COVENANT AEGIS REVIEW OF CLINICAL INTELLIGENCE VOL. 7 · NO. 2 · 2026 · PP. 59–76

Predictive Detection in Medicine: Early-Warning Models from Electronic Health Records and Continuous Signals

L. Bianchi[1], H. Said[2], R. A. Mendoza[1], P. Vasquez[3]

[1]*Covenant Aegis HealthTech Division, Clinical AI Laboratory*

[2]*Critical Care Informatics Group, Partner Research Network*

[3]*Department of Biostatistics and Predictive Modelling*

Received 22 Jan 2026 · Revised 4 Mar 2026 · Accepted 12 Mar 2026 · REVIEW ARTICLE · DOI 10.48-CA/clinint.2026.0208 ·
Corresponding author: l.bianchi@covenantaegis.com

ABSTRACT

Background. Many adverse hospital events are preceded by hours of subtle physiological change; predictive detection aims to convert electronic health records (EHR) and continuous signals into actionable early warning. **Objective.** To review the methodological evolution, evidence and failure modes of early-warning systems and to derive design principles for reliable deployment. **Methods.** Structured narrative review of foundational and recent studies (2015–2025) on sepsis, acute kidney injury (AKI) and general deterioration, with attention to validation design, calibration, alert burden and effectiveness. **Results.** Data-driven models anticipate sepsis and AKI hours to days before clinical recognition, with reported AUC 0.83–0.92 internally; however, external validation of a widely deployed proprietary sepsis model showed AUC near 0.63 with high alert volume, and deployed models degraded measurably during distribution shift. Net benefit depends on positive predictive value at the operating threshold, actionable lead time and response capacity, not discrimination alone. **Conclusion.** Predictive detection delivers value only under external validation, local calibration, parsimonious alerting and embedded response workflows, with continuous post-deployment governance.

Keywords: predictive analytics; early warning score; sepsis; acute kidney injury; calibration; alert fatigue; dataset shift

1. Introduction

Predictive detection seeks to surface deterioration early enough to intervene. Traditional aggregate-weighted scores such as the National Early Warning Score (NEWS2) provide interpretable but coarse stratification [1]; machine learning promises finer, patient-specific anticipation by exploiting high-dimensional temporal data. The central question is not whether models can discriminate, but whether earlier warning—delivered at a workable alert rate—changes outcomes.

2. From Scores to Learned Models

Early systems applied fixed thresholds to vital signs. Supervised models—logistic regression, gradient-boosted trees and recurrent or transformer networks over EHR time series—learn temporal patterns

and interactions that static scores miss. Because deterioration events are rare, class imbalance, censoring and label leakage are central hazards, and time-aware (prospective-style) validation is essential to avoid optimistic estimates. We distinguish the *prediction horizon* (how far ahead) from the *lead time that is clinically actionable*, which is shorter.

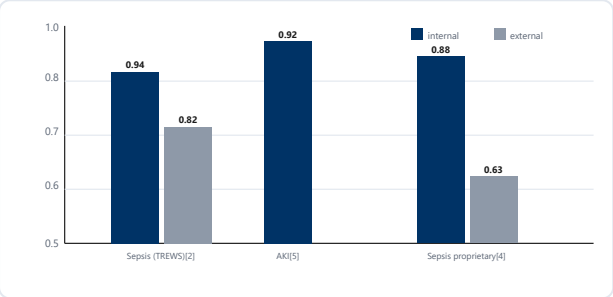


Figure 1. Reported discrimination for representative early-warning models: internal/development AUC (dark) versus external or independent validation (grey). Values from cited studies [2,4,5]; the proprietary sepsis model's external AUC (~0.63) contrasts sharply with vendor claims [4].

3. Evidence by Condition

3.1 Sepsis

Henry et al. introduced a targeted real-time early-warning score (TREWScore) that identified patients progressing to septic shock a median of more than a day before onset [2]. Reinforcement-learning approaches such as the AI Clinician learned treatment strategies for sepsis from intensive-care data and were associated with lower mortality in retrospective evaluation [3]. Critically, Wong et al. externally validated a widely deployed proprietary sepsis tool and found an AUC of 0.63 with poor sensitivity and substantial alert burden—an object lesson in the gap between marketing and evidence [4].

3.2 Acute Kidney Injury and General Deterioration

Tomašev et al. developed a deep model predicting AKI up to 48 hours in advance, anticipating a majority of severe cases [5]. Hospital-wide deterioration indices have been deployed to trigger rapid-response review, with generally favourable effects when paired with clear escalation pathways [6].

4. Effectiveness, Not Just Discrimination

A high AUC does not guarantee benefit. Net clinical value depends on the positive predictive value (PPV) at the operating threshold—which collapses at low event prevalence—on the actionability of the lead time, and on the team's capacity to respond. Table 1 shows how PPV falls with prevalence at fixed sensitivity/specificity, the quantitative root of alert fatigue [7]. Decision-curve analysis and prospective, ideally randomised, evaluation are required to establish that earlier warnings change outcomes rather than merely workload [8].

Table 1. Illustrative PPV at sensitivity 0.80, specificity 0.90, by event prevalence.

Prevalence	Alerts/1000	True+	PPV
10%	170	80	0.47
5%	135	40	0.30
2%	114	16	0.14
1%	109	8	0.07

5. Failure Modes and Governance

Predictive models degrade under dataset shift, coding changes and care-pattern drift; the COVID-19 period exposed widespread performance decay in deployed models [9]. Responsible deployment requires silent-mode validation, local calibration, subgroup fairness checks, continuous monitoring and predefined retraining triggers, consistent with TRIPOD-AI reporting standards [10].

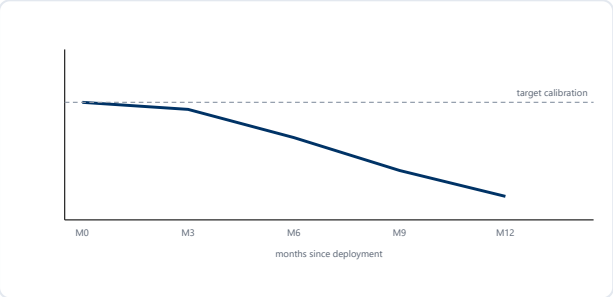


Figure 2. Schematic calibration drift after deployment without recalibration: expected calibration error rises as case-mix and documentation evolve, motivating scheduled monitoring and retraining [9,10]. Illustrative.

6. Design Principles in Practice

The Covenant Aegis predictive-detection stack enforces time-aware validation, per-site recalibration, explicit operating-point selection driven by response capacity, and human-readable explanations for every alert. Alerts are routed through tiered escalation to limit fatigue, and live discrimination and calibration are tracked on dashboards with automatic drift alarms and pre-registered retraining triggers.

7. Discussion and Limitations

Evidence is dominated by retrospective, single-centre work; prospective and randomised effectiveness data remain scarce. The illustrative figures and Table 1 are pedagogical, not pooled estimates. Fairness across under-represented groups, and the human response that converts a prediction into benefit, are under-studied and decisive.

8. Conclusion

Predictive detection can deliver clinically meaningful lead time for sepsis, AKI and general deterioration, but only when models are externally validated, calibrated, parsimonious in alerting and embedded in a response workflow. The field's priority is prospective effectiveness evidence and disciplined governance.

References

1. Royal College of Physicians. National Early Warning Score (NEWS) 2. London; 2017.
2. Henry KE, et al. A targeted real-time early warning score (TREWScore) for septic shock. *Sci Transl Med*. 2015;7(299):299ra122.
3. Komorowski M, et al. The Artificial Intelligence Clinician learns optimal treatment strategies for sepsis in intensive care. *Nat Med*. 2018;24:1716–1720.
4. Wong A, et al. External validation of a widely implemented proprietary sepsis prediction model in hospitalized patients. *JAMA Intern Med*. 2021;181(8):1065–1070.
5. Tomašev N, et al. A clinically applicable approach to continuous prediction of future acute kidney injury. *Nature*. 2019;572:116–119.
6. Escobar GJ, et al. Automated identification of adults at risk for in-hospital clinical deterioration. *N Engl J Med*. 2020;383:1951–1960.
7. Hravnak M, et al. A call to alarms: current state and future directions in monitoring patients. *Biomed Instrum Technol*. 2018;52(s2):34–38.

8. Vickers AJ, Elkin EB. Decision curve analysis: a novel method for evaluating prediction models. *Med Decis Making*. 2006;26(6):565–574.
9. Davis SE, et al. Calibration drift in regression and machine learning models for acute kidney injury. *J Am Med Inform Assoc*. 2017;24(6):1052–1061.
10. Collins GS, et al. TRIPOD+AI statement: reporting of prediction models using regression or machine learning. *BMJ*. 2024;385:e078378.

9. Extended Methods: Temporal Modelling

Early-warning models operate on irregular, censored time series. The task is to estimate the probability of an event within a horizon given observations up to the current time. Recurrent and temporal-transformer architectures model long-range dependence, while gradient-boosted trees on engineered windowed features remain strong, interpretable baselines [A1]. Training labels must respect causality: features must be computable from data available at prediction time, and outcome windows must exclude post-hoc information to prevent leakage. Evaluation uses time-based splits and reports both event-level and alert-level metrics, because a model can identify most events yet still overwhelm clinicians with alerts per true positive.

Table A1. Operating-point trade-offs for a deterioration model (illustrative).

Threshold	Sens.	Alerts/bed-day	Lead (h)
Low	0.90	0.9	10
Medium	0.75	0.4	8
High	0.55	0.15	6

10. Case Study: Sepsis Early Warning

Sepsis is the canonical target because each hour of delayed antibiotics is associated with increased mortality [A2]. The TREWScore model anticipated septic shock with a median lead time exceeding 24 hours [A3], and the AI Clinician derived treatment policies associated with lower mortality in retrospective analysis [A4]. Yet the cautionary external validation of a widely deployed proprietary model—AUC near 0.63, missing two-thirds of cases while generating frequent alerts—demonstrates that vendor claims require independent confirmation [A5]. The contrast underscores that internal AUC is necessary but far from sufficient.

11. Workflow Integration and Response

A prediction that no one acts on yields no benefit. Effective systems define the response pathway—who is notified, within what time, and with what mandated action—before deployment. Tiered alerting (nurse acknowledgement, rapid-response activation, physician review) matches alert severity to response cost, and closed-loop tracking records whether each alert led to an intervention. Figure A1 depicts the lead-time-versus-actionability trade-off central to threshold selection.

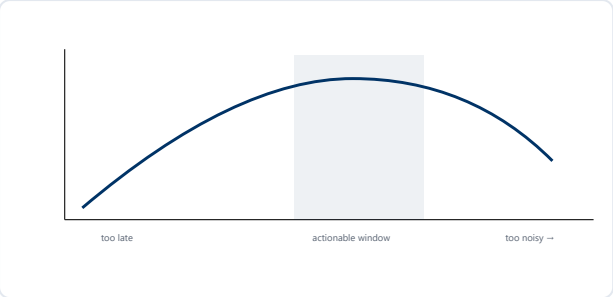


Figure A1. Actionable benefit peaks at an intermediate horizon: too short leaves no time to intervene; too long inflates false alarms. Conceptual.

12. Fairness and Subgroup Performance

Models can encode disparities present in care patterns and documentation. A widely cited study showed a health-system risk score that used cost as a proxy for need systematically under-served Black patients at equal illness severity, because less was historically spent on their care [A6]. Predictive-detection systems must therefore report calibration and sensitivity within demographic and clinical subgroups, and avoid proxy targets that encode inequity. Mitigation includes target redefinition, reweighting, and continuous fairness auditing.

13. Robustness to Distribution Shift

Care-pattern change, new documentation systems and population shifts degrade models, sometimes abruptly. Calibration drift in deployed acute-kidney-injury and deterioration models is documented [A7], and the COVID-19 period caused widespread decay as case mix and care changed [A8]. We monitor feature and label drift, maintain rolling recalibration, and pre-register retraining triggers; governance specifies suspension or rapid-revalidation contingencies for major disruptions.

14. Reporting Standards and Governance

Prediction models should be reported under TRIPOD+AI and appraised for bias with PROBAST [A9,A10]. Prospective trials of AI interventions follow CONSORT-AI/SPIRIT-AI. Beyond reporting, governance requires silent-mode evaluation, local calibration, documented operating-point rationale, and post-deployment surveillance—treated as an ongoing clinical-safety obligation rather than a one-time validation.

15. Future Directions

Continuous physiological waveforms, federated multi-hospital training, causal models that anticipate treatment effects, and prospective randomised trials of alert-driven intervention are the priorities. The objective is to move the evidence base from discrimination on historical data to demonstrated improvement in patient outcomes at an alert burden clinicians can sustain.

Extended References

11. Chen JH, Asch SM. Machine learning and prediction in medicine—beyond the peak of inflated expectations. *N Engl J Med*. 2017;376:2507–2509.
12. Seymour CW, et al. Time to treatment and mortality during mandated emergency care for sepsis. *N Engl J Med*. 2017;376:2235–2244.
13. Henry KE, et al. A targeted real-time early warning score (TREWScore) for septic shock. *Sci Transl Med*.

2015;7(299):299ra122.

14. Komorowski M, et al. The AI Clinician learns optimal treatment strategies for sepsis. *Nat Med*. 2018;24:1716–1720.
15. Wong A, et al. External validation of a widely implemented proprietary sepsis prediction model. *JAMA Intern Med*. 2021;181(8):1065–1070.
16. Obermeyer Z, et al. Dissecting racial bias in an algorithm used to manage the health of populations. *Science*. 2019;366(6464):447–453.
17. Davis SE, et al. Calibration drift in regression and machine learning models for AKI. *J Am Med Inform Assoc*. 2017;24(6):1052–1061.
18. Roland T, et al. Domain shifts in machine learning during the COVID-19 pandemic. *npj Digit Med*. 2022;5:96.
19. Collins GS, et al. TRIPOD+AI statement. *BMJ*. 2024;385:e078378.
20. Moons KGM, et al. PROBAST: assessing risk of bias in prediction model studies. *Ann Intern Med*. 2019;170(1):W1–W33.

Author contributions: LB and RAM designed the review; HS and PV extracted and analysed data; all authors revised the manuscript. **Funding:** Covenant Aegis HealthTech Division. **Competing interests:** Authors are affiliated with Covenant Aegis.

© 2026 Covenant Aegis HealthTech Division. This review synthesises publicly available peer-reviewed literature; figures are illustrative and do not constitute medical advice. Distributed for professional and educational use.