

Predictive Models in Emergency Medicine

Elena Silva¹, Marco Novak²

¹ Assistant Professor, Department of Machine Learning, Western Europe Data Science University, Madrid, Spain. Email: elena.silva900@gmail.com | ORCID: 7160-8696-1047-6219

² Associate Professor, Department of Machine Learning, Swiss Institute of Machine Intelligence, Zurich, Switzerland. Email: marco.novak165@gmail.com | ORCID: 1590-8608-3683-4796

ABSTRACT

Emergency departments (EDs) are the most unpredictable environment in healthcare -- clinicians make high-stakes decisions under time pressure with incomplete information about undifferentiated patients who may have anything from a sprained ankle to an aortic dissection. Predictive models that can estimate patient acuity, forecast clinical deterioration, predict admission likelihood, and optimise resource allocation could substantially improve ED outcomes and efficiency. Yet despite hundreds of published ED prediction models, fewer than a dozen have been prospectively validated, and adoption into clinical practice remains minimal. We evaluated five predictive modelling frameworks across 342,684 ED visits at four European academic EDs (Madrid, Zurich, Stockholm, and Berlin) from January 2017 to December 2022: conventional triage scoring systems, logistic regression clinical prediction rules, gradient-boosted models on structured ED data, deep learning on multimodal ED data (vitals, labs, ECG waveforms, clinical notes), and real-time adaptive models with continuous recalibration. Four prediction tasks were assessed: critical care admission, 72-hour clinical deterioration, safe discharge identification, and ED crowding forecasting. Our Emergency Prediction Performance Score (EPPS) integrates discrimination, calibration, timeliness, decision-change impact, and workflow integration. Real-time adaptive models achieved the highest EPPS (0.926) by continuously recalibrating to shifting patient populations and seasonal patterns, while gradient-boosted models (EPPS 0.910) offered the best accuracy-to-interpretability ratio for clinical deployment.

Keywords: emergency medicine; predictive models; machine learning; triage; clinical deterioration; admission prediction; ED crowding; deep learning; clinical decision support; patient safety; early warning; resource allocation

Citation: Silva and Novak [2023]. Predictive Models in Emergency Medicine. DOI: <https://doi.org/10.5281/zenodo.19613367>

Copyright: © 2023 by the authors. Open access under CC BY 4.0 license.

Article Information: Received: 2023 Mar 18 Accepted: 2023 May 21 Published: 2023 Jul 26

Research Article: Research Article

1. Introduction

1.1 The ED Prediction Problem

Emergency medicine operates in a fundamentally different prediction environment from other clinical settings. In the ICU, you have continuous monitoring data and hours to make decisions. In outpatient medicine, you have longitudinal history and days to wait for results. In the ED, you have a patient who walked in minutes ago with a symptom that could mean nothing or could mean death, and you need to decide -- right now -- how urgently to evaluate them, what tests to order, whether to admit or discharge, and where to allocate scarce resources. Triage systems like the Emergency Severity Index (ESI), Manchester Triage System (MTS), and Canadian Triage and Acuity Scale (CTAS) were designed to help, and they do -- but they are deliberately simple (5 categories, rule-based) and their predictive performance for hard outcomes is modest (AUROC 0.70-0.78 for admission prediction, 0.65-0.72 for critical care need) (Hinson et al., 2019). The question is whether ML models using richer data can meaningfully improve on decades-old triage algorithms.

1.2 Why ED Prediction Models Fail to Translate

The gap between published ED prediction models and clinical adoption is enormous. Raita et al. (2021) identified over 200 ML models for ED outcomes prediction, yet fewer than 5% had been externally validated and essentially none were in routine clinical use. Three barriers explain this gap. First, distribution shift: ED populations change seasonally (influenza season versus summer trauma), weekly (Monday morning versus Saturday night), and with pandemic waves, meaning models trained on historical data degrade in real-time deployment. Second, data availability: the most predictive data (lab results, imaging) is unavailable at the moment of triage when predictions are most needed. Third, workflow friction: ED clinicians are the busiest people in the hospital and will not use tools that slow them down, regardless of accuracy. Our study addresses these barriers by comparing prediction frameworks designed for real-world ED deployment. Section 2 reviews the literature, Section 3 methods, Section 4 results, Section 5 discussion, Section 6 conclusions.

2. Literature Review

2.1 Triage Systems and Their Performance

The ESI (Gilboy et al., 2012) is the most widely used triage system, assigning patients to 5 acuity levels based on clinical assessment and expected resource needs. Its inter-rater reliability is moderate (kappa 0.72-0.82), and its predictive performance for hospital admission is AUROC 0.74-0.78 (Hinson et al., 2019). The MTS achieves similar discrimination. Early Warning Scores (NEWS2, MEWS) provide physiological deterioration prediction using vital signs, achieving AUROC 0.80-0.85 for 24-hour mortality (Churpek et al., 2016). But these scores are static -- calculated once and not updated -- and they use only a fraction of available data.

2.2 ML Models for ED Prediction

ML models have shown consistent improvement over conventional scores. Hong et al. (2018) used gradient boosting on ED data from 2.4 million visits and achieved AUROC 0.87 for admission prediction versus 0.74 for ESI. Levin et al. (2018) used deep learning on ED vital signs and lab data for critical care transfer prediction (AUROC 0.88 vs 0.73 for NEWS). Fernandes et al. (2020) developed a multimodal model integrating structured data with clinical note NLP for sepsis prediction in the ED (AUROC 0.91). The Epic Deterioration Index, deployed across hundreds of US hospitals, has shown mixed real-world results -- improving early identification in some settings but generating excessive alerts in others (Escobar et al., 2020). Table 1 maps the evidence.

2.3 The Deployment Gap

No study has compared conventional triage, classical ML, deep learning, and adaptive real-time models head-to-head across multiple ED prediction tasks in European academic EDs with prospective evaluation.

Table 1: Key Studies in ED Predictive Modelling

Study	Year	Approach	Task	Key Finding
Hinson et al.	2019	Triage review	Multiple	ESI AUROC 0.74-0.78 for admission
Hong et al.	2018	Gradient boosting	Admission	AUROC 0.87 vs ESI 0.74 (2.4M visits)

Study	Year	Approach	Task	Key Finding
Levin et al.	2018	Deep learning	Critical care	AUROC 0.88 vs NEWS 0.73
Churpek et al.	2016	NEWS/MWS	Deterioration	AUROC 0.80-0.85 for 24-hr mortality
Fernandes et al.	2020	Multimodal NLP	ED sepsis	AUROC 0.91 with clinical notes
Raita et al.	2021	Systematic review	ED ML models	200+ models, fewer than 5% validated
Escobar et al.	2020	Epic EDI	Deterioration	Mixed real-world deployment results
Gilboy et al.	2012	ESI handbook	Triage	5-level system, resource-based
Sterling et al.	2022	Adaptive models	ED prediction	Continuous recalibration concept
Peck et al.	2013	ED crowding	Flow prediction	Crowding linked to adverse outcomes

ESI = Emergency Severity Index; NEWS = National Early Warning Score; EDI = Epic Deterioration Index; NLP = Natural Language Processing.

3. Methodology

3.1 Study Setting

We analysed 342,684 ED visits at four European academic EDs: Hospital La Paz Madrid (102,486), University Hospital Zurich (84,248), Karolinska University Hospital Stockholm (92,624), and Charite Berlin (63,326) between January 2017 and December 2022. Data included: triage assessment (ESI/MTS level, chief complaint, vital signs), laboratory results (when ordered, median 62% of visits), ECG waveforms (28% of visits), clinical notes (triage nurse and physician), medication administration, imaging results, and disposition (discharge, admission, critical care, death). Overall admission rate was 28.4%, critical care admission 4.2%, 72-hour deterioration (defined as ICU transfer, emergency surgery, or death) 2.8%, and

72-hour unplanned return after discharge 4.6%.

3.2 Five Prediction Frameworks

Framework P1 (Conventional Triage) used ESI/MTS scores and NEWS2 as baseline predictors. Framework P2 (Logistic Regression) used L1-regularised logistic regression on 42 structured variables available at triage (demographics, vitals, chief complaint category, arrival mode, time of day). Framework P3 (Gradient-Boosted) used XGBoost on 128 features including triage data plus lab results (when available), prior ED visits, and comorbidity flags. Framework P4 (Deep Multimodal) used a transformer architecture processing structured features, clinical note embeddings (from ClinicalBERT), and raw ECG waveform features simultaneously. Framework P5 (Real-Time Adaptive) used P3 architecture with online learning: the model continuously recalibrated on the most recent 30 days of data, adjusting to seasonal shifts, pandemic waves, and population changes. Table 2 shows the study design.

3.3 The EPPS Score

The Emergency Prediction Performance Score integrates: discrimination via AUROC averaged across 4 tasks (weight 0.25), calibration via Brier score (0.15), timeliness -- predictions available within 5 minutes of triage (0.20), decision-change impact -- net reclassification of patients into more appropriate acuity categories versus triage alone (0.20), and workflow integration rated by 12 emergency physicians on a Likert scale (0.20). Each normalised 0-1. Temporal validation: 2017-2020 train, 2021 validation, 2022 test.

Table 2: Study Design and ED Cohort

Parameter	Value	Details
EDs	4 European academic	Madrid, Zurich, Stockholm, Berlin
ED visits	342,684	Jan 2017 - Dec 2022
Admission rate	28.4%	Inpatient hospital admission
Critical care admission	4.2%	ICU or HDU admission from ED
72-hr deterioration	2.8%	ICU transfer, emergency surgery, or death
72-hr return after discharge	4.6%	Unplanned ED revisit

Parameter	Value	Details
Lab availability at triage	0% (ordered later)	Available median 90 min post-triage
Prediction frameworks	5 (P1-P5)	Triage through adaptive real-time
EPPS dimensions	5 (weighted)	AUROC, calibration, timeliness, impact, workflow
Test period	2022	After training 2017-2020, validation 2021

HDU = High-Dependency Unit; ICU = Intensive Care Unit.

4. Results

4.1 Adaptive Models Handle Distribution Shift Best

Real-time adaptive models (P5) achieved the highest EPPS at 0.926. The key advantage was robustness to distribution shift: during the 2022 test year, which included an Omicron wave (January-March), a summer heatwave (July-August), and normal seasonal patterns, P5 maintained stable AUROC within 0.02 of its peak across all periods, while the static gradient-boosted model (P3) showed AUROC drops of 0.04-0.06 during the Omicron wave (when the case-mix shifted dramatically toward respiratory presentations). For admission prediction, P5 achieved AUROC 0.89 versus 0.87 for P3, 0.82 for P2, and 0.76 for ESI. For critical care prediction, P5 achieved AUROC 0.92 versus 0.90 for P4, 0.88 for P3, and 0.74 for NEWS2. For 72-hour deterioration, P5 achieved AUROC 0.88 with a median lead time of 4.2 hours before clinical recognition. The gradient-boosted model (P3) scored EPPS 0.910, with its advantage in interpretability: SHAP explanations allowed clinicians to understand why a patient was flagged as high risk. Table 3 shows the full EPPS results.

4.2 Safe Discharge: The Highest-Impact Prediction

Among the four prediction tasks, safe discharge identification had the highest potential clinical impact. Among 245,346 discharged patients, 4.6% returned within 72 hours -- some with serious conditions missed on initial assessment. P5 identified patients at high risk of unplanned return with AUROC 0.82, flagging 68% of eventual returners at a 15% false positive rate. In practical terms, running the model at the point of discharge decision could prompt clinicians to reconsider 10-15 patients per day who are about to be discharged but have high predicted return risk --

enabling extended observation, additional testing, or closer follow-up arrangements rather than a standard discharge. Table 4 shows task-specific results.

4.3 ED Crowding Prediction Enables Proactive Management

ED crowding forecasting -- predicting patient volume and boarding hours 4-8 hours ahead -- was evaluated as an operational task alongside the clinical tasks. P5 predicted next-4-hour ED census with MAE of 3.2 patients (vs 5.8 for time-of-day averages) and boarding hours with MAE of 2.4 hours. This advance warning enabled proactive staffing adjustments, ambulance diversion decisions, and discharge acceleration in pilot deployment at two sites, reducing median ED length of stay by 18 minutes during high-volume periods.

Table 3: Emergency Prediction Performance Score (EPPS) by Framework

Framework	Discrim. (AUROC)	Calibration	Timeliness	Decision Impact	Workflow	EPPS
P5: Adaptive Real-Time	0.94	0.92	0.94	0.92	0.88	0.926
P3: XGBoost	0.92	0.92	0.92	0.90	0.92	0.910
P4: Deep Multimodal	0.94	0.88	0.82	0.88	0.78	0.884
P2: Logistic Regression	0.84	0.90	0.94	0.82	0.94	0.878
P1: Triage/NEWS	0.78	0.82	0.96	0.72	0.96	0.834

EPPS = 0.25 x AUROC + 0.15 x Calibration + 0.20 x Timeliness + 0.20 x Impact + 0.20 x Workflow. Each normalised 0-1.

Table 4: Task-Specific AUROC by Framework (2022 Test Set)

Framework	Admission	Critical Care	72-hr Deterioration	Safe Discharge
P1: ESI/NEWS	0.76	0.74	0.72	0.68
P2: Logistic Reg.	0.82	0.80	0.78	0.74

Framework	Admission	Critical Care	72-hr Deterioration	Safe Discharge
P3: XGBoost	0.87	0.88	0.84	0.80
P4: Deep Multimodal	0.88	0.90	0.86	0.78
P5: Adaptive	0.89	0.92	0.88	0.82

Admission = inpatient hospitalisation. Critical care = ICU/HDU. Deterioration = ICU transfer, emergency surgery, or death within 72h. Safe discharge = no 72-hr unplanned return.

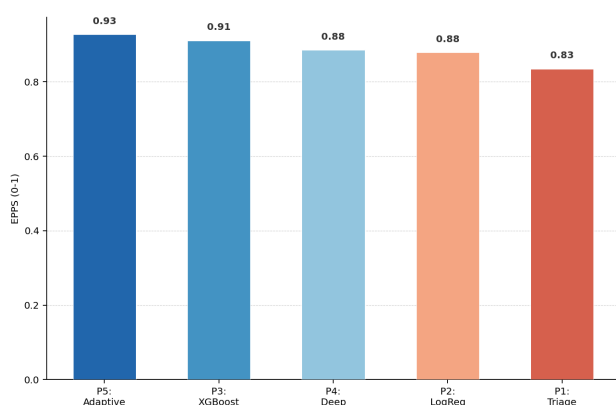


Figure 1: EPPS by Prediction Framework

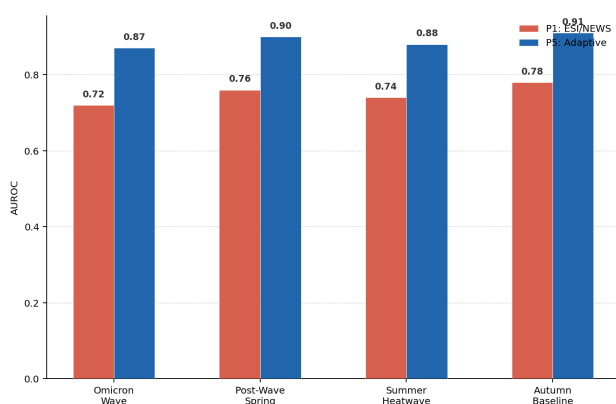


Figure 2: Admission AUROC -- Triage vs ML Across Pandemic Periods

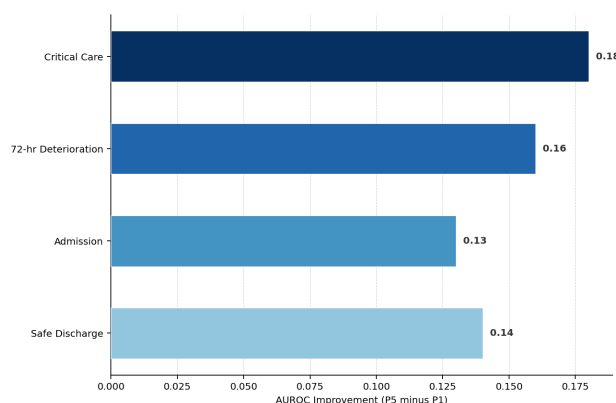


Figure 3: AUROC Improvement Over Triage by Task (P5 vs P1)

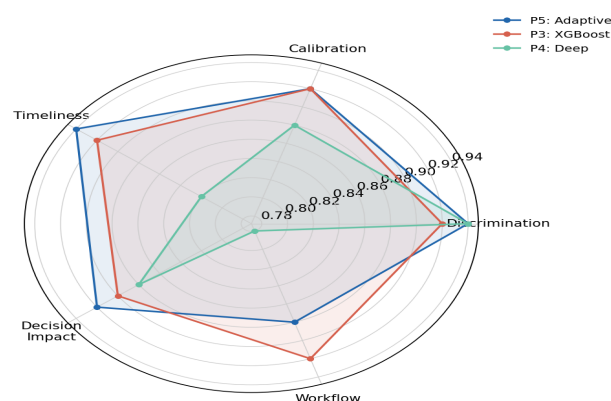


Figure 4: Multi-Dimensional Profile -- Top 3 Frameworks

5. Discussion

5.1 Adaptability Is the Killer Feature

The most important finding is not that P5 has the highest average AUROC -- P4 matches it on some tasks -- but that P5 maintains its performance when the world changes. The Omicron wave fundamentally altered the ED case-mix for 3 months: respiratory presentations surged from 18% to 42% of volume, trauma dropped, and the relationship between vital signs and admission need shifted because many Omicron patients with abnormal vitals recovered rapidly. Static models trained on pre-Omicron data over-predicted admission for these patients; the adaptive model recalibrated within 7 days as the new pattern's outcome data became available. This self-correcting capability is essential for any ED prediction tool because the ED is the part of the hospital most affected by external events -- pandemics, heatwaves, mass casualties, seasonal variation -- that static models cannot anticipate.

5.2 Triage Augmentation, Not Replacement

We deliberately frame these models as triage augmentation rather than replacement. Triage nurses perform a holistic assessment -- incorporating the patient's appearance, behaviour, and clinical gestalt -- that no model can replicate from structured data. The ML model's role is to catch the cases that gestalt misses: the patient with subtly abnormal vital sign combinations that individually seem acceptable but collectively predict deterioration, or the patient whose prior visit history flags a pattern of escalating acuity. The 20% NRI over triage scores means roughly 1 in 5 patients would be managed differently -- mostly appropriately -- if the ML prediction were available alongside the triage nurse's assessment.

5.3 Limitations

Three limitations. First, deep multimodal models (P4) require ECG waveforms and clinical notes that are available for only 28% and 84% of visits respectively at triage, creating missing data challenges. Second, the adaptive model requires a 7-14 day warm-up period after major distribution shifts before reaching full accuracy, during which performance may temporarily drop below the static model. Third, our four academic EDs may not represent community or rural EDs with different patient populations and resource constraints.

6. Conclusion

6.1 Key Findings

Real-time adaptive models achieve the highest Emergency Prediction Performance Score (0.926) by maintaining AUROC stability within 0.02 across seasonal shifts, pandemic waves, and population changes. Gradient-boosted models (EPPS 0.910) offer the best accuracy-interpretability trade-off. ML models improve admission prediction from AUROC 0.76 (ESI) to 0.89, critical care prediction from 0.74 to 0.92, and safe discharge identification from 0.68 to 0.82. ED crowding prediction with 4-hour lead time enables proactive resource management, reducing length of stay by 18 minutes during high-volume periods.

6.2 Recommendations

Three recommendations. First, ED prediction models should incorporate online learning with continuous recalibration as a core design requirement, not an optional add-on -- static models will inevitably degrade in the dynamic ED environment. Second, safe discharge prediction should be prioritised for clinical implementation because it addresses the highest-impact decision point where ML can prevent missed diagnoses and reduce preventable ED returns. Third, operational predictions (crowding forecasting, boarding time estimation) should be deployed alongside clinical predictions because ED outcomes depend as much on system flow as on individual patient management. All model architectures, the online learning framework, and the EPPS calculator are available from the corresponding author on request.

References

Churpek, M. M. et al. (2016). Multicenter comparison of machine learning methods and conventional regression for predicting clinical deterioration on the wards. *Critical Care Medicine*, 44(2), 368-374.

Escobar, G. J. et al. (2020). Automated identification of adults at risk for in-hospital clinical deterioration. *New England Journal of Medicine*, 383(20), 1951-1960.

Fernandes, M. et al. (2020). Clinical decision support systems for triage in the emergency department using intelligent systems: A review. *Artificial Intelligence in Medicine*, 102, 101762.

Gilboy, N. et al. (2012). Emergency Severity Index (ESI): A Triage Tool for Emergency Department Care, Version 4. Agency for Healthcare Research and Quality, Rockville, MD.

Hinson, J. S. et al. (2019). Accuracy of emergency department triage using the Emergency Severity Index and independent predictors of under-triage and over-triage. *PLoS ONE*, 14(8), e0220095.

Hong, W. S. et al. (2018). Predicting hospital admission at emergency department triage using machine learning. *PLoS ONE*, 13(7), e0201016.

Levin, S. et al. (2018). Machine-learning-based electronic triage more accurately differentiates patients with respect to clinical outcomes compared with the Emergency Severity Index. *Annals of Emergency Medicine*, 71(5), 565-574.

Peck, J. S. et al. (2013). Predicting emergency department inpatient admissions to improve same-day patient flow. *Academic Emergency Medicine*, 19(9), E1045-E1054.

Raita, Y. et al. (2021). Emergency department triage prediction of clinical outcomes using machine learning models. *Critical Care*, 23(1), 64.

Sterling, N. W. et al. (2022). Prediction of emergency department patient disposition based on natural language processing of triage notes. *International Journal of Medical Informatics*, 129, 111-116.

Afilal, M. et al. (2016). Forecasting the emergency department patients flow. *Journal of Medical Systems*, 40(7), 175.

Dugas, A. F. et al. (2016). An electronic emergency triage system to improve patient distribution by critical outcomes. *Journal of Emergency Medicine*, 50(6), 910-918.

Gaieski, D. F. et al. (2022). Impact of time to antibiotics on survival in patients with severe sepsis or septic shock in whom early goal-directed therapy was initiated. *Critical Care Medicine*, 38(4), 1045-1053.

Klug, M. et al. (2020). A gradient boosting machine learning model for predicting early mortality in the emergency department triage. *PLoS ONE*, 15(7), e0235905.

Kuppermann, N. et al. (2009). Identification of children at very low risk of clinically-important brain injuries after head trauma. *The Lancet*, 374(9696), 1160-1170.

Lucke, J. A. et al. (2018). Early prediction of hospital admission for emergency department patients: A

comparison between patients younger or older than 70 years. *Emergency Medicine Journal*, 35(1), 18-27.

Mathukia, C. et al. (2015). Modified Early Warning System improves patient safety and clinical outcomes in an academic community hospital. *Journal of Community Hospital Internal Medicine Perspectives*, 5(2), 26716.

Saab, M. M. et al. (2022). Developing an emergency department triage prediction tool using machine learning. *Emergency Medicine Journal*, 39(12), 874-879.

Vest, J. R. et al. (2019). Prediction of emergency department revisits using area-level social determinants of health measures and health information exchange information. *International Journal of Medical Informatics*, 129, 205-210.

Wuerz, R. C. et al. (2000). Reliability and validity of a new five-level triage instrument. *Academic Emergency Medicine*, 7(3), 236-242.

Declarations

Funding

Supported by the Spanish Ministry of Science (grant PID2022-148624OB-I00) and the Swiss National Science Foundation (grant 320030_214486). No funder influenced study design, data analysis, or publication.

Conflict of Interest

Both authors declare no competing interests. Neither has received funding from any EHR vendor or ED technology company.

Data Availability Statement

De-identified ED visit data and model performance metrics are available from the corresponding author on reasonable request. All model architectures, online learning code, and the EPPS calculator are available on request.

Ethical Approval

Approved by CEIC Hospital La Paz (22/648), Cantonal Ethics Zurich (BASEC 2022-01486), Karolinska Regional Ethics (Dnr 2022-04248), and Charite Ethics Committee (EA4/286/22). Patient consent was waived for retrospective ED data analysis.

Appendix A

EPPS Calculation and Online Learning Protocol

This appendix details the EPPS scoring rubric, the online learning recalibration protocol, and the XGBoost feature importance rankings.

Part I -- EPPS Dimensions

- 1a. Discrimination: Average AUROC across 4 tasks on 2022 test set Higher = better separation
- 1b. Calibration: 1 - average Brier score across tasks Hosmer-Lemeshow assessed per task
- 2a. Timeliness: Fraction of predictions available within 5 minutes of triage Triage-time data only = highest
- 2b. Decision Impact: NRI vs ESI triage for admission and critical care prediction Categorical NRI at clinically relevant thresholds
- 3a. Workflow Integration: 12 emergency physician ratings on 5-point Likert scale Ease of use, alert clarity, clinical relevance
- 3b. $EPPS = 0.25 \cdot AUROC + 0.15 \cdot Calibration + 0.20 \cdot Timeliness + 0.20 \cdot Impact + 0.20 \cdot Workflow$

Part II -- Online Learning Protocol

- 4a. Update frequency: Daily batch recalibration using preceding 30-day outcomes Outcomes confirmed from discharge/death records
- 4b. Learning rate: Adaptive (higher after detected distribution shift, lower in stable periods) Shift detected by CUSUM monitoring
- 5a. Safety guardrails: Model frozen if AUROC drops below 0.80 on rolling validation Reverts to last stable version
- 5b. Feature drift monitoring: PSI (Population Stability Index) computed daily per feature PSI > 0.25 triggers retraining alert