

# International Journal of **Health Sciences** (IJHS)

**Early Warning at Scale: Development and Validation of a Predictive  
Model for Patient Deterioration across an 8,000-Staff Regional  
Referral Network**



**CARI**  
**Journals**

## Early Warning at Scale: Development and Validation of a Predictive Model for Patient Deterioration across an 8,000-Staff Regional Referral Network

 Rajesh Babu Kodali,  Sailesh Kotha

<https://orcid.org/0009-0005-4327-2035> <https://orcid.org/0009-0008-4727-7839>

Accepted: 19<sup>th</sup> Sep, 2026, Received in Revised Form: 4<sup>th</sup> Sep, 2026, Published: 14<sup>th</sup> Oct, 2026



### Abstract

**Purpose:** This paper develops a conceptual framework and validation protocol for a machine learning based early warning system to detect patient deterioration across a regional referral network of approximately 8,000 staff.

**Methodology:** The study is conceptual and methodological. Drawing on the early warning and machine learning literature, it specifies data sources, feature engineering, three candidate models, and a tiered alerting workflow, together with a TRIPOD+AI and PROBAST aligned protocol covering retrospective derivation, silent-mode testing, staged implementation, fairness auditing, and implementation evaluation using the Consolidated Framework for Implementation Research and RE-AIM. The framework is scoped to a network of the scale represented by Vandalia Health and its CAMC Health System campuses in West Virginia.

**Findings:** Rule-based scores such as NEWS2 offer limited sensitivity and specificity relative to machine learning tools, but most evidence comes from large academic centers and integrated systems. The framework shows how such a system can be sized for a multi-campus referral network, routing escalating risk to nurses, rapid response teams, and physicians while limiting alarm fatigue. No empirical patient-level outcomes are reported.

**Unique Contribution to Theory, Policy and Practice:** For theory, the paper integrates technical model design, fairness evaluation, and implementation science into a single development framework. For practice, it provides a literature-grounded blueprint and evaluation protocol that mid-sized referral networks can use to plan deployment. For policy, it supports pre-specified, transparent validation and subgroup performance auditing before clinical prediction models are deployed.

**Keywords:** *Patient Deterioration, Early Warning Score, Machine Learning, Clinical Decision Support, Regional Referral Network, Implementation Science*

**JEL Codes:** *I10, I18, C45, C53*

## 1. Introduction

Hospitalized patients who deteriorate on general wards rarely do so without warning; abnormal vital signs are often documented hours before a cardiac arrest or unplanned intensive care transfer. The challenge is detecting a meaningful signal within routine observations, particularly in large systems where staff work across many units and shifts. Missed signs lead to failure to rescue; Ghaferi et al. showed that high- and low-mortality hospitals differed not in complication rates but in how reliably they recognized and responded to complications [12].

Aggregate weighted scores such as the Modified Early Warning Score [1] and NEWS2 [2] standardized escalation, but their reliance on a few manually entered vital signs and fixed thresholds limits accuracy and contributes to alarm fatigue [16]. Machine learning tools have since emerged, including eCART [3], [4], [5], the Rothman Index [9], [10], sepsis early warning systems [6], [7], and Kaiser Permanente's Advance Alert Monitor [8].

Most of this research originates from large academic centers or integrated delivery systems with mature data science infrastructure. Mid-sized regional referral networks, often the main source of tertiary care for rural catchment areas, face heterogeneous legacy systems, workforce shortages, and limited capacity to build and govern a model to current methodological standards [23], [24].

This paper addresses that gap. It develops a conceptual framework, comprising data architecture, feature engineering, candidate models, and a tiered alerting workflow, for a regional referral network of approximately 8,000 staff, a profile represented by Vandalia Health and its CAMC Health System campuses in southern West Virginia, and specifies a validation protocol built around TRIPOD+AI [23], PROBAST [24], the Consolidated Framework for Implementation Research [25], and RE-AIM [26]. The paper is explicitly conceptual; it reports no results from a deployed system, and all illustrative numbers are hypothetical.

## 2. Background and Literature Review

### 2.1 Evolution of Clinical Early Warning Systems

Track-and-trigger systems assign points to abnormal vital signs and specify a fixed escalation action once a threshold is crossed. Subbe et al.'s Modified Early Warning Score used five routine variables, with higher scores associated with intensive care admission and death [1]. NEWS2 added supplemental oxygen use and a pathway for hypercapnic respiratory failure, and is now the National Health Service standard and a common benchmark [2]. These scores are simple but cannot capture trajectories, laboratory trends, or interactions among variables, and they generate many low-value alerts that contribute to alarm fatigue [16].

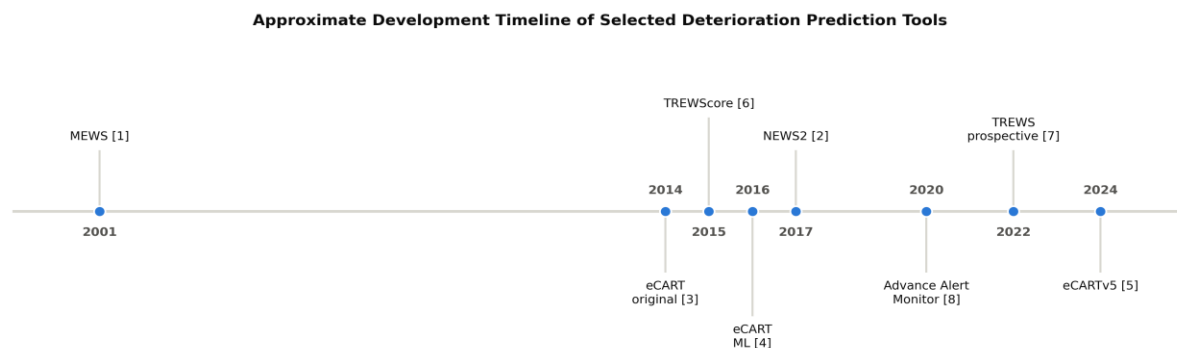


*Figure 1. Approximate timeline of major clinical early warning and deterioration prediction tools discussed in this review.*

## 2.2 Machine Learning Approaches to Deterioration Prediction

Second-generation tools use statistical and machine learning models trained on electronic health record data. Churpek et al.'s eCART used logistic regression over vital sign, laboratory, and demographic predictors and outperformed threshold-based scores [3]; random forests gave further gains [4], and eCARTv5 uses a prospectively validated gradient boosted model [5], [21], [22].

The Rothman Index aggregates nursing assessments, vital signs, and laboratory results into a continuous acuity score associated with deterioration, rapid response activation [10], unplanned intensive care readmission, and mortality [9].

For sepsis [11], the Targeted Real-Time Early Warning Score of Henry et al. [6] was evaluated prospectively by Adams et al., who found better outcomes when clinicians confirmed and responded promptly to alerts, showing that benefit depends on workflow as much as on the algorithm [7].

At system scale, Escobar et al. described Kaiser Permanente's Advance Alert Monitor, a deterioration index embedded in a structured escalation workflow across twenty-one hospitals using a stepped-wedge design, associated with improved recognition and favorable mortality outcomes. It demonstrates feasibility at a multi-facility scale broadly comparable to the network considered here [8].

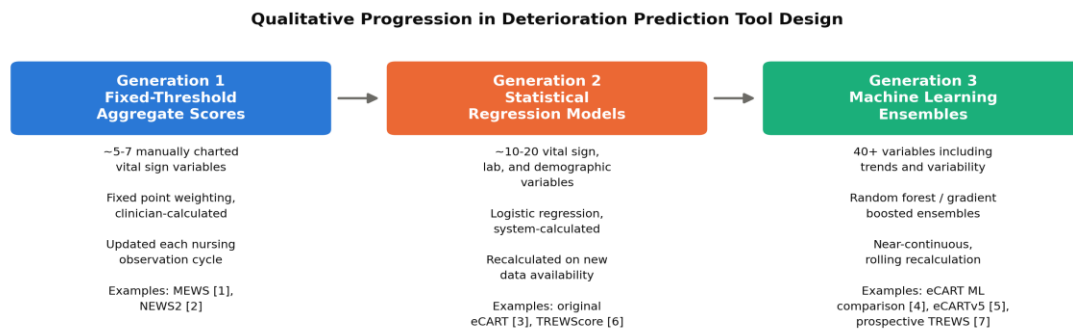


*Figure 2. Conceptual illustration of the generational progression in deterioration prediction tool design, from fixed-threshold scores through regression models to machine learning ensembles. Categories are qualitative and schematic; estimates across the cited studies are not directly comparable.*

## 2.3 Methodological, Fairness, and Implementation Considerations

TRIPOD+AI standardizes reporting of clinical prediction models, including machine learning methods [23], and PROBAST appraises risk of bias and applicability across participants, predictors, outcome, and analysis [24].

Algorithmic fairness is equally urgent. Obermeyer et al. showed that an algorithm using healthcare cost as a proxy for need underestimated the needs of Black patients relative to equally sick White patients [19]. Rudin argues for interpretable models in high stakes settings [20], and Topol and Bates et al. emphasize that benefit depends on workflow integration, clinician trust, and organizational readiness [17], [18].

The Consolidated Framework for Implementation Research organizes implementation determinants across the innovation, setting, individuals, and process [25], and RE-AIM evaluates Reach, Effectiveness, Adoption, Implementation, and Maintenance [26].

## 2.4 The Rural and Mid-Sized Regional Referral Network Gap

Most validation studies come from well-resourced centers such as Kaiser Permanente and the University of Chicago. Little guidance addresses mid-sized regional referral networks, which typically lack dedicated analytics teams, face workforce shortages, and operate heterogeneous legacy systems. This framework is scoped to these constraints.

*Table 1. Selected clinical early warning and deterioration prediction tools discussed in this review.*

| Tool / Study                                                            | Year | Modeling Approach                                                             | Care Setting                            | Reported Contribution                                                                                       |
|-------------------------------------------------------------------------|------|-------------------------------------------------------------------------------|-----------------------------------------|-------------------------------------------------------------------------------------------------------------|
| Modified Early Warning Score (Subbe et al.) [1]                         | 2001 | Aggregate weighted vital sign score                                           | Acute medical admissions                | Established feasibility of a simple bedside track-and-trigger score associated with ICU admission and death |
| National Early Warning Score 2, NEWS2 (Royal College of Physicians) [2] | 2017 | Aggregate weighted vital sign score with oxygen and hypercapnia adjustment    | General acute wards, NHS-wide           | Standardized acute illness severity assessment; adopted as UK national standard                             |
| eCART, original (Churpek et al.) [3]                                    | 2014 | Logistic regression over expanded vital sign, lab, and demographic predictors | Multicenter general wards               | Improved discrimination over threshold-based scores for cardiac arrest, ICU transfer, and death             |
| eCART, ML comparison (Churpek et al.) [4]                               | 2016 | Random forest vs. logistic regression comparison                              | Multicenter general wards               | Machine learning methods provided further discrimination gains over regression                              |
| eCARTv5 (Snyder et al.) [5]                                             | 2024 | Gradient boosted machine learning                                             | Multicenter , prospective               | Prospectively validated gradient boosted early warning score across multiple hospitals                      |
| Rothman Index variability (Wengerter et al.) [10]                       | 2018 | Proprietary composite index acuity                                            | Single-center ward                      | Score variability over time associated with deterioration and rapid response activation                     |
| Rothman Index readmission study (Stahel et al.) [9]                     | 2024 | Proprietary composite index acuity                                            | Propensity-matched cohort               | Index-based risk stratification associated with unplanned ICU readmission and longer stay                   |
| TREWScore (Henry et al.) [6]                                            | 2015 | Statistical early warning model for septic shock                              | Single-center proof of concept          | Demonstrated feasibility of targeted, condition-specific early warning modeling                             |
| TREWS (Adams et al.) [7]                                                | 2022 | Machine learning early warning system for sepsis                              | Prospective , multi-site implementation | Alert confirmation and timely clinician response associated with improved sepsis outcomes                   |

| Tool / Study                               | Year | Modeling Approach                        | Care Setting                                          | Reported Contribution                                                                     |
|--------------------------------------------|------|------------------------------------------|-------------------------------------------------------|-------------------------------------------------------------------------------------------|
| Advance Alert Monitor (Escobar et al.) [8] | 2020 | Continuously updated deterioration index | 21-hospital integrated delivery system, stepped-wedge | System-wide, structured deployment demonstrated feasibility at large multi-hospital scale |

### 3. Conceptual Framework and Proposed Model Architecture

#### 3.1 Institutional Context and Design Motivation

The framework is scoped to a nonprofit, academically affiliated referral network of approximately 8,000 staff, represented by Vandalia Health and its CAMC Health System campuses, a more than 1,100-bed referral and teaching center in southern West Virginia. Such networks combine a flagship tertiary hospital with smaller community campuses, so a network-wide system must accommodate site-level variation in data, staffing, and rapid response structures while producing a standardized risk output. No real patient data are used; the archetype provides a realistic scale for reasoning about design decisions.

#### 3.2 Data Sources and Integration Architecture

The architecture integrates four data categories (Table 2): vital signs at full monitor and documentation frequency [1], [2]; laboratory results as values and trends, as in eCART [3], [4], [5]; structured nursing assessments, reflecting the information underlying the Rothman Index [9], [10]; and static context such as diagnosis, comorbidity, and code status. Given the network's Cerner Millennium platform, data flow through Cerner Command Language and HL7 FHIR pathways to a central repository, and scores return to the clinical decision support layer for bedside display.

#### 3.3 Feature Engineering

For each vital sign, features include the most recent value, the rate of change over four and twelve hours, the number of abnormal threshold crossings, and variability over the observation window. This approach is motivated by the association between acuity score variability and deterioration described by Wengerter et al. [10]. Laboratory features include the latest value, recent trend, and time since last measurement. Missingness is treated as an explicit feature, and features are refreshed every fifteen minutes.

#### 3.4 Candidate Modeling Approaches

The framework compares three model families (Table 3): penalized logistic regression as an interpretable baseline [20]; a random forest, which Churpek et al. found superior to regression [4],

[21]; and a gradient boosted ensemble, as in eCARTv5 [5], [22]. All would be trained on the same cohort and compared per Section 4.5, so that selection reflects the observed trade-off between discrimination, calibration, and interpretability [23], [24].

### 3.5 Alerting and Clinical Workflow Integration

A model delivers benefit only when embedded in a clear workflow [7]. The score is therefore coupled to a three-tier workflow (Figure 3). At the lowest tier, a modest elevation increases reassessment frequency and shows a passive dashboard indicator, limiting alarm fatigue [16]. At the intermediate tier, a sustained or rising score notifies the bedside and charge nurse for structured reassessment and physician escalation as needed. At the highest tier, the score triggers the existing rapid response pathway, which reduces non-intensive-care-unit arrest although evidence on mortality is less consistent [13], [14]. Each tier routes to existing structures rather than parallel ones [25].

*Table 2. Proposed data source categories for the network-scale deterioration prediction architecture.*

| Data Category                           | Representative Elements                                                                              | Primary System                                                                           | Source | Proposed Interval                          | Refresh |
|-----------------------------------------|------------------------------------------------------------------------------------------------------|------------------------------------------------------------------------------------------|--------|--------------------------------------------|---------|
| Continuous and intermittent vital signs | Heart rate, respiratory rate, blood pressure, oxygen saturation, temperature, level of consciousness | Bedside monitors and nursing flowsheet documentation within the electronic health record |        | Every 15 minutes, or on new charting event |         |
| Laboratory results                      | Complete blood count, metabolic panel, lactate, inflammatory markers                                 | Laboratory information system, integrated via the electronic health record               |        | On result availability                     |         |
| Structured nursing assessment           | Mobility status, mental status, structured flowsheet items                                           | Nursing documentation module                                                             |        | On new charting event                      |         |
| Static and slowly changing context      | Admitting diagnosis, comorbidity burden, code status, time since admission                           | Admission, discharge, and transfer feed; problem list                                    |        | On admission and on update                 |         |

Table 3. Candidate modeling approaches proposed for structured comparison.

| Candidate Model                | Type                            | Principal Strength                                                                               | Principal Limitation                                                                              | Literature Precedent                                                    |
|--------------------------------|---------------------------------|--------------------------------------------------------------------------------------------------|---------------------------------------------------------------------------------------------------|-------------------------------------------------------------------------|
| Penalized logistic regression  | Interpretable statistical model | Transparent, clinically inspectable coefficients; established regulatory familiarity             | May underfit complex nonlinear interactions among predictors                                      | Original eCART model [3]; general clinical prediction modeling standard |
| Random forest ensemble         | Tree-based ensemble             | Captures nonlinear interactions; robust to outliers and missingness patterns                     | Reduced interpretability relative to regression; larger computational footprint                   | eCART machine learning comparison [4]; Breiman [21]                     |
| Gradient boosted tree ensemble | Sequential tree-based ensemble  | Typically strongest raw discrimination among the three candidates in comparable prediction tasks | Least interpretable of the three candidates; requires careful regularization to avoid overfitting | eCARTv5 [5]; XGBoost framework [22]                                     |

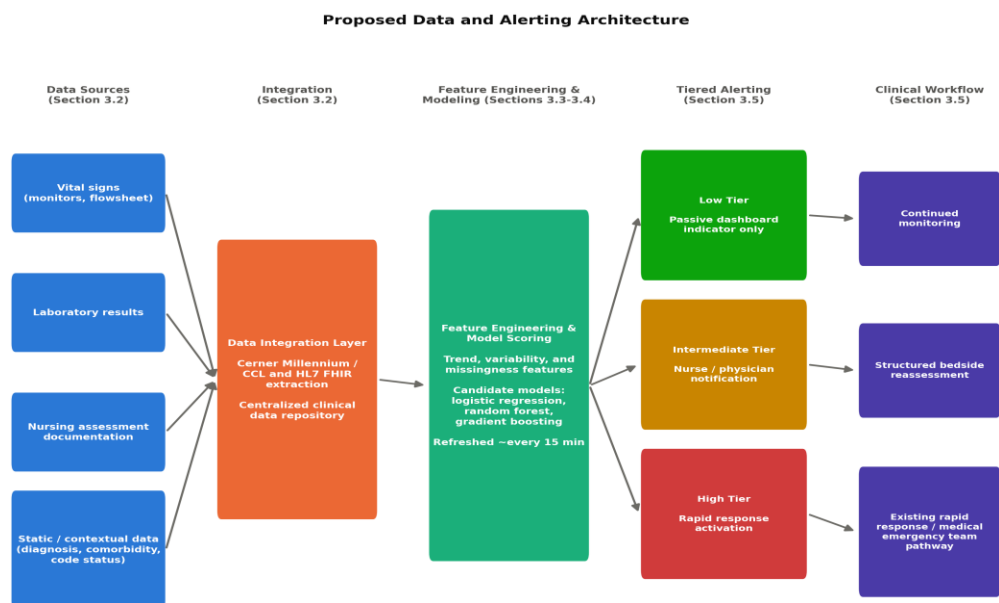


Figure 3. Proposed data and alerting architecture, showing data sources, the centralized feature and modeling layer, and the tiered escalation workflow.

#### **4. Proposed Validation Protocol and Methodology**

This section specifies the protocol required to move the framework from design to an evaluated clinical tool. It is structured around TRIPOD+AI [23] and PROBAST [24]; no results of executing it are presented.

##### **4.1 Study Design**

A four-phase design is proposed: (1) retrospective derivation and internal validation; (2) prospective silent-mode validation, with scores logged but not displayed; (3) staged implementation on pilot units; and (4) network-wide implementation with ongoing performance and equity monitoring, mirroring the Advance Alert Monitor rollout [8] and guidance on prospective validation [24].

##### **4.2 Population and Setting**

The cohort would include adult inpatients on general medical, surgical, and stepdown units across all campuses, excluding direct intensive care admissions, patients under eighteen, and patients with comfort-care-only or do-not-resuscitate-do-not-intubate status before admission. The high share of transferred patients should be described per TRIPOD+AI [23] and considered under PROBAST's applicability domain [24].

##### **4.3 Outcome Definition**

The primary outcome is a composite of unplanned intensive care transfer, rapid response activation, cardiac or respiratory arrest, or death, excluding comfort-care events, following eCART precedent [3], [4], [5]. Secondary outcomes include each component, length of stay, and Sepsis-3 bundle measures [6], [7], [11]. Twelve- and twenty-four-hour horizons would be compared.

##### **4.4 Model Development and Internal Validation Plan**

Development would use a temporally split cohort, with campus-level validation to assess the need for site-specific recalibration [24], and class weighting or resampling applied only within training folds.

##### **4.5 Performance Evaluation Plan**

Discrimination would be assessed by the areas under the receiver operating characteristic and precision-recall curves, calibration by plots, slope, and intercept [23], and clinical utility by decision curve analysis and the number needed to evaluate, an indicator of alert burden [16], all with bootstrapped confidence intervals.

##### **4.6 Fairness and Subgroup Performance Evaluation Plan**

Because an algorithm can encode racial bias without using race as an input [19], subgroup evaluation is mandatory before deployment. Discrimination and calibration would be assessed

across race and ethnicity, sex, language, insurance, and campus, and any prospectively defined meaningful disparity would require revision or recalibration [24].

#### 4.7 Ethical and Regulatory Considerations

Studies would require institutional review board approval or a quality improvement determination. Data governance would follow HIPAA-compliant policies with de-identified data where feasible, and clinician-facing materials would state the model's intended use, limitations, and the retained role of clinical judgment [17], [18], [23].

#### 4.8 Implementation Evaluation Plan

Because benefit depends on how reliably staff act on outputs [7], the Consolidated Framework for Implementation Research would structure assessment of adoption barriers [25], and RE-AIM would structure quantitative evaluation [26], as mapped in Table 4.

*Table 4. Mapping of the RE-AIM implementation evaluation framework to this protocol.*

| RE-AIM Dimension | Definition in This Protocol                                                         | Proposed Measure                                                                                         |
|------------------|-------------------------------------------------------------------------------------|----------------------------------------------------------------------------------------------------------|
| Reach            | Proportion of eligible inpatients and units covered by the alerting workflow        | Count of eligible patient-days scored by the model divided by total eligible patient-days, by campus     |
| Effectiveness    | Impact on the clinical outcomes defined in Section 4.3                              | Composite and component outcome rates, pre-versus post-implementation, adjusted for case mix             |
| Adoption         | Proportion of eligible units and clinical staff actively engaging with the workflow | Proportion of flagged alerts with a documented nursing reassessment within the specified response window |
| Implementation   | Fidelity to the specified three-tier escalation protocol                            | Proportion of highest-tier alerts resulting in rapid response team activation per protocol               |
| Maintenance      | Sustained use and stable performance beyond initial implementation                  | Trend in model calibration and alert response fidelity at 6 and 12 months post-implementation            |

### 5. Illustrative Application Scenario

The following scenario is a hypothetical composite illustration. It does not describe an actual patient or data, and all figures are illustrative.

Consider a hypothetical sixty-eight-year-old patient admitted to a community campus with pneumonia and placed in the lowest tier. Over eighteen hours, the respiratory rate rises from eighteen to twenty-four breaths per minute, oxygen saturation drifts downward, and lactate rises

modestly. None crosses a NEWS2 action threshold, but together they form a trajectory the trend-based features are designed to detect. The score enters the intermediate tier, prompting immediate nurse reassessment (Figure 4).

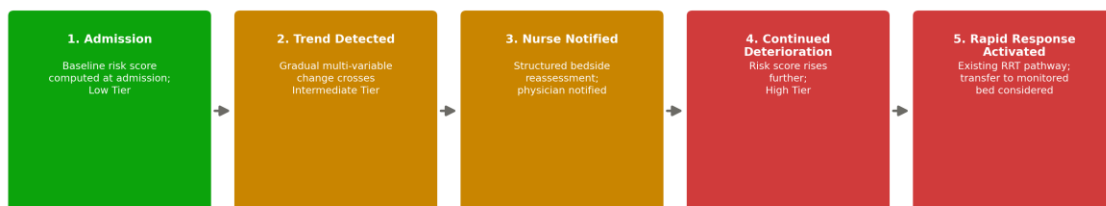
Suppose the physician orders a chest radiograph and cultures and adjusts antibiotics, but the patient continues to worsen. The score then enters the highest tier, activating the rapid response team [13], and transfer occurs hours earlier than under a NEWS2-based protocol, although any real benefit could only be established through Section 4. Early isolated changes produce only a passive indicator, limiting alarm burden [16].

Table 5 presents hypothetical tiers analogous to the NEWS2 color-banded structure [2]. Actual thresholds would be derived from the calibration analysis in Section 4.5 and validated across campuses.

*Table 5. Illustrative, hypothetical alert tier structure (conceptual placeholder values, not empirically derived).*

| Tier         | Illustrative Indicator                                   | Risk                                                                                                          | Clinical Action                                 | Design Rationale                                                 |
|--------------|----------------------------------------------------------|---------------------------------------------------------------------------------------------------------------|-------------------------------------------------|------------------------------------------------------------------|
| Low          | Mildly elevated, non-rising model score                  | Passive                                                                                                       | dashboard indicator only; no interruptive alert | Limit low-value interruptions and mitigate alarm fatigue [16]    |
| Intermediate | Sustained elevation or rapid upward trend in model score | Direct notification to bedside and charge nurse; structured reassessment; physician notification as indicated |                                                 | Prompt early clinical reassessment while trend is still evolving |
| High         | Score consistent with imminent deterioration risk        | Activation of existing rapid response or medical emergency team pathway                                       |                                                 | Route into established escalation infrastructure [13], [14]      |

**Illustrative Hypothetical Escalation Pathway (Section 5)**



*Hypothetical scenario for illustration only; not derived from real patient data.*

*Figure 4. Illustrative escalation pathway for the hypothetical scenario in Section 5, from passive monitoring through nurse notification to rapid response activation.*

## 6. Anticipated Organizational Impact Across an Eight-Thousand-Staff Referral Network

Deploying the framework across a network of this size, with nursing, physician, respiratory therapy, and rapid response staff across seven campuses, raises organizational considerations as demanding as model development. These are discussed in general terms; no implementation outcomes are reported.

Workforce capacity is a central constraint. Aiken et al. linked higher patient-to-nurse ratios to increased mortality, failure to rescue, and burnout [15]. Because alerts increase reassessment, benefit depends on staff capacity; in a network facing rural workforce shortages, alert volume calibration determines sustainability. Similarly, rapid response staffing at smaller campuses must be strengthened before highest-tier alerting is activated [13], [14], since an alert that cannot be answered erodes trust.

Training several thousand staff requires coordinated education on the score, its limitations, and the primacy of clinical judgment [17], [18], concentrated first on pilot units [25]. Centralized data engineering is also needed to maintain the pipeline and monitor drift, consistent with the RE-AIM Maintenance dimension [26]. The principal risk is therefore not statistical inadequacy but that alert volume, workforce capacity, and cross-campus consistency could undermine adoption.

## 7. Discussion

This paper addresses the scarcity of organizationally grounded guidance for deterioration early warning at the scale of a mid-sized referral network. It does not propose a novel algorithm [3], [4], [5], [21], [22]; its contribution is integrating model design, multi-campus data architecture, alarm-fatigue-aware alerting, and a protocol that treats technical performance, fairness, and implementation as equally necessary.

Fairness evaluation warrants emphasis because bias can enter through the outcome label, undetectable by reviewing inputs alone [19]. A rural referral network cannot assume aggregate performance holds across subgroups or campuses; Section 4.6 makes this testable. Including an interpretable model reflects Rudin's caution about opaque models where interpretable ones perform comparably [20].

Treating implementation as coequal with validation follows from the TREWS finding that benefit depended on prompt clinician response [7], [17], [18]. Compared with the Advance Alert Monitor [8], which benefited from an integrated system's data science capacity, this framework is designed for a smaller decision support function using standard interoperability tools.

Finally, the conceptual scoping is a deliberate methodological choice. Prediction model literature has been criticized for inadequate validation and overstated readiness, concerns that motivated TRIPOD+AI and PROBAST [23], [24]. Publishing an appraisable protocol before execution supports pre-specified evaluation and allows critique before resources are committed.

## 8. Limitations

First, this paper presents a framework and protocol rather than results; no model has been trained or deployed. Second, the scenario in Section 5 is hypothetical, the thresholds in Table 5 are placeholders, and Figure 2 is not a harmonized meta-analysis.

Third, the network archetype is drawn from public information and does not imply institutional endorsement; no institution-specific patient data were used. Fourth, the literature review is not exhaustive and is restricted to sources through 2024. Fifth, vendor, cost, and return-on-investment considerations are out of scope.

Finally, fairness safeguards reduce but do not eliminate the risk of uneven performance, so post-deployment monitoring would remain necessary.

## 9. Future Work

The most direct extension is execution of the protocol with approved data, reporting results per TRIPOD+AI [23], and assessment of generalizability to other networks [24]. Further work should refine alert tiers, grounded initially in NEWS2 banding [2] and the TREWS experience [7], through decision curve analysis and clinician feedback; specify drift monitoring and retraining triggers [26]; and conduct a formal cost-effectiveness analysis.

## 10. Conclusion

Machine learning tools have shown potential to improve on aggregate weighted scores, but most evidence comes from large academic centers, leaving a gap for mid-sized regional referral networks serving rural populations. This paper developed a framework spanning data architecture, feature engineering, candidate models, and tiered alerting, with a protocol grounded in TRIPOD+AI, PROBAST, the Consolidated Framework for Implementation Research, and RE-AIM, scoped to a network of approximately 8,000 staff. It reports no clinical outcomes; executing and transparently reporting this protocol is the necessary next step.

## References

- [1] C. P. Subbe, M. Kruger, P. Rutherford, and L. Gemmell, "Validation of a modified Early Warning Score in medical admissions," *QJM*, vol. 94, no. 10, pp. 521-526, 2001.
- [2] Royal College of Physicians, National Early Warning Score (NEWS) 2: Standardising the Assessment of Acute-Illness Severity in the NHS. Updated Report of a Working Party. London, U.K.: RCP, 2017.
- [3] M. M. Churpek, T. C. Yuen, C. Winslow, A. A. Robicsek, D. O. Meltzer, R. D. Gibbons, and D. P. Edelson, "Multicenter development and validation of a risk stratification tool for ward patients," *Am. J. Respir. Crit. Care Med.*, vol. 190, no. 6, pp. 649-655, 2014.

- [4] M. M. Churpek et al., "Multicenter comparison of machine learning methods and conventional regression for predicting clinical deterioration on the wards," *Crit. Care Med.*, vol. 44, no. 2, pp. 368-374, 2016.
- [5] A. Snyder et al., "Multicenter development and prospective validation of eCARTv5: A gradient-boosted machine-learning early warning score," *medRxiv*, Mar. 2024. doi: 10.1101/2024.03.18.24304462.
- [6] K. E. Henry, D. N. Hager, P. J. Pronovost, and S. Saria, "A targeted real-time early warning score (TREWScore) for septic shock," *Sci. Transl. Med.*, vol. 7, no. 299, art. 299ra122, 2015.
- [7] R. Adams et al., "Prospective, multi-site study of patient outcomes after implementation of the TREWS machine learning-based early warning system for sepsis," *Nat. Med.*, vol. 28, no. 7, pp. 1455-1460, 2022.
- [8] G. J. Escobar, V. X. Liu, A. Schuler, B. Lawson, J. D. Greene, and P. Kipnis, "Automated identification of adults at risk for in-hospital clinical deterioration," *N. Engl. J. Med.*, vol. 383, no. 20, pp. 1951-1960, 2020.
- [9] P. F. Stahel, K. W. Belk, S. J. McInnis, K. Holland, R. Nanz, J. Beals, J. Gosnell, O. Ogundele, and K. S. Mastriani, "The Rothman Index predicts unplanned readmissions to intensive care associated with increased mortality and hospital length of stay: a propensity-matched cohort study," *Patient Saf. Surg.*, vol. 18, art. 10, 2024.
- [10] B. C. Wengerter, K. Y. Pei, D. Asuzu, and K. A. Davis, "Rothman Index variability predicts clinical deterioration and rapid response activation," *Am. J. Surg.*, vol. 215, no. 1, pp. 37-41, 2018.
- [11] M. Singer et al., "The third international consensus definitions for sepsis and septic shock (Sepsis-3)," *JAMA*, vol. 315, no. 8, pp. 801-810, 2016.
- [12] A. A. Ghaferi, J. D. Birkmeyer, and J. B. Dimick, "Variation in hospital mortality associated with inpatient surgery," *N. Engl. J. Med.*, vol. 361, no. 14, pp. 1368-1375, 2009.
- [13] P. S. Chan, R. Jain, B. K. Nallamothu, R. A. Berg, and C. Sasson, "Rapid response teams: a systematic review and meta-analysis," *Arch. Intern. Med.*, vol. 170, no. 1, pp. 18-26, 2010.
- [14] B. D. Winters, S. J. Weaver, E. R. Pfoh, T. Yang, J. C. Pham, and S. M. Dy, "Rapid-response systems as a patient safety strategy: a systematic review," *Ann. Intern. Med.*, vol. 158, no. 5, pt. 2, pp. 417-425, 2013.
- [15] L. H. Aiken, S. P. Clarke, D. M. Sloane, J. Sochalski, and J. H. Silber, "Hospital nurse staffing and patient mortality, nurse burnout, and job dissatisfaction," *JAMA*, vol. 288, no. 16, pp. 1987-1993, 2002.
- [16] M. Cvach, "Monitor alarm fatigue: an integrative review," *Biomed. Instrum. Technol.*, vol. 46, no. 4, pp. 268-277, 2012.
- [17] D. W. Bates, S. Saria, L. Ohno-Machado, A. Shah, and G. Escobar, "Big data in health care: using analytics to identify and manage high-risk and high-cost patients," *Health Aff. (Millwood)*, vol. 33, no. 7, pp. 1123-1131, 2014.

- [18] E. J. Topol, "High-performance medicine: the convergence of human and artificial intelligence," *Nat. Med.*, vol. 25, no. 1, pp. 44-56, 2019.
- [19] Z. Obermeyer, B. Powers, C. Vogeli, and S. Mullainathan, "Dissecting racial bias in an algorithm used to manage the health of populations," *Science*, vol. 366, no. 6464, pp. 447-453, 2019.
- [20] C. Rudin, "Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead," *Nat. Mach. Intell.*, vol. 1, pp. 206-215, 2019.
- [21] L. Breiman, "Random forests," *Mach. Learn.*, vol. 45, no. 1, pp. 5-32, 2001.
- [22] T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system," in *Proc. 22nd ACM SIGKDD Int. Conf. Knowledge Discovery and Data Mining*, San Francisco, CA, USA, 2016, pp. 785-794.
- [23] G. S. Collins, K. G. M. Moons, P. Dhiman, R. D. Riley, A. L. Beam, B. Van Calster, M. Ghassemi, X. Liu, J. B. Reitsma, M. van Smeden, A.-L. Boulesteix, J. C. Camaradou, G. Celi, S. Denaxas, A. K. Denniston, B. Glocker, R. M. Golub, H. Harvey, G. Heinze, M. M. Hoffman, A. P. Kengne, E. Lam, N. Lee, E. W. Loder, L. Maier-Hein, B. A. Mateen, M. D. McCradden, L. Oakden-Rayner, J. Ordish, R. Parnell, S. Rose, K. Singh, L. Wynants, and P. Logullo, "TRIPOD+AI statement: updated guidance for reporting clinical prediction models that use regression or machine learning methods," *BMJ*, vol. 385, art. e078378, 2024.
- [24] R. F. Wolff, K. G. M. Moons, R. D. Riley, P. F. Whiting, M. Westwood, G. S. Collins, J. B. Reitsma, J. Kleijnen, and S. Mallett, "PROBAST: a tool to assess the risk of bias and applicability of prediction model studies," *Ann. Intern. Med.*, vol. 170, no. 1, pp. 51-58, 2019.
- [25] L. J. Damschroder, C. M. Reardon, M. A. Opra Widerquist, and J. Lowery, "The updated Consolidated Framework for Implementation Research based on user feedback," *Implement. Sci.*, vol. 17, art. 75, 2022.
- [26] R. E. Glasgow, T. M. Vogt, and S. M. Boles, "Evaluating the public health impact of health promotion interventions: the RE-AIM framework," *Am. J. Public Health*, vol. 89, no. 9, pp. 1322-1327, 1999.



2026 by the Authors. This Article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<http://creativecommons.org/licenses/by/4.0/>)