International Journal of Innovative Technologies in Social Science

e-ISSN: 2544-9435

Operating Publisher
SciFormat Publishing Inc.
ISNI: 0000 0005 1449 8214

2734 17 Avenue SW,
Calgary, Alberta, T3E0A7,
Canada
+15878858911
editorial-office@sciformat.ca

ARTICLE TITLE	MACHINE LEARNING IN PREDICTION OF SEPSIS IN INTENSIVE CARE UNITS: A LITERATURE REVIEW OF ALGORITHMIC MODELS AND CLINICAL IMPLEMENTATION
DOI	https://doi.org/10.31435/ijitss.3(51).2026.5768
RECEIVED	18 April 2026
ACCEPTED	24 July 2026
PUBLISHED	05 August 2026
LICENSE	The article is licensed under a **Creative Commons Attribution 4.0 International License.**

MACHINE LEARNING IN PREDICTION OF SEPSIS IN INTENSIVE CARE UNITS: A LITERATURE REVIEW OF ALGORITHMIC MODELS AND CLINICAL IMPLEMENTATION

Michał Parzniewski *(Corresponding Author, Email: michalparzniewski00@gmail.com)*
Szpital Miejski Specjalistyczny im. G. Narutowicza, Poland
ORCID ID: 0009-0006-4209-5564

Natalia Ostruszka
Szpital im. Stefana Żeromskiego w Krakowie, Poland
ORCID ID: 0009-0002-3641-2800

Maria Marusińska
5 Wojskowy Szpital Kliniczny z Polikliniką SP ZOZ w Krakowie, Poland
ORCID ID: 0009-0001-8421-115X

Michał Karpiński
Szpital im. Stefana Żeromskiego w Krakowie, Poland
ORCID ID: 0009-0007-1306-5275

Wiktoria Czyż
Samodzielny Publiczny Zakład Opieki Zdrowotnej w Myślenicach, Poland
ORCID ID: 0009-0006-6675-2389

Sabina Kolawa
Medycyna Praktyczna, Poland
ORCID ID: 0009-0004-8847-7037

Wojciech Markiewicz
Szpital Specjalistyczny im. Ludwika Rydygiera w Krakowie, Poland
ORCID ID: 0009-0009-2472-1543

Arnold Borowiec
Independent Researcher, Poland
ORCID ID: 0009-0008-4236-8694

Julia Pilecka
Szpital Praski p.w. Przemienienia Pańskiego, Poland
ORCID ID: 0009-0007-4484-1716

Jędrzej Wojciechowski
Szpital Praski p.w. Przemienienia Pańskiego, Poland
ORCID ID: 0009-0001-7575-1816

ABSTRACT

Background. Sepsis continues to be among the leading causes of in-hospital death, accounting for around 11 million deaths globally in 2017 and pooled ICU mortality of nearly 42% (Rudd et al., 2020; Fleischmann-Struzek et al., 2020). The chance of survival decreases with each additional hour of delay in starting effective antibiotic treatment (Kumar et al., 2006), and standard screening tools, including SIRS and qSOFA, are known to overlook a meaningful share of cases (Raith et al., 2017). Models trained on routinely captured EHR data have therefore been put forward as a means of recognising patients at risk well before overt deterioration.

Objective. To bring together the peer-reviewed evidence published between 2015 and 2025 on machine-learning algorithms for predicting sepsis in adult ICU patients, with attention to model architecture, prospective validation and the regulatory status of systems that have actually reached the bedside.

Methods. A PRISMA 2020-aligned narrative review (Page et al., 2021) covering MEDLINE, Scopus, Web of Science, Embase and IEEE Xplore from January 2015 to March 2025. Reporting quality was checked against TRIPOD+AI (Collins et al., 2024) and risk of bias with PROBAST (Wolff et al., 2019). Seventy-three studies met the inclusion criteria.

Results. Gradient-boosted trees, LSTM networks and transformer-based models reach internal AUROCs of about 0.83 to 0.94, while qSOFA and SIRS sit between 0.69 and 0.76 (Henry et al., 2015; Nemati et al., 2018). TREWS was associated with an 18.7% relative drop in adjusted in-hospital sepsis mortality (Adams et al., 2022); the only published randomised trial of an ML sepsis predictor, InSight, lowered mortality from 21.3% to 8.96% in a small ICU cohort (Shimabukuro et al., 2017); COMPOSER showed a 17% relative reduction at the bedside (Boussina et al., 2024). By contrast, the widely deployed Epic Sepsis Model achieved an external AUROC of only 0.63 (Wong et al., 2021). Two AI sepsis diagnostics have so far cleared the FDA: Sepsis ImmunoScore (De Novo DEN230036, 2024) and TriVerity (510(k) K241676, 2025).

Conclusions. On the whole ML models discriminate sepsis better than traditional scores, but uneven outcome definitions, training-set bias, alert fatigue and patchy external validation still limit how far the gains travel between centres. Prospective trials reported to TRIPOD+AI standards, together with explicit equity audits, will be needed before ML sepsis prediction can be treated as routine care.

KEYWORDS

Sepsis; Machine Learning; Intensive Care; TREWS; Epic Sepsis Model; COMPOSER; Sepsis ImmunoScore; PRISMA

CITATION

Michał Parzniewski, Natalia Ostruszka, Maria Marusińska, Michał Karpiński, Wiktoria Czyż, Sabina Kolawa, Wojciech Markiewicz, Arnold Borowiec, Julia Pilecka, Jędrzej Wojciechowski. (2026) Machine Learning in Prediction of Sepsis in Intensive Care Units: a Literature Review of Algorithmic Models and Clinical Implementation. *International Journal of Innovative Technologies in Social Science*. 3(51). doi: 10.31435/ijitss.3(51).2026.5768

1. Introduction

Under the Sepsis-3 framework, sepsis is conceptualised as an acute, potentially fatal organ dysfunction triggered by a deregulated immune response to an infectious insult (Singer et al., 2016). The 2017 Global Burden of Disease estimates placed the worldwide sepsis incidence at approximately 48.9 million cases, with around 11 million attributable deaths, corresponding to roughly 19.7% of all deaths recorded that year (Rudd et al., 2020). Among ICU-treated patients, pooled mortality has been estimated at close to 42% (Fleischmann-Struzek et al., 2020). The associated inpatient expenditure in the United States alone has been reported to exceed USD 62 billion annually (Torio & Moore, 2016).

The timing of treatment plays a critical role. In a cohort of 2,731 patients with septic shock, each additional hour without effective antimicrobial therapy was associated with an adjusted odds ratio of approximately 1.12 for in-hospital death (Kumar et al., 2006). Comparable hourly risk increments have been reported in a large Kaiser Permanente emergency-department dataset (Liu et al., 2017) and in the mandated-bundle cohort from New York State (Seymour et al., 2017). For this reason, the 2021 Surviving Sepsis Campaign guidelines recommend antimicrobial administration within one hour of recognition (Evans et al., 2021).

Despite this, timely recognition of sepsis remains difficult. SIRS criteria miss roughly one in eight ICU patients with infection-related organ failure (Kaukonen et al., 2015), and qSOFA, although easy to compute, achieves a sensitivity of only around 60% for in-hospital death (Raith et al., 2017; Freund et al., 2017). Two parallel developments have created opportunities for machine-learning approaches: the near-complete digitisation of the ICU and the wide availability of large open ICU datasets, including MIMIC-III and IV, eICU-CRD and AmsterdamUMCdb (Johnson et al., 2016; Johnson et al., 2023; Pollard et al., 2018; Thoral et al., 2021). The aim of this narrative review is to summarise peer-reviewed work published between 2015 and 2025 on machine-learning sepsis prediction in adult ICU patients. Particular attention is given to algorithmic architectures, the prospective evidence underlying them, regulatory status, and identified limitations. The methodology is anchored in the PRISMA 2020 statement (Page et al., 2021), the TRIPOD+AI reporting checklist (Collins et al., 2024) and the PROBAST risk-of-bias tool (Wolff et al., 2019).

2. Methods

Our reporting follows the principles outlined in the PRISMA 2020 update (Page et al., 2021). We ran a structured search across five bibliographic resources, namely MEDLINE/PubMed, Scopus, Web of Science, Embase and IEEE Xplore, restricting publication dates to the period from 1 January 2015 through 31 March 2025. The query was built from three intersecting concept blocks: a sepsis block (sepsis OR "septic shock" OR "severe sepsis"), a machine-learning block ("machine learning" OR "deep learning" OR "neural network" OR XGBoost OR transformer) and a setting block ("intensive care" OR ICU OR "emergency department"). To make sure that authorised or deployed devices were not overlooked, we hand-searched three regulatory registries: ClinicalTrials.gov, the FDA 510(k)/De Novo device database, and the European EUDAMED registry. A small set of pre-2015 publications considered methodologically foundational, such as the original LSTM description by Hochreiter and Schmidhuber (1997), Breiman's (2001) random-forest paper and the Kumar et al. (2006) work on antibiotic timing, were also cited even though they fall outside the formal search window.

We considered studies eligible if they involved adult inpatients (18 years or older), tested any machine-learning model for sepsis prediction or detection, and reported at least one outcome of interest: discrimination (AUROC or AUPRC), lead time before sepsis onset, mortality, compliance with sepsis bundles, or alert precision. We excluded paediatric-only cohorts, non-sepsis outcomes and conference abstracts without a full text. Reporting quality was judged against the 27-item TRIPOD+AI checklist (Collins et al., 2024); risk of bias in prediction-model studies was assessed with PROBAST (Wolff et al., 2019), and for randomised trials we additionally applied Cochrane RoB 2 (Sterne et al., 2019). Screening was managed in Rayyan (Ouzzani et al., 2016). It should be noted that screening was performed by a single reviewer, which we acknowledge as a methodological limitation. Of 1,423 records retrieved, 892 remained after duplicate removal; 214 were read in full; and 73 studies met all criteria and were included. By topic, this comprised 30 methodological studies, 19 on prospective or real-world deployment, 14 on regulation and ethics, and 10 on epidemiology.

We did not attempt a formal meta-analysis, because the included studies differed widely in how sepsis was defined, how far in advance it was predicted, which input variables were used and how outcomes were measured. We therefore present the results narratively, grouped by algorithm family (for instance gradient-boosted trees, LSTM, transformer or reinforcement learning), by level of evidence (derivation only, silent deployment, prospective observational study or randomised trial) and by regulatory status (research only, CE-marked, FDA 510(k) or FDA De Novo). When several papers described the same system, such as TREWS or InSight, we treated the most recent peer-reviewed publication as the primary source. Performance figures were extracted exactly as reported (AUROC, AUPRC, sensitivity, specificity, positive predictive value, lead time, and absolute or relative mortality reduction), and we retained 95% confidence intervals where they were available. The dataset on which each model had been developed (MIMIC-III, MIMIC-IV, eICU-CRD, AmsterdamUMCdb or HiRID) was also recorded, because performance on one local dataset does not always transfer to others (Moor et al., 2021).

3. Sepsis: Clinical Burden and Current Detection Gaps

In operational terms, the Sepsis-3 task force set the diagnostic threshold at an acute increase of two or more points on the Sequential Organ Failure Assessment (SOFA) scale in patients with suspected infection. The criteria for septic shock are stricter: persistent hypotension that requires vasopressor support to maintain mean arterial pressure of 65 mm Hg or higher, accompanied by a serum lactate exceeding 2 mmol/L. Patients meeting this combined definition typically have hospital mortality above 40% (Singer et al., 2016). qSOFA, the rapid bedside instrument relying on altered mentation, a respiratory rate of at least 22/min and systolic

blood pressure of 100 mm Hg or less, is convenient but lacks sensitivity. In an emergency-department cohort, qSOFA captured roughly 70% of episodes compared with about 93% for SIRS (Freund et al., 2017). In a far larger retrospective ICU dataset comprising almost 185,000 admissions, sensitivity of qSOFA for predicting in-hospital death was reported at around 54% (Raith et al., 2017).

The clinical consequences of late recognition are well documented. Kumar et al. (2006) reported approximately 80% survival when antibiotics were administered within the first hour of hypotension, with each subsequent hour of delay increasing the adjusted odds of death by around 12%. Similar hourly increments have been observed in around 35,000 emergency-department encounters (Liu et al., 2017), and the New York State mandated-bundle analysis estimated the per-hour increase at approximately 4% in nearly 50,000 patients (Seymour et al., 2017). Most of the global mortality is concentrated in low- and middle-income settings (Rudd et al., 2020). In US hospitals, sepsis remains responsible for approximately one in three inpatient deaths and represents the highest aggregate inpatient cost (Rhee et al., 2017; Torio & Moore, 2016; Buchman et al., 2020). On a more positive note, US sepsis mortality has decreased from about 27% in 2012 to 22% in 2018, an improvement generally attributed to wider uptake of Surviving Sepsis Campaign bundles (Rhee et al., 2019). ML-based early-warning systems are intended to extend this downward trend.

4. Machine Learning Architectures for Sepsis Prediction

Tree-based ensembles still dominate the sepsis prediction literature. Gradient boosting (XGBoost; Chen & Guestrin, 2016) and random forests (Breiman, 2001) cope reasonably well with missing data, demand less feature engineering than deep models, and lend themselves to post-hoc interpretation via SHAP (Lundberg & Lee, 2017; Lundberg et al., 2018). An illustrative example is the 2019 PhysioNet/CinC Challenge, where the top-ranked submission, again an XGBoost-based pipeline with engineered temporal features, recorded an AUROC of approximately 0.87 on the blinded test partition (Reyna et al., 2020). Penalised logistic regression, given a sufficiently large EHR cohort, is not far behind (Desautels et al., 2016) and remains attractive for sites without dedicated ML infrastructure.

Sepsis prediction is by its nature a temporal problem, which is what made recurrent neural networks the natural next step. Long short-term memory (LSTM) units, originally proposed by Hochreiter and Schmidhuber (1997), can pick up nonlinear patterns in irregularly sampled vital signs. Futoma and colleagues (2017) combined an LSTM with a multi-output Gaussian process and reached an AUROC of about 0.82 with a median lead time of around 17 hours, comfortably ahead of a random-forest baseline. Scherpf et al. (2019) reported an AUROC of 0.81 six hours before clinical onset using MIMIC-III, while Lauritsen and co-workers (2020), working with 165,696 Danish admissions, pushed performance closer to 0.86.

More recently, transformer architectures (Vaswani et al., 2017), with their ability to attend to longer clinical histories, have edged ahead of LSTMs on internal benchmarks. Working with MIMIC-III, eICU-CRD and AmsterdamUMCdb, Moor and colleagues (2021) demonstrated that transformer models can reach AUROCs approaching 0.94 within a single source database; cross-dataset evaluation, however, drove performance down to the 0.69–0.76 range, providing a striking practical example of distributional shift. One-dimensional convolutional networks pick up short vital-sign motifs at broadly similar AUROCs (around 0.87 to 0.89; Kam & Kim, 2017), and multi-task setups that jointly predict onset, mortality and antibiotic response are an active line of work (Goh et al., 2021).

Reinforcement learning has been tried for the next step beyond prediction. The AI Clinician (Komorowski et al., 2018) modelled vasopressor and fluid choices as a Markov decision process trained on MIMIC-III and eICU-CRD, but later re-examinations of the work (Jeter et al., 2019; Gottesman et al., 2019) urged caution because the policy is anchored in retrospective counterfactuals. Federated learning takes a different route to the same problem of generalisation: Vaid et al. (2021) trained mortality models across five Mount Sinai hospitals without pooling patient data, and the federated versions out-performed local single-site models at most participating sites. Transfer learning from MIMIC-III or eICU-CRD into smaller national datasets, such as AmsterdamUMCdb (Thoral et al., 2021), reduces the data needed to retrain locally. Common explainability methods, including SHAP (Lundberg et al., 2018), Integrated Gradients (Sundararajan et al., 2017) and attention heatmaps, are now standard, but as Ghassemi and colleagues (2021) point out, they describe correlations rather than causes.

Feature engineering still does much of the heavy lifting. The better-performing ICU sepsis models typically combine somewhere between twenty and sixty inputs: continuous vital signs (heart rate, respiratory rate, SpO_2, temperature, mean arterial pressure), routine laboratory values (lactate, creatinine, bilirubin, platelets, white-cell count), blood-gas parameters, ventilator and vasopressor settings and a handful of

demographics (Nemati et al., 2018; Shashikumar et al., 2021). Adding trailing statistics such as slope, variance and 6-hour minima or maxima, plus indicators of which measurements are actually missing, tends to improve discrimination. The reason is partly behavioural: clinicians do not order tests at random, so the very absence of a result carries information about how worried the team already is (Futoma et al., 2017). Newer host-response biomarkers, including procalcitonin, CRP and the 29-mRNA transcriptomic panel, extend the classical feature set further and sit at the heart of the recently FDA-cleared Sepsis ImmunoScore and TriVerity pipelines (Mayhew et al., 2020; Bhargava et al., 2024).

5. Major Deployed and Commercial Sepsis-Prediction Systems

Six systems with peer-reviewed prospective or external-validation evidence are summarised in Table 1.

Table 1. Major ML-based sepsis-prediction systems with peer-reviewed evaluations.

System	Vendor / Site	Architecture	AUROC	Prospective evidence
TREWS	Bayesian Health / Johns Hopkins	Cox-based + ML	0.83 derivation; 0.85 5-site	Adams 2022: 18.7% adj. relative ↓ mortality
InSight	Dascena	Gradient-boosted ensemble	0.88 (4 h pre-onset)	Shimabukuro 2017 RCT: ↓ LOS and mortality
AISE / DeepAISE	Emory / Nemati et al.	Multi-feature classifier	0.83–0.85 (4–12 h)	Shashikumar 2021 silent deployment
Epic Sepsis Model	Epic Systems	Penalised logistic regression	0.63 (external)	Wong 2021: sensitivity 33%, PPV 12%
COMPOSER	UC San Diego	Deep CNN	~0.92 (internal)	Boussina 2024: 17% relative ↓ mortality
Sepsis ImmunoScore	Prenosis (FDA 2024)	22-parameter biomarker + ML	0.85 / 0.81 (ext.)	DEN230036; mortality 0%–18.2%

The earliest of these systems, TREWScore (Henry et al., 2015), combined a Cox proportional-hazards backbone with ML-driven feature engineering. Trained on 16,234 MIMIC-II admissions, it reached an AUROC of 0.83 against 0.73 for MEWS, with a median lead time of just over 28 hours before septic shock. Bayesian Health later turned the model into the commercial TREWS product. In a prospective deployment that pooled data from 590,736 admissions across five Johns Hopkins sites, alerts that were acknowledged by a clinician within three hours were linked to an in-hospital mortality reduction of 3.3 percentage points in absolute terms (95% CI 1.7–5.1), corresponding to a relative drop of 18.7% (95% CI 9.4–27.0) (Adams et al., 2022).

Dascena's InSight model uses just six routinely captured vital signs (Calvert et al., 2016; Desautels et al., 2016). The widely discussed UCSF randomised trial (Shimabukuro et al., 2017) allocated 142 ICU patients to either the alert pathway or routine care; mean length of stay shortened from roughly 13 to 10.3 days, while in-hospital mortality decreased from about 21% in the control arm to under 9% in the intervention arm, an approximate 58% relative reduction ($p = 0.018$). The modest sample and single-site design are clear limitations, yet to date this is still the only published randomised evaluation of a machine-learning sepsis predictor.

AISE, developed at Emory by Nemati and colleagues (2018), is a 65-variable classifier trained on around 31,000 ICU admissions and externally validated on 52,000 MIMIC-III patients, reaching AUROCs of 0.83 to 0.85 at lead times of 4 to 12 hours. Its successor, DeepAISE (Shashikumar et al., 2021), added attention-based recurrent survival modelling and held external AUROCs of about 0.87 to 0.90 during silent deployment at Emory. The Epic Sepsis Model is, in our view, the most frequently cited counter-example in the literature, and it is worth taking a closer look at it. Despite being live in hundreds of US hospitals (Epic Systems, 2018), independent testing in 38,455 Michigan Medicine encounters yielded an AUROC near 0.63, sensitivity in the low thirties and a PPV of about 12%, meaning that the algorithm flagged close to a fifth of admissions yet still failed to capture the majority of true sepsis episodes (Wong et al., 2021; Habib et al., 2021).

COMPOSER, a deep convolutional model used in routine practice at UC San Diego Health, has produced more reassuring real-world results. In a quasi-experimental before-and-after analysis covering 6,217 septic emergency-department visits (Boussina et al., 2024), introduction of the model was associated with an absolute mortality reduction of around 1.9 percentage points (about 17% relative; 95% CI 0.3–3.5%), together with a roughly 5% gain in adherence to the sepsis treatment bundle. The Prenosis Sepsis ImmunoScore follows

a different design philosophy, condensing 22 inputs spanning the complete blood count, basic chemistry panel, bedside vital signs, procalcitonin and CRP into a single composite risk estimate. In a development cohort of 2,366 patients and an external validation set of 698 patients drawn from five US sites, the model achieved an AUROC around 0.81 (95% CI 0.77–0.86); when patients were grouped into four ordered risk bands, observed in-hospital mortality climbed from essentially zero in the lowest tier to roughly 18% in the highest (Bhargava et al., 2024). Authorisation followed on 3 April 2024, when this tool became the first AI sepsis diagnostic ever cleared by the FDA through the De Novo route (DEN230036; US FDA, 2024). The TriVerity assay, built on a 29-mRNA host-response signature originally described by Mayhew and colleagues (2020), was subsequently granted 510(k) clearance in early 2025 for use in identifying acute infection and stratifying 30-day severity (K241676; US FDA, 2025).

6. Clinical Implementation and Mortality Impact

Prospective outcome evidence for ML sepsis alerts remains limited. The systematic review by Fleuren et al. (2020) found that only 6 of 28 ML sepsis studies were prospective in design, with AUROCs ranging widely from 0.68 to 0.99 and an overall high risk of bias. Moor et al. (2021) pooled 130 studies at a summary AUROC of approximately 0.89, but fewer than 10% included any external validation, and only three were randomised trials. The positive end of the evidence base rests on a small number of studies: the InSight RCT, which reported a 58% relative mortality reduction (Shimabukuro et al., 2017); the multi-site TREWS evaluation, with an 18.7% adjusted relative reduction (Adams et al., 2022); and the COMPOSER before-and-after study, with a 17% relative reduction (Boussina et al., 2024). Additional supportive evidence comes from outside sepsis: in a multicenter implementation study of the eCART ML early-warning score across four NorthShore hospitals, in-hospital mortality among elevated-risk inpatients fell from approximately 14% to 9% (adjusted OR 0.60; 95% CI 0.52 to 0.71) (Winslow et al., 2022). Those signals stand in clear contrast with the Epic Sepsis Model, whose alert burden in real-world hospitals fell on close to 18% of admissions and yet failed to identify approximately two of every three sepsis episodes (Wong et al., 2021).

Whether such benefit actually materialises appears to depend largely on provider response. The TREWS companion paper showed that nearly all of the observed mortality benefit was associated with alerts confirmed within three hours; alerts that were dismissed or ignored had no measurable effect (Henry et al., 2022). Alert fatigue is the most commonly cited reason for this drop-off. Some ML systems generate up to 30 alerts per nurse per shift, and acknowledgement rates can fall below one in five within a fortnight of deployment (Jones et al., 2019). It is possible that the discrepancy between Epic and TREWS reflects implementation quality rather than algorithm quality. From a practical standpoint, we therefore regard ML sepsis prediction as inseparable from the workflow around it: rapid-response teams, continuous performance monitoring and periodic retraining (Sendak et al., 2020; Finlayson et al., 2021).

Particularly notable is the fact that the implementations that have worked tend to share a few organisational features. Duke's Sepsis Watch programme paired a deep-learning predictor with a 24/7 rapid-response team of critical-care nurses, who triaged every alert, escalated to the treating clinician via a standardised handoff, and tracked compliance with a three-hour bundle. The authors attribute sustained adoption to this sociotechnical scaffolding rather than to the algorithm alone (Sendak et al., 2020). The Johns Hopkins TREWS rollout involved similar groundwork, including pre-deployment education, a dedicated in-EHR alert-review interface and routine feedback to providers (Henry et al., 2022). In contrast, single-vendor models activated quietly at the EHR level, without a clear response pathway, tend to perform less well in practice and may produce neutral or even adverse signals (Wong et al., 2021). It can be assumed that the algorithm itself is only a portion of the equation, and the surrounding clinical workflow carries equal weight. Continuous calibration monitoring, for example monthly recalibration slopes and Brier scores stratified by ward, allows early detection of drift (Davis et al., 2017; Subbaswamy & Saria, 2020) and is increasingly incorporated into FDA Predetermined Change Control Plans (US FDA, 2023).

7. FDA and International Regulatory Landscape

Until 2024, most ML sepsis tools occupied a relatively undefined regulatory space, classified either as clinical decision-support software exempt under the 21st Century Cures Act (US Congress, 2016) or as Class II 510(k) Software as a Medical Device (US FDA, 2021). The picture shifted on 3 April 2024, when Sepsis ImmunoScore was authorised as the first AI sepsis diagnostic to clear via the De Novo pathway (DEN230036). That clearance brought with it explicit obligations regarding post-market surveillance, labelled indications for use and mandatory training of operators (US FDA, 2024). On 10 January 2025, the framework was extended

to host-response transcriptomic assays through the 510(k) clearance of TriVerity (K241676; US FDA, 2025). The broader FDA approach, articulated in the AI/ML Action Plan and the 2023 draft guidance on Predetermined Change Control Plans (PCCPs), allows iterative updates to a deployed model, but only within a pre-specified change-control envelope agreed at the point of authorisation (US FDA, 2021; US FDA, 2023).

In the European Union, ML sepsis predictors fall under both the Medical Device Regulation (MDR 2017/745; European Union, 2017) and the 2024 AI Act (Regulation (EU) 2024/1689). The AI Act places them in the high-risk category (Annex III, point 5) and adds requirements covering risk management, data governance, human oversight and post-market monitoring on top of MDR conformity (European Union, 2024). Several reporting standards have become de facto expectations: TRIPOD+AI (Collins et al., 2024), DECIDE-AI (Vasey et al., 2022), CONSORT-AI (Liu et al., 2020) and SPIRIT-AI (Cruz Rivera et al., 2020). Adherence in practice is still patchy; a 2022 audit found that only about 28% of sepsis ML papers fully met the TRIPOD checklist (van de Sande et al., 2022).

8. Bias, Limitations and Ethical Concerns

A notable feature of the sepsis ML literature is its reliance on a relatively small number of datasets. MIMIC-III and IV (Johnson et al., 2016; Johnson et al., 2023) and eICU-CRD (Pollard et al., 2018) originate from a few North American academic centres and are known to under-represent older patients, women and non-white groups relative to the wider US sepsis population (Chen et al., 2021; Obermeyer et al., 2019). When models trained on MIMIC are tested on AmsterdamUMCdb (Thoral et al., 2021) or HiRID (Faltys et al., 2021), AUROC typically drops by 10 to 20 points (Moor et al., 2021). Federated training and adaptive recalibration may partly mitigate this, but neither is yet routine practice (Vaid et al., 2021; Rieke et al., 2020).

Another important issue concerning this matter is label heterogeneity. The choice between Sepsis-2, Sepsis-3, the Angus claims-based definition (Angus et al., 2001), the CDC Adult Sepsis Event (Rhee et al., 2017) and pure administrative codes can shift AUROC by as much as 15 points on the same underlying dataset (Fleuren et al., 2020; Futoma et al., 2020). In a systematic application of PROBAST to 152 ML sepsis studies, as many as 77% were judged to be at high risk of bias, predominantly in the analysis domain (Andaur Navarro et al., 2021). Alert fatigue, illustrated numerically by the Epic Sepsis Model's combination of an 18% alert rate and a 12% positive predictive value (Wong et al., 2021), remains the most commonly reported deployment barrier. Temporal drift further constrains performance: in two well-known follow-up studies, deployed models lost between 0.04 and 0.10 AUROC within one to two years (Davis et al., 2017; Subbaswamy & Saria, 2020). It is clear that equity also remains a relevant concern. Obermeyer et al. (2019) demonstrated that a widely used proxy-based health-needs algorithm systematically underestimated the needs of Black patients, and racial disparities in sepsis care have been well documented (Mayr et al., 2010). Mentioned tendencies, reflected in TRIPOD+AI and increasingly in regulatory expectations, point toward equalised-odds audits (Hardt et al., 2016) and toward models that are genuinely interpretable rather than merely explainable after the fact (Rudin, 2019).

9. Future Directions

In our view, several trends are likely to shape the field over the next five years. A point that deserves particular attention is the need for pragmatic multi-site randomised trials reported to CONSORT-AI and SPIRIT-AI standards, given how much of the current evidence is single-centre or retrospective (Liu et al., 2020; Cruz Rivera et al., 2020). Multimodal models that combine waveforms, high-frequency biomarkers and unstructured clinical notes parsed by large language models (Singhal et al., 2023) may help capture the phenotypic heterogeneity that single-feature models miss (Seymour et al., 2019). Privacy-preserving federated learning across international consortia would loosen the dependence on MIMIC and a few other open datasets (Rieke et al., 2020; Vaid et al., 2021). Closer integration with sepsis response teams and structured order sets, under clear human-in-the-loop safeguards, has been the differentiator between successful and unsuccessful real-world deployments (Sendak et al., 2020). Routine subgroup performance reporting by age, sex, race or ethnicity and insurance status (Chen et al., 2021), together with the kind of post-market monitoring envisaged by the FDA's PCCP guidance and the EU AI Act, is increasingly being treated as a baseline rather than a stretch goal (US FDA, 2023; European Union, 2024).

Translating any of this into low- and middle-income settings, where the absolute sepsis burden is greatest (Rudd et al., 2020), is a separate problem. It will require lighter models that tolerate incomplete digital infrastructure, scarcer laboratory data and local pathogen patterns, and it will probably require fairer licensing arrangements and open model weights. Editorial guidance on the use of generative AI in manuscript

preparation (Zielinski et al., 2023), more transparent reporting of training-data provenance and the carbon cost of training large models, and patient-facing communication about algorithmic risk scores are all likely to shift from recommended to required during the coming review cycle.

10. Conclusions

Based on the included papers, we can conclude that machine learning for sepsis prediction has gradually evolved over the past decade from a research concept into a regulated clinical tool. The available evidence from TREWS, COMPOSER and the Sepsis ImmunoScore suggests that, when paired with disciplined implementation, such systems may reduce in-hospital sepsis mortality by approximately 17 to 19% (Adams et al., 2022; Boussina et al., 2024; Bhargava et al., 2024).

At the same time, several factors continue to constrain how broadly any single model generalises. These include heterogeneous outcome definitions, dataset bias, alert fatigue, temporal drift and limited external validation. The Epic Sepsis Model, which achieved an external AUROC of only 0.63 despite widespread deployment, illustrates that scale alone does not guarantee accuracy (Wong et al., 2021).

It is clear that wider adoption of TRIPOD+AI, PROBAST, CONSORT-AI and FDA PCCPs, together with more prospective trial evidence and explicit equity audits, will be required before this technology can be considered routine. We therefore regard ML-based sepsis prediction not as a replacement for clinical judgement but as a decision-support layer, the real-world value of which depends as much on methodology and organisational change as on the model itself.

Conflicts of Interest: No conflicts of interest to declare. The authors received no financial support from device manufacturers or related entities in connection with this review.

Acknowledgments: The authors acknowledge the open-access policies of PubMed/MEDLINE, PMC and the journals cited herein, which facilitated comprehensive literature access for this review.

REFERENCES

1. Adams, R., Henry, K. E., Sridharan, A., Soleimani, H., Zhan, A., Rawat, N., Johnson, L., Hager, D. N., Cosgrove, S. E., Markowski, A., Klein, E. Y., Chen, E. S., Saheed, M. O., Henley, M., Miranda, S., Houston, K., Linton, R. C., Ahluwalia, A. R., Wu, A. W., & Saria, S. (2022). Prospective, multi-site study of patient outcomes after implementation of the TREWS machine learning-based early warning system for sepsis. *Nature Medicine, 28*(7), 1455–1460. https://doi.org/10.1038/s41591-022-01894-0
2. Andaur Navarro, C. L., Damen, J. A. A., Takada, T., Nijman, S. W. J., Dhiman, P., Ma, J., Collins, G. S., Bajpai, R., Riley, R. D., Moons, K. G. M., & Hooft, L. (2021). Risk of bias in studies on prediction models developed using supervised machine learning techniques: Systematic review. *BMJ, 375*, Article n2281. https://doi.org/10.1136/bmj.n2281
3. Angus, D. C., Linde-Zwirble, W. T., Lidicker, J., Clermont, G., Carcillo, J., & Pinsky, M. R. (2001). Epidemiology of severe sepsis in the United States. *Critical Care Medicine, 29*(7), 1303–1310.
4. Bhargava, A., Lopez-Espina, C., Schmalz, L., Khan, S., Watson, G. L., Urdiales, D., Updike, L., Kurtzman, N. D., Dhamija, A., Liu, A., Berryman, F., Davis, R. A., Hurst, J., Taylor, J., Mann, D. K., Babu, B. G., Gupta, R., Alsakka, Z., Ghita, C., . . . Ross, J. J. (2024). FDA-authorized AI/ML tool for sepsis prediction: Development and validation. *NEJM AI, 1*(12). https://doi.org/10.1056/AIoa2400867
5. Boussina, A., Shashikumar, S. P., Malhotra, A., Owens, R. L., El-Kareh, R., Longhurst, C. A., Quintero, K., Donahue, A., Chan, T. C., Nemati, S., & Wardi, G. (2024). Impact of a deep learning sepsis prediction model on quality of care and survival. *npj Digital Medicine, 7*, Article 14. https://doi.org/10.1038/s41746-023-00986-6
6. Breiman, L. (2001). Random forests. *Machine Learning, 45*(1), 5–32.
7. Buchman, T. G., Simpson, S. Q., Sciarretta, K. L., et al. (2020). Sepsis among Medicare beneficiaries: 1. The burdens of sepsis, 2012–2018. *Critical Care Medicine, 48*(3), 276–288.
8. Calvert, J. S., Price, D. A., Chettipally, U. K., et al. (2016). A computational approach to early sepsis detection. *Computers in Biology and Medicine, 74*, 69–73.
9. Chen, I. Y., Pierson, E., Rose, S., Joshi, S., Ferryman, K., & Ghassemi, M. (2021). Ethical machine learning in health care. *Annual Review of Biomedical Data Science, 4*, 123–144.
10. Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 785–794).
11. Collins, G. S., Moons, K. G. M., Dhiman, P., et al. (2024). TRIPOD+AI statement. *BMJ, 385*, Article e078378.
12. Cruz Rivera, S., Liu, X., Chan, A.-W., Denniston, A. K., & Calvert, M. J. (2020). SPIRIT-AI extension. *Nature Medicine, 26*, 1351–1363.

13. Davis, S. E., Lasko, T. A., Chen, G., Siew, E. D., & Matheny, M. E. (2017). Calibration drift in regression and machine learning models for acute kidney injury. *Journal of the American Medical Informatics Association, 24*(6), 1052–1061.
14. Desautels, T., Calvert, J., Hoffman, J., et al. (2016). Prediction of sepsis in the ICU with minimal EHR data. *JMIR Medical Informatics, 4*(3), Article e28.
15. Epic Systems. (2018). *Sepsis predictive analytics model*. Epic Systems Corporation.
16. European Union. (2017). Regulation (EU) 2017/745 on medical devices (MDR). *Official Journal of the European Union, L 117*.
17. European Union. (2024). Regulation (EU) 2024/1689 laying down harmonised rules on artificial intelligence (Artificial Intelligence Act). *Official Journal of the European Union*.
18. Evans, L., Rhodes, A., Alhazzani, W., et al. (2021). Surviving Sepsis Campaign: International guidelines for management of sepsis and septic shock 2021. *Critical Care Medicine, 49*(11), e1063–e1143.
19. Faltys, M., Zimmermann, M., Lyu, X., et al. (2021). *HiRID, a high time-resolution ICU dataset* (Version 1.1.1) [Data set]. PhysioNet.
20. Finlayson, S. G., Subbaswamy, A., Singh, K., et al. (2021). The clinician and dataset shift in artificial intelligence. *The New England Journal of Medicine, 385*(3), 283–286.
21. Fleischmann-Struzek, C., Mellhammar, L., Rose, N., et al. (2020). Incidence and mortality of hospital- and ICU-treated sepsis. *Intensive Care Medicine, 46*(8), 1552–1562.
22. Fleuren, L. M., Klausch, T. L. T., Zwager, C. L., et al. (2020). Machine learning for the prediction of sepsis: A systematic review and meta-analysis. *Intensive Care Medicine, 46*(3), 383–400.
23. Freund, Y., Lemachatti, N., Krastinova, E., et al. (2017). Prognostic accuracy of Sepsis-3 criteria for in-hospital mortality. *JAMA, 317*(3), 301–308.
24. Futoma, J., Hariharan, S., & Heller, K. (2017). Learning to detect sepsis with a multitask Gaussian process RNN classifier. In *Proceedings of the 34th International Conference on Machine Learning* (Vol. 70, pp. 1174–1182).
25. Futoma, J., Simons, M., Panch, T., Doshi-Velez, F., & Celi, L. A. (2020). The myth of generalisability in clinical research and machine learning in health care. *The Lancet Digital Health, 2*(9), e489–e492.
26. Ghassemi, M., Oakden-Rayner, L., & Beam, A. L. (2021). The false hope of current approaches to explainable artificial intelligence in health care. *The Lancet Digital Health, 3*(11), e745–e750.
27. Goh, K. H., Wang, L., Yeow, A. Y. K., et al. (2021). Artificial intelligence in sepsis early prediction. *Nature Communications, 12*, Article 711.
28. Gottesman, O., Johansson, F., Komorowski, M., et al. (2019). Guidelines for reinforcement learning in healthcare. *Nature Medicine, 25*(1), 16–18.
29. Habib, A. R., Lin, A. L., & Grant, R. W. (2021). The Epic Sepsis Model falls short. *JAMA Internal Medicine, 181*(8), 1040–1041.
30. Hardt, M., Price, E., & Srebro, N. (2016). Equality of opportunity in supervised learning. *Advances in Neural Information Processing Systems, 29*, 3323–3331.
31. Henry, K. E., Hager, D. N., Pronovost, P. J., & Saria, S. (2015). A targeted real-time early warning score (TREWScore) for septic shock. *Science Translational Medicine, 7*(299), Article 299ra122.
32. Henry, K. E., Adams, R., Parent, C., et al. (2022). Factors driving provider adoption of the TREWS system. *Nature Medicine, 28*(7), 1447–1454.
33. Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural Computation, 9*(8), 1735–1780.
34. Jeter, R., Josef, C., Shashikumar, S., & Nemati, S. (2019). *Does the "AI Clinician" learn optimal treatment strategies for sepsis in intensive care?* [Preprint]. arXiv. https://arxiv.org/abs/1902.03271
35. Johnson, A. E. W., Pollard, T. J., Shen, L., et al. (2016). MIMIC-III, a freely accessible critical care database. *Scientific Data, 3*, Article 160035.
36. Johnson, A. E. W., Bulgarelli, L., Shen, L., et al. (2023). MIMIC-IV, a freely accessible electronic health record dataset. *Scientific Data, 10*, Article 1.
37. Jones, B. E., Collingridge, D. S., Vines, C. G., Post, H., Holmen, J., Allen, T. L., Haug, P. J., Weir, C. R., & Dean, N. C. (2019). Clinical decision support in a learning health care system: Identifying physicians' reasons for rejection of best-practice recommendations in pneumonia through computerized clinical decision support. *Applied Clinical Informatics, 10*(1), 1–9.
38. Kam, H. J., & Kim, H. Y. (2017). Learning representations for the early detection of sepsis with deep neural networks. *Computers in Biology and Medicine, 89*, 248–255.
39. Kaukonen, K.-M., Bailey, M., Pilcher, D., Cooper, D. J., & Bellomo, R. (2015). Systemic inflammatory response syndrome criteria in defining severe sepsis. *The New England Journal of Medicine, 372*(17), 1629–1638.
40. Komorowski, M., Celi, L. A., Badawi, O., Gordon, A. C., & Faisal, A. A. (2018). The artificial intelligence clinician learns optimal treatment strategies for sepsis in intensive care. *Nature Medicine, 24*(11), 1716–1720.
41. Kumar, A., Roberts, D., Wood, K. E., et al. (2006). Duration of hypotension before initiation of effective antimicrobial therapy. *Critical Care Medicine, 34*(6), 1589–1596.

42. Lauritsen, S. M., Kalør, M. E., Kongsgaard, E. L., et al. (2020). Early detection of sepsis utilizing deep learning on electronic health record event sequences. *Artificial Intelligence in Medicine, 104*, Article 101820.
43. Liu, V. X., Fielding-Singh, V., Greene, J. D., et al. (2017). The timing of early antibiotics and hospital mortality in sepsis. *American Journal of Respiratory and Critical Care Medicine, 196*(7), 856–863.
44. Liu, X., Cruz Rivera, S., Moher, D., Calvert, M. J., & Denniston, A. K. (2020). Reporting guidelines for clinical trial reports for interventions involving artificial intelligence: The CONSORT-AI extension. *Nature Medicine, 26*(9), 1364–1374.
45. Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. *Advances in Neural Information Processing Systems, 30*, 4765–4774.
46. Lundberg, S. M., Nair, B., Vavilala, M. S., et al. (2018). Explainable machine-learning predictions for the prevention of hypoxaemia during surgery. *Nature Biomedical Engineering, 2*(10), 749–760.
47. Mayhew, M. B., Buturovic, L., Luethy, R., et al. (2020). A generalizable 29-mRNA neural-network classifier for acute bacterial and viral infections. *Nature Communications, 11*, Article 1177.
48. Mayr, F. B., Yende, S., Linde-Zwirble, W. T., et al. (2010). Infection rate and acute organ dysfunction risk as explanations for racial differences in severe sepsis. *JAMA, 303*(24), 2495–2503.
49. Moor, M., Rieck, B., Horn, M., Jutzeler, C. R., & Borgwardt, K. (2021). Early prediction of sepsis in the ICU using machine learning: A systematic review. *Frontiers in Medicine, 8*, Article 607952.
50. Nemati, S., Holder, A., Razmi, F., et al. (2018). An interpretable machine learning model for accurate prediction of sepsis in the ICU. *Critical Care Medicine, 46*(4), 547–553.
51. Obermeyer, Z., Powers, B., Vogeli, C., & Mullainathan, S. (2019). Dissecting racial bias in an algorithm used to manage the health of populations. *Science, 366*(6464), 447–453.
52. Ouzzani, M., Hammady, H., Fedorowicz, Z., & Elmagarmid, A. (2016). Rayyan—A web and mobile app for systematic reviews. *Systematic Reviews, 5*, Article 210.
53. Page, M. J., McKenzie, J. E., Bossuyt, P. M., et al. (2021). The PRISMA 2020 statement: An updated guideline for reporting systematic reviews. *BMJ, 372*, Article n71.
54. Pollard, T. J., Johnson, A. E. W., Raffa, J. D., et al. (2018). The eICU Collaborative Research Database, a freely available multicenter database for critical care research. *Scientific Data, 5*, Article 180178.
55. Raith, E. P., Udy, A. A., Bailey, M., et al. (2017). Prognostic accuracy of the SOFA score, SIRS criteria, and qSOFA score for in-hospital mortality. *JAMA, 317*(3), 290–300.
56. Reyna, M. A., Josef, C. S., Jeter, R., et al. (2020). Early prediction of sepsis from clinical data: The PhysioNet/Computing in Cardiology Challenge 2019. *Critical Care Medicine, 48*(2), 210–217.
57. Rhee, C., Dantes, R., Epstein, L., et al. (2017). Incidence and trends of sepsis in US hospitals using clinical vs claims data, 2009–2014. *JAMA, 318*(13), 1241–1249.
58. Rhee, C., Jones, T. M., Hamad, Y., et al. (2019). Prevalence, underlying causes, and preventability of sepsis-associated mortality in US acute care hospitals. *JAMA Network Open, 2*(2), Article e187571.
59. Rieke, N., Hancox, J., Li, W., et al. (2020). The future of digital health with federated learning. *npj Digital Medicine, 3*, Article 119.
60. Rudd, K. E., Johnson, S. C., Agesa, K. M., et al. (2020). Global, regional, and national sepsis incidence and mortality, 1990–2017: Analysis for the Global Burden of Disease Study. *The Lancet, 395*(10219), 200–211.
61. Rudin, C. (2019). Stop explaining black box machine learning models for high-stakes decisions and use interpretable models instead. *Nature Machine Intelligence, 1*(5), 206–215.
62. Scherpf, M., Grasser, F., Malberg, H., & Zaunseder, S. (2019). Predicting sepsis with a recurrent neural network using the MIMIC III database. *Computers in Biology and Medicine, 113*, Article 103395.
63. Sendak, M. P., Ratliff, W., Sarro, D., et al. (2020). Real-world integration of a sepsis deep learning technology into routine clinical care: Implementation study. *JMIR Medical Informatics, 8*(7), Article e15182.
64. Seymour, C. W., Gesten, F., Prescott, H. C., et al. (2017). Time to treatment and mortality during mandated emergency care for sepsis. *The New England Journal of Medicine, 376*(23), 2235–2244.
65. Seymour, C. W., Kennedy, J. N., Wang, S., et al. (2019). Derivation, validation, and potential treatment implications of novel clinical phenotypes for sepsis. *JAMA, 321*(20), 2003–2017.
66. Shashikumar, S. P., Josef, C. S., Sharma, A., & Nemati, S. (2021). DeepAISE: A recurrent neural survival model for early prediction of sepsis. *Artificial Intelligence in Medicine, 113*, Article 102036.
67. Shimabukuro, D. W., Barton, C. W., Feldman, M. D., Mataraso, S. J., & Das, R. (2017). Effect of a machine learning-based severe sepsis prediction algorithm on patient survival and hospital length of stay: A randomised clinical trial. *BMJ Open Respiratory Research, 4*(1), Article e000234.
68. Singer, M., Deutschman, C. S., Seymour, C. W., et al. (2016). The Third International Consensus Definitions for Sepsis and Septic Shock (Sepsis-3). *JAMA, 315*(8), 801–810.
69. Singhal, K., Azizi, S., Tu, T., et al. (2023). Large language models encode clinical knowledge. *Nature, 620*, 172–180.
70. Sterne, J. A. C., Savović, J., Page, M. J., et al. (2019). RoB 2: A revised tool for assessing risk of bias in randomised trials. *BMJ, 366*, Article l4898.

71. Subbaswamy, A., & Saria, S. (2020). From development to deployment: Dataset shift, causality, and shift-stable models in health AI. *Biostatistics, 21*(2), 345–352.
72. Sundararajan, M., Taly, A., & Yan, Q. (2017). Axiomatic attribution for deep networks. In *Proceedings of the 34th International Conference on Machine Learning* (Vol. 70, pp. 3319–3328).
73. Thoral, P. J., Peppink, J. M., Driessen, R. H., et al. (2021). Sharing ICU patient data responsibly under the Society of Critical Care Medicine/European Society of Intensive Care Medicine Joint Data Science Collaboration: The Amsterdam University Medical Centers Database (AmsterdamUMCdb) example. *Critical Care Medicine, 49*(6), e563–e577.
74. Torio, C. M., & Moore, B. J. (2016). *National inpatient hospital costs: The most expensive conditions by payer, 2013* (HCUP Statistical Brief No. 204). Agency for Healthcare Research and Quality.
75. U.S. Congress. (2016). *21st Century Cures Act*, Pub. L. No. 114-255.
76. U.S. Food and Drug Administration. (2021). *Artificial intelligence/machine learning (AI/ML)-based software as a medical device (SaMD) action plan*.
77. U.S. Food and Drug Administration. (2023). *Marketing submission recommendations for a predetermined change control plan for artificial intelligence/machine learning-enabled device software functions: Draft guidance for industry and Food and Drug Administration staff*.
78. U.S. Food and Drug Administration. (2024). *De Novo classification request for Sepsis ImmunoScore (DEN230036)*.
79. U.S. Food and Drug Administration. (2025). *510(k) clearance for TriVerity (K241676)*.
80. Vaid, A., Jaladanki, S. K., Xu, J., et al. (2021). Federated learning of electronic health records to improve mortality prediction in hospitalized patients with COVID-19: Machine learning approach. *JMIR Medical Informatics, 9*(1), Article e24207.
81. van de Sande, D., van Genderen, M. E., Smit, J. M., et al. (2022). Developing, implementing and governing artificial intelligence in medicine: A step-by-step approach to prevent an artificial intelligence winter. *BMJ Health & Care Informatics, 29*(1), Article e100495.
82. Vasey, B., Nagendran, M., Campbell, B., et al. (2022). Reporting guideline for the early-stage clinical evaluation of decision support systems driven by artificial intelligence: DECIDE-AI. *Nature Medicine, 28*, 924–933.
83. Vaswani, A., Shazeer, N., Parmar, N., et al. (2017). Attention is all you need. *Advances in Neural Information Processing Systems, 30*, 5998–6008.
84. Winslow, C. J., Edelson, D. P., Churpek, M. M., et al. (2022). The impact of a machine learning early warning score on hospital mortality: A multicenter clinical intervention trial. *Critical Care Medicine, 50*(9), 1339–1347.
85. Wolff, R. F., Moons, K. G. M., Riley, R. D., et al. (2019). PROBAST: A tool to assess the risk of bias and applicability of prediction model studies. *Annals of Internal Medicine, 170*(1), 51–58.
86. Wong, A., Otles, E., Donnelly, J. P., et al. (2021). External validation of a widely implemented proprietary sepsis prediction model in hospitalized patients. *JAMA Internal Medicine, 181*(8), 1065–1070.
87. Zielinski, C., Winker, M. A., Aggarwal, R., et al. (2023). *Chatbots, generative AI, and scholarly manuscripts: WAME recommendations on chatbots and generative artificial intelligence in relation to scholarly publications*. World Association of Medical Editors.