level performance. The chance-level performance of multiple MIA methods confirmed that computational gains do not compromise patient confidentiality, a prerequisite for secure collaboration across institutions. Patient-level stratification and consistent pipeline design acted as effective safeguards. While residual risks remain inherent to secondary data use, our findings suggested that Parquet does not introduce additional vulnerabilities. For collaborative research networks—where privacy assurances are essential for sharing—these results provide practical reassurance.

Interpretation of Predictive Performance and Label Dependencies

Classification analyses revealed stable performance for common conditions such as hypertension, whereas imbalanced disorders such as diabetes and spinal disease yielded lower scores. This pattern highlights the need to account for dataset structure and interlabel dependencies when interpreting predictive outputs. The elevated AUROC values likely reflect the combination of high-prevalence conditions and extensive diagnostic feature sets. Consistent with the modeling strategy, defining codes were excluded from the feature space; however, correlated diagnostic features may still have influenced performance estimates.

While performance convergence at elevated levels constrained the sensitivity to subtle comparative differences, it underscored the robustness and reproducibility of the pipeline, supporting its suitability for large-scale clinical research, where stability and reliability are critical. Predictive reliability was not compromised by storage structure, supporting Parquet as a

dependable option for secondary use of EHR data in diverse research settings.

Model Reliability and Calibration

Calibration analyses further showed that hypertension models were well aligned with observed risks, whereas diabetes and spinal disorder tasks exhibited modest miscalibration. These results underscore that even high-performing classifiers may generate overconfident probability estimates, a limitation directly relevant to clinical decision support. Future studies should evaluate calibration-aware techniques, such as temperature scaling or isotonic regression, to improve the reliability of probability estimates (Multimedia Appendix 6).

Beyond technical validation, the findings demonstrated practical value for clinical research operations. Large-scale trials and observational studies often face delays in cohort assembly and iterative validation; Parquet-based pipelines shorten these steps, enabling more adaptive study designs. In multicenter collaborations, they promote secure and efficient data exchange, reducing delays that commonly impede cooperative studies.

Limitations

Several limitations should be acknowledged. First, the datasets analyzed were derived from specific institutional cohorts, which may constrain generalizability to other health care systems. Second, the elevated performance baseline of the predictive models limited the sensitivity to subtle differences between CSV and Parquet formats. Third, while MIA offered a widely accepted proxy for privacy risk, broader adversarial evaluations were not included. Fourth, the evaluation was primarily conducted in on-premise settings, so the scalability of findings in distributed or cloud-based environments remains to be established. Finally, despite the clear efficiency advantages of Parquet, its adoption may introduce a technical learning curve for clinical researchers who are less familiar with column-oriented storage formats or big-data frameworks. Transitioning from traditional row-oriented workflows to optimized columnar architectures requires initial investments in technical training and infrastructure adjustment. However, these upfront efforts are likely to be offset by the long-term gains in scalability and processing speed demonstrated in our results, ultimately supporting more efficient and reproducible clinical evidence generation.

Future Work

Future research should extend the evaluation of Parquet-based pipelines to distributed and cloud-based environments, focusing on both scalability and cost-effectiveness. In addition, validation across hybrid and federated architectures will be important to confirm applicability in multicenter collaborations where data cannot be centralized. Beyond structured EHR data, future studies should also investigate applications to unstructured modalities, such as clinical narratives, imaging metadata, and time-series signals. Such efforts would test the versatility of Parquet-based pipelines in supporting multimodal learning and integrative analyses. Privacy assessments should likewise be expanded to encompass attribute inference, linkage, and other adversarial scenarios, ideally within federated or privacy-preserving analytic frameworks. Moreover, interoperability across international coding standards should be validated using external multicenter datasets to ensure global applicability. Pursuing these directions will enhance the technical robustness of Parquet-based infrastructures while reinforcing their clinical relevance, ultimately enabling secure collaboration and more efficient integration of evidence into patient care.

Conclusions

This study demonstrated that Parquet-based end-to-end processing effectively alleviated the storage and computational bottlenecks inherent in traditional row-oriented and relational database workflows while maintaining predictive fidelity and privacy safeguards. By improving computational efficiency and enabling the rapid preparation of analysis-ready cohorts, Parquet reduces the resources required for large-scale clinical data handling. These efficiencies translate into tangible research benefits, enabling earlier study initiation and more adaptive study designs across diverse clinical research settings.

The integration of efficiency with privacy safeguards further ensures that accelerated workflows remain suitable for secure data sharing across institutions. In parallel, these capabilities support regulatory-grade analyses and health technology assessments, where timely and reproducible evidence is critical. For clinicians and researchers, the central implication is that Parquet-based pipelines reduce the manual effort and computational overhead required to transform raw EHR data into analysis-ready cohorts. This capability not only strengthens multicenter collaboration but also facilitates the timely translation of evidence into clinical decision support.

Ultimately, by aligning computational efficiency with privacy-conscious infrastructure, Parquet-based pipelines reinforce the evidence base that informs clinical decision-making, positioning this approach as a robust foundation for sustainable precision medicine.

Acknowledgments

This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (grant RS-2025-24683718). This research was supported by the Ministry of Trade, Industry and Energy, Republic of Korea, under the project titled “Research and Development on AI Standardization in the Healthcare Industry Based on Human Big Data and Artificial Intelligence” (grant RS-2025-13882968), funded through the Korea Evaluation Institute of Industrial Technology. This research was supported by a grant from the medical data-driven hospital support project through the Korea Health Information Service, funded by the Ministry of Health and Welfare, Republic of Korea. This research was supported by the Regional Innovation System and Education program through the Gangwon Regional Innovation System and Education Center, funded by the Ministry of Education and the Gangwon State, Republic of Korea (grant 2025-RISE-10-006).

Funding

This study was funded by the following grants: (1) the National Research Foundation of Korea (NRF) grant funded by the Korea government (grant RS-2025-24683718); (2) a grant from the medical data-driven hospital support project through the Korea Health Information Service, funded by the Ministry of Health and Welfare, Republic of Korea; and (3) the Regional Innovation System and Education program through the Gangwon Regional Innovation System and Education Center, funded by the Ministry of Education and the Gangwon State, Republic of Korea (grant 2025-RISE-10-006). The funders had no involvement in the study design, data collection, analysis, interpretation, or the writing of the manuscript.

Data Availability

The datasets generated or analyzed during this study are available from the corresponding author on reasonable request.

Authors' Contributions

Conceptualization: KHL
Data curation: JL
Formal analysis: JL, KHL
Methodology: SH
Project administration: KHL
Resources: JL
Writing – original draft: JL, SH
Writing – review and editing: KHL

Conflicts of Interest

None declared.

Multimedia Appendix 1

Column definitions for the structured clinical data table.
[XLSX File (Microsoft Excel File), 10 KB-Multimedia Appendix 1]

Multimedia Appendix 2

Synthetic example data used for scenario-based workload evaluation.
[XLSX File (Microsoft Excel File), 24 KB-Multimedia Appendix 2]

Multimedia Appendix 3

Resource use profiles during Parquet-based data processing across different chunk sizes.
[DOCX File , 252 KB-Multimedia Appendix 3]

Multimedia Appendix 4

Hyperparameter configurations and the formal statistical framework for equivalence testing.
[DOCX File , 16 KB-Multimedia Appendix 4]

Multimedia Appendix 5

Detailed demographic characteristics and visit statistics of the study cohort.
[PNG File , 235 KB-Multimedia Appendix 5]

Multimedia Appendix 6

Calibration metrics (Brier score and expected calibration error [ECE] and ECE) for each binary classification task.
[DOCX File , 14 KB-Multimedia Appendix 6]

Multimedia Appendix 7

Integrated statistical performance and equivalence analysis.
[DOCX File , 18 KB-Multimedia Appendix 7]

Multimedia Appendix 8

Membership inference results with shadow and Likelihood Ratio Attack attackers.
[DOCX File , 15 KB-Multimedia Appendix 8]

References

1. Jensen PB, Jensen LJ, Brunak S. Mining electronic health records: towards better research applications and clinical care. Nat Rev Genet. May 02, 2012;13(6):395-405. [doi: 10.1038/nrg3208] [Medline: 22549152]
2. Weiskopf NG, Weng C. Methods and dimensions of electronic health record data quality assessment: enabling reuse for clinical research. J Am Med Inform Assoc. Jan 01, 2013;20(1):144-151. [FREE Full text] [doi: 10.1136/amiajnl-2011-000681] [Medline: 22733976]
3. Hripcsak G, Albers DJ. Next-generation phenotyping of electronic health records. J Am Med Inform Assoc. Jan 01, 2013;20(1):117-121. [FREE Full text] [doi: 10.1136/amiajnl-2012-001145] [Medline: 22955496]
4. Birkhead GS, Klompas M, Shah NR. Uses of electronic health records for public health surveillance to advance public health. Annu Rev Public Health. Mar 18, 2015;36:345-359. [doi: 10.1146/annurev-publhealth-031914-122747] [Medline: 25581157]
5. Vohra D. Apache Parquet. In: Practical Hadoop Ecosystem: A Definitive Guide to Hadoop-Related Frameworks and Tools. Berkeley, CA. Apress; 2016:325-335.
6. Apache parquet documentation. Apache Software Foundation. URL: https://parquet.apache.org/ [accessed 2025-06-01]
7. Meystre SM, Friedlin FJ, South BR, Shen S, Samore MH. Automatic de-identification of textual documents in the electronic health record: a review of recent research. BMC Med Res Methodol. Aug 02, 2010;10(1):70. [FREE Full text] [doi: 10.1186/1471-2288-10-70] [Medline: 20678228]
8. Negash B, Katz A, Neilson CJ, Moni M, Nesca M, Singer A, et al. De-identification of free text data containing personal health information: a scoping review of reviews. Int J Popul Data Sci. Dec 12, 2023;8(1):2153. [FREE Full text] [doi: 10.23889/ijpds.v8i1.2153] [Medline: 38414537]
9. Kushida CA, Nichols DA, Jadrnicek R, Miller R, Walsh JK, Griffin K. Strategies for de-identification and anonymization of electronic health record data for use in multicenter research studies. Med Care. Jul 2012;50 Suppl(Suppl):S82-101. [FREE Full text] [doi: 10.1097/MLR.0b013e3182585355] [Medline: 22692265]
10. El-Hayek C, Barzegar S, Faux N, Doyle K, Pillai P, Mutch SJ, et al. An evaluation of existing text de-identification tools for use with patient progress notes from Australian general practice. Int J Med Inform. May 2023;173:105021. [doi: 10.1016/j.ijmedinf.2023.105021] [Medline: 36870249]
11. Liu J, Gupta S, Chen A, Wang CK, Mishra P, Dai HJ, et al. OpenDeID pipeline for unstructured electronic health record text notes based on rules and transformers: deidentification algorithm development and validation study. J Med Internet Res. Dec 06, 2023;25:e48145. [FREE Full text] [doi: 10.2196/48145] [Medline: 38055317]
12. Jabet A, Bérot V, Chiarabini T, Dellière S, Bosshard PP, Siguier M, et al. Trichophyton mentagrophytes ITS genotype VII infections among men who have sex with men in France: an ongoing phenomenon. J Eur Acad Dermatol Venereol. Feb 2025;39(2):407-415. [doi: 10.1111/jdv.20439] [Medline: 39587983]
13. Crane HM, Nance RM, Ruderman SA, Drumright LN, Mixson LS, Heckbert SR, et al. Smoking and type 1 versus type 2 myocardial infarction among people with HIV in the United States: results from the center for AIDS research network integrated clinical systems cohort. J Assoc Nurses AIDS Care. Sep 6, 2024;35(6):507-518. [doi: 10.1097/JNC.0000000000000492] [Medline: 39241219]
14. Solomon J, Dauber-Decker K, Richardson S, Levy S, Khan S, Coleman B, et al. Integrating clinical decision support into electronic health record systems using a novel platform (EvidencePoint): developmental study. JMIR Form Res. Oct 19, 2023;7:e44065. [FREE Full text] [doi: 10.2196/44065] [Medline: 37856193]
15. Trinkley KE, Kroehl ME, Kahn MG, Allen LA, Bennett TD, Hale G, et al. Applying clinical decision support design best practices with the practical robust implementation and sustainability model versus reliance on commercially available clinical decision support tools: randomized controlled trial. JMIR Med Inform. Mar 22, 2021;9(3):e24359. [FREE Full text] [doi: 10.2196/24359] [Medline: 33749610]
16. Yoo J, Lee J, Rhee PL, Chang DK, Kang M, Choi JS, et al. Alert override patterns with a medication clinical decision support system in an academic emergency department: retrospective descriptive study. JMIR Med Inform. Nov 04, 2020;8(11):e23351. [FREE Full text] [doi: 10.2196/23351] [Medline: 33146626]
17. Jing X, Min H, Gong Y, Biondich P, Robinson D, Law T, et al. Ontologies applied in clinical decision support system rules: systematic review. JMIR Med Inform. Jan 19, 2023;11:e43053. [FREE Full text] [doi: 10.2196/43053] [Medline: 36534739]
18. Sommers SW, Tolle HJ, Trinkley KE, Johnston CG, Dietsche CL, Eldred SV, et al. Clinical decision support to increase emergency department Naloxone Coprescribing: implementation report. JMIR Med Inform. Nov 06, 2024;12:e58276. [FREE Full text] [doi: 10.2196/58276] [Medline: 39504560]
19. Barton HJ, Maru A, Leaf MA, Hekman DJ, Wiegmann DA, Shah MN, et al. Academic detailing as a health information technology implementation method: supporting the design and implementation of an emergency department-based clinical

decision support tool to prevent future falls. JMIR Hum Factors. Apr 18, 2024;11:e52592. [FREE Full text] [doi: 10.2196/52592] [Medline: 38635318]
20. Dong J, Jin Z, Li C, Yang J, Jiang Y, Li Z, et al. Machine learning models with prognostic implications for predicting gastrointestinal bleeding after coronary artery bypass grafting and guiding personalized medicine: multicenter cohort study. J Med Internet Res. Mar 06, 2025;27:e68509. [FREE Full text] [doi: 10.2196/68509] [Medline: 40053791]
21. Jung C, Mamandipoor B, Fjølner J, Bruno RR, Wernly B, Artigas A, et al. Disease-course adapting machine learning prognostication models in elderly patients critically ill with COVID-19: multicenter cohort study with external validation. JMIR Med Inform. Mar 31, 2022;10(3):e32949. [FREE Full text] [doi: 10.2196/32949] [Medline: 35099394]
22. Alvarez-Romero C, Martinez-Garcia A, Ternero Vega J, Díaz-Jimènez P, Jimènez-Juan C, Nieto-Martín MD, et al. Predicting 30-day readmission risk for patients with chronic obstructive pulmonary disease through a federated machine learning architecture on Findable, Accessible, Interoperable, and Reusable (FAIR) data: development and validation study. JMIR Med Inform. Jun 02, 2022;10(6):e35307. [FREE Full text] [doi: 10.2196/35307] [Medline: 35653170]
23. Blasi L, Bordonaro R, Serretta V, Piazza D, Firenze A, Gebbia V. Virtual clinical and precision medicine tumor boards-cloud-based platform-mediated implementation of multidisciplinary reviews among oncology centers in the COVID-19 era: protocol for an observational study. JMIR Res Protoc. Sep 10, 2021;10(9):e26220. [FREE Full text] [doi: 10.2196/26220] [Medline: 34387553]
24. Ahmed A, Shahzad A, Naseem A, Ali S, Ahmad I. Evaluating the effectiveness of data governance frameworks in ensuring security and privacy of healthcare data: a quantitative analysis of ISO standards, GDPR, and HIPAA in blockchain technology. PLoS One. May 23, 2025;20(5):e0324285. [FREE Full text] [doi: 10.1371/journal.pone.0324285] [Medline: 40408373]
25. Bradford L, Aboy M, Liddell K. International transfers of health data between the EU and USA: a sector-specific approach for the USA to ensure an 'adequate' level of protection. J Law Biosci. 2020;7(1):lsaa055. [FREE Full text] [doi: 10.1093/jlb/lsaa055] [Medline: 34221424]
26. Gomase VS, Ghatule AP, Sharma R, Sardana S. Cybersecurity and compliance in clinical trials: the role of artificial intelligence in secure healthcare management. Rev Recent Clin Trials. Apr 25, 2025;20(4):332-351. [doi: 10.2174/0115748871366467250413070850] [Medline: 40277117]
27. Scheibner J, Raisaro JL, Troncoso-Pastoriza JR, Ienca M, Fellay J, Vayena E, et al. Revolutionizing medical data sharing using advanced privacy-enhancing technologies: technical, legal, and ethical synthesis. J Med Internet Res. Feb 25, 2021;23(2):e25120. [FREE Full text] [doi: 10.2196/25120] [Medline: 33629963]
28. Broen K, Trangucci R, Zelner J. Measuring the impact of spatial perturbations on the relationship between data privacy and validity of descriptive statistics. Int J Health Geogr. Jan 07, 2021;20(1):3. [FREE Full text] [doi: 10.1186/s12942-020-00256-8] [Medline: 33413390]
29. Chen PF, Wang SM, Liao WC, Kuo LC, Chen KC, Lin YC, et al. Automatic ICD-10 coding and training system: deep neural network based on supervised learning. JMIR Med Inform. Aug 31, 2021;9(8):e23230. [FREE Full text] [doi: 10.2196/23230] [Medline: 34463639]
30. Chen PF, Chen KC, Liao WC, Lai F, He TL, Lin SC, et al. Automatic international classification of diseases coding system: deep contextualized language model with rule-based approaches. JMIR Med Inform. Jun 29, 2022;10(6):e37557. [FREE Full text] [doi: 10.2196/37557] [Medline: 35767353]
31. Chen PF, He TL, Lin SC, Chu YC, Kuo CT, Lai F, et al. Training a deep contextualized language model for international classification of diseases, 10th revision classification via federated learning: model development and validation study. JMIR Med Inform. Nov 10, 2022;10(11):e41342. [FREE Full text] [doi: 10.2196/41342] [Medline: 36355417]
32. Wu P, Gifford A, Meng X, Li X, Campbell H, Varley T, et al. Mapping ICD-10 and ICD-10-CM codes to phecodes: workflow development and initial evaluation. JMIR Med Inform. Nov 29, 2019;7(4):e14325. [FREE Full text] [doi: 10.2196/14325] [Medline: 31553307]
33. Falissard L, Morgand C, Ghosn W, Imbaud C, Bounebache K, Rey G. Neural translation and automated recognition of ICD-10 medical entities from natural language: model development and performance assessment. JMIR Med Inform. Apr 11, 2022;10(4):e26353. [FREE Full text] [doi: 10.2196/26353] [Medline: 35404262]
34. Ghasemi P, Lee J. Unsupervised feature selection to identify important ICD-10 and ATC codes for machine learning on a cohort of patients with coronary heart disease: retrospective study. JMIR Med Inform. Jul 26, 2024;12:e52896. [FREE Full text] [doi: 10.2196/52896] [Medline: 39087585]
35. Dai HJ, Wang CK, Chen CC, Liou CS, Lu AT, Lai CH, et al. Evaluating a natural language processing-driven, AI-assisted international classification of diseases, 10th revision, clinical modification, coding system for diagnosis related groups in a real hospital environment: algorithm development and validation study. J Med Internet Res. Sep 20, 2024;26:e58278. [FREE Full text] [doi: 10.2196/58278] [Medline: 39302714]
36. DuckDB documentation. DuckDB Foundation. URL: https://duckdb.org/ [accessed 2025-05-25]
37. PostgreSQL documentation. PostgreSQL Global Development Group. URL: https://www.postgresql.org/docs/ [accessed 2025-06-10]
38. Tardío R, Maté A, Trujillo J. Beyond TPC-DS, a benchmark for Big Data OLAP systems (BDOLAP-Bench). Futur Gener Comp Syst. Jul 2022;132:136-151. [doi: 10.1016/j.future.2022.02.015]

39. Ding B, Chaudhuri S, Gehrke J, Narasayya V. DSB: a decision support benchmark for workload-driven and traditional database systems. Proc VLDB Endow. Oct 28, 2021;14(13):3376-3388. [doi: 10.14778/3484224.3484234]
40. Pedragosa M, Riera G, Casella V, Esteve-Codina A, Steuerman Y, Seth C, et al. Linking cell dynamics with gene coexpression networks to characterize key events in chronic virus infections. Front Immunol. May 3, 2019;10:1002. [FREE Full text] [doi: 10.3389/fimmu.2019.01002] [Medline: 31130969]
41. Pang Y, Zhang X, Gao R, Xu L, Shen M, Shi H, et al. Efficacy of web-based self-management interventions for depressive symptoms: a meta-analysis of randomized controlled trials. BMC Psychiatry. Aug 11, 2021;21(1):398. [FREE Full text] [doi: 10.1186/s12888-021-03396-8] [Medline: 34380440]
42. Bugiardini R, Cenko E. Sex differences in myocardial infarction deaths. Lancet. Jul 11, 2020;396(10244):72-73. [doi: 10.1016/S0140-6736(20)31049-7] [Medline: 32445694]
43. Kaul V, Gross S, Corbett FS, Malik Z, Smith M, Tofani C, et al. Clinical utility of wide-area transepithelial sampling with three-dimensional computer-assisted analysis (WATS3D) in identifying Barrett's esophagus and associated neoplasia. Dis Esophagus. Dec 07, 2020;33(12):32607543. [doi: 10.1093/dote/doaa069] [Medline: 32607543]
44. Rajkomar A, Oren E, Chen K, Dai AM, Hajaj N, Hardt M, et al. Scalable and accurate deep learning with electronic health records. NPJ Digit Med. May 8, 2018;1(1):18. [FREE Full text] [doi: 10.1038/s41746-018-0029-1] [Medline: 31304302]
45. Wu J, Roy J, Stewart WF. Prediction modeling using EHR data: challenges, strategies, and a comparison of machine learning approaches. Med Care. Jun 2010;48(6 Suppl):S106-S113. [doi: 10.1097/MLR.0b013e3181de9e17] [Medline: 20473190]
46. Rifkin R, Klautau A. In defense of one-vs-all classification. J Mach Learn Res. 2004;5:41. [doi: 10.5555/1005332.1005336]
47. Read J, Pfahringer B, Holmes G, Frank E. Classifier chains for multi-label classification. Mach Learn. Jun 30, 2011;85(3):333-359. [doi: 10.1007/S10994-011-5256-5]
48. Dunn OJ. Multiple comparisons among Means. J Am Stat Assoc. Mar 1961;56(293):52-64. [doi: 10.1080/01621459.1961.10482090]
49. Bland JM, Altman DG. Multiple significance tests: the Bonferroni method. BMJ. Jan 21, 1995;310(6973):170. [FREE Full text] [doi: 10.1136/bmj.310.6973.170] [Medline: 7833759]
50. Choi JE, Kang HW, Hong YM, Sohn S. C-reactive protein and N-terminal pro-brain natriuretic peptide discrepancy: a differentiation of adenoviral pharyngoconjunctival fever from Kawasaki disease. Korean J Pediatr. Jan 2018;61(1):12-16. [FREE Full text] [doi: 10.3345/kjp.2018.61.1.12] [Medline: 29441107]
51. Schuirmann DJ. A comparison of the two one-sided tests procedure and the power approach for assessing the equivalence of average bioavailability. J Pharmacokinet Biopharm. Dec 1, 1987;15(6):657-680. [doi: 10.1007/BF01068419] [Medline: 3450848]
52. Walker E, Nowacki AS. Understanding equivalence and noninferiority testing. J Gen Intern Med. Feb 21, 2011;26(2):192-196. [FREE Full text] [doi: 10.1007/s11606-010-1513-8] [Medline: 20857339]
53. Rockenschaub P, Akay EM, Carlisle BG, Hilbert A, Wendland J, Meyer-Eschenbach F, et al. External validation of AI-based scoring systems in the ICU: a systematic review and meta-analysis. BMC Med Inform Decis Mak. Jan 06, 2025;25(1):5. [FREE Full text] [doi: 10.1186/s12911-024-02830-7] [Medline: 39762808]
54. Glenn WB. Verification of forecasts expressed in terms of probability. Mon Weather Rev. Jan 1950;78(1):1-3. [doi: 10.1175/1520-0493(1950)078<0001:vofeit>2.0.co;2]
55. Dawood T, Chen C, Sidhu BS, Ruijsink B, Gould J, Porter B, et al. Uncertainty aware training to improve deep learning model calibration for classification of cardiac MR images. Med Image Anal. Aug 2023;88:102861. [FREE Full text] [doi: 10.1016/j.media.2023.102861] [Medline: 37327613]
56. Eelbode T, Bertels J, Berman M, Vandermeulen D, Maes F, Bisschops R, et al. Optimization for medical image segmentation: theory and practice when evaluating with dice score or Jaccard index. IEEE Trans Med Imaging. Nov 2020;39(11):3679-3690. [doi: 10.1109/tmi.2020.3002417]
57. Zhang Z, Yan C, Malin BA. Membership inference attacks against synthetic health data. J Biomed Inform. Jan 2022;125:103977. [FREE Full text] [doi: 10.1016/j.jbi.2021.103977] [Medline: 34920126]
58. Cobilean V, Mavikumbure HS, Drake D, Stuart M, Manic M. Investigating membership inference attacks against CNN models for BCI systems. IEEE J Biomed Health Inform. Nov 2025;29(11):8164-8174. [doi: 10.1109/jbhi.2025.3593443]
59. Chen J, Wang WH, Shi X. Differential privacy protection against membership inference attack on machine learning for genomic data. Pac Symp Biocomput. 2021;26:26-37. [FREE Full text] [Medline: 33691001]
60. Bohrson CL, Barton AR, Lodato MA, Rodin RE, Luquette LJ, Viswanadham VV, et al. Linked-read analysis identifies mutations in single-cell DNA-sequencing data. Nat Genet. Apr 18, 2019;51(4):749-754. [FREE Full text] [doi: 10.1038/s41588-019-0366-2] [Medline: 30886424]
61. Luquette LJ, Miller MB, Zhou Z, Bohrson CL, Zhao Y, Jin H, et al. Single-cell genome sequencing of human neurons identifies somatic point mutation and indel enrichment in regulatory elements. Nat Genet. Oct 26, 2022;54(10):1564-1571. [FREE Full text] [doi: 10.1038/s41588-022-01180-2] [Medline: 36163278]

Abbreviations

AI: artificial intelligence

AUC: area under the curve
AUPRC: area under the precision-recall curve
AUROC: area under the receiver operating characteristic curve
CC: classifier chains
CPU: central processing unit
ECE: expected calibration error
EHR: electronic health record
GPU: graphics processing unit
ICD-10: International Classification of Diseases, Tenth Revision
KCD-8: Korean Classification of Diseases, Eighth Revision
LiRA: Likelihood Ratio Attack
MIA: membership inference attack
OvR: one-vs-rest
TOST: two 1-sided tests
XGBoost: extreme gradient boosting

Edited by A Benis; submitted 03.Sep.2025; peer-reviewed by A Kalluchi, Y He; comments to author 19.Jan.2026; accepted 05.Feb.2026; published 04.Mar.2026

Please cite as:
Lee J, Hwang S, Lee KH
Scalable and Privacy-Conscious End-to-End Processing of Large-Scale Clinical Data for Precision Medicine: Empirical Evaluation Study
JMIR Med Inform 2026;14:e83487
URL: https://medinform.jmir.org/2026/1/e83487
doi: 10.2196/83487
PMID: