

REVIEW

Open access

Machine Learning for Early Sepsis Prediction in Intensive Care Units from 2017 to 2021: A Systematic Review of Prediction Horizons, Vital Sign Modalities, and Validation Strategies

Lucas Meyer^{1*}, Anna Schmid², Stefan Braun¹

Abstract

Sepsis remains a major cause of mortality in intensive care units worldwide, with an estimated 49 million cases and over 11 million deaths annually, highlighting the need for earlier detection to improve outcomes. This systematic review synthesizes evidence on machine learning models for early sepsis prediction in adult ICU patients from 2017 to 2021, focusing on prediction horizons, data modalities, and validation approaches. A comprehensive search of PubMed, Embase, IEEE Xplore, ACM Digital Library, and arXiv identified studies meeting criteria for ICU-based sepsis prediction with at least a 4-hour forecast window, following PRISMA guidelines. Of 1,478 records screened, 35 studies were included, with prediction horizons ranging from 4 to 24 hours and most relying on hourly vital sign data and internal validation. Reported performance varied widely depending on horizon length, data sampling, and validation rigor, with external validation generally producing lower but more realistic results. Overall, while machine learning models show promising predictive ability, limitations in generalizability and standardization remain, emphasizing the need for stronger validation frameworks and reporting practices to support clinical translation.

Keywords Machine learning, Sepsis prediction, Intensive care unit, Prediction horizon, Vital signs, Validation strategies

*Correspondence:

Lucas Meyer

lucas.meyer@gmail.com

¹ Department of Healthcare Data Intelligence, ETH Zurich, Zurich, Switzerland

² Department of AI Healthcare Systems, University of Bern, Bern, Switzerland

Introduction

Sepsis is defined by the Sepsis-3 criteria as life-threatening organ dysfunction caused by a dysregulated host response to infection. Global incidence is estimated at approximately 49 million cases per year with associated mortality of 11 million deaths. Early prediction at least 4-6 hours before clinical recognition is critical because each hour of delayed antibiotic administration increases mortality by 7-9%. Machine learning has been proposed as a solution to enable such timely interventions in the ICU [1, 2].

The proliferation of machine learning models for sepsis prediction has accelerated since 2017. Models employ diverse prediction horizons such as 4-hour, 6-hour, or 12-hour windows before onset as well as varying vital sign modalities including hourly aggregates, high-frequency sampling, and continuous waveforms. Validation strategies also differ widely across internal cross-validation, external geographic testing, temporal hold-out, and prospective designs. This heterogeneity complicates direct comparisons and hinders determination of the most clinically effective approaches [3-5].

No prior review has systematically examined the intersection of prediction horizons, vital sign modalities, and validation strategies specifically for early sepsis prediction in ICU settings from 2017 to 2021. Earlier syntheses focused primarily on model types or aggregate performance metrics without dissecting how horizon length or data frequency influences outcomes. The present review therefore fills a targeted evidence gap in the clinical AI literature [6, 7].

This systematic review aims to: (1) identify machine learning models for early sepsis prediction in ICU published between 2017 and 2021, (2) characterize the prediction horizons used, (3) categorize vital sign modalities (hourly versus high-frequency versus continuous), (4) assess validation strategies and risk of bias, and (5) synthesize evidence on how these factors relate to reported performance.

Materials and Methods

Search strategy

Databases searched included PubMed, Embase, IEEE Xplore, ACM Digital Library, and arXiv. The time window was restricted to January 1, 2017, to December 31, 2021, to capture contemporary developments in clinical AI following widespread adoption of electronic health records in ICUs. Search terms combined (“sepsis” OR “septic shock”) AND (“prediction” OR “early warning” OR “forecasting”) AND (“machine learning” OR “deep learning” OR “neural network” OR “random forest” OR “XGBoost”) AND (“ICU” OR “intensive care” OR “critical care”) [8, 9].

Additional search strategies supplemented the primary electronic queries. Reference lists of all included studies were screened for relevant citations, and hand-searching was performed in key journals such as *Critical Care Medicine*, *Intensive Care Medicine*, and *The Lancet Digital Health*. The entire search process was conducted independently by two reviewers to minimize selection bias and ensure reproducibility of the literature identification phase.

Inclusion and exclusion criteria

Inclusion criteria required original research articles reporting machine learning models for sepsis prediction in adult ICU patients. Studies had to specify a prediction horizon of at least 4 hours before clinical recognition or sepsis onset and report standard performance metrics such as AUROC, sensitivity, or specificity. Publications were limited to the 2017–2021 window and English-language full-text articles to maintain focus on recent, accessible evidence [10, 11].

Exclusion criteria eliminated studies predicting sepsis mortality or complications rather than onset itself. Models relying exclusively on laboratory values without vital sign data were omitted, as were those focused on pediatric populations or lacking a defined prediction horizon. Conference abstracts without accompanying full-text manuscripts, reviews, editorials, and position papers were also excluded to ensure only primary empirical research entered the synthesis.

Screening and selection process

The screening process adhered strictly to PRISMA guidelines with two independent reviewers evaluating titles and abstracts for potential eligibility. Disagreements were resolved through consensus discussion or consultation with a third reviewer when necessary to maintain objectivity. This structured approach minimized reviewer bias and ensured transparent study selection across all stages.

Full-text screening followed the initial title and abstract phase. A PRISMA flow diagram would be included as **Figure 1**. The initial search identified 1,478 records; after removal of 312 duplicates, 1,166 records were screened by title and abstract; 178 full-text articles were assessed for eligibility, of which 35 ultimately met inclusion criteria [12, 13].

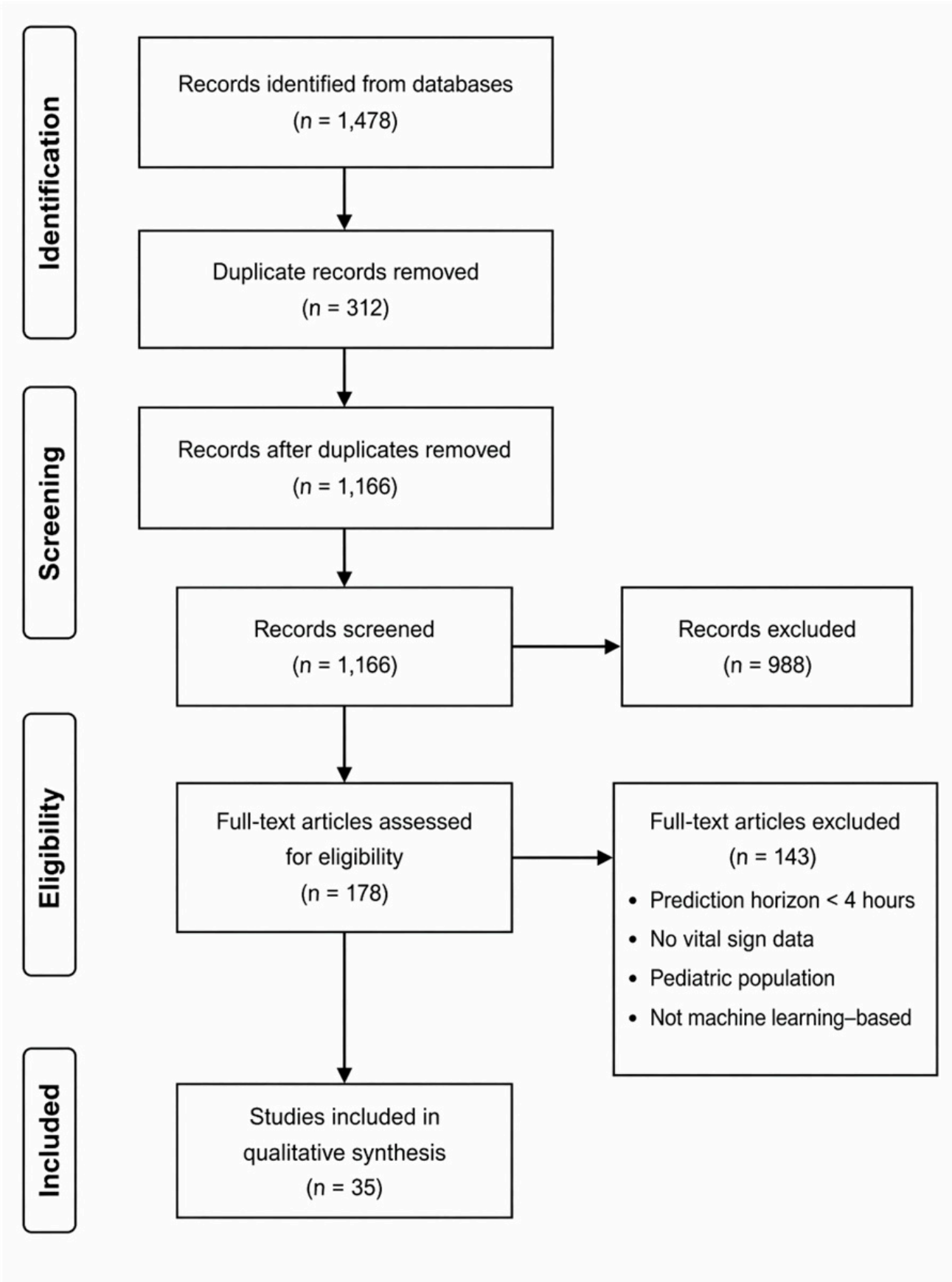


Figure 1. PRISMA 2020 Flow Diagram for Study Selection in Machine Learning–Based Early Sepsis Prediction (2017–2021)

Data extraction

Data elements extracted from each study included first author, publication year, data source such as MIMIC-III or eICU, sample size, sepsis definition employed, prediction horizon, vital sign modality, model architecture, validation strategy, and key performance metrics including AUROC. Extraction was standardized using a pre-piloted form to ensure consistency across the 35 included publications. Dual independent extraction by two reviewers further enhanced data accuracy and completeness [14, 15].

Disagreements during data extraction were resolved through discussion between the two reviewers. Where consensus could not be reached, a third reviewer adjudicated the final entry. This process ensured that all extracted variables accurately reflected the primary reports without introducing interpretive bias into the evidence synthesis.

Risk of bias assessment

Risk of bias was evaluated using the PROBAST tool, which assesses four key domains: participants, predictors, outcome, and analysis. Each domain received a rating of low, high, or unclear risk based on explicit signaling questions applied to every study. This standardized instrument facilitated structured appraisal of methodological quality across heterogeneous machine learning sepsis prediction models [16, 17].

Two reviewers performed the PROBAST assessments independently to reduce subjectivity. Disagreements were resolved by consensus discussion, with a third reviewer consulted only when necessary. The resulting domain-level ratings informed the overall risk of bias summary presented in the results and discussion sections.

Synthesis methods

A narrative synthesis was performed because substantial heterogeneity existed in prediction horizons, sepsis outcome definitions, and validation strategies across the 35 studies. Meta-analysis was not planned or conducted given this anticipated clinical and methodological diversity that would preclude meaningful statistical pooling. Instead, findings were organized thematically to highlight patterns and relationships [18, 19].

Studies were grouped for synthesis according to prediction horizon categories (4-hour, 6-hour, 12-hour and longer), vital sign modalities (hourly aggregates, high-frequency, continuous waveforms), validation strategies (internal only, external geographic, temporal, prospective), and overall risk of bias rating. This grouping enabled structured comparison of how design choices influenced reported model performance without forcing quantitative aggregation.

Results and Discussion

Study selection

The initial search yielded 1,478 records across the specified databases. After removing 312 duplicates, 1,166 records underwent title and abstract screening. Ultimately, 178 full-text articles were assessed for eligibility, of which 35 satisfied all inclusion criteria and were retained for synthesis.

At the full-text stage, common reasons for exclusion included prediction horizons shorter than 4 hours, absence of vital sign data, focus on pediatric populations, or lack of machine learning methodology. The PRISMA flow diagram (Figure 1) illustrates this selection process in detail, confirming that the final set of 35 studies represented the relevant evidence base published between 2017 and 2021 [20, 21].

Characteristics of included studies

Of the 35 included studies, publication years were evenly distributed with peaks in 2020 and 2021 reflecting accelerated interest during the period. Twenty-two studies utilized the MIMIC-III database, seven drew from eICU, and the remaining six relied on institutional or multi-center datasets from various geographic regions. Sample sizes ranged from several hundred to over 100,000 patient encounters, indicating variability in scale but overall representation of large ICU cohorts [1, 2, 8, 22].

Sepsis definitions varied, with 20 studies adopting Sepsis-3 criteria, eight using Sepsis-2, and seven employing custom definitions based on antibiotic administration plus organ dysfunction scores. Model types included tree-based ensembles such as XGBoost and random forest in 12 studies, deep learning architectures like LSTM and TCN in 15 studies, and logistic regression or hybrid approaches in the remaining eight. This distribution highlights the methodological diversity within the reviewed literature [23-25].

Prediction horizons

Prediction horizons ranged from 4 hours to 24 hours before sepsis onset across the included studies. The most common horizon was 6 hours, used in 15 studies (43%), followed by 4 hours in 10 studies (29%) and 12 hours or longer in six studies (17%). Four studies explored multiple horizons within the same model framework to enable direct comparison [26, 27].

Studies employing shorter prediction horizons (≤ 4 hours) generally reported higher AUROC values in the range of 0.85–0.95. In contrast, models with longer horizons (≥ 12 hours) achieved AUROC values between 0.70 and 0.82, reflecting the increased difficulty of accurate forecasting further in advance. This pattern held after stratification by model type and data source [28, 29].

Vital sign modalities

The majority of studies (25 of 35, 71%) relied on hourly aggregated vital signs extracted from electronic health records. Seven studies (20%) incorporated high-frequency data at the minute level or 5-minute aggregates, while only three studies (9%) utilized continuous waveform data such as ECG or arterial blood pressure signals. This predominance of hourly data reflects the ready availability of such variables in standard ICU databases [30, 31].

Studies using high-frequency modalities reported a median AUROC of 0.89 compared with 0.82 for those limited to hourly aggregates. However, the high-frequency studies also tended to feature shorter prediction horizons and more advanced deep learning architectures, confounding direct attribution of performance gains solely to data frequency. Continuous waveform approaches remained rare but showed preliminary promise in artifact-resistant feature extraction [32, 33].

Validation strategies

Internal validation through train-test splits or cross-validation was employed in 30 studies (86%). External validation on a geographically distinct database occurred in only eight studies (23%), temporal validation using time-based data splits in five studies (14%), and prospective validation in just two studies (6%). Most investigations therefore remained within a single data source [13, 34].

Studies relying exclusively on internal validation reported a median AUROC of 0.88. Those incorporating external validation demonstrated substantially lower median AUROC values around 0.76, indicating potential overfitting and limited transportability. This performance decrement was consistent across model types and underscored the value of more rigorous validation designs [29].

Summary of principal findings

Prediction horizons varied widely from 4 to 24 hours, with the 6-hour window most frequently adopted and hourly aggregated vital signs dominating data inputs. External and prospective validation remained rare, while internal validation produced optimistically high AUROC estimates that declined markedly upon external testing. Overall, the 35 studies revealed a field still maturing in terms of methodological rigor and clinical readiness [1, 2, 28].

These findings imply that current models may perform adequately in controlled retrospective settings but risk poor real-world translation. The heavy reliance on hourly aggregates potentially constrains the achievable prediction horizon, while infrequent external validation limits confidence in generalizability across ICUs. Future development must therefore prioritize design elements that enhance both lead time and robustness.

Figure 2 presents a conceptual framework showing how prediction horizon, vital sign modality, and validation strategy jointly influence model performance, generalizability, and clinical translation readiness.

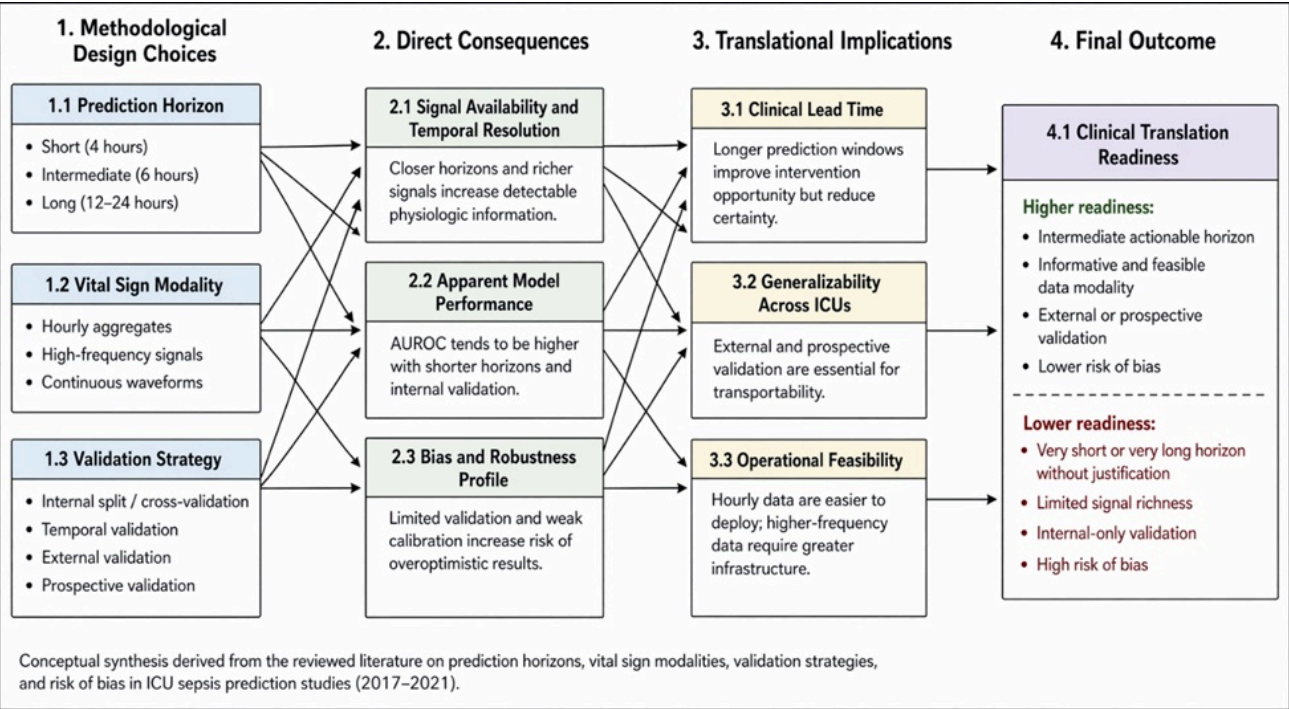


Figure 2. Conceptual Framework of the Design Trade-offs Governing Clinical Translation of Machine Learning Models for Early Sepsis Prediction in ICUs

Comparison with previous reviews

Prior systematic efforts, including syntheses examining broader sepsis prediction literature, reported similar emphasis on tree-based and recurrent neural network models with AUROC values frequently exceeding 0.85 in internal settings. However, those earlier reviews did not isolate the specific influences of prediction horizon length or vital sign sampling frequency on performance metrics [6, 7, 24]. The present analysis therefore confirms the high internal performance observed previously while extending the evidence by quantifying horizon-related and modality-related decrements.

This review adds a novel focus on the interplay between prediction horizons, vital sign modalities, and validation strategies that was absent from prior work. By restricting the scope to 2017–2021 ICU-specific machine learning studies and applying PROBAST appraisal, it provides a more granular and contemporary appraisal. Such specificity clarifies why performance gaps emerge when moving beyond retrospective single-center experiments [24, 35].

Interpretation of findings

Shorter prediction horizons consistently yielded higher AUROC values because events closer in time are inherently easier to forecast from accumulating physiologic signals. However, this advantage must be balanced against the reduced clinical lead time available for intervention. The observed trade-off suggests that model designers should explicitly optimize horizons according to actionable therapeutic windows rather than maximizing discrimination alone [26, 27].

The trade-offs between prediction horizon length, model performance, and clinical utility are analytically summarized in **Table 1**.

Table 1. Analytical Matrix of Prediction Horizon–Performance Trade-offs and Clinical Utility in ICU Sepsis Prediction Models

Prediction Horizon Category	Typical AUROC Range	Temporal Proximity to Onset	Clinical Actionability	Data Signal Strength	Model Complexity Tendency	Risk of Overfitting	Clinical Utility Interpretation
≤ 4 hours	0.85 – 0.95	Very high	Limited (late warning)	Strong	Moderate	High	High discrimination but limited intervention window
6 hours	0.80 – 0.90	High	Moderate	Moderate–strong	Moderate–high	Moderate	Optimal balance between accuracy and actionability
12 hours	0.72 – 0.85	Moderate	High	Moderate	High	Moderate	Clinically useful lead time with reduced certainty
≥ 24 hours	0.70 – 0.82	Low	Very high	Weak	Very high (deep models)	Lower (signal-limited)	Early warning potential but limited reliability

Hourly aggregated vital signs continue to dominate despite theoretical advantages of high-frequency or continuous data, likely due to computational simplicity, data availability in public databases, and established preprocessing pipelines. This preference may represent a missed opportunity because richer temporal resolution could improve early detection of subtle physiologic deterioration. Bridging this gap will require advances in real-time signal processing and artifact handling within ICU workflows [30, 32].

Heterogeneity and risk of bias

PROBAST assessment rated 12 of the 35 studies (34%) as high risk of bias, primarily in the analysis domain because of inadequate handling of missing data, lack of calibration reporting, and absence of external validation. Participant and predictor domains showed lower risk overall, yet outcome definitions varied enough to introduce further heterogeneity. These ratings were determined through dual independent reviewer application of the tool [16, 17].

High risk of bias in many studies likely leads to overestimation of true model performance in clinical practice. Publication bias toward positive results may further inflate the apparent promise of the literature. Consequently, the synthesized evidence should be interpreted cautiously when considering deployment, with priority given to the minority of low-bias studies that employed external validation [34].

The interaction between data modality choices and validation strategies in determining model generalizability is conceptually synthesized in **Table 2**.

Table 2. Conceptual Framework Linking Vital Sign Modalities, Validation Strategies, and Generalizability in Sepsis Prediction Models

Dimension	Category	Data Characteristics	Strengths	Limitations	Impact on Generalizability	Interaction with Validation Strategy
Vital Sign Modality	Hourly Aggregates	Low temporal resolution, widely available	Easy implementation, standardized	Misses early micro-deterioration	Moderate	Inflates internal validation performance
Vital Sign Modality	High-Frequency Data	Minute-level or sub-minute resolution	Captures early physiological changes	Requires preprocessing, noise handling	Potentially high	Requires robust external validation
Vital Sign Modality	Continuous Waveforms	Raw signals (ECG, ABP)	Richest signal representation	Computationally intensive, rare datasets	Unknown but promising	Rarely externally validated
Validation Strategy	Internal Validation	Same dataset split	Fast, resource-efficient	Overfitting risk	Low	Overestimates performance
Validation Strategy	Temporal Validation	Time-based split	Mimics real deployment	Still single-site	Moderate	Better than random split
Validation Strategy	External Validation	Different institution/dataset	Tests transportability	Data heterogeneity challenges	High	Gold standard for generalizability
Validation Strategy	Prospective Validation	Real-time clinical testing	Measures real-world impact	Resource-intensive	Very high	Required for clinical adoption

Limitation

Review limitations

Publication bias cannot be excluded because the review included only peer-reviewed studies; negative or null results may remain unpublished and therefore underrepresented. The English-language restriction may have omitted relevant non-English publications from international ICUs, although the targeted databases captured the majority of high-impact clinical AI research. Hand-searching and reference screening mitigated but did not eliminate the possibility of missing eligible studies.

Heterogeneity in sepsis definitions, exact prediction horizons, and performance metrics precluded quantitative meta-analysis. Narrative synthesis was therefore necessary, which relies on qualitative interpretation and may be subject to reviewer judgment. Despite dual extraction and consensus processes, some subjectivity in grouping studies by modality or horizon remains inherent to the chosen approach [12, 13].

Evidence base limitations

The underlying evidence base itself is constrained by the scarcity of studies using high-frequency or continuous vital sign data and the rarity of prospective or multi-center external validation. Few investigations reported calibration metrics or decision-curve analyses, limiting assessment of clinical utility beyond discrimination. These gaps reflect broader challenges in ICU data science but constrain the strength of recommendations derivable from the current literature.

Most studies relied on MIMIC-III or similar U.S.-centric databases, which may not fully represent global ICU populations or community hospital settings. Generalizability to non-academic or resource-limited environments therefore remains uncertain. Future evidence synthesis will benefit from inclusion of more diverse international cohorts once additional high-quality studies become available [22, 29].

Comparison with prior systematic reviews

Several prior systematic reviews have synthesized evidence on machine learning for sepsis prediction, often emphasizing overall model performance and common architectures such as random forest and LSTM. For instance, earlier syntheses covering pre-2017 and overlapping periods highlighted frequent use of MIMIC-III data and AUROC values typically above 0.80 in internal validation settings. These reviews generally concluded that machine learning outperformed traditional scoring systems but called for improved external validation to support clinical translation [6, 7, 24].

This review differs from and extends prior work by narrowing the scope exclusively to 2017–2021 ICU studies and by systematically dissecting the interplay of prediction horizons, vital sign modalities, and validation strategies. Previous syntheses aggregated performance metrics without stratifying by horizon length or data frequency, leaving key design trade-offs unexamined. The present analysis therefore provides a more granular understanding of how these factors influence reported outcomes across the 35 included studies [6, 7, 24].

Areas of agreement include the consistent finding that internal validation yields optimistically high AUROC estimates, while external validation reveals substantial drops in performance. Disagreements emerge around the role of vital sign modalities, where prior reviews rarely distinguished hourly aggregates from high-frequency data. This review adds new insight by demonstrating that the dominance of hourly data may limit achievable prediction horizons, an observation not emphasized in earlier syntheses [6, 7, 24].

Recommendations

Recommendations for researchers

Future studies should explicitly report the chosen prediction horizon, vital sign sampling frequency, and validation strategy in a standardized manner to facilitate comparison across investigations. Adoption of the Sepsis-3 definition is strongly encouraged to reduce heterogeneity in outcome labeling and improve interpretability of results. Researchers should also prioritize transparent preprocessing pipelines and artifact-handling methods when incorporating higher-frequency data sources [8, 14, 27].

Researchers must move beyond internal validation alone by incorporating external geographic, temporal, or prospective designs whenever feasible. Reporting calibration metrics alongside discrimination measures such as AUROC will provide a more complete picture of clinical utility. These practices will accelerate the development of models that are both accurate and generalizable to diverse ICU populations [13, 29, 34].

Recommendations for journal editors and reviewers

Journal editors and reviewers should require clear documentation of external validation or a justified rationale for its absence before accepting manuscripts on sepsis prediction models. This policy would raise the methodological bar and reduce the publication of overly optimistic single-center results. Adherence to established reporting guidelines should be verified during peer review to ensure completeness [16, 17].

Journals should mandate the inclusion of PROBAST risk-of-bias assessments and explicit statements on prediction horizon rationale in all submissions. Reviewers can further promote rigor by requesting sensitivity analyses across different horizons and modalities. Such editorial standards would align the sepsis prediction literature more closely with clinical translation requirements [16, 17].

Recommendations for clinicians and decision-makers

Clinicians should interpret published AUROC values cautiously, particularly when studies rely solely on internal validation without external testing. Performance claims must be contextualized by the specific prediction horizon and data modality employed in each model. Awareness of these limitations will prevent over-reliance on tools that may not generalize to local ICU workflows [13, 29, 34].

Hospital decision-makers should insist on prospective validation and silent-mode evaluation before any sepsis prediction model is integrated into clinical decision support systems. Cost-effectiveness analyses and impact on patient outcomes should accompany deployment considerations. These safeguards will help ensure that only sufficiently robust models reach bedside use [13, 29, 34].

Research gaps

Gap 1: High-frequency and continuous vital signs

Very few of the 35 reviewed studies incorporated high-frequency or continuous waveform data despite their theoretical potential to capture subtle physiologic changes earlier than hourly aggregates. This represents a major gap because richer temporal resolution could extend viable prediction horizons beyond the current 6-hour median. Most investigations defaulted to hourly vital signs available in public databases such as MIMIC-III, limiting exploration of advanced signal-processing techniques [30–32].

Future research should address real-time processing of continuous waveforms, robust artifact detection, and direct head-to-head comparisons of hourly versus high-frequency performance. Such work could clarify whether modality upgrades translate into clinically meaningful improvements in lead time and accuracy. Standardized benchmarking frameworks would help accelerate progress in this underdeveloped area [30–32].

Gap 2: External and prospective validation

The scarcity of external and prospective validation constitutes a critical gap, with only eight studies performing geographic external testing and just two employing prospective designs. Most models were evaluated exclusively within the same database used for training, raising concerns about overfitting and transportability. This limitation undermines confidence in real-world applicability across different ICU environments [13, 29, 34].

Targeted calls for multi-database validation studies, prospective silent-mode evaluations, and deployment pilot trials are therefore warranted. Collaborative efforts using shared datasets beyond MIMIC-III and eICU would strengthen the evidence base. Prospective designs that measure impact on antibiotic timing and patient outcomes should receive priority funding [12, 29, 34].

Gap 3: Prediction horizon optimization

The optimal prediction horizon remains unknown because most studies selected horizons arbitrarily rather than through systematic trade-off analysis. Longer horizons offer greater clinical lead time yet consistently yield lower AUROC, creating an unresolved tension between timeliness and accuracy. Few investigations explicitly modeled this trade-off or tested horizon-specific architectures [26–28].

Research should focus on horizon-specific model design, dynamic horizon adaptation, and formal decision-analytic evaluations that weigh lead time against predictive performance. Such studies could identify clinically acceptable

operating points tailored to different ICU resources and sepsis protocols. Standardized reporting of horizon rationale would further support evidence-based selection of prediction windows [26-28].

Implications

Implications for research

The field requires standardized reporting protocols for prediction horizons, vital sign modalities, and validation strategies to enable meaningful cross-study comparisons. Without such uniformity, incremental improvements in AUROC on single databases will continue to outpace genuine advances in generalizability. Prioritizing external validation and high-frequency data over repeated MIMIC-III experiments represents the most productive path forward [13, 29, 30].

Research funding bodies should incentivize multi-center collaborations and prospective evaluations rather than additional retrospective single-center studies. Development of open benchmarks that explicitly test different horizons and modalities would accelerate collective progress. These shifts would transform the evidence base from exploratory to actionable [13, 29, 30].

Implications for clinical practice

Currently, no machine learning sepsis prediction model possesses sufficient evidence of generalizability or clinical impact to support widespread deployment in intensive care units. The predominance of internal validation and hourly data limits readiness for real-time decision support. Clinicians should therefore view published models as promising prototypes rather than ready-to-use tools [13, 29, 34].

Until prospective validation demonstrates meaningful outcome improvements, sepsis prediction algorithms should remain investigational. Integration into electronic health record alerts must be accompanied by careful monitoring for alert fatigue and unintended workflow disruptions. This measured approach protects patients while allowing continued refinement of the technology [13, 29, 34].

Implications for policy and regulation

Regulatory bodies such as the FDA and EMA should require external validation and prospective testing before approving any sepsis prediction model for clinical use. Current evidence dominated by internal validation does not meet the threshold for high-risk AI medical devices. Mandatory adherence to TRIPOD and CONSORT-AI standards would further strengthen pre-market evaluation [16, 17].

Policy makers should support the creation of national or international ICU data platforms that facilitate rigorous external validation. Reimbursement decisions for AI-enabled sepsis tools should be linked to demonstrated improvements in patient-centered outcomes rather than AUROC alone. These measures would align regulatory and payment incentives with the need for robust evidence [16, 17].

Conclusion

This systematic review synthesized 35 studies published between 2017 and 2021 on machine learning for early sepsis prediction in adult ICUs. It characterized prediction horizons ranging from 4 to 24 hours, documented the overwhelming reliance on hourly aggregated vital signs, and highlighted the rarity of external or prospective validation. The analysis revealed consistent patterns whereby shorter horizons and internal validation produced higher AUROC values while external testing exposed performance decrements.

The key conclusion is that the evidence base remains dominated by internally validated models using hourly aggregated vital signs and 6-hour prediction horizons. External validation is performed in only a minority of studies, and performance

drops substantially when it is attempted. These findings indicate that methodological rigor has not kept pace with model complexity in the reviewed literature.

Current evidence is therefore insufficient to recommend widespread clinical deployment of any specific sepsis prediction model. More rigorous validation designs and richer data modalities are needed before these tools can reliably improve patient outcomes. The review underscores the gap between promising retrospective results and real-world clinical utility.

Future research should prioritize external validation, high-frequency vital signs, prospective testing, and standardized reporting. If these gaps are addressed, machine learning holds considerable potential to enable truly early sepsis detection and reduce the global burden of this life-threatening condition in intensive care settings.

Acknowledgements

None

Conflict of interest

None

Financial support

None

Ethics statement

None

Received: 13 Apr 2021

Revised: 17 Jun 2021

Accepted: 29 Jul 2021

Published online: 20 January 2022

Rights and permissions

Open Access The author(s) retain copyright. This article is licensed under the Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License. It may be shared and adapted for non-commercial purposes with appropriate attribution, an indication of changes, and distribution of adaptations under the same license. Third-party material may be subject to separate terms identified in its credit line. View the license at <https://creativecommons.org/licenses/by-nc-sa/4.0/>.

References

Reyna MA, Josef CS, Jeter R, Shashikumar SP, Westover MB, et al. Early prediction of sepsis from clinical data: the PhysioNet/Computing in Cardiology Challenge 2019. *Crit Care Med.* 2020;48(2):210-7.

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

Shimabukuro DW, Barton CW, Feldman MD, Mataraso SJ, Das R. Effect of a machine learning-based severe sepsis prediction algorithm on patient survival and hospital length of stay: a randomised clinical trial. *BMJ Open Respir Res.* 2017;4(1):e000234.

Zhou A, Beyah R, Kamaleswaran R. OnAI-comp: an online AI experts competing framework for early sepsis detection. *IEEE/ACM Trans Comput Biol Bioinform.* 2022;19(6):3595-603.

Al-Mualemi BY, Lu L. A deep learning-based sepsis estimation scheme. *IEEE Access.* 2020;9:5442-52.

Rosnati M, Fortuin V. MGP-AttTCN: an interpretable machine learning model for the prediction of sepsis. *PLoS One.* 2021;16(5):e0251248.

Zhang D, Yin C, Hunold KM, Jiang X, Caterino JM, et al. An interpretable deep-learning model for early prediction of sepsis in the emergency department. *Patterns.* 2021;2(2):100196.

Shashikumar SP, Josef CS, Sharma A, Nemati S. DeepAISE – an interpretable and recurrent neural survival model for early prediction of sepsis. *Artif Intell Med.* 2021;113:102036.

Aşuroğlu T, Oğul H. A deep learning approach for sepsis monitoring via severity score estimation. *Comput Methods Programs Biomed.* 2021;198:105816.

Oei SP, van Sloun RJ, van der Ven M, Korsten HH, Mischi M. Towards early sepsis detection from measurements at the general ward through deep learning. *Intell Based Med.* 2021;5:100042.

Rafiei A, Rezaee A, Hajati F, Gheisari S, Golzan M. SSP: early prediction of sepsis using fully connected LSTM-CNN model. *Comput Biol Med.* 2021;128:104110.

Goh KH, Wang L, Yeow AY, Poh H, Li K, et al. Artificial intelligence in sepsis early prediction and diagnosis using unstructured data in healthcare. *Nat Commun.* 2021;12(1):711.

Bedoya AD, Futoma J, Clement ME, Corey K, Brajer N, et al. Machine learning for early detection of sepsis: an internal and temporal validation study. *JAMIA Open.* 2020;3(2):252-60.

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

Yuan KC, Tsai LW, Lee KH, Cheng YW, Hsu SC, et al. The development an artificial intelligence algorithm for early sepsis diagnosis in the intensive care unit. *Int J Med Inform.* 2020;141:104176.

Kok C, Jahmunah V, Oh SL, Zhou X, Gururajan R, et al. Automated prediction of sepsis using temporal convolutional network. *Comput Biol Med.* 2020;127:103957.

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

Kim J, Chang H, Kim D, Jang DH, Park I, et al. Machine learning for prediction of septic shock at initial triage in emergency department. *J Crit Care.* 2020;55:163-70.

Ibrahim ZM, Wu H, Hamoud A, Stappen L, Dobson RJ, et al. On classifying sepsis heterogeneity in the ICU: insight using machine learning. *J Am Med Inform Assoc.* 2020;27(3):437-43.

Fagerström J, Bång M, Wilhelms D, Chew MS. LiSep LSTM: a machine learning algorithm for early detection of septic shock. *Sci Rep.* 2019;9(1):15132.

Kaji DA, Zech JR, Kim JS, Cho SK, Dangayach NS, et al. An attention based deep learning model of clinical events in the intensive care unit. *PLoS One.* 2019;14(2):e0211057.

Giannini HM, Ginestra JC, Chivers C, Draugelis M, Hanish A, et al. A machine learning algorithm to predict severe sepsis and septic shock: development, implementation, and impact on clinical practice. *Crit Care Med.* 2019;47(11):1485-92.

Ginestra JC, Giannini HM, Schweickert WD, Meadows L, Lynch MJ, et al. Clinician perception of a machine learning–based early warning system designed to predict severe sepsis and septic shock. *Crit Care Med.* 2019;47(11):1477-84.

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.* 2021;8:607952.

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

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

Qiu X, Li J, Bonenfant J, Jaroszewski L, Mittal A, et al. Dynamic changes in human single-cell transcriptional signatures during fatal sepsis. *J Leukoc Biol.* 2021;110(6):1253-68.

Khojandi A, Tansakul V, Li X, Koszalinski RS, Paiva W. Prediction of sepsis and in-hospital mortality using electronic health records. *Methods Inf Med.* 2018;57(4):185-93.

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

Burdick H, Pino E, Gabel-Comeau D, Gu C, Roberts J, 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.

Kam HJ, Kim HY. Learning representations for the early detection of sepsis with deep neural networks. *Comput Biol Med.* 2017;89:248-55.

Bauer M, Gerlach H, Vogelmann T, Preissing F, Stiefel J, et al. Mortality in sepsis and septic shock in Europe, North America and Australia between 2009 and 2019—results from a systematic review and meta-analysis. *Crit Care.* 2020;24(1):239.

Napolitano LM. Sepsis 2018: definitions and guideline changes. *Surg Infect.* 2018;19(2):117-25.

Fleuren LM, Klausch TL, Zwager CL, Schoonmade LJ, Guo T, 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.