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.

AI-derived Information Phenotypes for Early Risk Stratification in the Emergency Department: A Retrospective Study

Tae Hyek Yim
Chung-Ang University
Boram Yoo
Chung-Ang University
MinKyu Park
Chung-Ang University
Young Gyu Kwon
Chung-Ang University
Chan Woong Kim
whenever@cau.ac.kr

Chung-Ang University

Article

Keywords: Medical Information Mart for Intensive Care (MIMIC)-IV-ED, Emergency department, Machine learning, Sequential analysis, Value of information, Information phenotypes

Posted Date: November 20th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11531219/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

Machine learning models for emergency department (ED) risk stratification often assume that additional clinical information, such as laboratory results, benefits all patients equally. We evaluated how the predictive value of clinical information changes over time during early ED risk stratification and whether this differs across clinical presentations. Using 263,240 adult encounters from the Medical Information Mart for Intensive Care (MIMIC)-IV-ED database, we developed models to predict early critical illness. Free-text chief complaints were clustered using natural language processing and mapped to 13 Canadian Emergency Department Information System subgroups. Incremental predictive utility, measured as the change in the area under the receiver operating characteristic curve (AUROC) and the area under the precision-recall curve (AUPRC), was quantified across three sequential observation windows: 30 min, 1 h, and 2 h after arrival. Extreme gradient boosting performed best (AUROC, 0.9137 at 2 h). Four information phenotypes were identified: information-dependent subgroups gained substantially in both discrimination and precision, precision-dependent subgroups gained mainly in precision, triage-sufficient subgroups improved minimally after triage, and one subgroup showed a nonmonotonic trajectory. The value of extended observation varies across ED presentations, supporting a selective approach in which diagnostic evaluation is aligned with the expected informational benefit of each presentation.

Introduction

Emergency departments (EDs) must make disposition decisions rapidly, often before a patient's clinical trajectory is fully understood [1]. Early recognition of patients at risk of deterioration within 24 h of arrival is important for improving patient outcomes and reducing ED crowding [2–4]. Recent artificial intelligence (AI) studies using large-scale public datasets, such as the Medical Information Mart for Intensive Care (MIMIC)-IV and MIMIC-IV-ED, have generally followed two approaches [5, 6]. Triage models rely on information available at arrival and can be deployed immediately; however, their predictive performance may be limited compared with models incorporating additional clinical information [7, 8]. In contrast, multimodal models incorporate serial vital signs and laboratory results, generally achieving higher predictive performance at the cost of delayed prediction [9, 10].

A key unresolved issue is whether all ED presentations benefit equally from delayed diagnostic information. Most existing studies have evaluated average performance across the entire cohort, potentially overlooking differences in the value of additional data across patient subgroups. In practice, however, the usefulness of laboratory results and other time-dependent information may vary according to the presenting complaint. If so, applying a uniform observation period and data integration strategy to all patients may result in unnecessary data collection and prolonged observation for some groups without meaningful gains in predictive performance.

This study evaluated how the value of clinical information changes over time during early ED risk stratification. We hypothesized that certain patient subgroups would exhibit distinct patterns of dependence on additional clinical data and that these patterns could be used to define information phenotypes for early risk stratification. To examine this possibility, we developed a sequential framework that reflected the timeline of routine ED care. This framework allowed us to evaluate predictive performance across incremental observation windows and quantify the incremental predictive utility of newly available clinical data.

METHODS

Study design, data sources, and population

This retrospective cohort study used secondary, de-identified data from publicly available databases. The study was reviewed by the Institutional Review Board of Chung-Ang University (approval no. 1041078-20260623-HR-279), which determined it to be exempt from full review and waived the requirement for informed consent. All methods were carried out in accordance with relevant guidelines and regulations, including the Declaration of Helsinki. Reporting followed the Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis and Artificial Intelligence (TRIPOD + AI) recommendations (Supplementary Table S1) [11].

Data from the MIMIC-IV and MIMIC-IV-ED databases were used for the analysis [5, 6]. Time-stamped predictor variables were obtained from MIMIC-IV-ED, whereas outcome information was derived from MIMIC-IV database (Fig. 1). Predictor and outcome variables were obtained from temporally aligned data sources, with outcome information excluded from the predictor construction process to minimize the risk of information leakage. Adult patients (age ≥ 18 years) presenting to the ED were eligible. The inclusion criterion was the successful assignment of a free-text chief complaint to a Canadian Emergency Department Information System (CEDIS) category through the natural language processing (NLP)-based mapping process. Patients were excluded if arrival times were missing, temporal records were inconsistent, or essential triage variables were unavailable. Multiple visits from the same patient were retained and analyzed as independent encounters.

Outcome measure definition

The primary outcome was early critical illness, defined as severe clinical deterioration within 24 h of ED arrival. Early critical illness was defined as the occurrence of any of the following: intensive care unit (ICU) admission, initiation of invasive mechanical ventilation, or vasopressor administration within 24 hours of ED arrival, or in-hospital mortality. ICU admission and in-hospital mortality were included as indicators of critical illness, whereas invasive mechanical ventilation and vasopressor administration represented markers of respiratory and circulatory failure, respectively.

Predictor variables and observation-window strategy

Predictor variables were extracted across three sequential observation windows anchored to ED arrival: minimal ($T_{0.5}$ ≤ 30 min), intermediate ($T_{1.0}$ ≤ 1 h), and augmented ($T_{2.0}$ ≤ 2 h). The minimal observation window ($T_{0.5}$ ≤ 30 min) included demographic characteristics, mode of arrival, triage acuity, pain score, chief complaint, and summary measures of initial vital signs. In the intermediate observation window ($T_{1.0}$ ≤ 1 h), vital signs and laboratory results recorded within 1 hour were included. In augmented observation window ($T_{2.0}$ ≤ 2 h), available comprehensive laboratory results were added to the cumulative feature set. Missing continuous laboratory values were imputed using the median, with binary missingness indicators included [12–14]. Past medical history and comorbidity measures, including the Charlson Comorbidity Index, were excluded. This design was intended to reflect the information typically available during the early stages of ED evaluation and to assess the prognostic contribution of presenting physiology and chief complaints.

Chief complaint preprocessing and categorization

Free-text chief complaints were processed using an NLP pipeline. ED-specific abbreviations were expanded, and Health Insurance Portability and Accountability Act (HIPAA) identifiers and noninformative stopwords were removed while clinically relevant negations, such as “no” and “without,” were preserved. BERTopic [15] with the S-PubMedBert-MS-MARCO embedding model [16] was used for unsupervised topic modeling, uniform manifold approximation and projection (UMAP) was used for dimensionality reduction [17], and hierarchical density-based spatial clustering of applications with noise (HDBSCAN) was used for density-based clustering [18]. The resulting clusters were

systematically mapped to the CEDIS ontology [19]. Encounters containing predefined high-acuity keywords were reassigned according to prespecified rule-based criteria and subsequently categorized into 13 CEDIS subgroups.

Hyperparameter optimization and composite score evaluation

HDBSCAN hyperparameters [18] were optimized using a grid search across 169 parameter combinations. The search used the leaf cluster selection method [20], *min_cluster_size* ranging from 50 to 700 and *min_samples* ranging 3 to 100. Each candidate configuration was evaluated using a custom composite score. The score incorporated penalties for excessive noise (> 30%), unclassified encounters (> 10%), extreme topic fragmentation (< 30 or > 150 topics), and imbalanced cluster size distributions. The explicit penalty function, precise variable weights, the mathematical definition of cluster imbalance, and the top 15 configurations are presented in Supplementary Table S2.

Coverage and topic cohesion for 12 predefined sentinel ED presentations were also included in the evaluation: chest pain, back pain, dizziness, gastrointestinal bleeding, stroke, stroke-like symptoms, suicidal ideation, shortness of breath, seizure, syncope, fever, and abdominal pain. These sentinel presentations were prespecified before hyperparameter selection because they represent common, clinically important, or time-sensitive ED chief complaint domains that should remain identifiable within a clinically meaningful clustering framework.

Among the 169 configurations tested, *min_cluster_size* = 600 and *min_samples* = 10 achieved the highest composite score (65.75). This configuration generated 169 clinical topics and captured all 12 sentinel presentations. The selected parameters were subsequently used to assign chief complaints to clusters in all downstream analyses. The predefined sentinel ED presentations and their keyword targets are listed in Supplementary Table S3.

Model development and evaluation

The dataset was partitioned using de-identified anchor-year values into a training set (≤ 2180, N = 220,570) and a held-out test set (> 2180, N = 42,670). These anchor-year values did not correspond to actual calendar years. We trained and evaluated four algorithms: logistic regression (LR) [21], random forest (RF) [22], extreme gradient boosting (XGBoost) [23], and residual multilayer perceptron (ResMLP) implemented in PyTorch [24]. To address class imbalance, balanced class weights were applied to the LR and RF models, and class weighting was incorporated into the binary cross-entropy loss of the ResMLP; XGBoost was trained without class reweighting. The probabilities of LR and RF were calibrated using five-fold cross-validated sigmoid and isotonic calibration, respectively.

LR (L2 regularization, C = 1.0) and RF (300 trees, maximum depth 15) were used as baseline models. XGBoost was selected as the primary model for subgroup comparisons because of its robustness for nonlinear tabular data and was configured with 300 trees, a learning rate of 0.05, and a maximum depth of 6. The PyTorch ResMLP comprised two residual blocks (128 units) with batch normalization, Gaussian error linear unit (GELU) activation, and 30% dropout, and was trained using the AdamW optimizer [25] with early stopping based on the area under the precision-recall curve (AUPRC) in a 10% validation subset of the training set.

Subgroup definitions and value of information

Sequential model performance was evaluated across the 13 CEDIS subgroups using the 3 observation windows. The incremental contribution of laboratory data was quantified as the absolute change in AUPRC (ΔAUPRC) and the absolute change in the area under the receiver operating characteristic curve (AUROC; ΔAUROC) between $T_{0.5}$ and $T_{2.0}$. Subgroups were categorized into four information phenotypes based on changes in predictive performance across the sequential observation windows. Given that there was no established threshold for defining a meaningful increase in

AUPRC in early risk prediction in the ED, we prespecified a ΔAUPRC of 0.07 as a practical cutoff for identifying subgroups with a clear gain in precision-recall performance.

Similarly, a ΔAUROC of 0.04 was used to indicate a noticeable improvement in discrimination. These cut-offs were used to summarize subgroup-level performance trajectories and were not intended to represent clinical decision thresholds. To systematically assign these phenotypes, we applied an explicit classification algorithm based on ΔAUROC and ΔAUPRC. Subgroups were first assessed for nonmonotonic trajectories and classified as anomalous trajectory when this criterion was met. The remaining subgroups were classified according to ΔAUPRC and ΔAUROC as information dependent, precision dependent, or triage sufficient. (1) information dependent (ΔAUPRC ≥ 0.07, ΔAUROC ≥ 0.04): substantial gains in both precision and discrimination; (2) precision dependent (ΔAUPRC ≥ 0.07, ΔAUROC < 0.04): substantial gains in precision with limited changes in discrimination; (3) triage sufficient (ΔAUPRC < 0.07, ΔAUROC < 0.04): minimal improvement after laboratory data integration; and (4) anomalous trajectory: nonmonotonic performance patterns across sequential observation windows, irrespective of the AUPRC magnitude.

Statistical analysis and explainability

Model performance was evaluated using the AUROC, AUPRC, and Brier score. Specificity was calculated in the test set at the decision threshold corresponding to a sensitivity of 0.90. Model performance metrics were reported as point estimates in the test set. For pairwise comparisons for AUROC between the machine learning models, statistical significance ($P < 0.05$) was assessed using DeLong's test. The Shapley Additive exPlanations (SHAP) [26, 27] were used to estimate global feature importance and to facilitate the interpretation of XGBoost predictions.

Fairness audit

Model fairness was assessed by comparing the AUROC and AUPRC of the $T_{2.0}$ XGBoost model across predefined demographic strata: race (White, Black, Hispanic, Asian, and others), sex (male and female), and age (18–39, 40–64, and ≥ 65 years).

Use of generative artificial intelligence

Generative artificial intelligence (GAI) tools were used to optimize data processing scripts, debug deep learning architectures, and assist with academic proofreading and structural reformatting to improve readability. All GAI-generated suggestions were thoroughly vetted and curated by the authors, who maintained full accountability for the scientific integrity, data accuracy, and final outcomes of this study.

RESULTS

Study cohort and baseline characteristics

Among the 263,240 encounters, 14,415 (5.5%) met the definition of early critical illness. Baseline characteristics differed significantly between the critical and noncritical groups (all $P < 0.001$). The critical group was older than the noncritical group (median age: 64.0 years vs 50.0 years) and had greater physiological abnormalities at triage, including higher maximum heart rates (96.0 vs 83.0 beats per minute [bpm]), higher maximum respiratory rates (20.0 breaths/min vs 18.0 breaths/min), lower minimum oxygen saturation values (97.0% vs. 99.0%), and lower minimum systolic blood pressure (116.0 vs. 130.0 mmHg). Laboratory values also differed between the groups. The critical group had higher initial white blood cell counts (10.3 × 10^3/μL vs 8.0 × 10^3/μL) and higher lactate levels (2.3 vs 1.6 mmol/L). Meanwhile, the median pain scores were lower in the critical group (1.0 vs 5.0). Arrival by ambulance was more common in the

critical group (64.1% vs. 35.8%), as were high-acuity triage assignments (Emergency Severity Index levels 1–2: 85.6% vs. 38.8%) (Table 1) [28].

Table 1
Comparison of baseline patient characteristics between the noncritical and critical groups

Characteristics	Non-critical group (N = 248,825)[a]	Critical group (N = 14,415)[a]	*P* value[b]
Demographics			
Age	50.0 [32.0–65.0]	64.0 [51.0–76.5]	< 0.001
Sex: male	107,243 (43.1%)	7,624 (52.9%)	< 0.001
Triage and acuity			
ESI triage acuity			< 0.001
Level 1 (most urgent)	8,331 (3.3%)	3,862 (26.8%)	
Level 2	88,170 (35.4%)	8,482 (58.8%)	
Triage pain score	5.0 [0.0–8.0]	1.0 [0.0–8.0]	< 0.001
Arrival by ambulance	89,191 (35.8%)	9,237 (64.1%)	< 0.001
Initial vital signs			
Min systolic BP (mmHg)	130.0 [117.0–145.0]	116.0 [99.0–134.0]	< 0.001
Max heart rate (bpm)	83.0 [72.0–96.0]	96.0 [80.0–113.0]	< 0.001
Max respiratory rate	18.0 [16.0–18.0]	20.0 [18.0–25.0]	< 0.001
Min O_2 saturation (%)	99.0 [97.0–100.0]	97.0 [94.0–99.0]	< 0.001
Initial laboratory results			
Lactate (mmol/L)	1.6 [1.2–2.3]	2.3 [1.6–3.7]	< 0.001
WBC count	8.0 [6.2–10.4]	10.3 [7.4–14.5]	< 0.001

[a]Continuous variables are expressed as the median [IQR]; categorical variables, N (%).

[b]*P*-values from the Mann-Whitney U test (continuous) or chi-squared test (categorical).

IQR, interquartile range; ESI, Emergency Severity Index; BP, blood pressure; bpm, beats per minute; O_2, oxygen; WBC, white blood cell

Overall predictive performance across sequential time windows

XGBoost outperformed the PyTorch ResMLP (AUROC, 0.9051; AUPRC, 0.5321), RF (AUROC, 0.9015; AUPRC, 0.4953), and LR (AUROC, 0.8926; AUPRC, 0.4413) models (all $P < 0.001$; Table 2 and Fig. 2). XGBoost also achieved the lowest Brier score (0.0380), whereas PyTorch ResMLP had the highest score (0.1034). At the predefined sensitivity target of 0.90, the XGBoost model achieved a specificity of 0.7320 at $T_{2.0}$.

Table 2
Predictive performance of candidate AI models for critical illness identification (XGBoost primary cohort)

Model	AUROC	AUPRC	Brier score	Spec-Sens = 90%[a]	*P* value[b]
Logistic Regression	0.8926	0.4413	0.0433	0.6946	< 0.001
Random Forest	0.9015	0.4953	0.0412	0.7017	< 0.001
PyTorch (ResMLP)	0.9051	0.5321	0.1034	0.7200	< 0.001
XGBoost (primary)	**0.9137**	**0.5662**	**0.0380**	**0.7320**	**Reference**

[a]Spec-Sens: 90% = specificity at 90% sensitivity.

[b]*P*-values are calculated using DeLong's test for the AUROC comparison against the XGBoost model.

AI, artificial intelligence; AUROC, area under the receiver operating characteristic curve; AUPRC, area under the precision-recall curve; ResMLP, Residual Multi-Layer Perceptron; XGBoost, eXtreme Gradient Boosting

Model performance improved across the sequential observation windows ($T_{0.5} \rightarrow T_{1.0} \rightarrow T_{2.0}$). In the test set, the XGBoost model achieved an AUROC of 0.8854 at $T_{0.5}$ and 0.9007 at $T_{1.0}$ (Table 3). Performance was highest at $T_{2.0}$, with an AUROC of 0.9137 and an AUPRC of 0.5662. At the predefined sensitivity target of 0.90, this corresponded to the correct identification of 90% of patients who developed early critical illness and the correct classification of 73.2% of noncritical encounters as negative.

Table 3
Predictive performance across sequential time windows (XGBoost model)

Time window	AUROC	AUPRC	Brier score	Spec-Sens = 90%[a]
$T_{0.5}$[b] (Minimal triage)	0.8854	0.4758	0.0423	0.6609
$T_{1.0}$[c] (Intermediate)	0.9007	0.5222	0.0401	0.6943
$T_{2.0}$[d] (Augmented)	0.9137	0.5662	0.0380	0.7320

[a]Spec-Sens = 90%: specificity at 90% sensitivity.

[b]$T_{0.5}$: 30-min triage window (minimal data).

[c]$T_{1.0}$: 1-h intermediate window.

[d]$T_{2.0}$: 2-h augmented window including laboratory results.

AUROC, area under the receiver operating characteristic curve; AUPRC, area under the precision-recall curve

Heterogeneity in the temporal value of information

The incremental gain in predictive performance exhibited marked heterogeneity across the clinical subgroups (Fig. 3 and Table 4). Tracking sequential evolution ($T_{0.5} \rightarrow T_{1.0} \rightarrow T_{2.0}$) identified 4 distinct information phenotypes based on their reliance on 2-h laboratory data (Table 4). (1) Information dependent: Patients with substance/toxicology, endocrine/metabolic, and trauma/injury complaints showed substantial gains in predictive performance with the full sequential pipeline. Integrating 2-h labs yielded substantial concurrent gains in both discrimination (ΔAUROC ≥ 0.04) and precision (ΔAUPRC ≥ 0.07). Notably, the largest increase in AUPRC was observed in the substance/toxicology

subgroup. Although the overall ΔAUPRC from $T_{0.5}$ to $T_{2.0}$ was + 0.1502, AUPRC declined slightly between $T_{1.0}$ and $T_{2.0}$ (0.2793 to 0.2648), suggesting divergent AUROC and AUPRC trajectories in this subgroup. (2) Precision dependent: The models demonstrated high initial discrimination at triage for the genitourinary, gastrointestinal/abdominal, infectious, musculoskeletal, neurological, cardiac, and respiratory subgroups; for example, the genitourinary subgroup had an AUROC of 0.9051 at *T0.5*. This phenotype showed substantial precision gains (genitourinary ΔAUPRC = + 0.1466) despite marginal AUROC improvements (< 0.04). (3) Triage sufficient: The dermatological and psychiatric/mental health subgroups received minimal benefit from extended observation, plateauing early with negligible precision improvements (ΔAUROC < 0.04 and ΔAUPRC < 0.07) from $T_{2.0}$ data. (4) Anomalous trajectory: The hematology/oncology subgroup exhibited an irregular predictive pattern, characterized by a nonmonotonic trajectory where discrimination paradoxically decreased upon integration of the final laboratory results. The AUROC decreased from 0.8759 at $T_{1.0}$ to 0.8404 at $T_{2.0}$, although the overall ΔAUROC from $T_{0.5}$ to $T_{2.0}$ was + 0.0076.

Table 4
Sequential predictive performance by time window across the 13 CEDIS clinical subgroups

CEDIS clinical subgroup	N[a]	AUROC			AUPRC			ΔAUROC[e, f]	ΔAUPRC[e]
		$T_{0·5}$[b]	$T_{1·0}$[c]	$T_{2·0}$[d]	$T_{0·5}$[b]	$T_{1·0}$[c]	$T_{2·0}$[d]	$(T_{0·5} \rightarrow T_{2·0})$	$(T_{0·5} \rightarrow T_{2·0})$
Information dependent (ΔAUROC ≥ 0.04 and ΔAUPRC ≥ 0.07)									
Substance / toxicology	410	0.8125	0.8535	0.8560	0.1146	0.2793	0.2648	+ 0.0435	+ 0.1502
Endocrine / metabolic	487	0.8382	0.8778	0.9122	0.6618	0.7087	0.7744	+ 0.0740	+ 0.1126
Trauma / injury	4,247	0.8754	0.8985	0.9176	0.4847	0.5290	0.5780	+ 0.0422	+ 0.0933
Precision dependent (ΔAUROC < 0.04 and ΔAUPRC ≥ 0.07)									
Genitourinary	1,800	0.9051	0.9203	0.9257	0.2576	0.3816	0.4042	+ 0.0206	+ 0.1466
Gastrointestinal / abdominal	9,838	0.8736	0.8911	0.9074	0.3525	0.4139	0.4783	+ 0.0338	+ 0.1258
Infectious	1,566	0.8576	0.8774	0.8935	0.5089	0.5660	0.6263	+ 0.0359	+ 0.1174
Musculoskeletal	2,964	0.8630	0.8854	0.9022	0.1934	0.2450	0.2985	+ 0.0392	+ 0.1051
Neurological	7,443	0.8763	0.8890	0.9033	0.5292	0.5737	0.6119	+ 0.0270	+ 0.0827
Cardiac	6,740	0.8726	0.8878	0.8990	0.4283	0.4581	0.5049	+ 0.0264	+ 0.0766
Respiratory	4,283	0.8451	0.8687	0.8843	0.5555	0.5959	0.6292	+ 0.0392	+ 0.0737
Triage sufficient (ΔAUROC < 0.04 and ΔAUPRC < 0.07)									
Dermatological	779	0.8915	0.9003	0.9200	0.3148	0.3033	0.3468	+ 0.0285	+ 0.0320
Psychiatric / Mental Health	1,802	0.8221	0.8376	0.8620	0.1448	0.1440	0.1607	+ 0.0399	+ 0.0159
Anomalous trajectory (nonmonotonic performance)									
Hematology / oncology	311	0.8328	0.8759	0.8404	0.4649	0.6050	0.5579	+ 0.0076	+ 0.0930

[a]N reflects the test-set encounter count per subgroup.

[b]$T_{0·5}$: 30-min minimal triage window.

[c]$T_{1·0}$: 1-h intermediate window.

[d]$T_{2·0}$: 2-h augmented window with comprehensive laboratory results.

[e]ΔAUROC and ΔAUPRC represent the absolute change from $T_{0·5}$ to $T_{2·0}$.

[f]When AUROC exhibits a nonmonotonic trajectory, the subgroup is classified as anomalous regardless of the ΔAUPRC magnitude.

AUROC, area under the receiver operating characteristic curve; AUPRC, area under the precision-recall curve; CEDIS, Canadian Emergency Department Information System.

Explainable AI and global feature importance

The global SHAP analysis identified triage acuity, age, respiratory rate, and heart rate as the most influential predictors of early critical illness in the augmented XGBoost model (Fig. 4). White blood cell count, minimum systolic blood pressure, and glucose were also ranked among the important features that contributed to the model predictions. Subgroup-level analyses revealed variations in the magnitudes of performance improvements associated with laboratory data across the clinical phenotypes (Table 4).

Fairness audit results

The performance of the $T_{2.0}$ XGBoost model was broadly comparable across the evaluated demographic strata. Across racial groups, AUROC values ranged from 0.89 to 0.92 and AUPRC values ranged from 0.42 to 0.53. Performance was also comparable between male (AUROC, 0.90; AUPRC, 0.54) and female (AUROC, 0.91; AUPRC, 0.52) patients and across age groups (patients aged 18–39 years: AUROC, 0.94 and AUPRC, 0.54; patients aged 40–64 years: AUROC, 0.91 and AUPRC, 0.51; and patients aged ≥ 65 years: AUROC, 0.87 and AUPRC, 0.54). The detailed subgroup results are provided in Supplementary Table S4.

DISCUSSION

This study examined the contribution of sequential clinical information to the prediction of early critical illness in the ED by measuring the performance of predictive AI models. The XGBoost model achieved strong overall performance (AUROC = 0.9137). However, the magnitude of performance improvement associated with additional data varied across the clinical subgroups. Particularly, there were substantial gains after the incorporation of serial vital signs and laboratory results in some subgroups, including the substance/toxicology, endocrine/metabolic, and trauma/injury subgroups. These findings indicate that the value of extended observation and laboratory testing differs according to the presenting clinical context in the ED.

Our study addresses this heterogeneity through three methodological contributions. First, rather than reporting performance at a single time point, we quantified the incremental value of each observation interval ($T_{0.5} \rightarrow T_{1.0} \rightarrow T_{2.0}$) using a nested feature set design. Second, laboratory measurement availability was represented through interpretable binary “measured flag” features that indicated whether laboratory results were available instead of treating missingness solely as a bias to be corrected [12–14]. Third, heterogeneity was examined across 13 chief complaint subgroups derived from NLP and mapped to CEDIS categories, revealing differences in the contribution of delayed clinical information that would not be apparent from aggregate cohort-level metrics.

Previous studies have developed machine learning models for early ED risk stratification. Raita et al. [7] and Yun et al. [8] reported strong performance using information available at triage, with AUROC values approaching 0.86. However, they evaluated the predictions at a single time point and did not examine how model performance changed as additional clinical information became available. Xie et al. [29] evaluated model performance across multiple observation windows using MIMIC-IV-ED, but their results were primarily reported at the overall cohort level. Klug et al. [30] focused on short-term mortality prediction but did not assess the incremental contribution of extended observation. In our study, the superior performance of XGBoost relative to ResMLP is consistent with previous reports showing the strong performance of tree-based methods for structured clinical data [31, 32].

Most ED prediction studies focus on maximizing the overall model performance by incorporating additional clinical information. In our study, however, the predictive performance of the triage-sufficient phenotype changed minimally after the laboratory data became available, whereas larger gains were observed for the information- and precision-dependent phenotypes. These findings suggest that a single observational strategy may not be equally informative for

all patient presentations. Importantly, classification into the triage-sufficient phenotype should not be interpreted as a recommendation to omit clinically indicated testing. The findings instead highlight differences in the incremental predictive contributions of delayed diagnostic information. This information may help guide future approaches to workflow design and resource allocation in the ED.

Differences in ΔAUPRC across clinical phenotypes provide additional context for interpreting the contribution of delayed diagnostic information. For infectious and gastrointestinal/abdominal presentations, there were continued performance gains following the incorporation of serial vital signs and laboratory data. In contrast, there was relatively high discrimination at $T_{0.5}$ but larger improvements in precision after laboratory integration for genitourinary presentations. Within the information-dependent phenotype, the AUPRC of the substance/toxicology subgroup increased from $T_{0.5}$ to $T_{1.0}$ but decreased slightly at $T_{2.0}$ despite continued improvement in AUROC. One possible explanation for this is that in comparison to routine laboratory measurements, short-term physiological trends may more strongly contribute to risk prediction in toxicological presentations. However, additional investigations are required to confirm this observation.

There were limited performance gains after laboratory integration in the dermatological and psychiatric/mental health subgroups, whereas a larger relative increase in AUPRC (ΔAUPRC = 0.1051) despite a modest absolute $T_{2.0}$ AUPRC (0.2985) was observed in the musculoskeletal subgroup. This pattern may reflect differences in baseline event rates between the subgroups.

This study has some limitations. First, missing laboratory values were handled using median imputation combined with measurement indicators. Although this approach captures information associated with test ordering patterns, it may not fully capture the clinical trajectories of patients who undergo highly atypical evaluation pathways. Alternative approaches, such as dynamic imputation and time-varying missingness models, warrant further investigation. Further, the analysis was based on predefined observation windows. In clinical practice, data are generated asynchronously and may not align with the fixed temporal boundaries. Future studies should evaluate continuous-time modeling approaches to represent irregularly sampled physiological data [33].

Moreover, chief complaint categorization relied on density-based clustering methods. Consequently, uncommon or highly heterogeneous presentations may have been underrepresented in the final subgroup structure. Additional studies using large language model–based representations of clinical narratives may improve the characterization of these presentations [34, 35]. In addition, the evaluation used a held-out subset of the MIMIC-IV-ED data without an external validation dataset. External validation across multiple institutions is necessary to assess the generalizability of the information phenotypes identified in this study. Although variation in ΔAUPRC was observed across the phenotypes, further studies are still required to evaluate the statistical stability of these subgroup-specific estimates. Finally, the hematology/oncology subgroup demonstrated a nonmonotonic performance trajectory that was not readily explained by the proposed framework.

Conclusions

The contribution of sequentially available clinical information to early ED risk stratification varied across clinical presentations. Overall model performance improved as additional information became available from $T_{0.5}$ to $T_{2.0}$, while variation in performance trajectories was observed across the 13 CEDIS subgroups. These subgroup-specific trajectories provide a framework for characterizing heterogeneity in the incremental predictive value of clinical information across ED presentations and may serve as a basis for further investigation of presentation-specific risk modeling strategies. External and prospective validation is needed to assess the reproducibility and generalizability of the proposed information phenotypes before their potential clinical relevance can be determined.

Declarations

Competing interests:

The authors declare no competing interests.

Supplementary Information

Supplementary Table S1 provides TRIPOD + AI checklist.

Supplementary Table S2 provides the top 15 hyperparameter configurations ranked by the custom composite score for HDBSCAN clustering optimization.

Supplementary Table S3 provides the predefined sentinel emergency department presentations.

Supplementary Table S4 provides the fairness audit results.

Funding:

This research was supported by a grant from the Korea Health Technology R&D Project through the Korea Health Industry Development Institute (KHIDI), funded by the Ministry of Health and Welfare, Republic of Korea (grant No. RS-2025-02222326).

Author Contribution

Conceptualization: T.Y, B.Y; Data curation: T.Y, M.P; Formal analysis: T.Y, M.P, Y.K, C.K; Investigation: T.Y, M.P; Methodology: T.Y, B.Y; Supervision: Y.K, C.K. Validation: B.Y, M.POriginal draft: T.Y; Writing - review & editing: B.Y, M.P, Y.K, C.KAll authors reviewed the manuscript.

Data Availability

The datasets analyzed in this study are derived from the MIMIC-IV (version 2.2; https://doi.org/10.13026/6mm1-ek67) and MIMIC-IV-ED (version 2.2; https://doi.org/10.13026/5ntk-km72) databases, which are available on PhysioNet. Raw data cannot be directly shared or deposited in the supplementary files of this manuscript owing to strict patient privacy restrictions and data use agreements. However, researchers can freely access the datasets after completing PhysioNet credentialing. To ensure research reproducibility, all computational codes, natural language processing pipelines, and model evaluation scripts generated are publicly available and accessible via the following GitHub repository, https://github.com/tae1234/2026_MIMIC_IV_ED_experiment.

References

1. Kelen, G. D. et al. Emergency department crowding: the canary in the health care system. *NEJM Catal. Innov. Care Deliv*. **2** 10.1056/CAT.21.0217 (2021).
2. Bernstein, S. L. et al. The effect of emergency department crowding on clinically oriented outcomes. *Acad. Emerg. Med.* **16**, 1–10 (2009). PMID: 19007346.

3. Morley, C., Unwin, M., Peterson, G. M., Stankovich, J. & Kinsman, L. Emergency department crowding: a systematic review of causes, consequences and solutions. PLOS One. ;13:e0203316. PMID: 30161242. (2018).
4. Stang, A. S., Wingert, A. S., Hartling, L. & Plint, A. C. Adverse events related to emergency department care: a systematic review. *PLOS One*. **8**, e74214 (2013). PMID: 24069281.
5. Johnson, A. E. W. et al. MIMIC-IV, a freely accessible electronic health record dataset. Sci Data. ;10:1. PMID: 36596836. (2023).
6. Johnson, A. et al. MIMIC-IV-ED version 2.2; PhysioNet. https://physionet.org/content/mimic-iv-ed/2.2/. Accessed Jun 10, 2026. (2023). 10.13026/5ntk-km72
7. Raita, Y. et al. Emergency department triage prediction of clinical outcomes using machine learning models. Crit Care. ;23:64. PMID: 30795786. (2019).
8. Yun, H., Choi, J. & Park, J. H. Prediction of critical care outcome for adult patients presenting to emergency department using initial triage information: an XGBoost algorithm analysis. *JMIR Med. Inf.* **9**, e30770 (2021). PMID: 34346889.
9. Taylor, R. A. et al. Prediction of in-hospital mortality in emergency department patients with sepsis: a local big data–driven, machine learning approach. *Acad. Emerg. Med.* **23**, 269–278 (2016). PMID: 26679719.
10. Wardi, G. et al. Predicting progression to septic shock in the emergency department using an externally generalizable machine-learning algorithm. *Ann. Emerg. Med.* **77**, 395–406 (2021). PMID: 33455840.
11. Collins, G. S. et al. TRIPOD + AI statement: updated guidance for reporting clinical prediction models that use regression or machine learning methods. *BMJ* **385**, e078378 (2024). PMID: 38626948.
12. Lipton, Z. C., Kale, D. C. & Wetzel, R. Directly modeling missing data in sequences with RNNs: improved classification of clinical time series. *Proc. Mach. Learn. Res.* **56**, 253–270 (2016).
13. Che, Z., Purushotham, S., Cho, K., Sontag, D. & Liu, Y. Recurrent neural networks for multivariate time series with missing values. Sci Rep. ;8:6085. PMID: 29666385. (2018).
14. Groenwold, R. H. H. & Dekkers, O. M. Missing data: the impact of what is not there. *Eur. J. Endocrinol.* **183**, E7–9. 10.1530/EJE-20-0732 (2020).
15. Grootendorst, M. & BERTopic Neural topic modeling with a class-based TF-IDF procedure. arXiv preprint arXiv:2203.05794. (2022). 10.48550/arXiv.2203.05794
16. Gu, Y. et al. Domain-specific language model pretraining for biomedical natural language processing. *ACM Trans. Comput. Healthc.* **3**, 1–23. 10.1145/3458754 (2021).
17. McInnes, L., Healy, J., Melville, J. & UMAP Uniform manifold approximation and projection for dimension reduction. arXiv preprint arXiv:1802.03426. (2018).
18. Campello, R. J. G. B., Moulavi, D. & Sander, J. Density-based clustering based on hierarchical density estimates. In: (eds Pei, J., Tseng, V. S., Cao, L., Motoda, H. & Xu, G.) Advances in Knowledge Discovery and Data Mining. PAKDD 2013. Lecture Notes in Computer Science, vol 7819, pp. 160–172. Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-642-37456-2_14
19. Grafstein, E., Unger, B., Bullard, M. & Innes, G. Canadian emergency department information system (CEDIS) presenting complaint list (version 1.0). *CJEM* **5**, 27–34 (2003). PMID: 17659149.
20. McInnes, L. & Healy, J. Accelerated hierarchical density based clustering. In: 2017 IEEE international conference on data mining workshops (ICDMW). New York: Institute of Electrical and Electronics Engineers; pp. 33–42, (2017). 10.1109/ICDMW.2017.12
21. Cox, D. R. & Stat Soc, B. The regression analysis of binary sequences. J R ;20:215 – 32. (1958). 10.1111/j.2517-6161.1958.tb00292.x
22. Breiman, L. Random forests. *Mach. Learn.* **45**, 5–32. 10.1023/A:1010933404324 (2001).

23. Chen, T. & Guestrin, C. Xgboost: a scalable tree boosting system. In: Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining. New York, NY, USA: ACM; pp. 785–94, (2016). 10.1145/2939672.2939785
24. Paszke, A. et al. Pytorch: an imperative style, high-performance deep learning library. *Adv. Neural Inf. Process. Syst.* **32** 10.48550/arXiv.1912.01703 (2019).
25. Kingma, D. P. & Ba, J. Adam: a method for stochastic optimization. International Conference on Learning Representations; (2015). 10.48550/arXiv.1412.6980
26. Lundberg, S. M. & Lee, S-I. A unified approach to interpreting model predictions. *Adv. Neural Inf. Process. Syst.* **30** 10.48550/arXiv.1705.07874 (2017).
27. Lundberg, S. M. et al. Explainable machine-learning predictions for the prevention of hypoxaemia during surgery. *Nat. Biomed. Eng.* **2**, 749–760 (2018). PMID: 31001455.
28. Gilboy, N., Tanabe, P., Travers, D. & Rosenau, A. M. Emergency Severity Index (ESI): a triage tool for emergency department care, version 4. In: Implementation Handbook, 12–0014. Rockville, MD: AHRQ Publication. ISBN: 1-58763-212-8; (2012).
29. Xie, F. et al. Benchmarking emergency department prediction models with machine learning and public electronic health records. *Sci. Data*. **9**, 658 (2022). PMID: 36302776.
30. Klug, M. et al. A gradient boosting machine learning model for predicting early mortality in the emergency department triage: devising a nine-point triage score. *J. Gen. Intern. Med.* **35**, 220–227 (2020). PMID: 31677104.
31. Shwartz-Ziv, R. & Armon, A. Tabular data: deep learning is not all you need. *Inf. Fusion*. **81**, 84–90. 10.1016/j.inffus.2021.11.011 (2022).
32. Gorishniy, Y., Rubachev, I., Khrulkov, V. & Babenko, A. Revisiting deep learning models for tabular data. *Adv. Neural Inf. Process. Syst.* **34**, 18932–18943. 10.48550/arXiv.2106.11959 (2021).
33. Chen, R. T. Q., Rubanova, Y., Bettencourt, J. & Duvenaud, D. K. Neural ordinary differential equations. *Adv. Neural Inf. Process. Syst.* **31** 10.48550/arXiv.1806.07366 (2018).
34. Singhal, K. et al. Large language models encode clinical knowledge. *Nature* **620**, 172–180 (2023). PMID: 37438534.
35. Li, Y., Wehbe, R. M., Ahmad, F. S., Wang, H. & Luo, Y. A comparative study of pretrained language models for long clinical text. J Am Med Inform Assoc. ;30:340-7. PMID: 36451266. (2023).

Figures

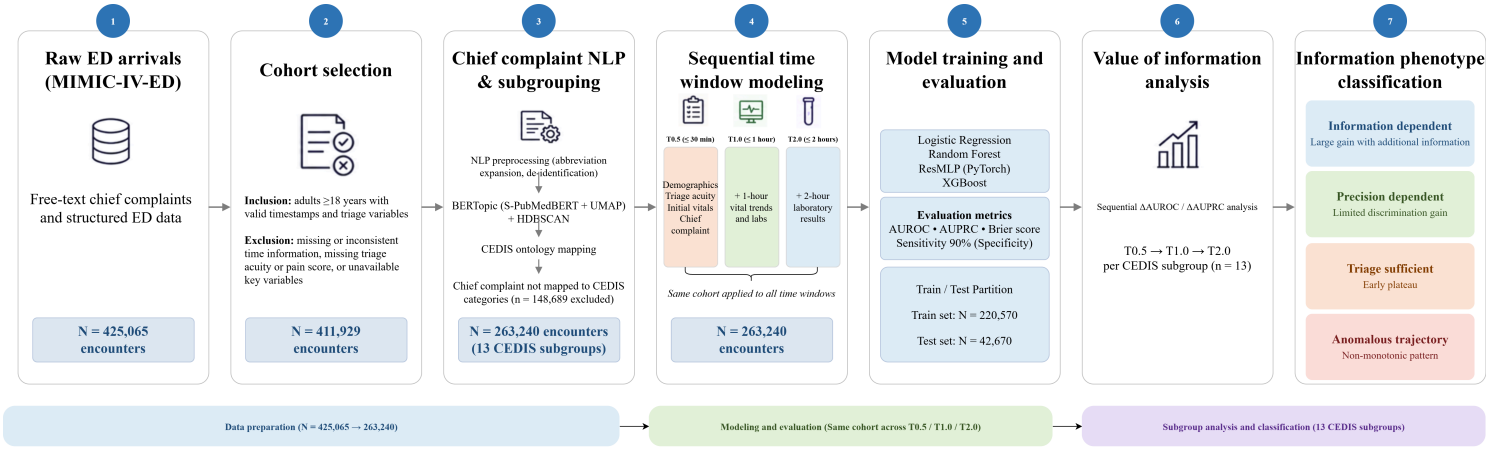


Figure 1

Flow diagram of the overall study design and sequential evaluation pipeline. The diagram illustrates the methodology applied to the MIMIC-IV-ED dataset, including natural language processing (NLP)-based chief complaint subgrouping, sequential time window modeling (T0.5, T1.0, and T2.0), and the final classification of information phenotypes.

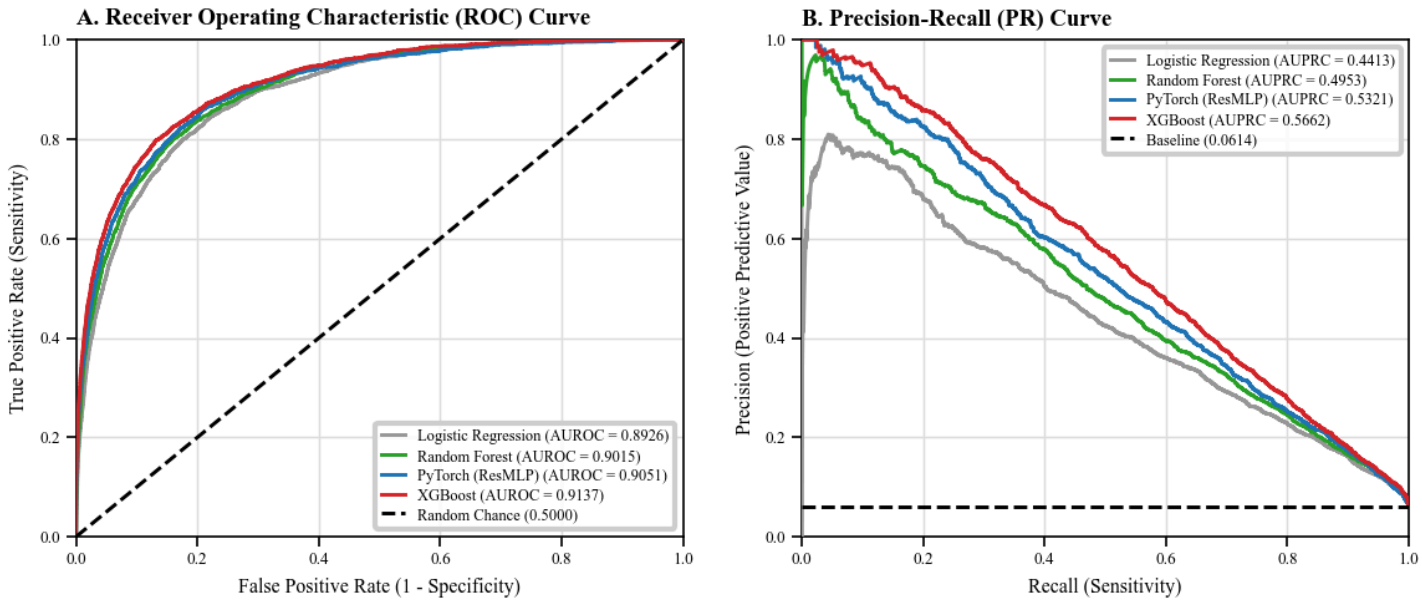


Figure 2

Predictive performance of the machine learning models at the augmented observation window ($T_{2.0}$). (**A**) Receiver operating characteristic (ROC) curves and (**B**) precision-recall (PR) curves for the prediction of early critical illness within 24 h of emergency department arrival.

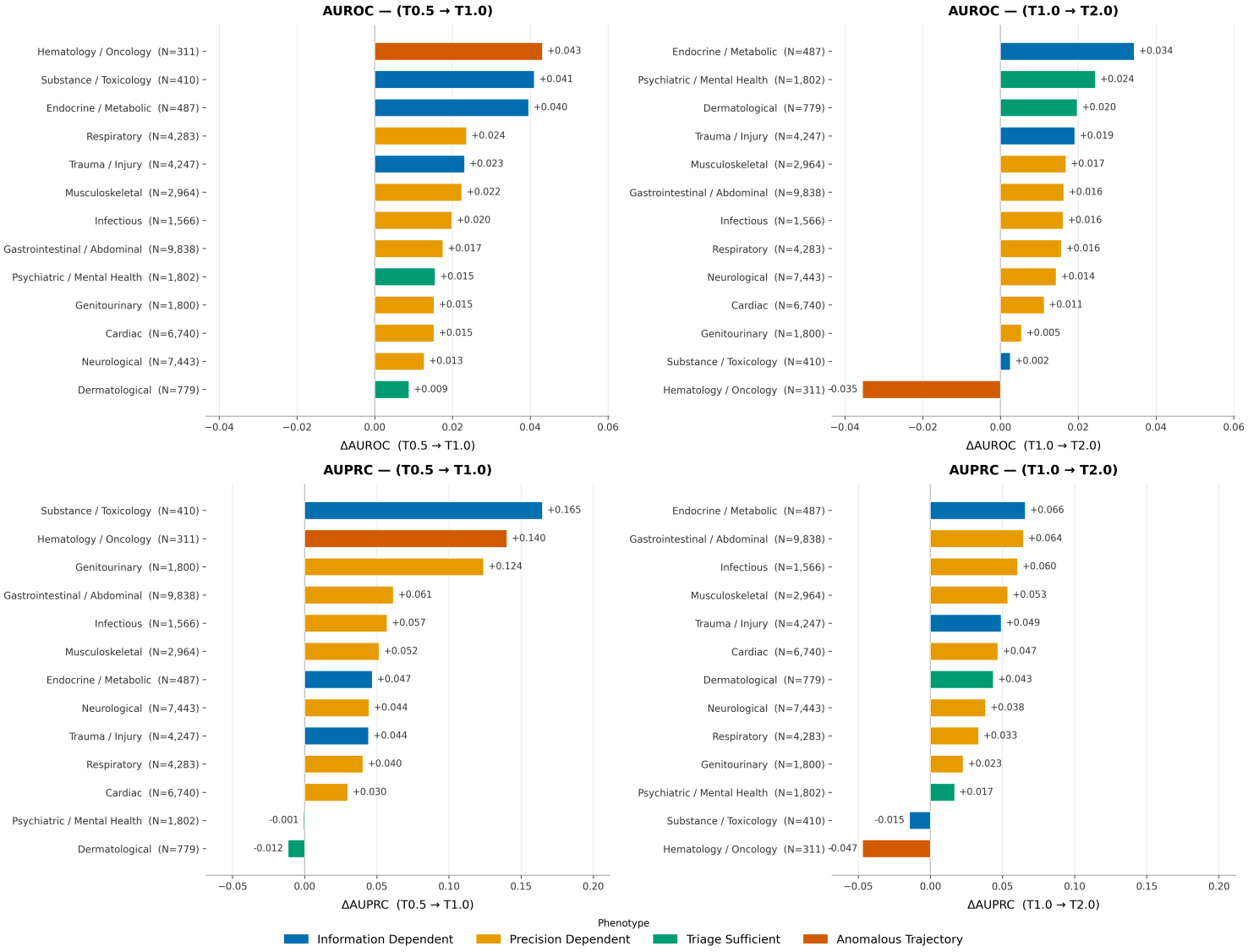


Figure 3

Stepwise ranking of subgroup performance changes across sequential observation windows. Horizontal bar charts rank the 13 CEDIS subgroups by change in AUROC and AUPRC for two sequential steps: $T_{0.5}$ to $T_{1.0}$, $T_{1.0}$ to $T_{2.0}$. Bars are colored by information phenotypes.

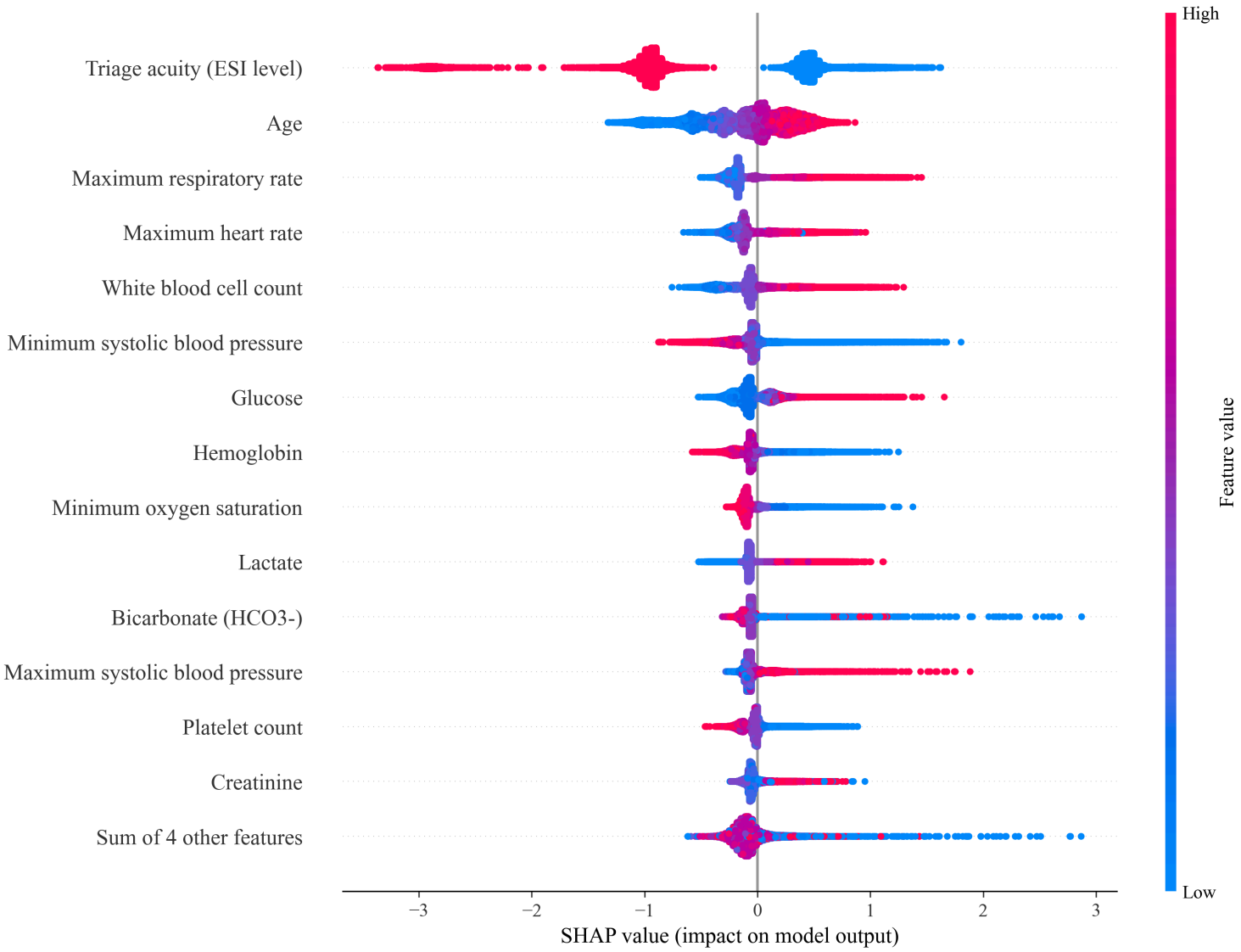


Figure 4

Global feature importance (SHAP analysis) in the XGBoost model at $T_{2\cdot0}$. This beeswarm plot displays the directional impact of clinical features on predicted risk of early critical illness, ranked vertically by mean absolute SHAP value. Each dot represents one emergency department encounter; the horizontal position indicates the direction and magnitude of impact on model output; and the color reflects the feature's numeric value (red = high, blue = low).

Supplementary Files

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

- SupplementaryTableS3EDPresentationsSR.docx
- SupplementaryTableS2SentinelHyperparameterSR.docx
- SupplementaryTableS4FairnessAuditSR.docx
- SupplementaryTableS1TRIPODAISR.docx