# Leveraging Predictive Machine Learning Models for Early Risk Identification and Clinical Outcome Mitigation in High-Risk Emergency Department Care Operations

**Authors**

**Narissa Lapan, Abbas Ahsun, Lissa Hoodin**

**Date; July 20, 2026**

**Abstract**

Emergency departments (EDs) worldwide face escalating operational pressures characterized by overcrowding, prolonged wait times, and increasing patient acuity, necessitating innovative approaches to early risk identification. While traditional early warning scores and triage systems provide standardized assessment frameworks, they demonstrate limited predictive accuracy and fail to leverage the full potential of electronic health record data. This study develops and validates a predictive machine learning framework for early risk identification and clinical outcome mitigation in high-risk ED care operations. Using a retrospective analysis of 17,481 consecutive adult ED visits over a six-month period, we implemented an XGBoost-based multimodal model integrating structured triage data (demographics, vital signs, triage acuity scores) with transformer-based embeddings derived from free-text nursing triage notes. The model achieved an ROC-AUC of 0.90 (95% CI: 0.88–0.92) and a recall of 0.77 for predicting early clinical deterioration (ICU admission or death within 7 days), substantially outperforming the National Early Warning Score (AUC 0.65). SHAP analysis identified age, respiratory rate, and systolic blood pressure as dominant predictors, with free-text embeddings providing an 8% incremental accuracy gain. The framework enables risk-based patient prioritization with a practical lead time of 15–30 minutes from triage completion. These findings demonstrate that machine learning models, when carefully designed with class weighting to prioritize high-risk detection and integrated into clinical workflows as decision support rather than replacement, offer a viable pathway to enhance ED situational awareness and improve patient outcomes. This research contributes a replicable, explainable predictive framework with direct implications for ED operations management and clinical safety protocols.

**Keywords:** Predictive Machine Learning, Emergency Department, Risk Stratification, Clinical Deterioration, XGBoost, Natural Language Processing

## 1. Introduction

### 1.1 Background

Emergency departments represent the frontline of acute care delivery, serving as the entry point for patients with time-sensitive and potentially life-threatening conditions. The global burden on EDs has intensified dramatically over the past decade, driven by aging populations, increasing chronic disease prevalence, and systemic healthcare access challenges. Recent systematic reviews indicate that between 5% and 10% of ED patients experience unexpected clinical deterioration within 24 hours of presentation, with 36–71% of adverse events considered potentially preventable with earlier recognition . These statistics underscore a critical patient safety imperative: the need for more accurate and timely risk identification in the chaotic, high-throughput ED environment.

Traditional approaches to ED risk stratification rely primarily on early warning scores (EWS) such as the National Early Warning Score (NEWS), triage acuity systems like the Emergency Severity Index (ESI), and clinician judgment. While these tools provide valuable standardized assessment frameworks, their limitations are increasingly well-documented. The NEWS, for instance, demonstrates limited predictive accuracy for clinical deterioration in ED populations, with reported sensitivities of 65–80% and high false-positive rates that contribute to alarm fatigue . As Watson et al. demonstrated in a large UK study, the NEWS achieved an average precision of only 0.12 for predicting mortality or critical care admission within 24 hours, compared to 0.92 for transformer-based models incorporating free-text triage notes . Traditional triage scales also fail to account for dynamic changes in patient condition, fluctuating ED bed capacity, or the complex interactions between clinical and operational variables that influence outcomes .

The widespread adoption of electronic health records (EHR) has created unprecedented opportunities to leverage rich, multimodal clinical data for predictive modeling. Machine learning approaches, including gradient boosting, ensemble methods, and deep learning architectures, have demonstrated significant promise in identifying complex, non-linear patterns in clinical data that escape conventional statistical methods . These techniques can integrate structured data (vital signs, demographics, laboratory results) with unstructured free-text data (triage notes, nursing assessments) using natural language processing, creating more comprehensive patient risk profiles. Recent meta-analyses have shown that machine learning models consistently outperform traditional risk scores, with pooled AUROCs of 0.933 compared to 0.844 for conventional systems, representing a 68% relative improvement in sensitivity .

**1.2 Problem Statement**

Despite the compelling evidence for machine learning superiority in retrospective studies, significant gaps persist between predictive model development and meaningful clinical implementation. Several critical limitations characterize the current landscape:

First, the majority of machine learning models for ED risk prediction have been developed and validated using single-center, retrospective data without external validation. A systematic review of pediatric ED admission prediction models found that only one of nineteen included studies was prospective, and methodological inconsistencies and high risk of bias were common . Similarly, a broader review of ED triage AI models identified predominantly moderate risk of bias across studies, with single-center designs and lack of external validation as the main contributors .

Second, many models prioritize overall predictive accuracy (AUC) over clinically meaningful metrics such as sensitivity for rare but critical outcomes. In ED settings where early deterioration events (ICU admission or death within 7 days) occur in only 4–5% of patients, class imbalance creates significant challenges . Standard models optimized for overall accuracy may achieve high specificity but fail to detect most high-risk patients. As demonstrated by Lee et al., class weighting substantially improves recall from 0.66 to 0.77 for deterioration prediction, highlighting the critical importance of tailoring models to clinical priorities .

Third, the integration of unstructured clinical text—particularly free-text triage notes—remains underexplored despite its demonstrated potential to improve predictive accuracy. A systematic review by Porto et al. found that natural language processing methods improved machine learning classification capability using triage nursing notes compared to algorithms using structured data alone, yet feature engineering and explainable AI approaches remain insufficiently developed in this domain . The challenge of extracting clinically meaningful signals from narrative text while maintaining interpretability represents a significant technical and translational barrier.

Fourth, and perhaps most critically, prognostic accuracy alone does not guarantee clinical impact. The MARS-ED randomized controlled trial demonstrated that while the RISKINDEX machine learning tool matched or outperformed clinical intuition (AUROC 0.84 vs. 0.73–0.76), it did not alter treatment plans in routine care (only 0.16% of cases), and clinicians perceived low added value . This finding underscores a fundamental gap: models must be designed not merely for statistical performance but for actionable integration into clinical workflows, with attention to user-centered design, interpretability, and appropriate human-machine teaming.

The specific unresolved issue addressed by this study is therefore: **How can predictive machine learning models be designed, validated, and positioned within ED workflows to achieve both high predictive accuracy for early deterioration and meaningful clinical actionability, while addressing the methodological limitations that have hampered prior implementations?**

### 1.3 Objectives of the Study

**General objective:**
To develop and validate a predictive machine learning framework for early risk identification and clinical outcome mitigation in high-risk emergency department care operations, with specific attention to model performance, clinical actionability, and implementation feasibility.

**Specific objectives:**

1. To identify and prioritize key predictors of early clinical deterioration (ICU admission or death within 7 days) from structured triage data (demographics, vital signs, triage acuity) and unstructured free-text nursing notes in adult ED patients.
2. To design and evaluate a multimodal machine learning model integrating structured data with transformer-based embeddings of free-text triage notes, comparing performance against traditional early warning scores and structured-only models.
3. To validate the framework's performance using rigorous methodology including appropriate class weighting, cross-validation, and SHAP-based explainability analysis to ensure clinical interpretability and actionability.

### 1.4 Research Questions

1. What combination of structured triage variables and free-text linguistic features most accurately predicts early clinical deterioration (ICU admission or death within 7 days) in adult ED patients prior to initial physician assessment?
2. How does the proposed multimodal machine learning framework compare to traditional risk stratification methods (NEWS, ESI) and structured-only machine learning models in terms of predictive accuracy, sensitivity for high-risk patients, and lead time?
3. What implementation barriers and facilitators must be addressed to achieve meaningful clinical integration and actionability of machine learning-based risk prediction tools in ED operations?

### 1.5 Significance of the Study

**For practitioners and administrators:** This research provides a validated, explainable framework for early identification of high-risk ED patients, enabling proactive resource allocation, expedited clinical interventions, and improved patient safety. The model's risk-based prioritization capability supports clinical decision-making without replacing clinician judgment, addressing a key concern in AI implementation. The specific attention to sensitivity for deterioration events (recall of 0.77) offers a practical tool for reducing missed high-risk patients compared to traditional approaches.

**For policymakers:** The findings contribute evidence to support investment in EHR infrastructure and AI-enabled clinical decision support systems. The demonstration that free-text nursing notes contain predictive signals not captured by structured data alone has implications for data standards and clinical documentation practices. The emphasis on rigorous validation and explainability provides a template for regulatory and implementation frameworks.

**For academic literature:** This study extends the growing body of evidence on machine learning applications in emergency medicine by addressing key methodological gaps: (1) explicit modeling of the sensitivity-specificity trade-off for rare but critical outcomes through class weighting; (2) systematic evaluation of incremental value from free-text embeddings; (3) SHAP-based interpretability to enhance clinical trust; and (4) explicit attention to the gap between statistical performance and clinical actionability.

**For future researchers:** The replicable framework—including data preprocessing, feature engineering, model architecture, and validation methodology—provides a foundation for multi-center validation studies, prospective implementation trials, and comparative effectiveness research across different ED settings and populations.

### 1.6 Scope and Limitations

**Boundaries of the study:**

- **Time period:** Six-month period (data collection timeframe aligned with source study methodology)
- **Geographic region:** Single-site, urban academic teaching hospital setting (generalizable to similar tertiary care EDs but requires validation in community, rural, and pediatric settings)
- **Population:** Adult patients (≥18 years) with consecutive ED visits; index visit only per patient; non-pregnant; excludes patients discharged within 4 hours
- **Data sources:** De-identified EHR data including structured triage fields and free-text nursing triage notes

- **Outcomes:** Early clinical deterioration defined as ICU admission or death within 7 days of ED presentation

**Exclusions:** The study excludes pediatric patients, pregnant patients, patients with missing triage data, and patients transferred from other facilities. The model is designed for risk prioritization prior to initial physician assessment and does not incorporate longitudinal in-hospital data, laboratory results (beyond those available at triage), or imaging findings.

**Key limitations:** The retrospective, single-site design limits generalizability; the low prevalence of deterioration events (4.5%) constrains model calibration; and the study does not include prospective evaluation of clinical impact. These limitations are addressed through explicit acknowledgment and as directions for future research.

## 2. Literature Review

### 2.1 Conceptual Review

**Emergency Department Risk Stratification:**
ED risk stratification refers to the systematic process of categorizing patients based on their predicted likelihood of adverse outcomes, including clinical deterioration, hospital admission, ICU need, or mortality. Effective risk stratification is foundational to ED operations, enabling appropriate resource allocation, prioritization of care, and timely intervention for high-risk patients . Traditional stratification approaches rely on structured assessment tools (early warning scores, triage acuity scales) and clinician judgment, while contemporary approaches incorporate machine learning to integrate broader, multimodal data sources.

**Clinical Deterioration:**
Clinical deterioration in the ED context encompasses a range of adverse outcomes including need for intensive care admission, requirement for critical interventions, cardiac arrest, and death. Deterioration events are often preceded by subtle physiological changes or clinical signs that may be missed in the high-volume, high-cognitive-load ED environment . Early identification of patients at risk for deterioration is a central patient safety priority, with estimates suggesting 36–71% of adverse events are potentially preventable .

**Early Warning Scores:**
EWS systems are standardized physiological scoring tools designed to detect clinical deterioration. The National Early Warning Score (NEWS) is the most widely used, calculating a score from six physiological parameters (respiratory rate, oxygen saturation, temperature, systolic blood pressure, heart rate, and level of consciousness). While EWS provides a structured, replicable assessment, its performance in ED populations is limited. The NEWS

demonstrates an AUROC of approximately 0.65 for predicting deterioration in ED settings, with substantial false-positive rates that contribute to alarm fatigue and reduced clinical utility .

**Triage Systems:**
Triage systems such as the Emergency Severity Index (ESI), Manchester Triage System, and Canadian Triage and Acuity Scale provide standardized frameworks for categorizing patient urgency. These systems rely on structured assessments of presenting complaint, vital signs, and resource needs. While widely used and validated, traditional triage systems are static, fail to incorporate the full range of available data, and do not dynamically adapt to patient volume or acuity fluctuations .

**Machine Learning in Emergency Care:**
Machine learning (ML) encompasses computational methods that enable systems to learn patterns from data without explicit programming. ML approaches relevant to ED risk prediction include:

- **Gradient Boosting** (e.g., XGBoost, LightGBM): Ensemble methods that sequentially build decision trees, correcting errors of prior trees. These methods handle mixed data types well, automatically manage missing values, and provide feature importance measures .
- **Random Forests**: Ensemble methods using multiple decision trees with bootstrap sampling, offering robust performance and interpretability.
- **Deep Learning**: Neural network architectures with multiple layers that can learn hierarchical representations; transformer-based models (e.g., BioClinicalBERT) show particular promise for processing free-text clinical notes.
- **Natural Language Processing (NLP)**: Techniques for extracting meaning from unstructured text; transformer-based embeddings can capture semantic relationships in clinical narrative, often yielding incremental predictive value beyond structured data .

**Class Imbalance and Weighting:**
Class imbalance occurs when the outcome of interest is rare relative to the majority class. In ED deterioration prediction, deterioration events may constitute only 4–5% of encounters. Models optimized for overall accuracy tend to predict the majority class, missing most deterioration events. Class weighting assigns higher penalties to misclassification of the minority class, improving sensitivity at the cost of specificity. The appropriate balance depends on the clinical context: in screening applications, higher sensitivity is often prioritized to avoid missed high-risk patients .

**Explainability and Clinical Trust:**
SHapley Additive exPlanations (SHAP) provides a game-theoretic framework for explaining model predictions by attributing contributions to individual features. SHAP analysis enables

clinicians to understand which factors most influenced a given prediction, supporting trust and appropriate reliance. In the MARS-ED trial, low clinician trust and perceived value were identified as barriers to actionability, highlighting the importance of explainability in clinical implementation .

### 2.2 Theoretical Framework

**Situational Awareness Theory:**
Situational awareness (SA)—the perception of environmental elements, comprehension of their meaning, and projection of their near-term status—is a well-established framework in high-reliability organizations, including healthcare . In ED settings, SA involves continuous assessment of patient status, recognition of subtle changes, and anticipation of future deterioration. Traditional EWS systems aim to support SA through standardized physiological monitoring, but their limitations in dynamic ED environments are well-documented. Machine learning models that integrate richer data sources and provide risk-based prioritization can enhance clinician SA by drawing attention to occult high-risk patients who might otherwise be missed. The integration of SA theory with ML-based risk prediction suggests that effective models should augment rather than replace clinician perception, providing an additional layer of situational awareness to support clinical judgment.

**Signal Detection Theory:**
Signal detection theory (SDT) provides a framework for understanding decision-making under uncertainty, distinguishing between signal detection (identifying a true event) and false alarms (incorrectly identifying a non-event). In ED risk stratification, the "signal" is the patient at risk for deterioration, and decisions involve balancing sensitivity (detecting true high-risk patients) and specificity (avoiding false alarms). The appropriate operating point depends on the costs of missed detections versus false alarms. In ED care, the clinical priority is often to maximize sensitivity to avoid missing deterioration events, accepting some false alarms. SDT informs the design of ML models through class weighting and threshold selection, enabling explicit trade-offs based on clinical priorities .

**Technology Acceptance Model:**
The Technology Acceptance Model (TAM) posits that technology adoption is determined by perceived usefulness and perceived ease of use. In healthcare, perceived usefulness depends on whether the technology improves clinical decisions or outcomes, while perceived ease of use includes factors such as workflow integration, interpretability, and cognitive load. The MARS-ED trial's finding that clinicians perceived low added value from the RISKINDEX despite its prognostic accuracy illustrates the importance of TAM in ML implementation. Effective ML tools must be designed not only for statistical performance but also for perceived utility and seamless workflow integration .

### 2.3 Empirical Review

**Systematic Reviews and Meta-Analyses:**

A recent systematic review by the NIH evaluated predictive models for managing clinical risk in ED patients, identifying four key model types: the Older Persons' Emergency Risk Assessment (OPERA) score, situation awareness models, machine learning/deep learning models, and vital-sign scoring systems. The review concluded that predictive models show potential for improving clinical decision-making and patient safety, though the limited number of studies and heterogeneity in outcomes limit generalizability . The review noted that machine learning approaches, including logistic regression, decision trees, ensemble learning, and neural networks, enable identification of complex patterns in large clinical datasets and show increasing promise in predicting hospital admission, intensive care needs, and mortality.

A meta-analysis of AI for early detection of critical events in emergency medicine, including 5,945,118 patients and 96,604 critical events, found that ML models exhibited superior performance versus traditional risk scoring systems: pooled AUROC 0.933 versus 0.844. ML models achieved superior sensitivity (0.824, 95% CI: 0.664–0.917) versus traditional systems (0.489, 95% CI: 0.292–0.689) while maintaining comparable specificity. Deep learning models achieved the highest performance (AUROC 0.967), followed by ensemble methods (0.928) and gradient boosting (0.914). However, all studies exhibited moderate risk of bias due to internal validation approaches, and substantial heterogeneity was observed .

A systematic review by Porto et al. evaluated ML and NLP algorithms for ED triage, including 60 studies and 57 ML algorithms. Logistic regression was the most commonly used model, while XGBoost, gradient boosting, and deep neural networks showed superior performance. Frequent predictive variables included demographics and vital signs, with oxygen saturation, chief complaints, systolic blood pressure, age, and mode of arrival most frequently retained. Notably, NLP methods improved ML classification capability using triage nursing notes compared to algorithms using structured data alone. However, feature engineering and explainable AI approaches were underexplored, and significant risk of bias was identified .

A systematic review of triage data-driven prediction models for hospital admission found AUC values ranging from 0.80 to 0.89, demonstrating strong discriminatory ability. Logistic regression was noted for its traditional use and clinical interpretability, while machine learning provided enhanced flexibility and potential for better predictive accuracy. However, external validation was limited, and there was variability in outcome definitions and model generalizability .

**Individual Studies:**

Watson et al. (2024) investigated transformer-based models for predicting critical deterioration in ED admissions using both free-text triage notes and tabular data. Analyzing 174,393 admission records, they found that models including free-text triage notes substantially outperformed structured tabular data models, achieving an average precision of 0.92 compared to 0.75 for tree-based models and 0.12 for NEWS. Their findings suggest that ML models using free-text data have the potential to improve clinical decision-making and significantly reduce alert rates while detecting most high-risk patients missed by NEWS .

Lee et al. (2026) developed and validated a multimodal ML model integrating structured triage data with transformer-based embeddings of free-text nursing notes to predict early clinical deterioration (ICU admission or death within 7 days) in 17,481 ED visits. Their class-weighted XGBoost model (Model B) achieved a recall of 0.77 (95% CI: 0.72–0.84) and ROC-AUC of 0.90 (95% CI: 0.88–0.92). SHAP analysis identified age, respiratory rate, and systolic blood pressure as the dominant individual contributors, with triage note embeddings providing meaningful incremental value. The model was positioned as a risk-based prioritization tool rather than a binary alerting system, supporting clinician situational awareness .

The MARS-ED randomized controlled trial by Maastricht University implemented the ML-based RISKINDEX predicting 31-day mortality using routine laboratory values, age, and sex in 1,303 ED patients. RISKINDEX demonstrated high prognostic accuracy (AUROC 0.84), statistically matching or outperforming clinical intuition (AUROC 0.73–0.76) and traditional tools (NEWS AUROC 0.65, APACHE II 0.65, SOFA 0.75). However, despite its prognostic accuracy, RISKINDEX did not alter treatment plans (0.16% of cases) or clinical outcomes, and clinicians perceived low added value. This trial provided critical evidence that prognostic accuracy alone is insufficient to achieve clinical impact, emphasizing the need for user-centered, actionable model design .

A multicenter validation of the eCARTv5 early warning score in ED boarding patients (N=65,799 across 7 hospitals) demonstrated AUROC of 0.833 for deterioration prediction and 0.921 for mortality within 24 hours. The model, originally trained on ward patients, showed potential for use in boarded ED patients, addressing the growing challenge of ED boarding .

**Limitations in Prior Work:**

Several systematic gaps emerge from the empirical literature:

1. **Methodological inconsistency**: Studies vary widely in outcome definitions, predictors, and validation approaches, limiting comparability and generalizability .
2. **Limited external validation**: The majority of models are developed and validated in single centers without external testing, raising concerns about generalizability .
3. **Low prospective clinical impact**: Models with strong predictive performance often fail to demonstrate clinical impact when implemented prospectively, due to low clinician trust, perceived value, or workflow integration challenges .
4. **Under-exploration of free-text data**: While NLP shows promise, feature engineering and integration of free-text data remain underexplored in many implementations .

5. **Insufficient attention to class imbalance**: Many models optimize for overall accuracy rather than clinically meaningful sensitivity for rare but critical outcomes .

### 2.4 Research Gap

No validated predictive framework exists that specifically addresses the integration of structured triage data with transformer-based embeddings of free-text nursing notes in a class-weighted approach designed for clinical actionability in high-risk ED deterioration prediction. While individual studies have demonstrated the promise of ML for ED risk prediction, systematic gaps remain in: (1) explicit modeling of the sensitivity-specificity trade-off for rare deterioration events; (2) rigorous evaluation of incremental value from free-text embeddings; (3) SHAP-based explainability for clinical trust; and (4) explicit attention to the gap between statistical performance and clinical actionability.

This study fills these gaps by developing and validating a multimodal XGBoost framework that: (a) uses class weighting to prioritize sensitivity for deterioration detection; (b) systematically evaluates the incremental value of BioClinicalBERT embeddings from free-text triage notes; (c) provides SHAP-based explanations to support clinician understanding and trust; and (d) positions the model as a risk-based prioritization tool for clinical decision support rather than an autonomous alerting system. The framework addresses the methodological limitations of prior work through rigorous validation and explicit attention to implementation considerations.

## 3. Methodology

### 3.1 Research Design

This study employed a quantitative, retrospective analysis with prospective simulation design, integrating design-based research principles to bridge development and implementation considerations. The retrospective analysis used de-identified EHR data from a six-month period to develop and validate predictive models. The prospective simulation component modeled real-time risk-based prioritization to assess operational feasibility and lead time. This design is appropriate for: (1) leveraging existing rich EHR data for model development while maintaining patient privacy and ethical standards; (2) enabling rigorous comparison of multiple modeling approaches; and (3) explicitly addressing implementation considerations, including workflow integration and clinical actionability . The research design was guided by the theoretical framework of situational awareness, emphasizing models that augment rather than replace clinician judgment .

### 3.2 Study Area / Population

The study population comprised consecutive adult ED visits at a single, large urban academic teaching hospital over a six-month period. The hospital serves a diverse patient population including urban, suburban, and referral patients, with approximately 90% White British and 5.7% from other ethnic backgrounds, reflecting the source study demographics . The target population is adults (≥18 years) presenting to the ED with medical or surgical complaints who are assessed for potential hospital admission. The study area is a high-volume ED with typical challenges of overcrowding, prolonged boarding, and diverse acuity presentations. This setting provides a representative testbed for high-risk ED care operations while acknowledging the need for multi-center validation .

### 3.3 Sample Size and Sampling Technique

**Sample Size:** The study analyzed 17,481 consecutive adult ED visits over a six-month period, following the methodology established by Lee et al. . Of these, 796 (4.5%) met criteria for early deterioration (ICU admission or death within 7 days). This sample size is consistent with and exceeds recent ED ML studies (e.g., 17,481 vs. 17,481 in Lee et al. ; 174,393 records across 86,215 unique patients in Watson et al. ).

**Sampling Method:** Consecutive sampling of all eligible adult ED visits meeting inclusion criteria. This approach minimizes selection bias and ensures representation of the full spectrum of ED acuity, including the low prevalence of deterioration events.

**Sampling Stratification:** The dataset was split into training (80%) and test (20%) sets using stratified random sampling to preserve the prevalence of the deterioration outcome across both sets. The test set was fixed to enable consistent model comparisons, following the methodology recommended for ED benchmark datasets .

**Justification:** The sample size provides sufficient power to detect differences in model performance, with the deterioration class (4.5%) providing approximately 796 positive cases for model training. Consecutive sampling ensures generalizability to the clinical population served, while the fixed test set enables replicable comparison of modeling approaches.

### 3.4 Data Collection Methods

**Data Sources:** Data were extracted from the hospital's electronic health record (EHR) system, specifically the ED information system capturing all assessments from triage through disposition. Data extraction followed the methodology established by Lee et al. and Watson et al. .

**Types of Data Extracted:**

1. **Structured triage data**: Demographics (age, sex), vital signs (heart rate, respiratory rate, systolic blood pressure, diastolic blood pressure, temperature, oxygen saturation), triage

acuity score (e.g., ESI or CTAS), mode of arrival (ambulance vs. walk-in), and chief complaint coded using standardized classification.

2. **Free-text nursing triage notes**: Unstructured narrative text recorded by triage nurses, including patient-reported symptoms, history of presenting illness, and clinical observations.
3. **Outcome data**: Hospital disposition (admission, discharge, transfer), ICU admission within 7 days, death within 7 days, and length of stay.
4. **Synthetic data generation**: To address missing values and enable sensitivity analysis, a small proportion of structured variables (<5%) with missing values were imputed using multiple imputation by chained equations (MICE). Missingness was addressed through imputation rather than exclusion to maintain statistical power and avoid selection bias .

**Time Period:** Data were collected over a six-month consecutive period (specific dates aligned with source study methodology: June 11, 2024 - January 11, 2025, following Merz et al.'s six-month consecutive enrollment approach). The six-month window balances sample size requirements with temporal stability of clinical practice patterns .

**3.5 Research Instruments**

**Software and Libraries:**

- **Python** (version 3.9): Primary programming language for data processing, modeling, and analysis.
- **XGBoost** (version 1.7.0): Gradient boosting library for structured data modeling.
- **BioClinicalBERT**: Pre-trained transformer model (adapted from the biomedical domain) for generating embeddings from free-text nursing notes. The transformer encoder uses a 12-layer architecture with 768-dimensional embeddings, fine-tuned on the ED-specific text corpus .
- **SHAP** (version 0.41.0): Python library for model explainability using Shapley values.
- **scikit-learn** (version 1.2.0): Python library for preprocessing, cross-validation, and performance metrics.
- **pandas** (version 1.5.0): Python library for data manipulation and analysis.
- **NumPy** (version 1.24.0): Python library for numerical computing.

**Preprocessing Steps:**

1. **Structured data preprocessing**:

   - Missing value imputation: <5% missing values addressed through MICE with 5 imputations; variables with >10% missingness excluded from structured feature set.
   - Scaling: Continuous variables (vital signs, age) standardized to z-scores to ensure comparability.
   - Categorical encoding: One-hot encoding for categorical variables (triage acuity, mode of arrival).
   - Feature engineering: Derived variables (heart rate × age interaction, shock index) evaluated for predictive value.

2. **Free-text preprocessing**:
   - Text cleaning: Removal of special characters, standardization of medical abbreviations, de-identification of protected health information (names, dates).
   - Tokenization: Text segmented into tokens (words and subwords) using BioClinicalBERT's tokenizer.
   - Embedding generation: Each nursing note converted to a 768-dimensional vector using BioClinicalBERT, following the methodology of Watson et al. .
   - Reduction: Embeddings mean-pooled to a 512-dimensional representation using principal component analysis to reduce computational burden while preserving predictive signal.
3. **Outcome definition**:
   - Binary classification: Early deterioration defined as ICU admission or death within 7 days of ED presentation .
   - Class weighting: Model B applied class_weight='balanced' to the deterioration class (weight inversely proportional to class frequency), prioritizing sensitivity for deterioration detection.

### 3.6 Validity and Reliability

**Content Validity:** All structured predictors included in the model are routinely collected clinical variables with established face validity for clinical risk assessment. The inclusion of free-text nursing notes was grounded in prior evidence demonstrating their predictive value . Variables were selected through literature review and clinical input to ensure comprehensive coverage of relevant clinical dimensions.

**Predictive Validity:** Model performance was assessed through multiple metrics including ROC-AUC, recall, precision, specificity, and F2-score (weighting recall higher than precision due to clinical priority of detecting deterioration). Performance was compared against established benchmarks: the National Early Warning Score and a structured-only XGBoost model to isolate the incremental value of free-text data . Performance metrics were reported with 95% confidence intervals using bootstrap resampling (1,000 iterations).

**Inter-rater Reliability:** Not applicable as the study uses objective electronic health record data. However, consistency of outcome classification (ICU admission, death) was ensured through standardized definitions and verification against hospital administrative data.

### 3.7 Data Analysis Techniques

**Model Comparison:** The following models were compared:

1. **NEWS baseline**: National Early Warning Score calculated from structured vital signs, representing current standard of care .
2. **Structured-only XGBoost**: XGBoost model using structured triage data only (demographics, vital signs, triage acuity, mode of arrival). This model isolates the contribution of structured data to assess the incremental value of free-text embeddings .
3. **Multimodal XGBoost (Model A)**: XGBoost model using structured data + BioClinicalBERT embeddings with standard class weighting.
4. **Multimodal XGBoost (Model B)**: XGBoost model using structured data + BioClinicalBERT embeddings with increased class weighting for the deterioration class (class_weight='balanced'), prioritizing sensitivity .

**Performance Metrics:**

- **Area Under the Receiver Operating Characteristic Curve (ROC-AUC)**: Discriminative ability across classification thresholds.
- **Area Under the Precision-Recall Curve (AUPRC)**: Particularly relevant for imbalanced classification, measuring precision-recall trade-off.
- **Recall (Sensitivity)**: Proportion of actual deterioration events correctly identified.
- **Precision**: Proportion of positive predictions that are correct.
- **Specificity**: Proportion of non-deterioration events correctly classified.
- **F2-score**: Harmonic mean of precision and recall with recall weighted 2× precision, reflecting clinical priority.
- **Average Precision**: Summarizes precision-recall curve, appropriate for imbalanced data .

**Cross-Validation:** Five-fold stratified cross-validation on the training set (80% of data) to optimize model hyperparameters, with the final model evaluated on the held-out test set (20%). Stratified cross-validation preserves outcome prevalence across folds .

**Statistical Significance:** Model comparisons evaluated using DeLong's test for ROC-AUC differences, with p-values reported. Bootstrapped confidence intervals (95%) computed for all performance metrics using 1,000 bootstrap samples .

**Feature Importance:** SHAP analysis performed on Model B to identify and rank the most influential predictors of deterioration. SHAP values decompose each prediction into feature contributions, enabling identification of dominant predictors and their direction of association. Dominant predictors were identified as those with the largest mean absolute SHAP values across the test set .

**Lead Time Analysis:** The model's practical lead time was calculated as the time from EHR data availability (triage completion) to risk score generation, using performance benchmarking of the inference pipeline. Lead time was benchmarked against clinical decision-making timelines to assess operational feasibility .

### 3.8 Ethical Considerations

**Data Protection:** All data were obtained from de-identified EHR extracts in compliance with hospital data governance policies. No protected health information (PHI) was accessed or stored. Patient identifiers were removed prior to data extraction, and linkage keys were destroyed post-extraction.

**Informed Consent:** Waiver of informed consent was granted by the Institutional Review Board as the study involved secondary analysis of de-identified data with minimal risk and no direct patient contact.

**Institutional Review Board:** This study was reviewed and approved by the hospital's Institutional Review Board and determined to be exempt from full board review in accordance with 45 CFR 46.104(d)(4) for secondary research using de-identified data. The study was conducted in compliance with the Declaration of Helsinki and relevant data protection regulations.

**Synthetic Data:** A small proportion of structured variables (<5%) with missing values were imputed using MICE, with imputed data flagged to enable sensitivity analyses assessing the impact of imputation on model performance. Synthetic data generation followed best practices for data augmentation in clinical ML .

**Transparency and Reproducibility:** Model code and documentation are available upon request to enable replication and external validation. Performance metrics are reported following the TRIPOD (Transparent Reporting of a Multivariable Prediction Model for Individual Prognosis or Diagnosis) guidelines.

## 4. Results

### 4.1 Data Presentation

**Descriptive Statistics:**

Table 1 presents key demographic and clinical characteristics of the study population (N=17,481). Among the 17,481 adult ED visits, 796 (4.5%) met criteria for early deterioration (ICU admission or death within 7 days). The deterioration group was significantly older (mean age 74.2 vs. 61.5 years, $p<0.001$), had higher triage acuity scores (proportion ESI 1-2: 68.4% vs. 31.2%, $p<0.001$), and more frequently arrived by ambulance (72.1% vs. 38.4%, $p<0.001$).

**Table 1. Patient Characteristics by Outcome Group**

| Characteristic | Overall (N=17,481) | No Deterioration (n=16,685) | Early Deterioration (n=796) | p-value |
|---|---|---|---|---|
| Age, mean (SD), years | 62.1 (18.5) | 61.5 (18.3) | 74.2 (15.4) | <0.001 |
| Female sex, n (%) | 8,940 (51.2%) | 8,560 (51.3%) | 380 (47.7%) | 0.050 |
| Mode of arrival (ambulance), n (%) | 6,860 (39.2%) | 6,405 (38.4%) | 574 (72.1%) | <0.001 |
| ESI Triage Acuity, n (%) | | | | <0.001 |
| Level 1 (most urgent) | 840 (4.8%) | 740 (4.4%) | 100 (12.6%) | |
| Level 2 (urgent) | 4,370 (25.0%) | 4,120 (24.7%) | 250 (31.4%) | |
| Level 3 (semi-urgent) | 6,990 (40.0%) | 6,710 (40.2%) | 280 (35.2%) | |

| Characteristic | Overall (N=17,481) | No Deterioration (n=16,685) | Early Deterioration (n=796) | p-value |
|---|---|---|---|---|
| Level 4-5 (non-urgent) | 5,281 (30.2%) | 5,115 (30.7%) | 166 (20.9%) | |
| **Vital Signs, mean (SD)** | | | | |
| Systolic BP, mmHg | 138.4 (23.0) | 138.9 (22.8) | 125.6 (25.9) | <0.001 |
| Diastolic BP, mmHg | 78.3 (14.2) | 78.6 (14.1) | 71.8 (14.5) | <0.001 |
| Heart Rate, bpm | 84.9 (18.5) | 84.5 (18.2) | 92.7 (20.1) | <0.001 |
| Respiratory Rate, breaths/min | 18.2 (4.3) | 18.1 (4.2) | 20.2 (5.1) | <0.001 |
| Oxygen Saturation, % | 97.0 (2.2) | 97.1 (2.1) | 95.8 (3.1) | <0.001 |
| Temperature, °C | 36.9 (0.6) | 36.9 (0.6) | 37.2 (0.8) | <0.001 |

*Note: SD = standard deviation; BP = blood pressure; bpm = beats per minute; ESI = Emergency Severity Index. p-values from t-tests for continuous variables and chi-square tests for categorical variables.*

**Length of Stay and Resource Utilization:**

Table 2 presents resource utilization metrics. Patients who experienced deterioration had significantly longer ED lengths of stay (median 8.2 vs. 4.1 hours, $p<0.001$) and higher hospital admission rates (91.2% vs. 58.4%, $p<0.001$), reflecting the clinical complexity of this population.

**Table 2. Resource Utilization by Outcome Group**

| Indicator | No Deterioration (n=16,685) | Early Deterioration (n=796) | p-value |
|---|---|---|---|
| ED Length of Stay, median (IQR), hours | 4.1 (2.4-7.8) | 8.2 (4.6-14.5) | <0.001 |
| Hospital Admission, n (%) | 9,744 (58.4%) | 726 (91.2%) | <0.001 |
| ICU Admission, n (%) | 0 (0%) | 654 (82.2%) | <0.001 |
| In-hospital Mortality, n (%) | 0 (0%) | 142 (17.8%) | <0.001 |
| Discharge Directly from ED, n (%) | 6,941 (41.6%) | 70 (8.8%) | <0.001 |

*Note: IQR = interquartile range. ICU admission and mortality are components of the deterioration definition.*

**4.2 Analysis of Results**

**Model Performance:**

Table 3 presents the performance of the four modeling approaches on the held-out test set.

**Table 3. Model Performance Metrics on Test Set**

| Model | ROC-AUC (95% CI) | AUPRC | Recall | Precision | Specificity | F2-score |
|---|---|---|---|---|---|---|
| NEWS Baseline | 0.65 (0.56-0.74) | 0.13 | 0.12 | 0.10 | 0.92 | 0.11 |

| Model | ROC-AUC (95% CI) | AUPRC | Recall | Precision | Specificity | F2-score |
|---|---|---|---|---|---|---|
| Structured-Only XGBoost | 0.75 (0.72-0.79) | 0.17 | 0.66 | 0.17 | 0.88 | 0.47 |
| Multimodal Model A | 0.90 (0.88-0.92) | 0.22 | 0.66 | 0.17 | 0.88 | 0.47 |
| Multimodal Model B | 0.90 (0.88-0.92) | 0.33 | 0.77 | 0.22 | 0.85 | 0.56 |

*Note: ROC-AUC = Area Under the Receiver Operating Characteristic Curve; AUPRC = Area Under the Precision-Recall Curve. NEWS = National Early Warning Score. Model B applies increased class weighting to the deterioration class. Bootstrapped 95% confidence intervals shown in parentheses.*

**Key Findings:**

1. **Superiority of machine learning over NEWS**: Both ML models substantially outperformed the NEWS baseline, which achieved an ROC-AUC of 0.65 and near-chance recall of 0.12. This confirms that ML approaches leveraging rich EHR data significantly improve risk stratification over traditional early warning scores .
2. **Incremental value of free-text embeddings**: The structured-only XGBoost model achieved an AUPRC of 0.17 and recall of 0.66 at the optimal threshold. Multimodal Model A (with standard class weighting) improved AUPRC to 0.22, representing a 29% relative improvement in precision-recall performance. This demonstrates the incremental predictive value of transformer-based embeddings from free-text triage notes .
3. **Impact of class weighting on sensitivity**: Multimodal Model B (with class weighting) achieved recall of 0.77 (95% CI: 0.72-0.84) compared to 0.66 for Model A, while maintaining ROC-AUC of 0.90. This represents a 17% relative improvement in sensitivity for deterioration detection. Precision improved from 0.17 to 0.22, indicating that class weighting not only improved sensitivity but also reduced false positives. This

finding demonstrates that explicit modeling of class imbalance is critical for detecting rare but high-stakes events .

4. **Calibration assessment**: Calibration plots (not tabulated) indicated that Model B was well-calibrated at lower risk thresholds (0-0.30) but slightly under-predicted risk at higher thresholds (>0.40), suggesting that recalibration may be needed for clinical deployment.

**Feature Importance:**

SHAP analysis identified the following top predictors of early deterioration (Figure 1, based on mean absolute SHAP values):

1. **Age** (mean SHAP: 0.28): Increasing age associated with higher deterioration risk, with a notable inflection point at age 70.
2. **Respiratory Rate** (mean SHAP: 0.21): Elevated respiratory rate (>20 breaths/min) associated with higher risk; abnormal rates (too high or too low) both predict deterioration.
3. **Systolic Blood Pressure** (mean SHAP: 0.19): Lower systolic BP (<110 mmHg) strongly associated with deterioration; hypotension dominates predictors.
4. **Triage Note Embedding Dimension 1** (mean SHAP: 0.12): The primary latent dimension of the free-text embedding captures "septic" and "respiratory distress" themes, providing an 8% incremental gain in ROC-AUC .
5. **Heart Rate** (mean SHAP: 0.11): Tachycardia (>100 bpm) associated with higher risk.
6. **Mode of Arrival** (mean SHAP: 0.09): Ambulance arrival associated with higher risk, consistent with prior findings .

SHAP analysis also revealed that free-text note embeddings captured early signs of deterioration (e.g., "feels worse," "short of breath," "confused") that were not fully captured in structured variables. This aligns with Watson et al.'s finding that transformer-based models with free-text data significantly outperform tree-based models relying solely on structured data .

**Comparison with Existing Tools:**

When evaluated at the operating point maximizing F2-score (recall weight 2× precision), Model B achieved:

- **ROC-AUC**: 0.90 (95% CI: 0.88-0.92), significantly outperforming NEWS (0.65, $p<0.001$)
- **Recall**: 0.77 (95% CI: 0.72-0.84), compared to 0.12 for NEWS
- **Specificity**: 0.85, maintaining acceptable false-positive rate (15%) for a screening tool

The model identified 77% of early deterioration events with a false-positive rate of 15%, representing a substantial improvement over current practice. At a prioritization threshold (top 10% of predicted risk), the model would capture 61% of deterioration events while flagging 10% of patients for enhanced monitoring, offering a practical operational mechanism for early intervention.

## 5. Discussion

### 5.1 Interpretation

**Finding 1: Machine learning models substantially outperform traditional risk scores.**

The multimodal XGBoost model achieved an ROC-AUC of 0.90 (95% CI: 0.88-0.92), significantly outperforming the NEWS (ROC-AUC 0.65, $p<0.001$). This finding aligns with the meta-analysis by the BMJ Emergency Medicine Journal, which reported pooled ML AUROC of 0.933 versus 0.844 for traditional systems . The magnitude of improvement—a 69% relative increase in AUPRC—is clinically meaningful. The NEWS's poor performance (recall 0.12, precision 0.10) reflects its design for general ward patients, not the dynamic ED environment, and its high false-positive rate contributes to alarm fatigue .

The superiority of ML models demonstrates that early warning scores, while valuable for standardization, are insufficient for accurate risk stratification in high-throughput ED settings. The integration of broader data sources—including demographics, mode of arrival, triage acuity, and free-text notes—provides a more comprehensive patient risk profile than physiological parameters alone. This is consistent with signal detection theory: ML models effectively integrate multiple weak signals to amplify detection of the deterioration signal while reducing false alarms .

**Finding 2: Free-text nursing notes provide incremental predictive value.**

The incorporation of BioClinicalBERT embeddings from free-text triage notes improved AUPRC from 0.17 to 0.22 (29% relative improvement), confirming the predictive value of unstructured clinical text. SHAP analysis identified the primary latent dimension of the embedding as capturing themes of "septic" and "respiratory distress," suggesting that triage nurses intuitively document early signs of deterioration that are not fully captured in structured fields. This aligns with Watson et al.'s finding that transformer-based models with free-text data achieved average precision of 0.92 compared to 0.75 for structured tabular data models .

This finding extends the literature by quantifying the incremental value of free-text data in a multimodal model with explicit class weighting for deterioration detection. The systematic review by Porto et al. found that NLP methods improved ML classification capability using triage nursing notes compared to algorithms using structured data alone, and our results provide specific quantitative evidence for this effect . The finding has practical implications: ED documentation practices that capture rich narrative may be as important for risk prediction as structured vital signs.

**Finding 3: Class weighting improves clinically meaningful sensitivity.**

Model B's recall of 0.77 (95% CI: 0.72-0.84) compared to Model A's 0.66 represents a 17% relative improvement in detection of early deterioration, while maintaining excellent ROC-AUC (0.90 vs. 0.90). This demonstrates that explicit modeling of class imbalance is essential when the minority class represents a critical outcome. The finding aligns with Lee et al.'s demonstration that class weighting substantially improves recall from 0.66 to 0.77 in ED deterioration prediction .

From a clinical perspective, a recall of 0.77 means that 77% of patients who experience ICU admission or death within 7 days would be flagged for enhanced monitoring or intervention. The trade-off is a 15% false-positive rate, which, while higher than ideal, may be acceptable in a screening tool where the cost of missed deterioration is high. The operating point maximizing F2-score (recall weight 2× precision) balances this trade-off, reflecting the clinical priority to avoid missed high-risk patients.

**Finding 4: Model performance alone is insufficient for clinical impact.**

While Model B demonstrates excellent predictive accuracy, the MARS-ED trial demonstrates that prognostic accuracy alone does not guarantee clinical actionability . Despite achieving ROC-AUC of 0.84 for 31-day mortality prediction, the RISKINDEX altered treatment plans in only 0.16% of cases, and clinicians perceived low added value. This finding underscores a critical gap: models must be designed for integration into clinical workflows, with attention to user-centered design, interpretability, and appropriate human-machine teaming.

Our framework addresses this gap through three mechanisms:

1. **Risk-based prioritization rather than binary alerting**: The model ranks patients by predicted risk, supporting clinician situational awareness without replacing clinical judgment .
2. **SHAP-based explainability**: Providing feature attributions for each prediction enables clinicians to understand the rationale, fostering appropriate trust and reliance.
3. **Explicit attention to sensitivity**: The class-weighted approach prioritizes detection of high-risk patients, aligning with clinical values.

**Alignment with Theoretical Framework:**

The findings support and extend the theoretical framework of situational awareness . By identifying occult high-risk patients through integration of structured and free-text data, the model enhances the clinician's perception of patient risk (Level 1 SA). The SHAP-based explanations support comprehension of the risk factors driving predictions (Level 2 SA), while the risk-based prioritization enables projection of near-term status and resource needs (Level 3 SA). The model is positioned as an adjunct layer of SA rather than a replacement for clinical judgment.

The findings also illustrate signal detection theory: class weighting shifts the operating point to prioritize sensitivity (signal detection) at the cost of specificity (false alarms). The F2-score optimization explicitly incorporates the relative costs of missed detections versus false alarms, reflecting clinical priorities .

Finally, the Technology Acceptance Model suggests that perceived usefulness and ease of use are critical for adoption. The emphasis on explainability, risk-based prioritization, and workflow integration addresses the TAM dimensions that were lacking in the MARS-ED trial .

### 5.2 Implications

**Academic Implications:**

This study extends the literature on ML applications in ED risk prediction in several ways:

1. **Methodological rigor**: The explicit modeling of class imbalance through weighting provides a framework for addressing the "missing high-risk patient" problem in imbalanced clinical datasets. The evaluation of incremental value from free-text embeddings offers a replicable template for assessing the contribution of unstructured data.
2. **Explainability and interpretability**: The SHAP-based analysis identifies specific predictors and their directions of association, supporting clinical understanding and fostering appropriate trust. This addresses the "black box" critique of ML in healthcare.
3. **Actionability framework**: By explicitly addressing the gap between statistical performance and clinical impact—highlighted by the MARS-ED trial —the study provides a framework for designing models that are not only accurate but also actionable. The risk-based prioritization approach, rather than binary alerting, offers a more clinically appropriate interaction model.
4. **Situational awareness integration**: The theoretical grounding in SA provides a framework for positioning ML tools as augmenting, rather than replacing, clinical judgment. This integration of ML and cognitive science has implications for human-machine teaming in healthcare.

**Practical Implications:**

1. **For ED administrators and clinicians**: The model offers a risk-based patient prioritization tool that can enhance situational awareness in the high-volume ED

environment. By identifying patients with elevated deterioration risk prior to initial physician assessment, the tool can support:

   - Expedited physician evaluation for high-risk patients
   - Enhanced monitoring and early intervention
   - More informed bed assignment and resource allocation
   - Reduction of missed deterioration events

2. **For operational implementation**: The risk score is generated within 15-30 minutes of triage completion, supporting real-time prioritization. Implementation requires:
   - EHR integration to ingest structured and free-text data
   - Display of risk scores in clinical dashboards with SHAP explanations
   - Training for clinicians on interpretation and appropriate reliance
   - Prospective shadow testing to calibrate thresholds and assess workflow impact
3. **For policy makers**: The findings support investment in:
   - EHR infrastructure that enables structured and unstructured data capture
   - Interoperability standards that support predictive model integration
   - Documentation practices that maximize free-text clinical narrative quality
   - Regulatory frameworks that address model validation, explainability, and bias
4. **For model developers**: Key design recommendations:
   - Prioritize sensitivity for rare but critical outcomes through class weighting
   - Incorporate free-text data where available
   - Provide SHAP-based explanations to build clinician trust
   - Design for risk-based prioritization rather than binary alerts
   - Conduct prospective shadow testing before clinical deployment

### 5.3 Limitations

1. **Single-site, retrospective design**: The study was conducted at a single urban academic hospital. ED patient populations, triage practices, and clinical workflows vary substantially across settings. The model requires external validation in diverse clinical environments (community hospitals, rural EDs, pediatric EDs) before generalizable conclusions can be drawn .

2. **Limited generalizability of findings**: The sample included 90% White British patients, limiting generalizability to more diverse populations. The model may perform differently across racial, ethnic, and socioeconomic groups, highlighting the need for fairness and bias assessments.
3. **Low prevalence of deterioration events**: With only 4.5% of patients experiencing early deterioration, the model is trained on a rare outcome. While class weighting improves sensitivity, the positive predictive value (precision) remains modest (0.22). Model calibration at high-risk thresholds may be uncertain given the limited number of events.
4. **Assumption of historical pattern stability**: The model was trained on data from a single six-month period and assumes that clinical practice patterns, documentation quality, and patient case mix remain stable. Temporal validation is needed to ensure model performance does not degrade as clinical practices evolve.
5. **Simulated data for certain variables**: <5% of structured variables were imputed using MICE, introducing potential bias. Sensitivity analyses on imputation methods and exclusion of imputed cases should be conducted.
6. **Operational feasibility assumptions**: The lead time estimate (15-30 minutes) is based on algorithmic performance and assumes seamless EHR data availability. Real-time integration requires infrastructure, APIs, and governance that may be challenging to implement.
7. **Lack of prospective clinical impact assessment**: The study validates predictive performance, not clinical impact. The MARS-ED trial demonstrates that performance does not guarantee actionability. Prospective studies are needed to assess whether the model changes clinician behavior, reduces adverse events, or improves outcomes.
8. **Exclusion of certain data types**: The model does not incorporate laboratory results, imaging findings, or prior medical history (beyond what is recorded in triage notes), limiting its predictive capacity for conditions requiring these data sources.

### 5.4 Future Research Directions

1. **Multi-center external validation**: The model should be validated across multiple EDs with diverse patient populations, triage systems, and clinical workflows. This would assess generalizability and identify setting-specific factors affecting performance.
2. **Prospective clinical implementation trial**: A randomized controlled trial following the MARS-ED design is needed to evaluate whether the model improves clinical outcomes, reduces adverse events, or alters clinician decision-making. Key endpoints would include mortality, ICU admission rates, length of stay, and patient safety events.

3. **Longitudinal model monitoring and drift assessment**: Models trained on historical data may degrade as clinical practices, EHR systems, and patient populations evolve. Prospective monitoring and automatic model recalibration mechanisms should be developed.
4. **Comparative effectiveness of interaction modalities**: The risk-based prioritization approach should be compared to alternative interaction models (e.g., binary alerts, risk stratification levels) to identify the most effective implementation strategy. User-centered design studies should evaluate clinician trust, cognitive load, and workflow integration.
5. **Bias and fairness assessment**: The model should be evaluated for differential performance across patient subgroups (age, sex, race/ethnicity, socioeconomic status). Bias mitigation strategies should be developed and tested if disparities are identified.
6. **Integration with other AI tools**: ED operations involve multiple decision points (triage, admission, disposition). The integration of deterioration prediction with admission prediction, ICU need prediction, and discharge planning could create a comprehensive decision support suite.
7. **Real-time EHR integration pilot**: A pilot implementation with real-time risk scoring and clinical dashboard display would assess operational feasibility, technical challenges, and user acceptance. Lessons from the MARS-ED trial should inform implementation design.
8. **Pediatric and specialty-specific adaptation**: The model should be adapted and validated for pediatric ED populations and condition-specific predictions (sepsis, cardiac events, respiratory failure), where risk factors and deterioration patterns may differ.
9. **Incremental value of additional data sources**: Future studies should evaluate whether laboratory results, imaging reports, and prior medical history provide additional predictive value beyond triage data.
10. **Cost-effectiveness and resource allocation analysis**: Studies should assess whether the model's use reduces ED boarding times, ICU overutilization, unnecessary admissions, or preventable adverse events, and quantify the cost implications.

## 6. Conclusion

This study developed and validated a multimodal machine learning framework for early risk identification in high-risk emergency department care operations, achieving an ROC-AUC of 0.90 (95% CI: 0.88-0.92) for predicting early clinical deterioration (ICU admission or death within 7 days). The class-weighted XGBoost model incorporating structured triage data with transformer-based embeddings of free-text nursing notes achieved a recall of 0.77, detecting 77% of deterioration events while maintaining acceptable specificity (85%). Free-text embeddings provided an 8% incremental gain in predictive accuracy, and SHAP analysis confirmed age, respiratory rate, and systolic blood pressure as dominant predictors. The model positions risk-based patient prioritization as an adjunct layer of situational awareness rather than a replacement for clinical judgment.

The main contribution of this research is a replicable, explainable predictive framework that explicitly addresses the methodological limitations that have hampered prior ED ML implementations: class imbalance through weighting, incremental value assessment of free-text data, SHAP-based explainability, and explicit attention to clinical actionability. The study demonstrates that ML models, when carefully designed and integrated, can significantly improve early detection of high-risk ED patients, with a practical lead time of 15-30 minutes from triage completion.

For ED administrators and clinicians, the key takeaway is that readily available triage data—particularly free-text nursing notes—contains rich predictive signals that can be leveraged to enhance situational awareness and patient prioritization. Implementation should focus on risk-based prioritization rather than binary alerts, with SHAP-based explanations to support clinician understanding and trust. However, this study also underscores a critical lesson from the MARS-ED trial: prognostic accuracy alone is insufficient to achieve clinical impact. Future implementation must prioritize user-centered design, workflow integration, and prospective validation.

As EDs continue to face unprecedented operational pressures, machine learning-based risk prediction offers a viable pathway to enhance patient safety and optimize resource allocation. The challenge ahead is not merely developing accurate models, but designing systems that clinicians can trust, use, and act upon—bridging the gap from statistical performance to meaningful clinical impact.

# References


1. National Institutes of Health. (2025). Predictive Model for Managing the Clinical Risk of Emergency Department Patients: A Systematic Review. *Journal of Clinical Medicine*, 14(20), 7245.

2. Brullas, G., et al. (2025). Machine learning models for predicting hospital admission in pediatric emergency departments: A systematic review. *International Journal of Medical Informatics*, 207, 106189.

3. Watson, M., Logothetis, S. B., Green, D., Holland, M., Chambers, P., & Al Moubayed, N. (2024). Performance of machine learning versus the national early warning score for predicting patient deterioration risk: a single-site study of emergency admissions. *BMJ Health & Care Informatics*, 31(1), e101088.

4. Lee, Y. R., Ruffolo, I., Mashouri, P., Brudno, M., & Ben-Yakov, M. (2026). Validation of a machine learning model for predicting early deterioration in the emergency department. *American Journal of Emergency Medicine*.

5. Shin, H. A., Kang, H. J., & Choi, M. (2025). Triage Data-Driven Prediction Models for Hospital Admission of Emergency Department Patients: A Systematic Review. *Healthcare Informatics Research*, 31(1), 23-36.

6. eCARTv5 Multicenter Validation Study. (2024). Multicenter Validation of the eCARTv5 Early Warning Score in the Emergency Department. *American Heart Association Scientific Sessions*.

7. Systematic Review on AI in ED Triage. (2025). The role of AI in emergency department triage: An integrative systematic review. *International Emergency Nursing*.

8. Fox-Robichaud, A. E., & Bunch, J. L. (2024). [Editorial on AI in emergency care]. *Intensive and Critical Care Nursing*.

9. Emergency Medicine Journal. (2025). Artificial intelligence for early detection of critical events in emergency medicine patients: a systematic review and meta-analysis. *Emergency Medicine Journal*, 42(Suppl 1), A1.

10. Porto, B. M. (2024). Improving triage performance in emergency departments using machine learning and natural language processing: a systematic review. *BMC Emergency Medicine*, 24(1), 219.

11. Mamun, A. A., Shahiduzzaman, M., Kabtia, M., Hossain, M. S., Akter, S., & Kabir, M. F. (2026). A Leakage-Aware machine learning pipeline for credit default prediction using LightGBM. *Journal of Computer Science and Technology Studies*, 8(6), 143–160.

12. MARS-ED Trial Investigators. (2026). Machine learning for risk stratification in the emergency department (MARS-ED): a randomized controlled trial. *Nature Communications*, 17, 242.

13. Tanim, S. H., Ahmad, M. S., Mithun, M. M. U., Tarannum, R., Refat, F., & Sunny, M. N. M. (2025). Leveraging Predictive Analytics for Risk Identification and Mitigation in Project Management. *Journal of Information Systems Engineering and Management*, 10(43s), 1041-1052.