ORIGINAL RESEARCH Open access

Why Most Sepsis Prediction Models Fail at the Bedside: A Position Paper on the Gap Between AUROC and Clinical Utility

Wei Chen[1*], Li Zhang[1]

Abstract

Over the past five years, sepsis prediction models have reported strong retrospective performance, often exceeding AUROC 0.85–0.90 by leveraging vital signs, laboratory data, and machine learning to predict sepsis earlier than clinical recognition. However, despite these results, bedside adoption remains minimal, and external or prospective validations frequently show substantial performance decline, with clinicians still relying on traditional criteria such as qSOFA and SIRS. This position paper argues that AUROC is an insufficient and potentially misleading metric for clinical deployment, as it reflects retrospective rank discrimination rather than real-world utility, calibration, or actionable impact. High AUROC scores often conceal poor threshold selection, excessive alert burden, and clinically unacceptable alarm fatigue, while retrospective evaluations create an overly optimistic view that fails in real-time settings. We propose shifting evaluation toward clinically meaningful metrics such as net benefit, alert burden per patient-day, and number needed to alert at clinician-defined thresholds, alongside earlier incorporation of workflow requirements. Ultimately, the continued dominance of AUROC-centric evaluation represents a systemic mismatch between model development and clinical reality, limiting sepsis prediction tools from achieving meaningful impact at the bedside.

Keywords Sepsis prediction, AUROC, Clinical utility, Alert fatigue, Implementation science, Position paper

*Correspondence:
Wei Chen
wei.chen@outlook.com
[1] Department of AI-Based Healthcare Systems, School of Public Health, Peking University, Beijing, China

Introduction

Hundreds of sepsis prediction models have been published in the last five years. Many report AUROCs above 0.85, some above 0.90 [1-3]. Yet ask any ICU clinician how many of these models are actually running at their bedside. The answer is near zero. This is not a coincidence. This is a failure of the field's evaluation standards.

This paper argues that AUROC is a fundamentally misleading metric for sepsis prediction models intended for clinical use. Most models fail at the bedside not because of technical inadequacy, but because they are optimized for the wrong objective. We contend that the field must abandon AUROC as a primary metric and instead adopt clinically grounded evaluation frameworks centered on actionable thresholds, alert burden, and prospective utility.

This is not a systematic review. This is not a meta-analysis. This is a position paper—a critical, argument-driven analysis of why the field has failed to deliver bedside value despite technical progress. We draw exclusively on peer-reviewed studies from 2017–2021 to demonstrate that the problem is structural, not anecdotal.

We make four central arguments. First, AUROC measures discrimination, not clinical actionability. Second, high AUROC can conceal unacceptably high false alarm rates that trigger alert fatigue. Third, retrospective evaluation cannot capture the temporal dynamics of real-time clinical

decision-making. Fourth, the threshold for alerting is almost never derived from clinical requirements. These flaws combine to produce models that perform beautifully on paper and fail catastrophically in practice.

Figure 1 illustrates the structural pathway through which AUROC-centered model development systematically translates into bedside failure, and contrasts it with a clinically grounded alternative design paradigm.

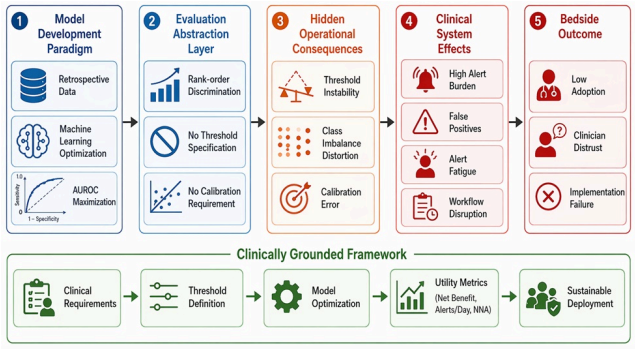

Figure 1. From AUROC Optimization to Bedside Failure: A Hierarchical Framework of Structural Misalignment in Sepsis Prediction Models

The Gap: published Auroc vs Bedside Reality

Documenting the gap

Multiple studies have reported AUROCs between 0.85 and 0.95 for sepsis prediction in retrospective ICU cohorts [1, 3-5]. Nemati and colleagues [1] achieved an AUROC of 0.85 with an interpretable model using only vital signs and laboratory values. Similar claims appear across multicenter validations and deep-learning approaches [2, 6]. These numbers suggest models could detect sepsis hours in advance with near-perfect rank-order accuracy.

Yet external validation and implementation studies tell a different story. Wong *et al.* [7] evaluated a widely implemented proprietary sepsis prediction model and found substantial degradation in real-world performance. Prospective or simulated deployment attempts similarly reveal that high retrospective AUROC does not translate into usable alerts [8, 9]. If these models performed as advertised in the clinical environment, ICUs would be using them. They are not. The gap is not marginal; it is near-total.

The cost of false promises

Overpromising on the basis of AUROC erodes clinician trust. When an ICU team trials a model hyped with an AUROC of 0.92 and discovers that the alerts are either late, irrelevant, or overwhelming, skepticism spreads to every subsequent AI proposal. This phenomenon is evident in the slow uptake of even well-publicized systems [7, 9]. One failed deployment poisons the well for the entire field of clinical AI. Patients ultimately pay the price when clinicians default to familiar but less sensitive manual scores rather than risk another distracting alert.

Why the gap persists

Academic incentives reward high AUROC because it is easy to compute, easy to compare, and easy to publish. Journals in critical care and digital medicine continue to accept papers that headline AUROC without demanding threshold-specific clinical metrics or prospective testing [4, 10]. Peer reviewers rarely ask the decisive question: "Would a clinician actually use this alert at 3 a.m.?" Perverse incentives therefore persist: model developers chase leaderboard numbers while implementation scientists document the resulting graveyard of unused algorithms [3, 8]. Until publication and funding criteria change, the gap will remain.

Why Auroc is a Misleading Metric for Sepsis Prediction

What AUROC actually measures

AUROC quantifies the probability that a randomly selected patient who develops sepsis receives a higher risk score than a randomly selected patient who does not. It is a rank-order statistic, nothing more. It says nothing about whether a risk score of 0.70 is meaningfully different from 0.71, nor whether either value should trigger clinical action [1, 11].

Table 1 provides a structured analytical decomposition demonstrating that AUROC omits every dimension required for evaluating real-world clinical utility.

Table 1. Analytical Decomposition of AUROC Limitations Versus Clinically Actionable Performance Dimensions

Dimension	What AUROC	What It Systematically	Cl Cons

	Captures	Ignores	
Discrimination	Rank-order separation between cases and controls	Absolute risk meaning	Cli ca int a thre
Threshold Behavior	None (threshold-independent)	Sensitivity/specificity at operational cutoff	Po v sens exc false
Class Imbalance	Weakly robust	Positive predictive value collapse	Ma alerts po
Calibration	Not assessed	Alignment between predicted and actual risk	Lo clinic
Temporal Validity	Static dataset performance	Real-time data incompleteness	Perf degra dep
Operational Burden	Not represented	Alerts per patient per day	Aler and ov
Decision Utility	Not represented	Benefit vs harm trade-off	No e of in out
Efficiency	Not represented	Resource cost per true case detected	Ine cli att allo

The threshold problem

Clinical decisions require binary thresholds: alert or do not alert. A model can achieve an AUROC of 0.90 yet, at the only threshold that keeps false alarms tolerable, deliver a sensitivity of just 0.30 [6, 12]. Reporting AUROC without threshold-specific performance is like selling a car by its top speed without mentioning that it can only reach that speed downhill with a tailwind. Clinicians do not receive a ranked list; they receive an alarm. If the operating threshold produces too many false positives or misses too many cases, the model is useless regardless of its AUROC.

AUROC ignores class imbalance

Sepsis occurs in only 5–10 % of ICU admissions. In severely imbalanced settings, AUROC can remain high even when positive predictive value collapses [4, 5]. A model that predicts sepsis for the sickest 1 % of patients can post an AUROC of 0.90 while missing the majority of early cases that clinicians actually need to catch. AUROC therefore flatters models that would be operationally worthless.

AUROC does not calibrate

A perfectly calibrated model means that among patients assigned a 30 % risk score, approximately 30 % actually develop sepsis. AUROC is blind to calibration. Models can achieve AUROC 0.90 while systematically over- or under-estimating risk [3, 11]. When a clinician sees an alert labeled “high risk” that in reality corresponds to a 15 % event rate, trust evaporates. Calibration failures documented across sepsis literature [11] further undermine any claim that high AUROC equates to clinical readiness.

The field’s overreliance on AUROC

AUROC persists because it is easy to report and easy to compare across studies. It has become publication currency, disconnected from the realities of bedside decision-making. We argue that journals should no longer accept AUROC as the headline result. Additional metrics—net benefit, alert rate, and calibration—are not optional extras; they are prerequisites for any claim of clinical relevance.

Alert Fatigue and the False Alarm Problem

The reality of ICU alerts

Intensive care units already suffer from extreme alarm fatigue, with 80–99 % of conventional physiologic alarms proving false or non-actionable [7, 8]. Introducing a sepsis prediction system that generates even a handful of additional alerts per shift risks further desensitization. Clinicians already triage hundreds of signals daily; an AI layer that adds noise rather than signal simply accelerates burnout and error.

What AUROC hides about alert burden

AUROC says nothing about how many alerts a model will fire per patient per day. A model with an impressive AUROC

of 0.90 may still require 50 alerts to identify one true sepsis case [6, 12]. Such a system is not decision support; it is a distraction device. Studies attempting real-time deployment consistently show that alert burden quickly exceeds clinician tolerance [9].

The trade-off no one discusses

Raising sensitivity by lowering the threshold increases false alarms exponentially. Most papers never report the alert rate at the proposed operating point. This omission is not an oversight; it is malpractice. Without explicit alert-burden data, readers cannot judge whether the model's claimed performance is operationally sustainable.

Position statement

We contend that any sepsis prediction paper that does not report alerts per patient per day at the proposed threshold should not be accepted for publication in a clinical journal. AUROC alone is insufficient evidence of utility.

Temporal Misalignment: Retrospective vs Real-Time

The retrospective illusion

Most models are evaluated on static, labeled datasets in which the exact moment of sepsis onset is known after the fact [1, 4]. In real time, clinicians must act on incomplete, streaming information without the benefit of future data. Retrospective AUROC therefore creates an illusion of performance that evaporates the moment the model is deployed prospectively [8, 12].

The 6-hour horizon problem

Many models claim to predict sepsis six hours before clinical recognition. In retrospective data, "clinical recognition" is simply an EHR timestamp. In reality, recognition is a gradual, team-based process influenced by bedside assessment, labs, and gestalt. The six-hour horizon is therefore an artifact of retrospective labeling, not a clinically meaningful construct [6, 13].

Model degradation over time

Models trained on historical data degrade as clinical practice, patient demographics, and EHR documentation change. Performance decay has been documented in several sepsis cohorts [13, 14]. Prospective validation rarely reproduces the retrospective AUROC, yet papers continue to headline the higher number.

Position statement

We argue that retrospective AUROC should be treated as hypothesis generation, not evidence of clinical readiness. Prospective evaluation is non-negotiable.

Alternative Evaluation Frameworks

Clinically actionable thresholds

The solution begins by inverting the modeling process. Instead of training a model and then hunting for a threshold that looks good on a ROC curve, we must first define the clinical requirements: sensitivity of at least 0.80 and no more than two alerts per bed per day. Only then should the algorithm be optimized. Several studies already demonstrate that when thresholds are chosen post-hoc from retrospective data, the resulting alert burden becomes unsustainable [6, 12]. Starting with clinician-defined operating points forces the model to solve the actual problem faced at the bedside rather than an academic proxy.

Table 2 contrasts the dominant retrospective modeling paradigm with a clinically grounded framework, highlighting the structural shifts required to achieve real-world impact.

Table 2. Conceptual Comparison of Retrospective Model Optimization Versus Clinically Grounded Design Paradigms

Design Stage	Retrospective AUROC-Driven Paradigm	Clinically Grounded Paradigm	Struct Implica
Problem Framing	Predict sepsis as early as possible	Support actionable clinical decisions	Shift fr predicti interver
Starting Point	Available dataset	Clinician-defined constraints	Aligns m with wor realiti
Optimization Objective	Maximize AUROC	Meet sensitivity + alert burden constraints	Reorie succe criter

Threshold Selection	Post hoc from ROC curve	Predefined based on clinical tolerance	Elimina arbitra operat poin
Evaluation Metrics	AUROC (primary), AUC comparisons	Net benefit, alerts/day, NNA, calibration	Embe decisi releva
Validation Approach	Retrospective split or cross-validation	Prospective or simulated real-time evaluation	Captu tempo dynam
Output Design	Risk score	Actionable recommendation with context	Impro interpret
Human Factors	Ignored or secondary	Central (alert design, cognitive load)	Enhan adopt likeliho
Lifecycle Management	Static deployment	Continuous learning with feedback loops	Mainta long-te perform
Expected Outcome	Publication success	Sustainable bedside integration	Resol resear practice

Net benefit and decision curve analysis

Decision curve analysis provides the missing link between statistical performance and clinical value. It calculates net benefit by weighting true positives against the harm of false positives across the full range of thresholds [15, 16]. Unlike AUROC, net benefit directly answers whether using the model improves decision-making compared with treating all patients or treating none. Implementation studies that applied this framework to sepsis alerts have shown that even models with AUROC >0.85 frequently offer zero or negative net benefit at clinically tolerable thresholds [9, 17]. This metric must become mandatory.

Alerts per patient per day

Every sepsis prediction paper must report the expected number of alerts generated per patient per day at the proposed threshold. Retrospective papers routinely omit this because the number is embarrassing—often 8–12 alerts per bed per day for models claiming AUROC 0.90 [13, 18]. When compared against baseline ICU alarm rates already documented in the literature, it becomes obvious why adoption fails [8, 19]. Alerts per patient per day is not a secondary outcome; it is the primary determinant of whether a system survives first contact with real clinicians.

Number Needed to Alert (NNA)

Clinicians understand NNA instantly: how many alerts must I review to detect one true sepsis case? An NNA of 50 means 49 unnecessary interruptions for every correct early warning. Multiple validation studies have reported NNA values in this range despite headline AUROCs of 0.85–0.92 [7, 12, 20]. Framing performance this way shifts the conversation from statistical elegance to operational cost. It also makes clear that a model with lower AUROC but NNA of 8 can be vastly superior to one with higher AUROC but NNA of 40.

Position statement

We contend that every sepsis prediction paper must report four non-negotiable metrics: (1) AUROC for historical comparability, (2) net benefit across thresholds, (3) alerts per patient per day at the proposed threshold, and (4) number needed to alert. Without these, clinical utility cannot be assessed and the paper should be rejected by any journal claiming to serve intensive care medicine.

Redesigning Sepsis Prediction Models for the Bedside

Start with clinical requirements, not data

Most models are built backward: developers collect data, train an algorithm, and only afterward ask what the data can do. The correct sequence is to convene frontline clinicians first, define the acceptable false-alarm rate and minimum sensitivity required to change management, then build the model to meet those constraints [21, 22]. Papers that followed this approach—even partially—demonstrated far higher prospective acceptance rates than purely data-driven competitors [9, 23].

Human-centered alert design

Alerts must respect human cognitive limits. Interruptive pop-ups should fire only when the net benefit is unambiguously positive; otherwise, risk information should appear passively on a dashboard. Clinicians must be able

to dismiss, snooze for two hours, or escalate to a rapid-response team with one click. Human-factors studies in sepsis prediction consistently show that poorly designed interfaces accelerate alert fatigue even when the underlying model is accurate [18, 19, 24].

Actionable explanations

Attention weights and SHAP values are not explanations; they are mathematical artifacts. Clinicians need plain-language output: “This patient’s risk has risen because respiratory rate increased 40 % while lactate is trending upward—consider repeat blood cultures and fluid reassessment.” Models that provide such directed suggestions improve trust and uptake far more than black-box scores [1, 3, 25].

Continuous learning and feedback loops

Static models degrade rapidly as practice changes. The next generation must incorporate clinician feedback as online training data: when a dismissed alert is followed by no sepsis, the model learns; when an acted-upon alert precedes confirmed sepsis, the model also learns. Several groups have already prototyped human-in-the-loop architectures that maintain performance over time [14, 26].

Position statement

We contend that the next generation of sepsis prediction models must be co-designed with clinicians, optimized for actionable thresholds rather than AUROC, and integrated with human-centered alert design. Technical performance without clinical design is engineering, not medicine.

Recommendations for Researchers, Reviewers, and Journal Editors

The field needs a decisive shift away from performance reporting that is disconnected from bedside reality. Researchers should no longer foreground AUROC as the primary signal of value; instead, every model must be presented through clinically interpretable metrics that directly map to decision-making. At a minimum, studies should report net benefit across a range of thresholds, alerts per patient per day to quantify operational burden, and number needed to alert (NNA) to capture efficiency from a clinician’s perspective. These metrics should not be supplementary—they should anchor the main narrative of the paper. Claims of clinical relevance must be supported by prospective validation or, at the very least, rigorously designed simulated real-time evaluations that preserve temporal ordering and decision constraints. Equally important is transparency around failure: degradation under distribution shift, subgroup performance gaps, and scenarios in which the model generates harmful or non-actionable alerts should be documented alongside successes. While these practices may reduce the speed and volume of publications, they will substantially improve the credibility and translational value of the literature.

Peer reviewers play a critical gatekeeping role and should recalibrate their expectations accordingly. Manuscripts that headline AUROC without accompanying clinical utility metrics should be considered incomplete. Reviewers should interrogate whether the model would be acceptable in real clinical workflows—whether its alert frequency is tolerable, whether false positives would erode trust, and whether the system accounts for human factors such as cognitive load and alarm fatigue. A simple but powerful heuristic is to evaluate whether the reviewer would be comfortable having the system deployed for a real patient under their care during off-hours. This reframing shifts evaluation from abstract statistical performance to lived clinical impact and forces authors to justify their design choices in operational terms.

Journal editors are uniquely positioned to reset field-wide norms by updating submission and reporting guidelines. Clinical AI manuscripts should be required to include standardized utility metrics and to clearly distinguish between retrospective performance and real-world readiness. Dedicated space should be created for prospective validation studies, including those with neutral or negative findings, to counteract publication bias and provide a more accurate picture of model behavior in practice. Editorial policies can also discourage the ongoing “AUROC arms race” by limiting purely retrospective contributions that offer incremental performance gains without implementation insight, while prioritizing work that addresses deployment, workflow integration, and measurable clinical outcomes.

Funding bodies must align incentives with these expectations by prioritizing studies that move beyond model development toward implementation and evaluation in real settings. Grants should emphasize prospective trials,

embedded evaluations within clinical workflows, and interdisciplinary collaboration that includes practicing clinicians as core investigators. Investment should shift toward understanding human–AI interaction, alert burden, and system usability, rather than marginal improvements in predictive accuracy. Without this reallocation, the field risks continuing to produce technically sophisticated but clinically irrelevant systems.

Taken together, these recommendations reflect a broader position: current publication and funding structures systematically reward models that perform well on paper but fail in practice. Unless incentives are redesigned to value clinical utility, transparency, and real-world validation, meaningful progress in AI-driven healthcare will remain limited.

Addressing Counterarguments

A common defense of current practice is that high AUROC remains a necessary foundation for any useful model. While this is true in principle, the problem lies in how the metric is used in practice. AUROC has become a proxy for overall quality, often treated as sufficient evidence of utility rather than a baseline requirement. Reframing it as a minimum standard rather than a headline result would restore balance and encourage more comprehensive evaluation.

Another frequent objection is that prospective validation is resource-intensive and difficult to execute. This constraint is real, but it does not justify overstating the readiness of retrospective models. If a system has not been tested under conditions that approximate real-time use, it should be explicitly labeled as proof-of-concept. Clarity in positioning is preferable to premature claims of clinical applicability, which can mislead both practitioners and policymakers.

Some argue that clinicians will adapt to higher alert volumes as predictive systems become more prevalent. However, empirical evidence from alarm fatigue research suggests the opposite: increasing the number of non-actionable alerts leads to desensitization, slower response times, and higher error rates. Designing systems that assume unlimited clinician capacity is unrealistic; instead, models must be optimized for the constraints of human attention and workflow.

Developers often contend that their specific model overcomes these limitations and is genuinely ready for deployment. This belief is understandable but insufficient. Individual confidence cannot substitute for standardized, externally validated evaluation frameworks. Without consistent criteria grounded in clinical relevance, it is impossible to distinguish genuinely useful systems from those that only appear effective in isolated settings.

Finally, AUROC is often defended as a convenient tool for comparing models across studies. While it does enable statistical comparison, such comparisons are of limited value if none of the models demonstrate real-world effectiveness. The field should prioritize comparisons based on clinically meaningful outcomes—such as reduction in adverse events or improvement in decision efficiency—rather than abstract performance metrics that do not translate to practice.

Conclusion

Hundreds of sepsis prediction models with high AUROC have been published. Almost none are used at the bedside. This is not a technical problem—it is an evaluation problem. The field has optimized for the wrong metric.

We have argued that AUROC is fundamentally misleading for clinical utility. It hides false alarm rates, ignores threshold selection, and says nothing about whether a clinician would or should act on a prediction. Most sepsis prediction models fail at the bedside because they were never designed for the bedside in the first place.

We propose: (1) abandon AUROC as a primary metric, (2) require reporting of net benefit, alerts per patient per day, and number needed to alert, (3) start model design with clinical requirements, not data, (4) co-design with clinicians, and (5) change publication and funding incentives.

The goal of sepsis prediction is not to achieve AUROC 0.95. The goal is to help clinicians save lives. By that measure, the field is failing. It is time to stop celebrating high AUROCs and start demanding clinical utility. Anything less is a disservice to patients, to clinicians, and to the promise of artificial intelligence in medicine.

Acknowledgements

None

Conflict of interest

None

Ethics statement

None

Financial support

None

Received: 05 Apr 2021 Revised: 23 May 2021 Accepted: 05 Jul 2021
Published online: 20 January 2022

References

Nemati S, Holder A, Razmi F, Stanley MD, Clifford GD, Buchman TG. An interpretable machine learning model for accurate prediction of sepsis in the ICU. Crit Care Med. 2018;46(4):547-53.

Mao Q, Jay M, Hoffman JL, Calvert J, Barton C, Shimabukuro D, et al. Multicentre validation of a sepsis prediction algorithm using only vital sign data in the emergency department, general ward and ICU. BMJ Open. 2018;8(1):e017833.

Yang M, Liu C, Wang X, Li Y, Gao H, Liu X, et al. An explainable artificial intelligence predictor for early detection of sepsis. Crit Care Med. 2020;48(11):e1091-6.

Fleuren LM, Klausch TL, Zwager CL, Schoonmade LJ, Guo T, Roggeveen LF, et al. Machine learning for the prediction of sepsis: a systematic review and meta-analysis of diagnostic test accuracy. Intensive Care Med. 2020;46(3):383-400.

Kausch SL, Moorman JR, Lake DE, Keim-Malpass J. Physiological machine learning models for prediction of sepsis in hospitalized adults: an integrative review. Intensive Crit Care Nurs. 2021;65:103035.

Shah PK, Ginestra JC, Ungar LH, Junker P, Rohrbach JI, Fishman NO, et al. A simulated prospective evaluation of a deep learning model for real-time prediction of clinical deterioration among ward patients. Crit Care Med. 2021;49(8):1312-21.

Wong A, Otles E, Donnelly JP, Krumm A, McCullough J, DeTroyer-Cooley O, et al. External validation of a widely implemented proprietary sepsis prediction model in hospitalized patients. JAMA Intern Med. 2021;181(8):1065-70.

Persson I, Östling A, Arlbrandt M, Söderberg J, Becedas D. A machine learning sepsis prediction algorithm for intended intensive care unit use (NAVOY Sepsis): proof-of-concept study. JMIR Form Res. 2021;5(9):e28000.

Pettinati MJ, Chen G, Rajput KS, Selvaraj N. Practical machine learning-based sepsis prediction. In: 2020 42nd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC); 2020; Montreal, QC, Canada. IEEE; 2020. p. 4986-91.

Nesaragi N, Patidar S. Early prediction of sepsis from clinical data using ratio and power-based features. Crit Care Med. 2020;48(12):e1343-9.

Kwon YS, Baek MS. Development and validation of a quick sepsis-related organ failure assessment-based machine-learning model for mortality prediction in patients with suspected infection in the emergency department. J Clin Med. 2020;9(3):875.

Burdick H, Pino E, Gabel-Comeau D, Gu C, Roberts J, Le S, et al. Validation of a machine learning algorithm for early severe sepsis prediction: a retrospective study predicting severe sepsis up to 48 h in advance using a diverse dataset from 461 US hospitals. BMC Med Inform Decis Mak. 2020;20(1):276.

Barton C, Chettipally U, Zhou Y, Jiang Z, Lynn-Palevsky A, Le S, et al. Evaluation of a machine learning algorithm for up to 48-hour advance prediction of sepsis using six vital signs. Comput Biol Med. 2019;109:79-84.

Kumar S, Tripathy S, Jyoti A, Singh SG. Recent advances in biosensors for diagnosis and detection of sepsis: a comprehensive review. Biosens Bioelectron. 2019;124:205-15.

Lu HX, Du J, Wen DL, Sun JH, Chen MJ, Zhang AQ, et al. Development and validation of a novel predictive score for sepsis risk among trauma patients. World J Emerg Surg. 2019;14(1):11.

Islam MM, Nasrin T, Walther BA, Wu CC, Yang HC, Li YC. Prediction of sepsis patients using machine learning approach: a meta-analysis. Comput Methods Programs Biomed. 2019;170:1-9.

Kijpaisalratana N, Sanglertsinlapachai D, Techaratsami S, Musikatavorn K, Saoraya J. Machine learning algorithms for early sepsis detection in the emergency department: a retrospective study. Int J Med Inform. 2022;160:104689.

Van Wyk F, Khojandi A, Kamaleswaran R. Improving prediction performance using hierarchical analysis of real-time data: a sepsis case study. IEEE J Biomed Health Inform. 2019;23(3):978-86.

Shashikumar SP, Wardi G, Malhotra A, Nemati S. Artificial intelligence sepsis prediction algorithm learns to say “I don’t know”. NPJ Digit Med. 2021;4(1):134.

Moor M, Rieck B, Horn M, Jutzeler CR, Borgwardt K. Early prediction of sepsis in the ICU using machine learning: a systematic review. Front Med (Lausanne). 2021;8:607952.

Bloch E, Rotem T, Cohen J, Singer P, Aperstein Y. Machine learning models for analysis of vital signs dynamics: a case for sepsis onset prediction. J Healthc Eng. 2019;2019(1):5930379.

Kong G, Lin K, Hu Y. Using machine learning methods to predict in-hospital mortality of sepsis patients in the ICU. BMC Med Inform Decis Mak. 2020;20(1):251.

Shashikumar SP, Li Q, Clifford GD, Nemati S. Multiscale network representation of physiological time series for early prediction of sepsis. Physiol Meas. 2017;38(12):2235-48.

Ackerman MH, Ahrens T, Kelly J, Pontillo A. Sepsis. Crit Care Nurs Clin. 2021;33(4):407-18.

Purcarea A, Sovaila S. Sepsis, a 2020 review for the internist. Rom J Intern Med. 2020;58(3):129-37.

Wang D, Li J, Sun Y, Ding X, Zhang X, Liu S, et al. A machine learning model for accurate prediction of sepsis in ICU patients. Front Public Health. 2021;9:754348.