Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Risk Stratification for 7-Day Mortality After ICU Admission Using General-Purpose Large Language Models and Structured Electronic Health Records: Predictive Performance and Reasoning Alignment Across Three Cohorts

Sungmin Cho
Yonsei University
Soojung Choi
Yonsei University
Yewon Choi
Yonsei University
Minsun Kim
Catholic University of Korea
Anupol Panitchote
Khon Kaen University
Natdanai Ketdao
Khon Kaen University
Anawat Ratchatorn
Khon Kaen University
Hyun Do Jung
Yonsei University
Changdai Gu
Yonsei University
Taeyun Park
Yonsei University
Dongin Moon
Yonsei University
Sangwoo Park
Seoul National University
Hyung-Chul Lee
Seoul National University Hospital
Hwiyoung Kim
hykim@hallym.or.kr

Hallym University Chuncheon Sacred Heart Hospital

Research Article

Keywords: Critical care, Intensive care units, Mortality prediction, Large language models, Electronic health records, Machine learning, Clinical prediction models, External validation, Explainable artificial intelligence, Clinical reasoning

Posted Date: November 18th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11428570/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License. Read Full License

Additional Declarations: No competing interests reported.

Abstract

Background

Early risk stratification after intensive care unit (ICU) admission may inform treatment intensity, goals-of-care discussions, and resource allocation. Developing and maintaining supervised machine learning models requires large labelled cohorts and technical infrastructure that may not be available in all settings. General-purpose large language models (LLMs) can generate risk estimates with human-readable rationales, but their performance and reasoning quality across independent healthcare systems remain uncertain.

Methods

We analysed three ICU cohorts: eICU-CRD (208 US hospitals), MIMIC-IV (US), and a Korean tertiary-hospital subset of K-MIMIC (Korea). Seven-day mortality after ICU admission was predicted from routinely collected demographic, laboratory, and vital sign variables recorded during the first 24 hours. Six supervised machine learning and deep learning models were developed in eICU-CRD and externally validated in MIMIC-IV and K-MIMIC. GPT-5 and Gemini-2.5-Pro were evaluated with four prompting strategies in 2,000-stay samples per cohort, without cohort-specific training or task-specific fine-tuning. Clinical concepts mentioned in GPT-5 rationales were compared with patient-level predictor importance from an XGBoost reference model; two clinicians rated 150 rationales for clinical trustworthiness.

Results

Analyses included 124,020, 84,926, and 6,673 ICU stays. Among supervised models, the highest area under the receiver operating characteristic curve (AUROC) was 0.936 for XGBoost in the held-out eICU-CRD set, 0.858 for the multilayer perceptron in MIMIC-IV, and 0.957 for XGBoost in K-MIMIC. In the corresponding 2,000-stay LLM evaluation samples, GPT-5 achieved AUROCs of 0.925, 0.822, and 0.926, respectively, but generally had a lower area under the precision–recall curve (AUPRC) and, at evaluated operating thresholds, lower specificity. Overlap with the five most influential concepts was partial (precision, 0.47; recall, 0.55), but mentioned concepts had greater mean patient-level importance than nonmentioned concepts in 98.1% of patients ($P < 0.001$). Clinician-rated trustworthiness averaged 4.03 ± 0.63 on a 5-point scale, with ratings within one point in 83% of cases.

Conclusions

Supervised models generally achieved the highest discrimination for 7-day mortality after ICU admission. GPT-5 achieved high discrimination without cohort-specific training or task-specific fine-tuning, and its patient-specific rationales preferentially referenced influential predictors and received favorable clinician ratings. Prospective evaluation of reasoning faithfulness and clinical utility is required before clinical use.

Background

Demand for intensive care is rising worldwide, driven by population aging, a growing burden of multimorbidity, broader use of life-sustaining therapies, and increasing admissions of older adults to intensive care units (ICUs) [1, 2]. These growing demands place increasing pressure on critical care systems, highlighting the need for timely and accurate prognostic assessment to support clinical decision-making. Early mortality prediction can help clinicians guide treatment intensity, inform goals-of-care discussions with patients and families, and allocate scarce critical care resources more equitably [3]. The prognostic relevance of acute illness severity evolves over an ICU stay and is most pronounced early after admission [4]. Therefore, risk stratification within this early period carries direct clinical relevance.

Traditional severity scores, including the Acute Physiology and Chronic Health Evaluation (APACHE) and Simplified Acute Physiology Score (SAPS), have long been used to estimate ICU mortality risk [5, 6]. More recently, machine learning (ML) and deep learning (DL) approaches have improved predictive performance by capturing high-dimensional and nonlinear patterns in electronic health record (EHR) data [7, 8]. However, these models generally depend on institution-specific training and may require recalibration or retraining when case mix, clinical protocols, or measurement conventions differ across settings. These development and maintenance requirements may limit implementation in institutions without sufficiently large labelled cohorts or technical infrastructure [9].

Large language models (LLMs) are increasingly being evaluated for patient-level clinical prediction in addition to medical knowledge and clinical reasoning tasks [10–14]. General-purpose LLMs can generate predictions without cohort-specific training and may therefore reduce the development burden that limits local implementation of supervised models. However, LLM-based clinical prediction has largely relied on unstructured clinical notes [14, 15], which provide rich contextual information [16] but depend on language, documentation practices, and institutional workflows [17, 18]. Routinely collected structured EHR variables are less dependent on local narrative conventions and may support more consistent evaluation across healthcare systems. Recent studies have generated predictions from structured EHR data through task-specific adaptation or direct prompting, although conventional ML and DL models have generally retained superior discrimination [19–22]. In critical care, LLM-based approaches have been evaluated for mortality, delirium, and postoperative acute kidney injury [19, 23–25], but their performance when common structured inputs and prompting strategies are applied across independent healthcare systems, and their sensitivity to prompt formulation, remain uncertain.

A further question concerns the reasoning accompanying these predictions. LLM-generated rationales may appear clinically plausible without necessarily reflecting influential patient-level factors [26, 27]. Expert review can establish clinical coherence, but coherence alone does not indicate correspondence with

independently derived patient-level predictor importance. Although prior studies have compared LLM-derived rankings with cohort-level feature importance or used LLMs to translate precomputed attributions into clinical narratives [28, 29], whether independently generated reasoning aligns with patient-level predictor importance across healthcare systems remains unclear.

To address these gaps, we evaluated general-purpose LLMs for 7-day ICU mortality prediction using structured EHR data across three cohorts in the United States and South Korea. We benchmarked GPT-5 and Gemini-2.5-Pro against established ML and DL models and assessed four prompting strategies without cohort-specific training or task-specific fine-tuning. We further evaluated whether LLM-generated reasoning aligned with influential patient-level predictors and whether clinicians judged the same rationales to be clinically trustworthy.

Methods

Data sources and study setting

This study analyzed three de-identified critical care databases: the eICU Collaborative Research Database (eICU-CRD) [30], the Medical Information Mart for Intensive Care IV (MIMIC-IV) [31], and the Korea Medical Information Mart for Intensive Care (K-MIMIC) [32]. For K-MIMIC, we analyzed the K-MIMIC SNUH Subset, comprising ICU admissions from Seoul National University Hospital (SNUH), South Korea. The eICU-CRD includes 200,859 ICU stays from 208 US hospitals between 2014 and 2015; MIMIC-IV includes 94,458 ICU stays from Beth Israel Deaconess Medical Center between 2008 and 2019; and the K-MIMIC SNUH Subset includes 8,726 ICU admissions in 2023. The study was designed and reported in accordance with the Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis+Artificial Intelligence (TRIPOD+AI) statement [33], and the overall workflow is shown in Figure 1.

Study population and Design

The eICU-CRD served as the development cohort for model training and internal validation, whereas the MIMIC-IV and K-MIMIC were used for external validation. Uniform exclusion criteria were applied across all cohorts. ICU stays were excluded if patients were younger than 18 years, had an ICU length of stay exceeding 60 days, had a documented do-not-resuscitate order, or had missing age, sex, or mortality status. In K-MIMIC, Korean nationals were classified as Asian; non-Korean patients were excluded because ethnicity was not reliably available. The final analytic cohorts comprised 124,020 ICU stays in eICU-CRD, 84,926 in MIMIC-IV, and 6,673 in K-MIMIC. Cohort selection is shown in Figure 2.

Predictors and outcome definition

The outcome was mortality within seven calendar days of ICU admission. Predictors were selected by clinical experts based on routine collection and consistent availability across the three cohorts. Demographic predictors included age, sex, and ethnicity, and neurological status was represented by the Glasgow Coma Scale (GCS) at ICU admission. Laboratory predictors included alanine aminotransferase (ALT), aspartate aminotransferase (AST), blood urea nitrogen (BUN), hemoglobin, prothrombin time (PT), activated partial thromboplastin time (aPTT), white blood cell count (WBC), albumin, chloride, creatinine, glucose, platelet count, potassium, sodium, and total bilirubin. Vital sign predictors included heart rate, systolic and diastolic blood pressure, respiratory rate, peripheral oxygen saturation, and body temperature. Laboratory and vital sign measurements recorded from ICU admission until the earlier of 24 hours or death were used. For patients who died within 24 hours, summary measures were calculated from observations available before death. For each laboratory variable, the first recorded measurement was retained, whereas vital signs were summarized as the minimum, maximum, and mean.

Data Pre-processing

A uniform preprocessing pipeline was applied across all three cohorts. Invasive and non-invasive blood pressure measurements were combined, with invasive measurements prioritized when both were available at the same time point. Values outside predefined physiologic ranges were treated as missing (Supplementary Table 1). Features without a valid measurement within the prespecified 24-hour window were imputed using medians derived exclusively from the eICU-CRD training set, and these fixed values were applied to the validation and external cohorts.

Conventional model development and validation

Six supervised models were developed to predict 7-day mortality after ICU admission: logistic regression, random forest, and XGBoost as ML models, and a convolutional neural network (CNN), multilayer perceptron (MLP), and transformer as DL models. The eICU-CRD cohort was first stratified by 7-day mortality status and divided into development and test sets in an 80:20 ratio. Within the development set, 20% of the data were reserved for validation and hyperparameter selection. Hyperparameters were optimized by grid search, with the search spaces and selected values reported in Supplementary Table 2. Final model performance was evaluated on the held-out eICU-CRD test set and externally validated in MIMIC-IV and K-MIMIC.

Large language model evaluation

Two general-purpose LLMs, GPT-5 (gpt-5-2025-08-07; OpenAI) [34] and Gemini-2.5-Pro (gemini-2.5-pro; Google) [35], were assessed using four prompting strategies: zero-shot [36], chain-of-thought (CoT) [37], instruction-based CoT incorporating explicit reasoning constraints and structured output requirements [38], and agentic CoT, which simulated sequential multidisciplinary ICU roles [39,40]. Both models were applied to the same random sample of 2,000 ICU stays from each cohort, proportionally stratified by 7-day mortality status to preserve the outcome prevalence of the source cohort. Each model was queried using its respective default generation settings. GPT-5 sampling parameters were not user-configurable. To assess run-to-run consistency, GPT-5 was evaluated in three independent runs for each prompting strategy and cohort. The first run was used for the primary performance comparison, and the additional runs were used to assess run-to-run consistency. Full prompt specifications are provided in Supplementary Figure 1. All four prompting strategies were included in the performance evaluation of both LLMs. Detailed reasoning alignment, run-to-run consistency, and clinician evaluation analyses were

conducted for GPT-5, with the instruction-based CoT strategy used for the reasoning-focused analyses to facilitate standardized comparison across patients and cohorts.

SHAP-based model explainability and SHAP-LLM reasoning alignment

To evaluate whether LLM-generated reasoning reflected influential patient-level predictors, Shapley additive explanations (SHAP) values from a reference model selected based on predictive performance were compared with clinical concepts identified in the generated reasoning [41]. Structured EHR features and reasoning terms were mapped to a shared set of 22 clinical concepts reviewed and refined by two board-certified intensivists. Each concept was linked to its constituent EHR variables and predefined reasoning terms (Supplementary Table 3), and concept mentions were identified using rule-based text matching (Supplementary Table 4). Concept-level importance was calculated for each patient by summing the absolute SHAP values of the corresponding variables. This concept-level comparison was consistent with prior clinical text evaluation approaches based on medical concept or entity overlap [42,43]. Alignment with the top-k SHAP-ranked concepts was quantified using precision, recall, F1-score, and the Jaccard index.

We also calculated D, defined as the difference between the mean concept-level SHAP importance of mentioned and non-mentioned concepts for each patient; D>0 indicated preferential reference to more influential concepts. Analyses were based on GPT-5 reasoning generated with the instruction-based CoT prompt, which provided a single standardized rationale based on the supplied patient variables.

Clinician trustworthiness evaluation

For the clinician evaluation, 50 GPT-5 reasoning outputs generated using the instruction-based CoT prompt were randomly sampled from each cohort, yielding 150 cases. Two medical experts independently assigned a single global rating to each output using a predefined 5-point scale. The rating jointly reflected consistency with the provided patient data, recognition and prioritization of major risk factors, coherence of the clinical interpretation, the presence of hallucinated or clinically important omitted findings, and internal consistency between the reasoning and the model's predicted 7-day mortality risk. Minor inaccuracies were considered in the context of the overall reasoning. Full rating anchors are provided in Supplementary Table 5.

Statistical Analysis

Baseline characteristics were summarized as mean ± standard deviation for continuous variables and counts (%) for categorical variables. 95% confidence intervals (CIs) for model performance metrics were estimated using 1,000 bootstrap resamples. Optimal classification thresholds for the ML and DL models were selected using the Youden index in the eICU-CRD validation set and fixed for internal and external evaluation. For the LLMs, threshold-dependent metrics were estimated by 5-fold cross-validation in eICU-CRD, and full-sample Youden thresholds were fixed for external evaluation.

For the SHAP–LLM alignment analysis, patient-level D values were compared with zero using the Wilcoxon signed-rank test. For each clinical concept, concept-level SHAP importance was compared according to mention status using the Mann–Whitney U test. P values from the concept-level analyses were adjusted for multiple comparisons using the Benjamini–Hochberg false discovery rate procedure. Clinician trustworthiness ratings were summarized as mean ± standard deviation for each rater and for pooled ratings across both raters. Agreement between the two raters was summarized as exact agreement and agreement within one scale point. Inter-rater agreement was assessed using Gwet's AC2 [44] and Cohen's κ [45], both with quadratic weights. All statistical tests were two-sided, and P values or false discovery rate-adjusted q values ≤0.05, as applicable, were considered statistically significant. All analyses were performed using Python version 3.10.16.

Results

Baseline Characteristics

Baseline characteristics are summarized in Table 1. The final analysis included 124,020 ICU stays from eICU-CRD, 84,926 from MIMIC-IV, and 6,673 from K-MIMIC, with 7-day mortality rates of 2.5%, 5.8%, and 2.9%, respectively. Across the three cohorts, non-survivors were older than survivors, with mean ages of 66.6 versus 61.0 years in eICU-CRD, 70.3 versus 63.9 years in MIMIC-IV, and 67.3 versus 63.2 years in K-MIMIC. Non-survivors also had lower GCS scores, with the largest differences observed in eICU-CRD and MIMIC-IV. The proportions of male and female patients were generally comparable between outcome groups. The ethnic composition differed across cohorts, with Caucasian patients comprising the majority in the two US cohorts, whereas all patients included in K-MIMIC were Asian.

Laboratory profiles also differed between survivors and non-survivors. BUN levels were higher in non-survivors than in survivors (36.8-40.0 vs 22.0-24.4 mg/dL). Non-survivors also had higher creatinine and bilirubin levels, lower albumin levels, and prolonged PT and aPTT. Hemoglobin and platelet counts were generally lower, whereas glucose levels were higher among non-survivors.

Vital signs also differed between survivors and non-survivors. Mean systolic blood pressure was 10–19 mmHg lower in non-survivors, whereas mean heart rate and respiratory rate were consistently higher. Minimum peripheral oxygen saturation was lower in non-survivors than in survivors (81.6%-88.5% vs 88.6%-94.2%), and minimum body temperature was also consistently lower.

Table 1. Baseline characteristics of the three ICU cohorts by 7-day mortality status.

Characteristic	eICU-CRD (n = 124,020)		MIMIC-IV (n = 84,926)		K-MIMIC (n = 6,673)	
	Survivors	Non-survivors	Survivors	Non-survivors	Survivors	Non-survivors
Number of participants, n (%)	120,914 (97.5%)	3,106 (2.5%)	79,971 (94.2%)	4,955 (5.8%)	6,481 (97.1%)	192 (2.9%)
Age, years, mean (SD)	61.0 (16.8)	66.6 (16.0)	63.9 (16.6)	70.3 (15.3)	63.2 (15.4)	67.3 (13.7)
Sex, n (%)						
Male	66,773 (55.2%)	1,728 (55.6%)	45,583 (57.0%)	2,694 (54.4%)	4,094 (63.2%)	108 (56.2%)
Female	54,141 (44.8%)	1,378 (44.4%)	34,388 (43.0%)	2,261 (45.6%)	2,387 (36.8%)	84 (43.8%)
Ethnicity, n (%)						
Caucasian	91,728 (75.9%)	2,284 (73.5%)	53,446 (66.8%)	2,953 (59.6%)	0 (0.0%)	0 (0.0%)
African American	14,595 (12.1%)	450 (14.5%)	9,038 (11.3%)	482 (9.7%)	0 (0.0%)	0 (0.0%)
Hispanic	4,922 (4.1%)	126 (4.1%)	3,255 (4.1%)	139 (2.8%)	0 (0.0%)	0 (0.0%)
Asian	1,761 (1.5%)	52 (1.7%)	2,434 (3.0%)	147 (3.0%)	6,481 (100.0%)	192 (100.0%)
Native American	872 (0.7%)	17 (0.5%)	281 (0.4%)	27 (0.5%)	0 (0.0%)	0 (0.0%)
Other/Unknown	7,036 (5.8%)	177 (5.7%)	11,517 (14.4%)	1,207 (24.4%)	0 (0.0%)	0 (0.0%)
Glasgow Coma Score, mean (SD)	13.3 (3.1)	9.6 (4.9)	11.6 (4.6)	9.3 (5.0)	14.3 (1.5)	13.0 (3.0)
Laboratory results, mean (SD)						
ALT (units/L)	73.5 (263.1)	263.9 (622.1)	98.5 (347.2)	227.1 (557.1)	60.2 (167.6)	225.6 (401.2)
AST (units/L)	104.4 (434.5)	479.5 (1143.8)	158.3 (646.4)	460.9 (1201.1)	79.0 (195.1)	426.6 (725.6)
Albumin (g/dL)	3.0 (0.7)	2.5 (0.7)	3.2 (0.6)	2.8 (0.8)	3.3 (0.6)	2.7 (0.5)
BUN (mg/dL)	23.7 (19.2)	36.8 (24.9)	24.4 (20.1)	38.0 (26.4)	22.0 (16.6)	40.0 (22.9)
Bilirubin (mg/dL)	1.1 (2.0)	1.9 (4.0)	1.6 (3.5)	3.0 (6.4)	1.4 (2.0)	3.2 (4.2)
Chloride (mmol/L)	104.4 (6.3)	104.2 (7.6)	103.9 (6.4)	102.5 (7.9)	104.7 (5.0)	101.8 (7.2)
Creatinine (mg/dL)	1.4 (1.6)	2.1 (1.8)	1.4 (1.5)	2.0 (1.6)	1.3 (1.4)	1.9 (1.2)
Glucose (mg/dL)	149.2 (75.0)	167.2 (91.0)	142.7 (63.0)	167.9 (91.9)	146.2 (56.0)	182.2 (98.5)
Hemoglobin (g/dL)	11.2 (2.3)	10.7 (2.6)	10.7 (2.3)	10.3 (2.5)	10.9 (2.0)	9.2 (1.8)
Prothrombin time (INR)	1.5 (0.8)	2.0 (1.2)	1.4 (0.7)	1.9 (1.1)	1.3 (0.4)	1.9 (0.7)
Platelets (10^3/μL)	205.9 (93.0)	185.0 (103.9)	208.7 (104.3)	193.3 (121.1)	176.4 (84.5)	113.9 (97.9)
Potassium (mmol/L)	4.1 (0.7)	4.4 (1.0)	4.2 (0.7)	4.4 (0.9)	4.0 (0.6)	4.2 (1.0)
Sodium (mmol/L)	138.2 (4.9)	138.6 (6.6)	137.8 (5.0)	137.9 (6.6)	137.5 (4.9)	136.8 (7.3)
WBC (10^3/μL)	11.7 (6.3)	14.9 (10.3)	11.5 (6.5)	14.9 (10.0)	11.5 (6.2)	11.9 (11.2)
aPTT (sec)	38.7 (22.9)	48.8 (30.2)	36.8 (22.4)	49.1 (33.5)	32.8 (8.3)	42.7 (13.2)
Vital signs, mean (SD)						
Heart rate (/min)						
Minimum	68.2 (15.2)	64.3 (23.6)	70.4 (14.8)	73.8 (21.0)	65.0 (14.8)	71.8 (23.5)
Maximum	107.6 (21.2)	122.3 (25.9)	102.2 (20.1)	114.1 (24.6)	97.3 (24.9)	135.0 (27.5)
Mean	83.6 (15.9)	93.8 (19.9)	84.2 (15.5)	92.9 (19.1)	78.6 (16.8)	102.1 (18.2)
Respiratory rate (/min)						
Minimum	11.1 (3.9)	12.0 (5.0)	12.4 (3.4)	14.0 (4.8)	12.2 (3.0)	14.1 (4.9)
Maximum	31.0 (8.5)	33.7 (8.6)	27.5 (6.1)	30.4 (6.7)	26.8 (7.0)	33.5 (8.0)
Mean	19.1 (4.2)	22.0 (5.5)	19.0 (3.7)	22.0 (4.8)	17.8 (3.1)	22.1 (4.3)
Systolic blood pressure (mmHg)						
Minimum	89.9 (20.3)	69.3 (22.1)	93.8 (16.8)	80.7 (20.5)	99.5 (20.7)	67.2 (19.4)
Maximum	158.6 (26.2)	160.3 (31.7)	147.2 (22.2)	142.7 (27.5)	156.7 (24.5)	158.5 (34.9)

Characteristic	eICU-CRD (n = 124,020)		MIMIC-IV (n = 84,926)		K-MIMIC (n = 6,673)	
	Survivors	Non-survivors	Survivors	Non-survivors	Survivors	Non-survivors
Mean	122.5 (18.3)	109.7 (19.7)	119.0 (16.4)	109.6 (18.4)	126.3 (16.4)	107.3 (18.5)
Diastolic blood pressure (mmHg)						
Minimum	46.5 (12.4)	35.8 (13.0)	48.5 (11.5)	42.3 (12.4)	57.1 (14.0)	37.9 (13.1)
Maximum	95.0 (17.4)	95.8 (19.7)	87.5 (16.8)	85.3 (19.0)	93.2 (14.5)	92.9 (18.9)
Mean	65.6 (11.6)	59.9 (12.1)	64.5 (11.4)	60.4 (11.7)	72.5 (11.0)	62.0 (11.1)
SpO_2 (%)						
Minimum	88.6 (6.8)	81.6 (9.6)	92.2 (4.3)	88.5 (7.9)	94.2 (4.8)	84.4 (9.6)
Maximum	99.6 (0.9)	99.4 (2.2)	99.4 (1.1)	99.0 (2.9)	99.7 (0.7)	99.3 (2.8)
Mean	96.9 (2.0)	95.2 (4.5)	96.9 (1.9)	95.6 (4.1)	98.2 (1.5)	95.3 (4.8)
Body temperature (°C)						
Minimum	36.3 (0.7)	35.6 (1.5)	36.4 (0.6)	36.0 (1.1)	36.0 (0.6)	35.2 (1.1)
Maximum	37.3 (0.7)	37.2 (1.3)	37.3 (0.6)	37.2 (1.1)	37.2 (0.8)	37.1 (1.1)
Mean	36.8 (0.5)	36.5 (1.2)	36.8 (0.4)	36.6 (0.9)	36.6 (0.5)	36.2 (0.9)

Data are presented as mean (standard deviation) for continuous variables and n (%) for categorical variables.

ALT, alanine aminotransferase; aPTT, activated partial thromboplastin time; AST, aspartate aminotransferase; BUN, blood urea nitrogen; eICU-CRD, eICU Collaborative Research Database; GCS, Glasgow Coma Scale; INR, international normalized ratio; K-MIMIC, Korea Medical Information Mart for Intensive Care; MIMIC-IV, Medical Information Mart for Intensive Care IV; SD, standard deviation; SpO_2, peripheral oxygen saturation; WBC, white blood cell count.

Model Performance

Model performance across the three cohorts is presented in Table 2. Conventional ML and DL models generally showed the highest discrimination. In the held-out eICU-CRD cohort, XGBoost achieved the highest area under the receiver operating characteristic curve (AUROC) and area under the precision–recall curve (AUPRC), at 0.936 (95% CI, 0.923–0.946) and 0.514 (95% CI, 0.465–0.561), respectively. Among the LLMs, GPT-5 achieved its highest AUROC with agentic CoT at 0.925 (95% CI, 0.881–0.959), whereas Gemini-2.5-Pro achieved its highest AUROC with instruction-based CoT at 0.906 (95% CI, 0.862–0.944).

In MIMIC-IV, the MLP achieved the highest AUROC at 0.858 (95% CI, 0.853–0.862), with similar values for XGBoost and the transformer at 0.857 and 0.856, respectively. Across prompting strategies, GPT-5 achieved AUROCs ranging from 0.814 to 0.822, compared with 0.784 to 0.800 for Gemini-2.5-Pro. In K-MIMIC, XGBoost achieved the highest AUROC at 0.957 (95% CI, 0.921–0.985). GPT-5 achieved AUROCs ranging from 0.902 to 0.926, whereas Gemini-2.5-Pro achieved AUROCs ranging from 0.856 to 0.884.

Across all three cohorts, the highest AUROC achieved by GPT-5 was higher than that achieved by Gemini-2.5-Pro. The prompting strategy associated with the highest AUROC varied across models and cohorts, with no single strategy consistently yielding the highest value. Figure 3 shows receiver operating characteristic curves for selected models in the held-out eICU-CRD cohort. Across three independent GPT-5 runs, AUROC estimates remained stable, with limited run-to-run variation across prompting strategies and cohorts (Supplementary Table 6).

Table 2. Performance of supervised models and large language models for 7-day mortality prediction across three ICU cohorts.

Model	Approach	eICU-CRD					MIMIC-IV				
		Threshold	AUROC	AUPRC	Sensitivity	Specificity	Threshold	AUROC	AUPRC	Sensitivity	Specificit
Machine learning											
Logistic Regression		0.494	0.917 (0.902–0.930)	0.430 (0.381–0.479)	0.812 (0.776–0.846)	0.864 (0.860–0.869)	0.494	0.819 (0.814-0.825)	0.232 (0.221-0.243)	0.870 (0.861-0.879)	0.589 (0.586-0.593)
Random Forest		0.034	0.931 (0.917–0.942)	0.486 (0.440–0.532)	0.845 (0.807–0.876)	0.877 (0.873–0.881)	0.033	0.836 (0.831-0.841)	0.241 (0.229-0.252)	0.821 (0.811-0.832)	0.709 (0.706-0.712)
XGBoost		0.010	0.936 (0.923–0.946)	0.514 (0.465–0.561)	0.864 (0.833–0.894)	0.847 (0.842–0.851)	0.010	0.857 (0.852-0.861)	0.293 (0.280-0.306)	0.878 (0.869-0.886)	0.670 (0.667-0.674)
Deep learning											
CNN		0.339	0.918 (0.903–0.932)	0.399 (0.351–0.448)	0.819 (0.784–0.851)	0.882 (0.878–0.886)	0.339	0.811 (0.806-0.816)	0.189 (0.181-0.199)	0.782 (0.771-0.793)	0.718 (0.715-0.721)
MLP		0.458	0.927 (0.914–0.940)	0.483 (0.435–0.531)	0.827 (0.791–0.859)	0.868 (0.863–0.872)	0.458	0.858 (0.853-0.862)	0.320 (0.307-0.335)	0.869 (0.859-0.878)	0.675 (0.672-0.679)
Transformer		0.641	0.932 (0.918–0.944)	0.499 (0.452–0.547)	0.838 (0.803–0.871)	0.891 (0.887–0.894)	0.641	0.856 (0.852-0.861)	0.291 (0.279-0.304)	0.868 (0.859-0.878)	0.686 (0.683-0.690)
Large language models											
GPT-5	Zero-Shot	0.340	0.901 (0.852-0.941)	0.269 (0.146-0.449)	0.811 (0.684-0.920)	0.753 (0.734-0.771)	0.340	0.819 (0.787-0.850)	0.212 (0.159-0.284)	0.957 (0.918-0.991)	0.472 (0.449-0.494)
	CoT	0.340	0.908 (0.866-0.943)	0.262 (0.139-0.412)	0.811 (0.676-0.932)	0.827 (0.809-0.843)	0.340	0.822 (0.789-0.852)	0.214 (0.162-0.287)	0.932 (0.882-0.973)	0.540 (0.517-0.563)
	Instruction-based CoT	0.380	0.912 (0.866-0.950)	0.342 (0.201-0.513)	0.784 (0.643-0.909)	0.823 (0.807-0.840)	0.380	0.814 (0.779-0.846)	0.200 (0.150-0.266)	0.906 (0.848-0.956)	0.558 (0.535-0.580)
	Agentic Approach CoT	0.320	0.925 (0.881-0.959)	0.376 (0.233-0.531)	0.865 (0.738-0.969)	0.832 (0.816-0.849)	0.320	0.822 (0.789-0.853)	0.230 (0.169-0.299)	0.923 (0.871-0.966)	0.526 (0.502-0.549)
Gemini-2.5-Pro	Zero-Shot	0.620	0.900 (0.851-0.936)	0.202 (0.110-0.333)	0.919 (0.812-1.000)	0.719 (0.699-0.738)	0.620	0.800 (0.764-0.835)	0.176 (0.136-0.225)	0.906 (0.848-0.957)	0.463 (0.439-0.487)
	CoT	0.650	0.874 (0.816-0.925)	0.211 (0.101-0.350)	0.735 (0.571-0.882)	0.776 (0.757-0.794)	0.650	0.798 (0.762-0.832)	0.174 (0.132-0.230)	0.930 (0.875-0.973)	0.474 (0.450-0.497)
	Instruction-based CoT	0.700	0.906 (0.862-0.944)	0.215 (0.129-0.347)	0.865 (0.744-0.967)	0.767 (0.748-0.785)	0.700	0.800 (0.763-0.833)	0.170 (0.131-0.224)	0.872 (0.810-0.932)	0.530 (0.508-0.552)
	Agentic Approach CoT	0.400	0.888 (0.840-0.932)	0.187 (0.103-0.326)	0.686 (0.524-0.833)	0.782 (0.764-0.799)	0.400	0.784 (0.745-0.820)	0.184 (0.137-0.243)	0.921 (0.870-0.967)	0.428 (0.405-0.450)

Values are point estimates with 95% confidence intervals. Bold indicates the highest AUROC, AUPRC, sensitivity, or specificity within each cohort.
AUPRC, area under the precision–recall curve; AUROC, area under the receiver operating characteristic curve; CNN, convolutional neural network; CoT, chain-of-thought; eICU-CRD, eICU Collaborative Research Database; K-MIMIC, Korea Medical Information Mart for Intensive Care; MIMIC-IV, Medical Information Mart for Intensive Care IV; MLP, multilayer perceptron.

Alignment Between LLM Reasoning and SHAP-Based Predictor Importance

In the eICU-CRD cohort, the most influential XGBoost predictors based on mean absolute SHAP values were minimum systolic blood pressure, GCS, mean heart rate, mean respiratory rate, and mean systolic blood pressure. Other highly ranked predictors included heart-rate extremes, peripheral oxygen saturation, body temperature, and BUN, indicating that measures of early physiologic and neurologic status predominated among the highest-ranked predictors (Figure 4a). Illustrative survivor and non-survivor cases show that GPT-5 reasoning referenced several highly ranked predictors across the four

prompting strategies (Figure 4b-c). Additional reasoning outputs for Gemini-2.5-Pro in eICU-CRD and for both LLMs in MIMIC-IV and K-MIMIC are provided in Supplementary Figures 2–4.

Patient-level alignment between GPT-5 reasoning and XGBoost-derived concept-level SHAP importance is summarized in Table 3. Across the three cohorts, GPT-5 mentioned a mean of 6.2 ± 2.0 clinical concepts per patient. In the pooled analysis, precision@5, recall@5, F1@5, and Jaccard@5 were 0.47, 0.55, 0.49, and 0.34, respectively. The median difference in mean SHAP importance between mentioned and non-mentioned concepts was 0.206 (interquartile range, 0.131–0.305), with D greater than zero in 98.1% of patients. Similar patterns were observed in each cohort, with positive D values in 98.0% of eICU-CRD, 98.9% of MIMIC-IV, and 97.4% of K-MIMIC patients. All corresponding Wilcoxon signed-rank tests yielded P<0.001.

At the concept level, peripheral oxygen saturation, neurological status, and blood pressure were each mentioned in more than 90% of GPT-5 rationales. Associations between mention status and SHAP importance varied across concepts, while sex and ethnicity were not explicitly mentioned (Supplementary Table 7).

Table 3. Patient-level alignment between GPT-5 reasoning and concept-level SHAP importance across three cohorts.

Cohort	Patients (n)	Mentioned concepts, mean ± SD	Precision@5	Recall@5	F1@5	Jaccard@5	Median D (IQR)	D>0, %	Wilcoxon p
eICU	2,000	6.9 ± 1.9	0.42	0.56	0.47	0.32	0.175 (0.110–0.249)	1,961 (98.0)	<0.001
MIMIC-IV	2,000	6.6 ± 2.0	0.45	0.56	0.49	0.33	0.211 (0.142–0.310)	1,977 (98.9)	<0.001
K-MIMIC	2,000	5.2 ± 1.8	0.52	0.52	0.51	0.36	0.237 (0.151–0.368)	1,947 (97.4)	<0.001
Pooled	6,000	6.2 ± 2.0	0.47	0.55	0.49	0.34	0.206 (0.131–0.305)	5,885 (98.1)	<0.001

D was defined as the mean concept-level SHAP importance of mentioned concepts minus that of non-mentioned concepts.
IQR, interquartile range; SD, standard deviation; SHAP, SHapley Additive exPlanations.

Clinician-Rated Trustworthiness of LLM Reasoning

Among 150 evaluated cases, the combined mean trustworthiness rating was 4.03 ± 0.63 on a 5-point scale (Table 4). Cohort-specific mean ratings were 4.15 ± 0.52 in eICU-CRD, 4.04 ± 0.64 in MIMIC-IV, and 3.89 ± 0.70 in K-MIMIC. The two raters assigned overall mean scores of 4.19 ± 0.84 and 3.87 ± 0.80, respectively.

Exact agreement was observed in 40% of cases, while ratings differed by no more than one point in 83%. Overall inter-rater agreement was 0.80 (95% CI, 0.75–0.85) according to Gwet's AC2 and 0.16 (95% CI, 0.02–0.30) according to quadratic-weighted κ.

Table 4. Clinician-rated trustworthiness of GPT-5 reasoning by cohort.

Cohort	Patients, n	Rater 1, mean ± SD	Rater 2, mean ± SD	Combined, mean ± SD	Exact agreement, %	Agreement within ±1, %	Gwet's AC2 (95% CI)	Quadratic-weighted κ (95% CI)
eICU-CRD	50	4.46 ± 0.73	3.84 ± 0.62	4.15 ± 0.52	32	88	0.82 (0.74-0.89)	0.11 (-0.05-0.31)
MIMIC-IV	50	4.16 ± 0.74	3.92 ± 0.97	4.04 ± 0.64	50	78	0.77 (0.64-0.88)	0.10 (-0.17-0.37)
K-MIMIC	50	3.94 ± 0.96	3.84 ± 0.79	3.89 ± 0.70	38	84	0.81 (0.73-0.88)	0.28 (0.04-0.49)
Overall	**150**	**4.19 ± 0.84**	**3.87 ± 0.80**	**4.03 ± 0.63**	**40**	**83**	0.80 (0.75-0.85)	0.16 (0.02, 0.30)

Combined values represent ratings pooled across both raters. Agreement within one point indicates that the two ratings differed by no more than one scale point.

AC2, agreement coefficient 2; CI, confidence interval; eICU-CRD, eICU Collaborative Research Database; K-MIMIC, Korea Medical Information Mart for Intensive Care; MIMIC-IV, Medical Information Mart for Intensive Care IV; SD, standard deviation.

Discussion

In this multicohort study, we evaluated general-purpose LLMs for 7-day mortality prediction using routinely collected structured EHR data from the first 24 hours of ICU admission. Although supervised ML and DL models generally achieved the highest discrimination, GPT-5 showed competitive discrimination through prompting alone, without cohort-specific training or task-specific fine-tuning. Beyond risk prediction, it generated patient-specific reasoning in human-readable language that clinicians could directly review. This reasoning preferentially referenced clinical concepts with greater patient-level importance in the reference prediction model and was rated as clinically trustworthy by the evaluating clinicians. Together, these findings support a broader evaluation framework for LLMs that considers predictive performance, reasoning alignment, and the perceived trustworthiness of generated reasoning.

Predictive performance and prompting strategies

Supervised ML and DL models generally achieved higher AUROC than directly prompted LLMs across all three cohorts, consistent with previous clinical prediction benchmarks [21, 22]. Developed exclusively in eICU-CRD, these models also achieved higher AUROC when applied without retraining to MIMIC-IV and K-MIMIC. Nevertheless, GPT-5 achieved competitive discrimination without cohort-specific training or task-specific fine-tuning, consistent with reports of substantial zero-shot predictive capability on structured EHR data [20]. Its highest AUROC was within 0.011–0.036 of the best supervised model in each cohort. However, LLMs generally showed lower AUPRC and high sensitivity accompanied by lower specificity at thresholds derived separately for each model family, indicating that competitive AUROC did not consistently extend to other performance measures. This may partly reflect LLM-generated probability outputs, which may require operating thresholds different from those of outcome-trained models and vary across prompting strategies [28].

No prompting strategy was consistently superior, and the best-performing strategy varied across models and cohorts. Agentic CoT ranked highest for GPT-5 in eICU-CRD but lowest in K-MIMIC, while ranking lowest for Gemini-2.5-Pro in both MIMIC-IV and K-MIMIC. These findings indicate that prompt design should be validated for the intended model, task, and clinical setting, consistent with prior reports of prompt-sensitive behavior in ICU mortality prediction [28].

Reasoning alignment with model-derived predictor importance

The most influential XGBoost predictors, including blood pressure, neurological status, heart rate, and respiratory rate, reflected domains central to established ICU severity scores [5, 6], supporting the plausibility of the reference model. Extending prior work based on cohort-level feature rankings or attribution-to-text generation [28, 29], we compared independently generated GPT-5 reasoning with patient-specific attributions across three ICU cohorts. Although overlap with the highest-ranked SHAP concepts was incomplete, GPT-5 preferentially referenced concepts with greater patient-level SHAP importance in more than 97% of patients in each cohort, suggesting that the reasoning emphasized factors more influential to estimated risk rather than merely summarizing the supplied variables.

Nevertheless, alignment was partial and varied across concepts. Cross-cohort differences in precision and recall may partly reflect the number of concepts articulated. Frequently mentioned concepts, such as blood pressure, did not necessarily show greater patient-specific importance, whereas selective emphasis was more apparent for concepts articulated in only a subset of patients. Concept-level associations were not uniform, with a small number of laboratory concepts showing the opposite direction. Sex and ethnicity were among the least important concepts in the reference model, yet neither was mentioned in any generated reasoning, whereas other low-importance concepts were referenced occasionally. However, their absence from the generated reasoning does not establish that they did not influence the predictions. Overall, the findings indicate measurable alignment with influential patient factors, but not complete reproduction of the reference model's feature hierarchy or faithful representation of the LLM's internal computation [26, 27].

Clinician-rated trustworthiness

Clinician evaluation provided a complementary assessment of the reasoning's clinical accuracy, coherence, grounding in the supplied patient data, and consistency with the predicted outcome. The mean trustworthiness rating was 4.03 ± 0.63 on a 5-point scale and was broadly similar across cohorts, indicating that clinicians generally considered the reasoning clinically trustworthy, consistent with a previous study in which clinicians rated LLM-generated narrative explanations of an ICU prediction model as helpful and safe [29]. Exact agreement was 40%, while ratings differed by no more than one point in 83% of cases, indicating that most disagreements were limited in magnitude. Some variation is expected in open-ended clinical text evaluation, where judgments of relevance, coherence, and clinical importance may remain partly subjective even among medical experts [46]. Chance-corrected agreement differed substantially between Gwet's AC2 and Cohen's κ, both computed with quadratic weights (0.80 and 0.16, respectively). The concentration of ratings near the upper end of the scale may have attenuated κ through its sensitivity to imbalanced marginal distributions, whereas AC2 is less sensitive to this effect [44, 47]. The observed agreement proportions should therefore be considered alongside both chance-corrected coefficients. These findings support the perceived clinical trustworthiness of the generated reasoning but do not establish externally validated reasoning quality or clinical utility.

Clinical implications

For clinical use, the value of general-purpose LLMs may lie less in outperforming locally developed models, as they generally did not in this study, and more in providing clinically informative risk stratification where local model development is not feasible. Prompt-based estimation required no cohort-specific model fitting or retraining, although structured data extraction would still be necessary and prompt selection remains a form of local tuning [28]. The lower AUPRC, together with reduced specificity at the evaluated thresholds, indicates that automated alerting could generate a substantial false-positive burden, and operating thresholds would therefore need to be selected locally before deployment. A more plausible near-term role is to accompany risk estimates with patient-specific reasoning that clinicians can inspect during admission review or handover, potentially directing attention to salient physiological,

neurological, and laboratory abnormalities. The preliminary GPT-5 alignment and clinician-rating findings support prospective testing of this role, including whether such rationales improve recognition of important clinical information without inappropriately influencing clinicians' judgments or confidence [13].

Limitations and strengths

Several limitations of this study should be considered. Due to data limitations, several clinically informative variables, including pre-existing comorbidities, were not included, which may have limited the available prognostic information. Missing values were imputed before LLM evaluation without being identified as such, although real-world clinical data are often incomplete and missingness may itself carry clinical information [48].

Potential exposure to the publicly available eICU-CRD and MIMIC-IV datasets during LLM pretraining could not be excluded. However, the lower discrimination in MIMIC-IV than in the non-public K-MIMIC cohort suggests that public-data exposure alone does not fully explain the observed performance. Furthermore, the use of proprietary models also limited reproducibility and transparency because their training data, architectures, and updates are not fully disclosed [49]. Future studies should compare proprietary and open-source models, including locally deployable options, with respect to performance, reproducibility, and data governance.

Moreover, SHAP-based alignment reflected concordance with patient-specific XGBoost attributions rather than clinical ground truth. XGBoost was used as the reference because it showed strong discrimination and enabled exact patient-level attribution through TreeSHAP, within the widely used SHAP framework for model explanation [50, 51]. The observed concordance should therefore be interpreted as convergence between language-based and data-driven prioritization, rather than as direct clinical validation [27].

In addition, clinician evaluation was based on two clinicians, 150 GPT-5 rationales, and a single prompting strategy, limiting the generalizability of the ratings. The rating scale may also have provided limited differentiation among outputs, given the concentration of scores near the upper end of the scale [52]. Moreover, the scores reflected perceived trustworthiness rather than externally validated reasoning quality. Broader evaluation using larger clinician panels and a wider range of models and prompting strategies is warranted.

Despite these limitations, this study has several notable strengths. The same set of structured predictors and prompting strategies was applied across three independent ICU cohorts from two countries without cohort-specific training or task-specific fine-tuning. LLM performance was benchmarked against six conventional ML and DL models developed within the same study framework, with additional assessment of GPT-5 run-to-run variability across three independent runs. Reasoning was evaluated through complementary quantitative and clinical approaches, combining patient-level concordance between LLM reasoning and model-derived predictor importance with clinician-rated trustworthiness. To our knowledge, this is the first multicohort study to integrate predictive benchmarking, patient-level reasoning alignment, and clinician-rated trustworthiness within a single framework for early ICU mortality prediction.

Conclusion

In this multicohort study, supervised ML and DL models generally achieved the highest discrimination for early mortality after ICU admission, whereas GPT-5 achieved competitive discrimination from routinely collected structured EHR data without cohort-specific model development. Its patient-specific reasoning preferentially referenced clinical concepts with greater patient-level importance in the reference prediction model and received generally favorable clinician ratings. Prospective evaluation of reasoning faithfulness, clinical utility, and implementation safety is needed before such models are used to support decision-making in critical care.

Abbreviations

ALT: Alanine aminotransferase; APACHE: Acute Physiology and Chronic Health Evaluation; aPTT: Activated partial thromboplastin time; AST: Aspartate aminotransferase; AUPRC: Area under the precision–recall curve; AUROC: Area under the receiver operating characteristic curve; BUN: Blood urea nitrogen; CI: Confidence interval; CNN: Convolutional neural network; CoT: Chain-of-thought; DL: Deep learning; EHR: Electronic health record; eICU-CRD: eICU Collaborative Research Database; GCS: Glasgow Coma Scale; ICU: Intensive care unit; K-MIMIC: Korea Medical Information Mart for Intensive Care; LLM: Large language model; MIMIC-IV: Medical Information Mart for Intensive Care IV; ML: Machine learning; MLP: Multilayer perceptron; PT: Prothrombin time; SAPS: Simplified Acute Physiology Score; SHAP: Shapley additive explanations; SNUH: Seoul National University Hospital; TRIPOD+AI: Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis–Artificial Intelligence; WBC: White blood cell count

Declarations

Ethics approval and consent to participate

This study was approved by the Institutional Review Board (IRB) of Seoul National University Hospital (IRB No. E-2508-001-1661) and conducted with institutional guidelines and applicable regulations. All data were analyzed in a de-identified manner. The requirement for informed consent to participate was waived by the Institutional Review Board of Seoul National University Hospital. All procedures performed in this study were in accordance with the ethical standards of the institutional and national regulations and with the 1964 Helsinki Declaration and its later amendments.

Clinical trial number

Not applicable.

Consent for publication

Not applicable.

Availability of data and materials

The datasets analyzed during the current study include MIMIC-IV and eICU-CRD, which are available from PhysioNet under the applicable access requirements, and K-MIMIC SNUH Subset, which is available only upon separate approval through the KHDP platform and is therefore not publicly available.

Competing interests

The authors declare that they have no competing interests.

Funding

No funding was received for this study.

Authors' contributions

Cho S and Choi S conceptualized the study, curated the data, developed the methodology, performed the analyses, prepared the visualizations, administered the project, and wrote the original manuscript. Choi S, Choi Y, Jung H.D., and Gu C contributed to software development and validation. Choi Y, Kim M, Panitchote A, Ketdao N, and Ratchatorn A contributed to data curation, validation, investigation, and methodological refinement. Park T, Moon D, and Park S contributed to formal analysis and validation. Gu C and Lee H.C. provided resources. Lee H.C. and Kim H supervised the study and contributed to project administration. Kim H acquired funding. All authors reviewed and edited the manuscript and approved the final version.

Acknowledgements

Not applicable.

References

1. Adhikari NKJ, Fowler RA, Bhagwanjee S, Rubenfeld GD. Critical care and the global burden of critical illness in adults. Lancet. 2010;376:1339–46. https://doi.org/10.1016/S0140-6736(10)60446-1.
2. Flaatten H, de Lange DW, Artigas A, Bin D, Moreno R, Christensen S, et al. The status of intensive care medicine research and a future agenda for very old patients in the ICU. Intensive Care Med. 2017;43:1319–28. https://doi.org/10.1007/s00134-017-4718-z.
3. Ferrante LE. Prognostication during Critical Illness: Moving the Field Forward. Ann Am Thorac Soc. 2017;14:1510–1. https://doi.org/10.1513/AnnalsATS.201706-496ED.
4. Iwashyna TJ, Hodgson CL, Pilcher D, Bailey M, van Lint A, Chavan S, et al. Timing of onset and burden of persistent critical illness in Australia and New Zealand: a retrospective, population-based, observational study. Lancet Respir Med. 2016;4:566–73. https://doi.org/10.1016/S2213-2600(16)30098-4.
5. Knaus WA, Draper EA, Wagner DP, Zimmerman JE. APACHE II: a severity of disease classification system. Crit Care Med. 1985;13:818–29. https://doi.org/10.1097/00003465-198603000-00013.
6. Le Gall JR, Lemeshow S, Saulnier F. A new Simplified Acute Physiology Score (SAPS II) based on a European/North American multicenter study. JAMA. 1993;270:2957–63. https://doi.org/10.1001/jama.270.24.2957.
7. Dhami A, Onyeukwu KA, Sattar S, Batra A, Mostafa Y, Haris M et al. The Prognostic Performance of Artificial Intelligence and Machine Learning Models for Mortality Prediction in Intensive Care Units: A Systematic Review. Cureus 2025;17. https://doi.org/10.7759/cureus.90465
8. Barboi C, Tzavelis A, Muhammad LN. Comparison of Severity of Illness Scores and Artificial Intelligence Models That Are Predictive of Intensive Care Unit Mortality: Meta-analysis and Review of the Literature. JMIR Med Inf. 2022;10:e35293. https://doi.org/10.2196/35293.
9. de Hond AAH, Kant IMJ, Fornasa M, Cinà G, Elbers PWG, Thoral PJ, et al. Predicting Readmission or Death After Discharge From the ICU: External Validation and Retraining of a Machine Learning Model. Crit Care Med. 2023;51:291–300. https://doi.org/10.1097/CCM.0000000000005758.
10. Singhal K, Azizi S, Tu T, Mahdavi SS, Wei J, Chung HW, et al. Large language models encode clinical knowledge. Nature. 2023;620:172–80. https://doi.org/10.1038/s41586-023-06291-2.
11. Workum JD, Volkers BWS, van de Sande D, Arora S, Goeijenbier M, Gommers D, et al. Comparative evaluation and performance of large language models on expert level critical care questions: a benchmark study. Crit Care. 2025;29:72. https://doi.org/10.1186/s13054-025-05302-0.
12. McDuff D, Schaekermann M, Tu T, Palepu A, Wang A, Garrison J, et al. Towards accurate differential diagnosis with large language models. Nature. 2025;642:451–7. https://doi.org/10.1038/s41586-025-08869-4.
13. Goh E, Gallo R, Hom J, Strong E, Weng Y, Kerman H, et al. Large Language Model Influence on Diagnostic Reasoning: A Randomized Clinical Trial. JAMA Netw Open. 2024;7:e2440969. https://doi.org/10.1001/jamanetworkopen.2024.40969.
14. Jiang LY, Liu XC, Nejatian NP, Nasir-Moin M, Wang D, Abidin A, et al. Health system-scale language models are all-purpose prediction engines. Nature. 2023;619:357–62. https://doi.org/10.1038/s41586-023-06160-y.
15. Nazih W, Abuhmed T, Alharbi M, El-Sappagh S. Mortality prediction for ICU patients with mental disorders using large language models ensemble and unstructured medical notes. PLoS ONE. 2025;20:e0332134. https://doi.org/10.1371/journal.pone.0332134.

16. Weissman GE, Hubbard RA, Ungar LH, Harhay MO, Greene CS, Himes BE, et al. Inclusion of Unstructured Clinical Text Improves Early Prediction of Death or Prolonged ICU Stay. Crit Care Med. 2018;46:1125–32. https://doi.org/10.1097/CCM.0000000000003148.
17. Wu J, Gu B, Zhou R, Xie K, Snyder D, Jiang Y et al. BRIDGE: benchmarking large language models for understanding real-world clinical practice texts. Nat Biomed Eng 2026:1–16. https://doi.org/10.1038/s41551-026-01719-2
18. Holmgren AJ, Downing NL, Bates DW, Shanafelt TD, Milstein A, Sharp CD, et al. Assessment of Electronic Health Record Use Between US and Non-US Health Systems. JAMA Intern Med. 2021;181:251–9. https://doi.org/10.1001/jamainternmed.2020.7071.
19. Contreras M, Kapoor S, Zhang J, Davidson A, Ren Y, Guan Z, et al. A large language model for delirium prediction in the intensive care unit using structured electronic health records. Sci Rep. 2025;15:38890. https://doi.org/10.1038/s41598-025-22634-7.
20. Zhu Y, Gao J, Wang Z, Liao W, Zheng X, Liang L, et al. ClinicRealm: Re-evaluating large language models with conventional machine learning for non-generative clinical prediction tasks. NPJ Digit Med. 2026;9:319. https://doi.org/10.1038/s41746-026-02539-z.
21. Brown KE, Yan C, Li Z, Zhang X, Collins BX, Chen Y, et al. Large language models are less effective at clinical prediction tasks than locally trained machine learning models. J Am Med Inf Assoc. 2025;32:811–22. https://doi.org/10.1093/jamia/ocaf038.
22. Chen C, Yu J, Chen S, Liu C, Wan Z, Bitterman D et al. ClinicalBench: Can LLMs Beat Traditional ML Models in Clinical Prediction? 2024. https://doi.org/10.48550/arXiv.2411.06469
23. Shi T, Ma J, Yu Z, Xu H, Yang R, Xiong M, et al. Large Language Models in Critical Care Medicine: Scoping Review. JMIR Med Inf. 2025;13:e76326. https://doi.org/10.2196/76326.
24. Oh N, Cha WC, Seo JH, Choi S-G, Kim JM, Chung CR, et al. ChatGPT Predicts In-Hospital All-Cause Mortality for Sepsis: In-Context Learning with the Korean Sepsis Alliance Database. Healthc Inf Res. 2024;30:266–76. https://doi.org/10.4258/hir.2024.30.3.266.
25. Shah-Mohammadi F, Millar A, Facelli JC, Gouripeddi R. Understanding uncertainty in large language model predictions of early death in critically ill patients: a conformal prediction approach. JAMIA Open. 2026;9:ooag108. https://doi.org/10.1093/jamiaopen/ooag108.
26. Turpin M, Michael J, Perez E, Bowman SR. Language models don't always say what they think: unfaithful explanations in chain-of-thought prompting. In: Proceedings of the 37th International Conference on Neural Information Processing Systems; 2023. Art. no.: 3275. https://doi.org/10.5555/3666122.3669397
27. Ghassemi M, Oakden-Rayner L, Beam AL. The false hope of current approaches to explainable artificial intelligence in health care. Lancet Digit Health. 2021;3:e745–50. https://doi.org/10.1016/S2589-7500(21)00208-9.
28. Kuo J-R, Chen G-Y, Yap X-HV, Li C-C, Kuo Y-D, Liu C-F. Prompt-Sensitive Decision Behavior of Large Language Models in Intensive Care Unit Mortality Prediction for Spontaneous Intracerebral Hemorrhage: Comparative Benchmarking Study. J Med Internet Res. 2026;28:e29701. https://doi.org/10.2196/29701.
29. Lee S, Cho WI, Lee Y, Kim DJ, Nam KH, Lee S, et al. A prompt framework for enhancing LLM-based explainability of medical machine learning models: an intensive care unit application. BMC Med Inf Decis Mak. 2025;25:430. https://doi.org/10.1186/s12911-025-03239-6.
30. Pollard TJ, Johnson AEW, Raffa JD, Celi LA, Mark RG, Badawi O. The eICU Collaborative Research Database, a freely available multi-center database for critical care research. Sci Data. 2018;5:180178. https://doi.org/10.1038/sdata.2018.178.
31. Johnson AEW, Bulgarelli L, Shen L, Gayles A, Shammout A, Horng S, et al. MIMIC-IV, a freely accessible electronic health record dataset. Sci Data. 2023;10:1. https://doi.org/10.1038/s41597-022-01899-x.
32. Kim Y-G, Shin J, Won SM, Lee S-M, Ryu HG, Lee G, et al. K-MIMIC: a nationwide Korean multi-institutional Multimodal intensive care dataset. Korean J Anesthesiol. 2026. https://doi.org/10.4097/kja.25752.
33. Collins GS, Moons KGM, Dhiman P, Riley RD, Beam AL, Calster BV, et al. TRIPOD + AI statement: updated guidance for reporting clinical prediction models that use regression or machine learning methods. BMJ. 2024;385:e078378. https://doi.org/10.1136/bmj-2023-078378.
34. Introducing GPT-5. 2025. https://openai.com/index/introducing-gpt-5/ (accessed November 23, 2025).
35. Comanici G, Bieber E, Schaekermann M, Pasupat I, Sachdeva N, Dhillon I et al. Gemini 2.5: Pushing the Frontier with Advanced Reasoning, Multimodality, Long Context, and Next Generation Agentic Capabilities 2025. https://doi.org/10.48550/arXiv.2507.06261
36. Radford A, Wu J, Child R, Luan D, Amodei D, Sutskever I. Language models are unsupervised multitask learners. OpenAI; 2019.
37. Wei J, Wang X, Schuurmans D, Bosma M, Ichter B, Xia F et al. Chain-of-thought prompting elicits reasoning in large language models. In: Proceedings of the 36th International Conference on Neural Information Processing Systems; 2022. Art. no.: 1800. https://doi.org/10.5555/3600270.3602070
38. Guo G, Zhang K, Hoo B, Cai Y, Lu X, Peng N et al. Structured Outputs Enable General-Purpose LLMs to be Medical Experts 2025. https://doi.org/10.48550/arXiv.2503.03194
39. Du Y, Li S, Torralba A, Tenenbaum JB, Mordatch I. Improving factuality and reasoning in language models through multiagent debate. In: Proceedings of the 41st International Conference on Machine Learning; 2024. Art. no.: 467. https://doi.org/10.5555/3692070.3692537
40. Sonoda Y, Kurokawa R, Hagiwara A, Asari Y, Fukushima T, Kanzawa J, et al. Structured clinical reasoning prompt enhances LLM's diagnostic capabilities in diagnosis please quiz cases. Jpn J Radiol. 2025;43:586–92. https://doi.org/10.1007/s11604-024-01712-2.
41. Lundberg SM, Lee SI. A unified approach to interpreting model predictions. In: Proceedings of the 31st International Conference on Neural Information Processing Systems; 2017. pp. 4768-77. https://doi.org/10.5555/3295222.3295230
42. Yim W, Fu Y, Ben Abacha A, Snider N, Lin T, Yetisgen M. Aci-bench: a Novel Ambient Clinical Intelligence Dataset for Benchmarking Automatic Visit Note Generation. Sci Data. 2023;10:586. https://doi.org/10.1038/s41597-023-02487-3.
43. Zhao W, Wu C, Zhang X, Zhang Y, Wang Y, Xie W. RaTEScore: A metric for radiology report generation. In: Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing; 2024. pp. 15004-19. https://doi.org/10.18653/v1/2024.emnlp-main.836

44. Gwet KL. Handbook of Inter-Rater Reliability, 4th Edition: The Definitive Guide to Measuring The Extent of Agreement Among Raters. Advanced Analytics, LLC; 2014.
45. Cohen J. Weighted kappa: nominal scale agreement with provision for scaled disagreement or partial credit. Psychol Bull. 1968;70:213–20. https://doi.org/10.1037/h0026256.
46. Savkov A, Moramarco F, Papadopoulos Korfiatis A, Perera M, Belz A, Reiter E. Consultation checklists: standardising the human evaluation of medical note generation. In: Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing: Industry Track; 2022. pp. 111 – 20. https://doi.org/10.18653/v1/2022.emnlp-industry.10
47. Feinstein AR, Cicchetti DV. High agreement but low kappa: I. The problems of two paradoxes. J Clin Epidemiol. 1990;43:543–9. https://doi.org/10.1016/0895-4356(90)90158-l.
48. Tan ALM, Getzen EJ, Hutch MR, Strasser ZH, Gutiérrez-Sacristán A, Le TT, et al. Informative missingness: What can we learn from patterns in missing laboratory data in the electronic health record? J Biomed Inf. 2023;139:104306. https://doi.org/10.1016/j.jbi.2023.104306.
49. Busch F, Hoffmann L, Rueger C, van Dijk EH, Kader R, Ortiz-Prado E, et al. Current applications and challenges in large language models for patient care: a systematic review. Commun Med (Lond). 2025;5:26. https://doi.org/10.1038/s43856-024-00717-2.
50. Lundberg SM, Erion G, Chen H, DeGrave A, Prutkin JM, Nair B, et al. From Local Explanations to Global Understanding with Explainable AI for Trees. Nat Mach Intell. 2020;2:56–67. https://doi.org/10.1038/s42256-019-0138-9.
51. Allgaier J, Mulansky L, Draelos RL, Pryss R. How does the model make predictions? A systematic literature review on the explainability power of machine learning in healthcare. Artif Intell Med. 2023;143:102616. https://doi.org/10.1016/j.artmed.2023.102616.
52. Terwee CB, Bot SDM, de Boer MR, van der Windt DAWM, Knol DL, Dekker J, et al. Quality criteria were proposed for measurement properties of health status questionnaires. J Clin Epidemiol. 2007;60:34–42. https://doi.org/10.1016/j.jclinepi.2006.03.012.

Figures

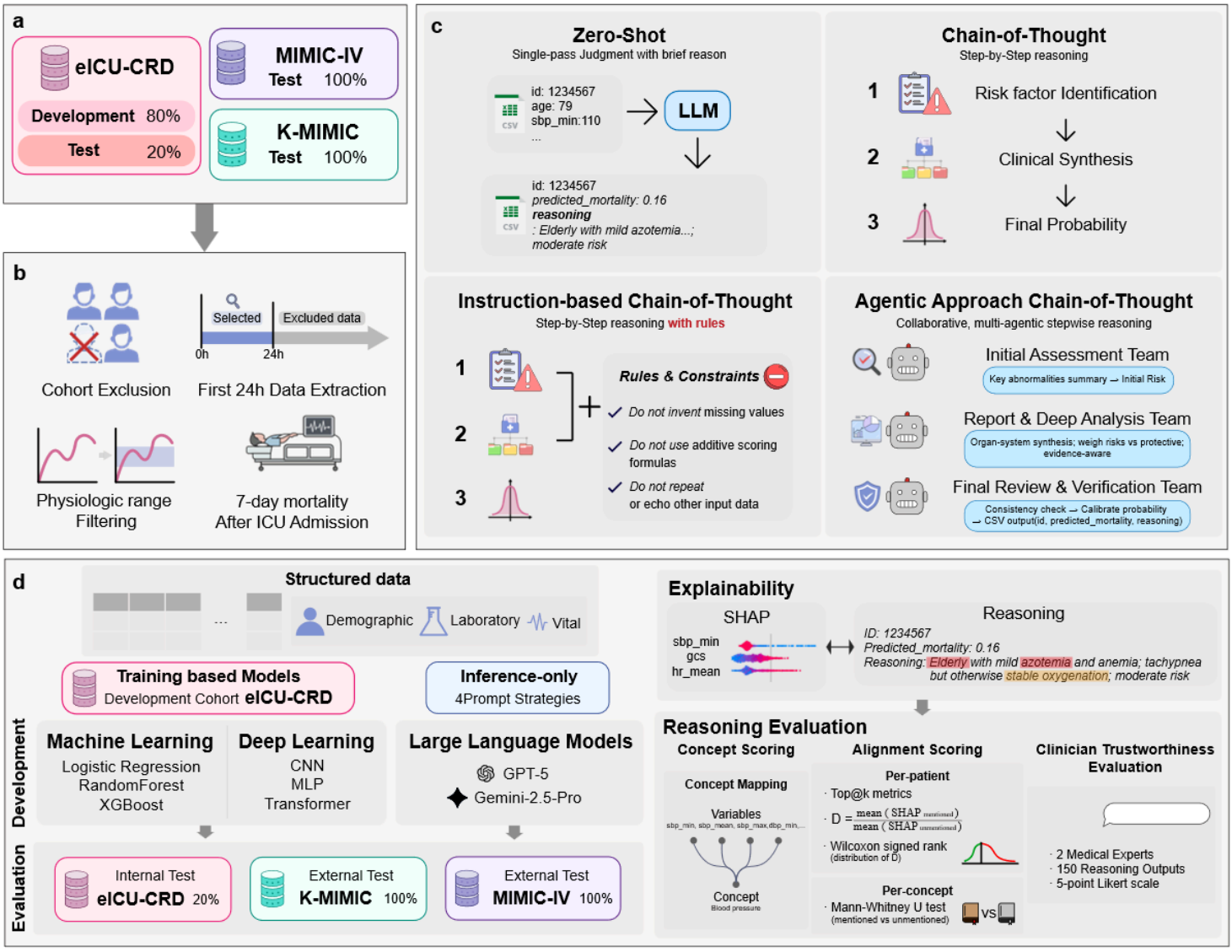


Figure 1

Overview of the study design and analytical workflow.

a, Cohort preparation and data partitioning. **b**, Data preprocessing. **c**, Four prompting strategies. **d**, Model development and evaluation. **e**, SHAP–LLM reasoning alignment and clinician evaluation.
CNN, convolutional neural network; CoT, chain-of-thought; eICU-CRD, eICU Collaborative Research Database; ICU, intensive care unit; K-MIMIC, Korea Medical Information Mart for Intensive Care; LLM, large language model; MIMIC-IV, Medical Information Mart for Intensive Care IV; MLP, multilayer perceptron; SHAP, SHapley Additive exPlanations.

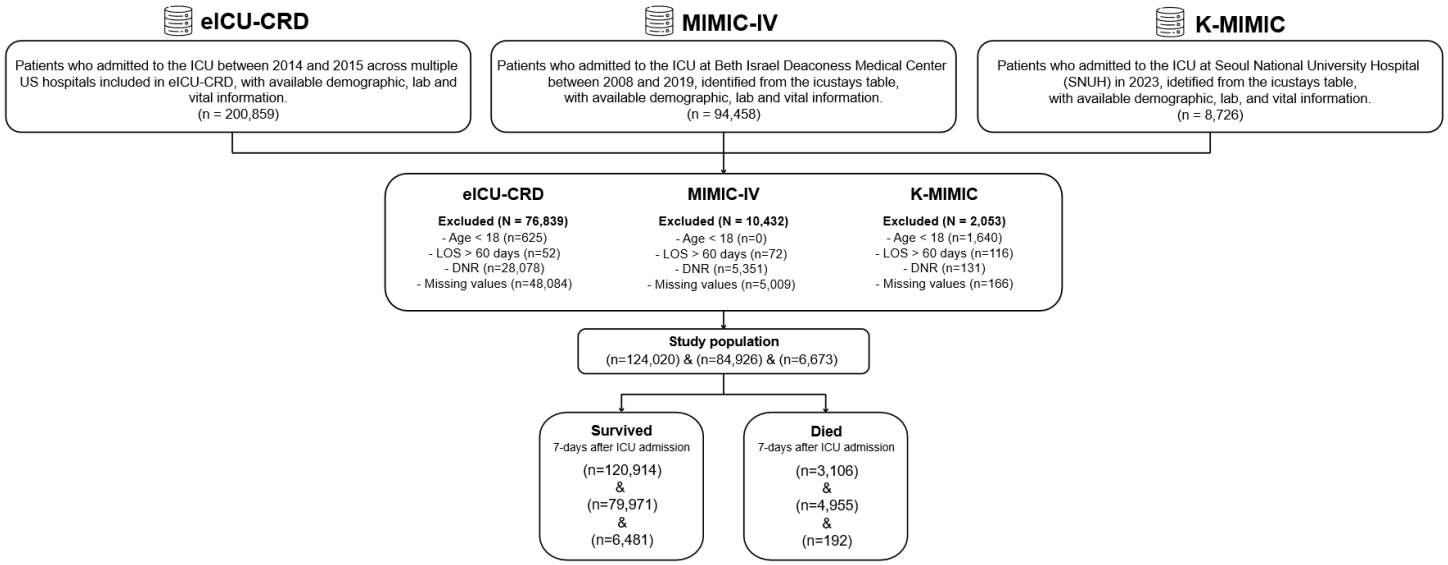


Figure 2

Cohort selection across the three ICU databases.

Flow diagram of cohort selection in eICU-CRD, MIMIC-IV, and K-MIMIC.
DNR, do not resuscitate; eICU-CRD, eICU Collaborative Research Database; ICU, intensive care unit; K-MIMIC, Korea Medical Information Mart for Intensive Care; LOS, length of stay; MIMIC-IV, Medical Information Mart for Intensive Care IV.

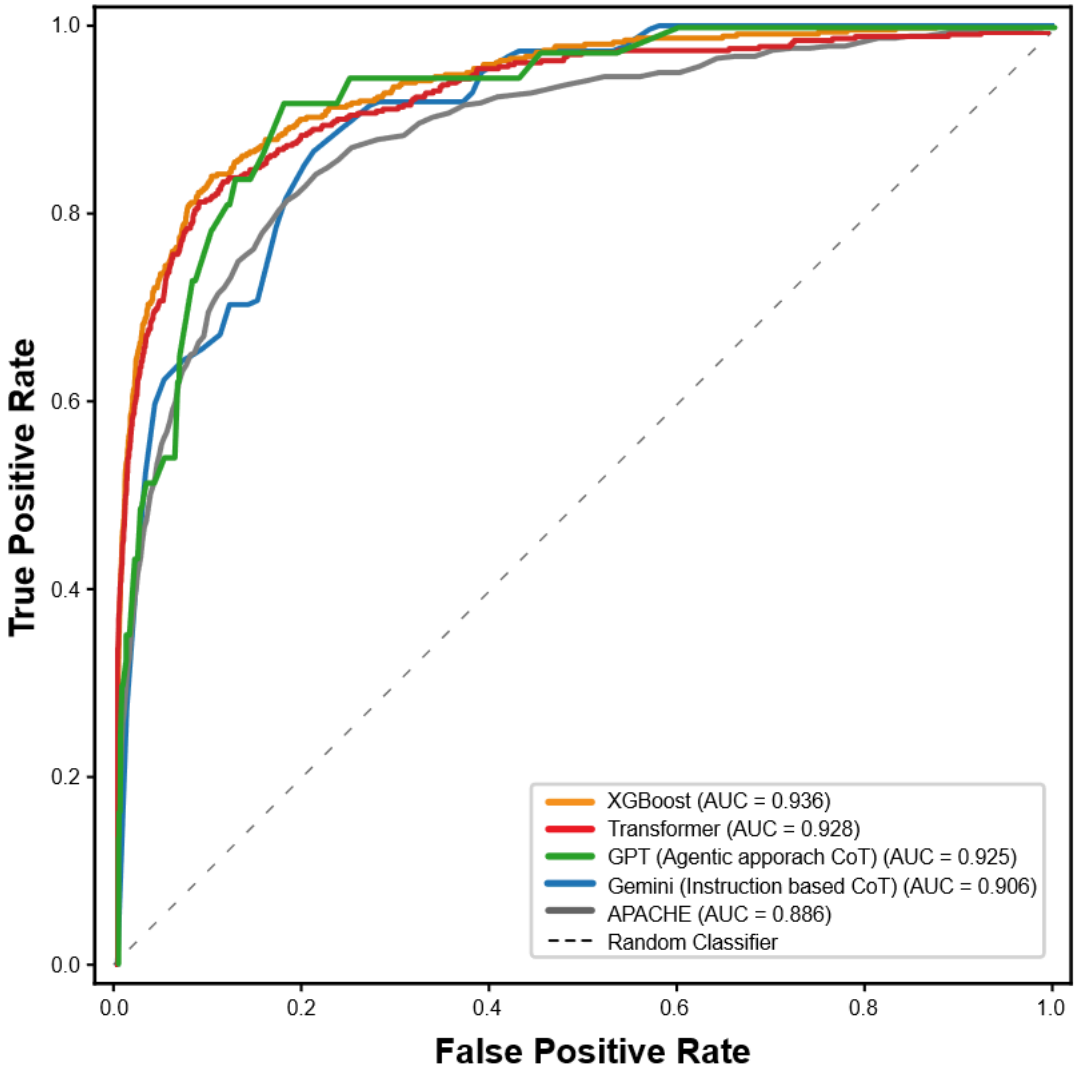


Figure 3

Cohort selection across the three ICU databases.

Flow diagram of cohort selection in eICU-CRD, MIMIC-IV, and K-MIMIC.
DNR, do not resuscitate; eICU-CRD, eICU Collaborative Research Database; ICU, intensive care unit; K-MIMIC, Korea Medical Information Mart for Intensive Care; LOS, length of stay; MIMIC-IV, Medical Information Mart for Intensive Care IV.

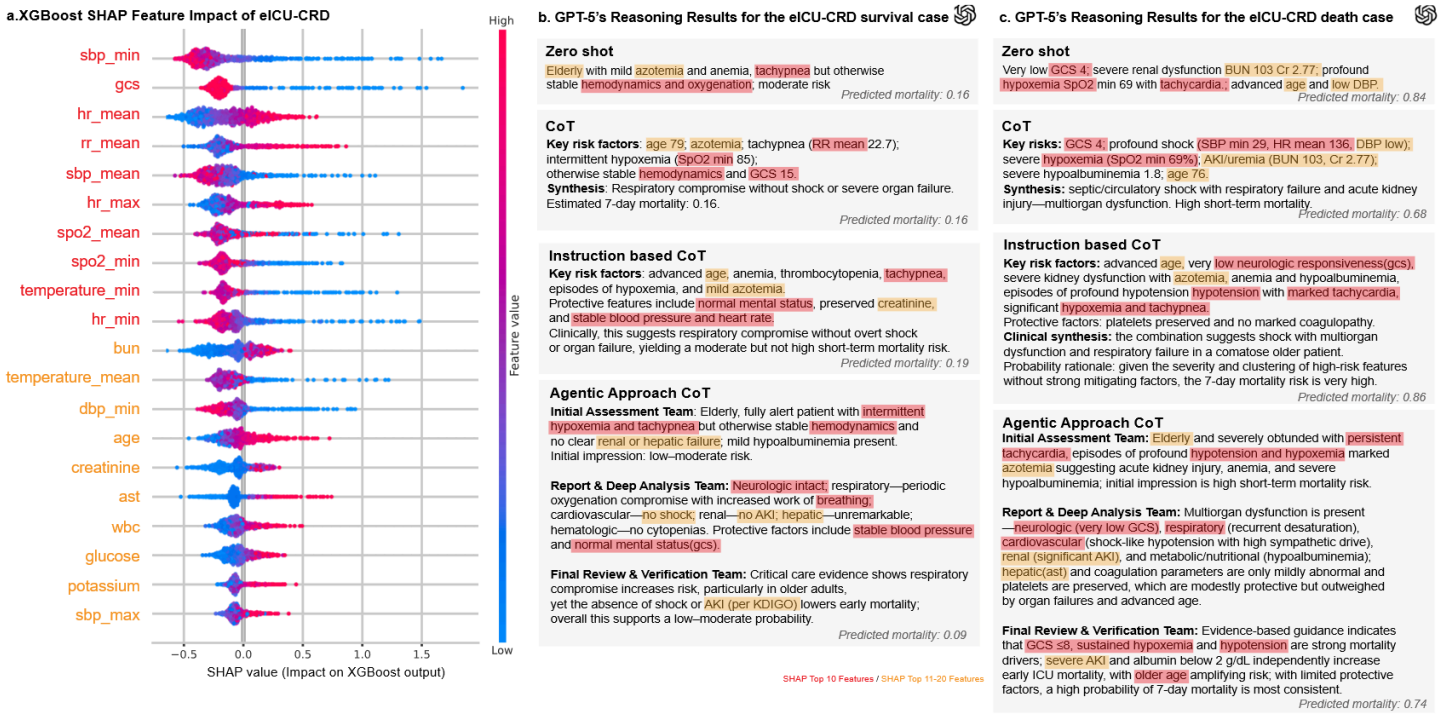


Figure 4

XGBoost feature importance and representative GPT-5 reasoning in the eICU-CRD cohort.

a, SHAP summary plot of the 20 highest-ranked XGBoost features. **b–c**, Representative GPT-5 reasoning outputs for a survivor (**b**) and a non-survivor (**c**) across the four prompting strategies. Terms corresponding to features ranked within the top 10 in panel **a** are highlighted in red, and those ranked 11–20 are highlighted in orange.
CoT, chain-of-thought; eICU-CRD, eICU Collaborative Research Database; GCS, Glasgow Coma Scale; SHAP, SHapley Additive exPlanations.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- SupplementaryMaterialsRiskStratificationfor7DayMortalityAfterICUAdmissionUsingGeneralPurposeLargeLanguageModelsandStructuredElectronicHealth