Clinical Decision Support Approaches for Neonatal Sepsis: A Systematic Review

Leila Shahmoradi[1,2], Shahrbanoo Pahlevanynejad[3,4], Erfan Esmaeeli[1,5], Alireza Banaye Yazdipour[6*] and Usama Mohammed Kasem[7, 8]

[1]Department of Health Information Management and Medical Informatics, School of Allied Medical Sciences, Tehran University of Medical Sciences, Tehran, Iran.

[2]Health Information Management Research Center, Tehran University of Medical Sciences, Tehran, Iran.

[3]Social Determinants of Health Research Center, Semnan University of Medical Sciences, Semnan, Iran.

[4]Department of Health Information Technology, School of Allied Medical Sciences, Semnan University of Medical Sciences, Semnan, Iran.

[5]Students' Scientific Research Center (SSRC), Tehran University of Medical Sciences, Tehran, Iran.

[6]Department of Health Information Technology, School of Paramedical and Rehabilitation Sciences, Mashhad University of Medical Sciences, Mashhad, Iran.

[7]Department of Health Information Management and Medical Informatics, School of Allied Medical Sciences, Tehran University of Medical Sciences, International Campus (TUMS-IC), Tehran, Iran.

[8]Head of Informatics, Planning and Development Sector, Ministry of Health, Gaza, Palestine.

Received: 13 February 2026

Accepted: 9 November 2026

*Corresponding author: alirezayazdipour20@gmail.com

DOI 10.5001/omj.2029.34

Abstract

Objectives: According to the World Health Organization (WHO), neonatal sepsis is the third most common cause of infant mortality. Clinical decision-support approaches have been investigated as potential tools for improving neonatal sepsis recognition and management. The purpose of this systematic review was to identify and synthesize clinical decision-support approaches for the detection, prediction, diagnosis, and management of neonatal sepsis.

Methods: A systematic search was conducted on PubMed, Web of Science, and Scopus between 1 January 2000 and 19 April 2025. Studies were included if they were original research, written in English, conducted in neonatal populations, and focused on diagnostic, predictive, computational monitoring, or clinically implemented decision-support approaches for neonatal sepsis. Quality and risk of bias were appraised with tools matched to each study design: PROBAST+AI, QUADAS-2, EPHPP, QuADS, and MMAT.

Results: Of the 10,617 records identified, 13 studies met the eligibility criteria. The findings suggest that clinical decision-support approaches have potential applications in neonatal sepsis recognition and management. However, most included studies focused on model development, diagnostic performance, or technical feasibility, and evidence of improvements in clinical outcomes remains limited. The included studies involved heterogeneous neonatal populations, including preterm infants of different gestational ages and broader neonatal intensive-care populations. The included approaches comprised various computerized methods, including artificial intelligence and machine-learning techniques, applied to the early detection and management of neonatal sepsis.

Conclusion: This systematic review identified substantial heterogeneity in study design, outcomes, validation approaches, and system characteristics across studies of neonatal sepsis decision support. Although several decision-support approaches showed promising diagnostic or predictive performance, evidence regarding independent external validation, prospective clinical evaluation, and clinical effectiveness remains limited.

Keywords: Neonatal Sepsis, Infants, CDSS, Artificial Intelligence, Information Technology.

Introduction

Neonatal sepsis remains a major cause of morbidity and mortality in newborns. Preterm infants are particularly vulnerable because of physiological immaturity, frequent exposure to invasive procedures, and prolonged hospitalization in neonatal intensive care units (NICUs)(1). Early recognition of neonatal sepsis remains challenging because its clinical signs are often nonspecific and overlap with other neonatal conditions (2). Recent estimates from the Global Burden of Disease (GBD) Study put the global incidence of newborn sepsis at 1.3 million cases per year, , with approximately 203,000 associated deaths annually(3). Neonatal sepsis is a major contributor to neonatal mortality(4). The term "neonatal sepsis" refers to a clinical condition characterized by generalized signs and symptoms induced by pathogen invasion that occurs during this period of life (5).

Neonatal sepsis has long been divided into two categories based on timing: early-onset sepsis (EOS), which appears within the first 72 hours after birth, and late-onset sepsis (LOS), which develops later. This division originally reflected an assumption about how the infection was acquired—vertically from the mother in EOS, and postnatally in LOS (6). That assumption, however, has become difficult to maintain. A review of 80 randomized clinical trials found no fewer than 128 distinct definitions of neonatal sepsis, which means that studies often describe and classify what appears to be the same condition in quite different ways(7).

For EOS, maternal and perinatal factors carry particular weight. A multicenter cohort in Hong Kong, for example, showed that universal screening for group B Streptococcus (GBS) together with intrapartum antibiotic prophylaxis was followed by a fall in EOS incidence, from 3.25 to 2.26 cases per 1,000 live births (8). In contrast, LOS in hospitalized preterm infants is associated with postnatal exposures, including central venous catheters, mechanical ventilation, prolonged parenteral nutrition, and extended hospitalization. The organisms responsible also vary by region. A systematic review and meta-analysis reported that Klebsiella species caused 31.7% of EOS cases in low- and lower-middle-income countries, a pattern that does not fit the traditional picture of GBS as the dominant early pathogen (9). In very preterm infants with LOS in a large US cohort, the two most frequent isolates were coagulase-negative staphylococci (29.3%) and Staphylococcus aureus (23.0%) (10). Taken together, these findings suggest that prevention and treatment strategies developed in high-income settings cannot simply be transferred to other populations. Diagnosing neonatal sepsis at the bedside remains difficult. The clinical signs are often subtle or nonspecific, and conventional laboratory tests do not always provide a clear answer (11). Respiratory manifestations, including tachypnea and respiratory distress, are among the clinical signs evaluated in newborns with suspected EOS, although these manifestations are not specific to neonatal infection (12). In the multicountry NeoOBS study, clinical variables were used to develop severity and recovery scores for predicting mortality among infants with sepsis. Importantly, these scores were designed for prognostic assessment after sepsis diagnosis rather than for the initial diagnosis of neonatal sepsis. The variables chosen for prediction models reflect these differences in population and setting (13). The updated EOS risk calculator incorporates gestational age, maternal temperature, maternal GBS colonization status, duration of membrane rupture, intrapartum antibiotic exposure, and the newborn's clinical condition to estimate sepsis risk among infants born at ≥35 weeks of gestation (14). A separate multicenter study of extremely preterm infants produced a LOS prediction model based on postmenstrual age, ventilator status, apnea, periodic breathing, intermittent hypoxemia, and bradycardia; that model reached an AUROC of 0.783(15). A 2024 systematic review of machine-learning models for neonatal sepsis which excluded studies relying on automated vital-sign data found substantial methodological weaknesses and no external validation across its 19 included studies (16). The gap between reported model performance and demonstrated clinical usefulness is therefore not a minor issue; it is the central problem this review sets out to examine.

Despite improvements in neonatal medicine, neonatal sepsis remains a significant source of morbidity and mortality (17). WHO announced in a report that 75% of neonatal fatalities occur within the first seven days of life, and sepsis is the second most prevalent cause (18). In high-income countries, the incidence of sepsis ranges from 1 to 4 cases per 1000 live births, whereas it ranges from 49 to 170 cases in low- and middle-income countries, with a fatality rate of up to 24% (2). In the United States, the prevalence of early sepsis is around 0.77 cases per 1000 live births, and when only infants over 34 weeks of gestational age are considered, it is around 0.5 cases per 1000 live

births. Furthermore, the implementation of intrapartum antibiotic therapy for pregnant women colonized with Streptococcus agalactiae has led to a significant decrease in the incidence of early neonatal sepsis in the United States of America (USA) (19). These diagnostic challenges have increased interest in computerized methods capable of supporting earlier recognition and more timely clinical decision-making. Recent reviews have also emphasized the need for improved prediction and diagnostic approaches for neonatal sepsis, particularly in high-risk neonatal populations (20).

With the increasing use of electronic health records (EHRs), patient data management systems (PDMS), physiologic monitoring systems, and other forms of digitalization in health care, computerized systems capable of supporting automated sepsis detection have increasingly been developed (21). Clinical decision support systems (CDSSs) can support clinical and diagnostic decision-making, treatment and test ordering, and patient monitoring and triage. . A CDSS is a computerized system or functionality that integrates patient-specific clinical information with clinical knowledge, predefined rules, or data-driven algorithms to generate assessments, predictions, alerts, or recommendations intended to support clinical decision-making. CDSSs may be knowledge-based, such as rule-based expert systems, or non-knowledge-based, including systems incorporating machine-learning approaches (22). In hospital settings, CDSSs can support various clinical activities, including medication safety, antimicrobial stewardship, laboratory testing, and other aspects of clinical management. However, their evaluation should consider not only clinical consequences but also implementation-related and economic outcomes (23). Furthermore, CDSSs may support screening and early disease detection, although their clinical value depends on appropriate evaluation and integration into clinical practice (24).

Importantly, the development or validation of a statistical or artificial intelligence/machine-learning model alone does not necessarily constitute a clinically implemented CDSS. Such models may instead serve as potential components of future decision-support systems (22). CDSS can aid clinicians in making timely and accurate choices despite complex circumstances, thereby reducing the burden on medical staff. EHR-based CDSS may facilitate earlier sepsis recognition and more timely clinical intervention. CDSS use automated sepsis-risk warning tools to notify clinicians of the possibility of sepsis (25). More recent reviews have also identified a growing number of computerized and machine-learning approaches for pediatric and neonatal sepsis, although substantial heterogeneity remains in system design, predictors, validation strategies, and clinical implementation (26, 27).

In recent years, an increasing number of studies have investigated computerized approaches, including CDSS and machine-learning methods, for neonatal sepsis. . Honoré et al. (2023) examined whether a machine-learning model built on high-frequency physiological data and demographic variables could predict late-onset sepsis (LOS) in neonates. Their study was carried out at a single center and included 325 neonates, of whom 20 developed sepsis. The model used a Naïve Bayes classifier trained on heart rate, respiratory rate, and oxygen saturation. It reached an AUROC of 0.82 for predicting sepsis up to 24 hours before clinicians first suspected the infection. It is worth noting, however, that the algorithm was developed and tested on observational data only. It was never integrated into routine care as a decision-support tool, so its performance in real clinical settings remains unknown (28). Li et al. (2020) outlined a protocol for a scoping review that would examine how knowledge-based computerized clinical decision support systems are used and evaluated for early sepsis detection in hospitals. Their proposed scope did include neonatal populations, but the protocol itself does not present the results of a completed synthesis. It therefore cannot serve as evidence of clinical effectiveness, and any reference to it should be treated as a description of planned work rather than a finished finding (27).

The evidence available today spans a heterogeneous set of approaches: diagnostic biomarkers, statistical and machine-learning prediction models, computerized monitoring systems, and CDSS that have been implemented in clinical practice. These approaches differ from one another in their intended clinical use, how they were validated, and how far they have been implemented. Evidence of diagnostic or predictive performance should therefore be distinguished from evidence of successful clinical implementation and improved patient outcomes. A systematic synthesis that accounts for these differences is needed to clarify the current evidence and to identify methodological and implementation gaps in decision support for neonatal sepsis.

Accordingly, this systematic review aimed to identify, classify, and critically synthesize decision-support approaches for neonatal sepsis according to their computational approach, validation stage, clinical implementation, and reported clinical or process outcomes. The review addressed the following questions:

1. What types of diagnostic, predictive, monitoring, and clinical decision-support approaches have been evaluated for neonatal sepsis?
2. What neonatal populations, sepsis phenotypes, data sources, predictors, algorithms, and validation approaches have been studied?
3. To what extent have these approaches progressed from model development and validation to clinical integration, implementation, and evaluation of clinical or process outcomes?
4. What methodological, validation, and implementation limitations remain in the current evidence base?

Methods

This systematic review was written and reported according to the Preferred Reporting Items for Systematic Review and Meta-Analysis (PRISMA) statement (29).

Information sources

We conducted a systematic search in the PubMed, Web of Science (WOS), and Scopus databases between 1 January 2000 and 19 April 2025.A combination of Medical Subject Headings (MeSH) and terms related to clinical decision support system, premature infant, and sepsis were applied in the search strategy. The search strategy was intentionally broad to identify not only clinically implemented CDSS but also potentially relevant diagnostic, predictive, and computational approaches involving neonatal sepsis. The full search strategy is provided in supplementary file [Table S1].

Operational definition and classification of CDSS-related approaches

In this review, a clinical decision support system (CDSS) was defined as a computerized system that processes patient-specific information using clinical knowledge, predefined rules, or data-driven algorithms to generate assessments, alerts, predictions, or recommendations intended to support clinical decision-making. Because clinical decision support may involve several stages of technological development and evaluation, the review distinguished clinically implemented CDSS from diagnostic or predictive methods that could potentially support future CDSS development.

Studies were classified according to their primary objective, computational or diagnostic approach, clinical target, and stage of development or implementation. The categories included:

- Diagnostic biomarker studies and clinical decision rules;
- Statistical and artificial intelligence/machine-learning prediction models;
- Computerized monitoring systems and technical decision-support architectures;
- Clinically implemented alert-based or protocol-based CDSS;

Clinical and process-related outcomes were assessed as a separate dimension across the categories rather than as an independent type of decision-support technology. Prognostic or composite-outcome models indirectly related to neonatal sepsis were identified separately, subject to the review's eligibility criteria.

The relationship of each study to neonatal sepsis was assessed separately, distinguishing sepsis-specific approaches from studies addressing broader neonatal outcomes or infection-related clinical processes.

Studies limited to biomarker evaluation, diagnostic-rule development, prediction-model development, or technical feasibility were not considered evidence of clinically implemented CDSS. Likewise, reported diagnostic or predictive performance was not interpreted as evidence of improved clinical outcomes.

Where studies evaluated broader or composite clinical outcomes, their original target outcomes were explicitly identified, and the reported performance measures were not interpreted as estimates of sepsis-specific diagnostic or predictive performance.

Selection criteria

Studies were eligible if they (1) reported original research, (2) were published in English, (3) included neonatal populations, (4) evaluated approaches relevant to neonatal sepsis focused on evaluating diagnostic biomarkers, clinical decision rules, statistical prediction or prognostic models, artificial intelligence/machine learning, computerized monitoring systems, technical decision support architectures, or clinically implemented CDSS; and (5) were published in or after 2000. Older studies were retained to reflect the developmental trajectory of the field and were interpreted according to their technological and clinical implementation stage rather than as directly equivalent to contemporary systems. Studies were considered relevant when neonatal sepsis was explicitly included in the assessed approach as a target condition, a diagnostic or prognostic predictor, a predefined component of a composite clinical outcome, or a condition addressed by a clinical decision support intervention. Technical monitoring and decision support architectures were also considered when neonatal sepsis was explicitly identified as an intended clinical application. Studies addressing clinical management related to neonatal infection, including antibiotic prescribing and neonatal infection protocols, were included where their relevance to the review objectives could be demonstrated. Studies of diagnostic biomarkers were eligible when they reported diagnostic performance relevant to neonatal sepsis-related clinical decision-making. To account for methodological and technological heterogeneity, included studies were further distinguished according to the role of neonatal sepsis in the assessed approach and the stage of model development, technical implementation, or clinical evaluation. The exclusion criteria were (1) reviews, meta-analysis, dissertation, reports, conference abstracts, letter to the editor, commentaries, systematic review protocol, (2) language other than English, (3) unavailability of full text for data extraction, (4) studies unrelated to the aim of the study.

Study selection

1) Imported all identified studies into the reference manager software EndNote X9 and removed any duplicates; 2) After deduplication, imported the references to the Rayyan platform (30); 3) Three independent researchers (SP, ES, and UMK) evaluated the titles and abstracts of the remaining studies and compared them with the inclusion and exclusion criteria through the Rayyan platform; 4) Retrieved the full-texts of the remaining relevant studies; 5) Three researchers (SP, ES, and UMK) independently screened these full-texts and made a final selection of relevant studies based on the inclusion and exclusion criteria; 6)Disagreements were resolved through discussion and consensus among the reviewers.

Data extraction

Data were extracted from the included studies using a structured form developed in Microsoft Excel. Three researchers (SP, ES, and UMK) carried out the extraction independently, and any disagreements were resolved through discussion until consensus was reached. For each study, we recorded bibliographic details (authors, publication year, journal, and country), the study aims and design, clinical setting, neonatal population, gestational age, sample size, methods, research tools, data sources, predictors, algorithms, target outcomes, and results. We also noted the strengths, limitations, and recommendations reported by the original authors. Sepsis phenotype was classified as early-onset sepsis (EOS), late-onset sepsis (LOS), both EOS and LOS, or neonatal sepsis without a specified onset category, following the definitions used in each original study. We further recorded how sepsis featured in each study, whether as a target condition, a predictor, one component of a composite outcome, or a condition addressed by an intervention. For diagnostic and prediction-model studies, we extracted the evaluation and validation methods that were reported. These included apparent performance, cross-validation, train–test splitting, temporal validation, external validation, and prospective clinical evaluation, where these were applicable. Performance measures were recorded where available, including sensitivity, specificity, accuracy, area under the

receiver operating characteristic curve (AUROC), positive predictive value (PPV), negative predictive value (NPV), calibration, decision thresholds, and false-positive rates. For clinically implemented decision-support systems, we additionally collected information on workflow integration, alert burden, alert fatigue, clinician response, and any reported process or clinical outcomes, such as antibiotic exposure, morbidity, and mortality. Sample sizes and units of analysis were kept distinct where relevant. These included the numbers of infants, monitoring episodes, patient-days, and sepsis events. Information that was not reported in the original studies was distinguished from information that did not apply or had not been evaluated. Limitations and strengths reported by the original authors were recorded separately from the reviewers' own methodological appraisal. The characteristics and findings extracted from each study are presented in Table 1 and Supplementary Tables S2 and S3. The independent risk-of-bias assessment was carried out separately, as described in Section 2-7.

Risk of bias and quality assessment

Three independent researchers (SP, ES, and UMK) assessed the methodological quality, risk of bias, and applicability of the included studies using appraisal tools selected according to the study design and assessment objectives. The Prediction Model Risk of Bias Assessment Tool plus Artificial Intelligence (PROBAST+AI) is designed to evaluate the methodological quality of prediction-model development, the risk of bias in model performance evaluation, and applicability, irrespective of whether conventional statistical or artificial intelligence methods are used (31). PROBAST+AI was applied to eight studies: Helguera-Repetto et al. (2020), Pokhylko et al. (2020), Stanculescu et al. (2014), Alvi et al. (2021), Blue et al. (2022), Shaw et al. (2023), Honoré et al. (2023), and Singh et al. (2019). The assessment covered four domains: participants and data sources, predictors, outcome, and analysis. Model development quality and risk of bias in model performance evaluation were assessed separately, where evaluable, while applicability concerns were assessed in relation to the review question. Distinct models were considered separately, including the two prediction models developed by Blue et al. and the CNN and LSTM-RNN architectures examined by Alvi et al. across two datasets. For Singh et al., the PROBAST+AI assessment was restricted to model development quality because the available reporting was insufficient for a reliable assessment of model performance evaluation. The Quality Assessment of Diagnostic Accuracy Studies-2 (QUADAS-2) is designed to assess risk of bias and applicability in diagnostic accuracy studies across four domains: patient selection, index test, reference standard, and flow and timing (32). QUADAS-2 was applied to Adib et al. (2006) and Ebenebe et al. (2021). For Ebenebe et al., IL-6 alone and IL-6 combined with CRP were assessed separately because the incorporation of CRP into the diagnostic reference definition raised additional concerns for the combination approach. The Effective Public Health Practice Project (EPHPP) Quality Assessment Tool evaluates the methodological quality of quantitative intervention studies across six components: selection bias, study design, confounders, blinding, data collection methods, and withdrawals and drop-outs (33). It was applied to the implementation study by Amoakoh et al. (2020). The global rating was classified as Strong when no component was rated Weak, Moderate when one component was rated Weak, and Weak when two or more components were rated Weak. The Quality Assessment with Diverse Studies (QuADS) tool assesses methodological and reporting quality across heterogeneous study designs using 13 criteria, each scored from 0 to 3 (34). QuADS was applied to the technical and decision-support system studies by Stacey et al. (2007) and Singh et al. (2019). The Mixed Methods Appraisal Tool (MMAT), version 2018, provides design-specific criteria for critically appraising qualitative, quantitative, and mixed-methods studies (35). The quantitative descriptive study criteria were applied to Sheehan et al. (2009), with individual items rated as Yes, No, or Can't tell. Singh et al. (2019) was assessed using both PROBAST+AI and QuADS to address distinct methodological components of the study: prediction-model development quality and the broader methodological and reporting characteristics of the technical system.

Because the appraisal tools assess different methodological dimensions and use different rating systems, their findings were reported separately without calculating a combined quality score. All assessments were conducted independently, and disagreements were resolved through discussion and consensus among the reviewers. Detailed findings are presented in Supplementary Tables S4 (PROBAST+AI), S5 (QUADAS-2), S6 (EPHPP), S7-A (QuADS), and S7-B (MMAT).

Data synthesis and analysis

Due to considerable heterogeneity across study designs, neonatal populations, sepsis phenotypes, prediction targets, and reported outcomes, a narrative synthesis was carried out. Studies were categorized based on their main decision-support approach and their stage of development or clinical implementation. The synthesis of findings was organized around clinical setting, population characteristics, data sources, algorithms, validation methods, predictive or diagnostic performance, methodological quality, and reported clinical or process outcomes. A quantitative meta-analysis was not performed, as the outcomes and evaluation methods lacked sufficient comparability. Study characteristics and methodological appraisal results are presented in the main and supplementary tables.

Descriptive bibliometric analyses were also carried out to examine the geographical distribution and keyword patterns of the included publications. Geographical distribution was illustrated on a map using Microsoft Excel. Keyword co-occurrence analysis was conducted with VOSviewer software (version 1.6.18) to investigate relationships among the keywords of the included publications. Risk of bias due to missing results and certainty of evidence were not formally assessed. Given the substantial heterogeneity in study designs, prediction targets, and clinical outcomes, a structured narrative synthesis was conducted, and study-design-specific appraisal tools were used to evaluate the methodological quality and risk of bias of individual studies.

Results

Search Output

Overall, 10617 relevant articles were identified from the three databases (PubMed, Web of Science, and Scopus.After removing 4,312 duplicates, the remaining 6,305 records were screened by title and abstract. Of these, 6,250 records were excluded, and 55 full-text articles were assessed for eligibility. Following full-text assessment, 42 articles were excluded because they did not meet the inclusion criteria. Finally, 13 articles were included in the systematic review. Figure 1 presents the PRISMA flow diagram illustrating the study selection process.

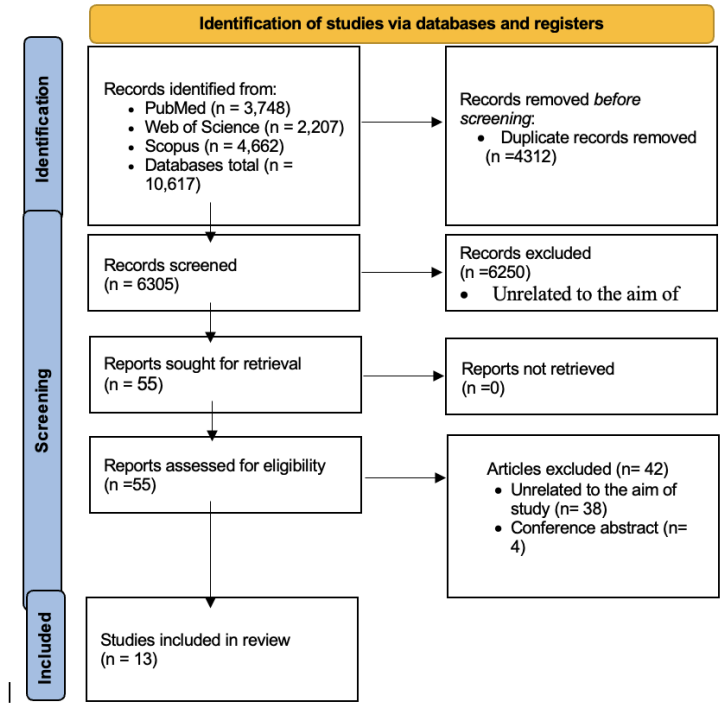

Figure 1. Study selection flow diagram adapted from PRISMA 2020, showing the identification, screening, eligibility assessment, and inclusion of studies

Characteristics of the included studies

Table 1 summarizes the 13 studies included in this review. Grouped by their primary objective, seven studies developed or evaluated prediction models (Stanculescu et al., Pokhylko et al., Helguera-Repetto et al., Alvi et al., Blue et al., Shaw et al., and Honoré et al.), two examined diagnostic biomarkers or clinical decision rules (Adib et al. and Ebenebe et al.), two described monitoring systems or technical prototypes (Stacey et al. and Singh et al.), and two evaluated decision-support systems that had already been deployed in clinical practice (Sheehan et al. and Amoakoh et al.). This grouping helped separate evidence on model development and technical feasibility from evidence on clinical implementation. Clinical setting, neonatal population, gestational-age criteria, sepsis type, and data source varied widely across studies. Ebenebe et al. and Pokhylko et al. enrolled preterm neonates specifically, Stanculescu et al. analyzed monitoring data from predominantly very preterm neonates, and Singh et al. evaluated infants born before 35 weeks of gestation. Other studies included broader neonatal populations or provided only limited gestational-age stratification. Blue et al. focused on small-for-gestational-age infants, while Alvi et al. did not report gestational-age stratification. Early-onset sepsis (EOS) was the focus of Ebenebe et al. and Pokhylko et al., and it was the intended prediction target in Alvi et al., although the EOS-specific applicability of its second dataset was unclear. Stanculescu et al. and Honoré et al. addressed late-onset sepsis (LOS), while Helguera-Repetto et al. covered both EOS and LOS. Adib et al. examined neonatal sepsis without distinguishing between EOS and LOS. In the remaining studies, sepsis formed part of a composite morbidity outcome (Blue et al.) or served as a predictor of mortality (Shaw et al.). Data sources included laboratory biomarkers, maternal and neonatal characteristics, physiological monitoring data, electronic health records, and previously collected datasets. The two studies examining operational clinical decision-support systems reported different types of evidence. Sheehan et al. characterized the frequency and delivery of antibiotic-prescribing alerts within an operational NICU system, whereas Amoakoh et al. evaluated adherence to neonatal care protocols during implementation of an mHealth decision-support intervention. Neither study established improvements in sepsis-related patient outcomes. Neither study demonstrated improvements in sepsis-related patient outcomes. The included studies were published between 2006 and 2023. The geographical distribution of the included studies and the keyword co-occurrence patterns were also examined as supplementary descriptive analyses. The corresponding country distribution map and keyword network and density visualizations are presented in Supplementary Figures S1 and S2, respectively.

Table 1. Characteristics of the 13 studies.

Author, Year	Aim of the study	Decision support category	Research Method	Research Tool	Research Sample (n)	Sepsis type	Algorithm	Result	Key limitations
Adib et al. Iran,2006 (36)	Evaluate neutrophil CD64 for early diagnosis of neonatal sepsis	Combined diagnostic decision-rule study	Observational diagnostic-accuracy study in a NICU.	Flow cytometry (FACS) with CellQuest analysis	77 neonates: 8 culture-confirmed sepsis, 57 suspected sepsis, and 12 healthy controls.	General neonatal sepsis (not stratified)	CD64-based diagnostic classification; no multivariable prediction model	CD64: sensitivity 100%, specificity 92.3%, PPV 88%, and NPV 100%	A: Not explicitly reported. R: Small sample (n=77); selected healthy controls; limited applicability to computerized CDSS.
Stacey et al. Australia,2007 (37)	Develop a multidimensional temporal-abstraction architecture for NICU monitoring.	Technical/ prototype study without clinical effectiveness evaluation	Technical architecture development and prototype testing using stored NICU data.	Java + Java Expert System Shell (JESS).	Prototype tested on stored NICU data; no clinical validation cohort.	Sepsis as an intended application; not clinically evaluated	Rule-based Event Calculus with multidimensional temporal abstraction	Prototype processed 1-Hz data with approximately 1-s reasoning cycles; no clinical accuracy evaluation.	A: Not explicitly reported. R: Technical prototype without a sepsis-specific clinical validation cohort or patient-outcome evaluation.
Sheehan et al USA,2009 (38)	Characterize antibiotic-related prescribing alerts in a neonatal CPOE system.	Clinical implementation and process-related evidence	Descriptive retrospective analysis of operational CPOE alert data	CPOE (Eclipsys XA4.5); operational antibiotic-alert records.	6,863 NICU alerts, including 1,660 antibiotic-related alerts.	Neonatal antibiotic management; not sepsis-specific	Rule-based CPOE alerts; Kuperman framework for alert classification	1,660 of 6,863 alerts (24%) were antibiotic-related; laboratory history 54%, renal dosing 37%, and gestational age 8.8%.	A: Clinician acceptance, utilization, and perceived usefulness of alerts were not established. R: Non-sepsis-specific alerts; prescribing changes and patient outcomes not evaluated.

Stanculescu et al. UK,2014 (39)	Develop a physiological event-based model for early detection of late-onset neonatal sepsis.	Predictive/ model-development study using real-time monitoring data	Retrospective time-series prediction-model development.	Continuous physiologic monitoring with expert event annotation	36 monitoring samples from 24 neonates (18 sepsis and 18 control samples).	LOS	Autoregressive hidden Markov model (AR-HMM)	Sepsis detected in 12/18 samples within 6 h before sampling; in 10/12, at least 3 h earlier.	A: Uncertain sepsis-onset timing; dependence on expert-annotated events. R: Small cohort (24 infants); no independent external or prospective clinical validation.
Ebenebe et al. Germany,2021 (40)	Assess IL-6 and perinatal factors for EOS diagnosis in preterm neonates.	Combined diagnostic decision-rule study	Retrospective nested case-control diagnostic-accuracy study	IL-6 electrochemiluminescence assay (Cobas e 411); statistical analysis in SPSS v24	182 preterm neonates: 67 EOS cases and 115 controls.	EOS	ROC/Youden-based IL-6 cut-off and IL-6–perinatal factor decision rules	IL-6 at 40 ng/L: sensitivity 75.0%, specificity 72.8%, and AUROC 0.804. Combination with perinatal factors increased either sensitivity or specificity, but not overall accuracy.	A: Blood culture was not the sole diagnostic reference; prophylactic antibiotic use in infants born before 32 weeks. R: Potential incorporation bias in IL-6/CRP combination rules.
Singh et al. India,2019 (41)	Develop and technically evaluate the Neo/iNICU bedside monitoring system.	Computerized monitoring/ decision-support system	NICU device-development study with real-time physiological evaluation.	Neo bedside device integrated with the cloud-based iNICU platform	92 infants from two NICUs (GA <35 weeks; birth weight <2,000 g); reported mean GA 31.4 ± 3.6 weeks; reported birth weight 1,528.2 g.	Multiple neonatal morbidity targets; no sepsis-specific evaluation	Real-time physiological monitoring and morbidity scoring; no validated sepsis-specific algorithm	Real-time data acquisition and visualization demonstrated; no sepsis-specific performance estimate.	A: Device-integration challenges, heterogeneous equipment, and unstable network connectivity. R: Limited technical evaluation (n=92); no sepsis-specific validation or patient-outcome assessment.

Amoakoh et al. Netherlands,2020 (42)	Evaluate neonatal protocol adherence during mHealth CDSS implementation.	Clinical implementation and process-related evidence	Repeated cross-sectional medical-record reviews conducted before and during a cluster-randomized implementation trial.	mCDMSS; structured medical-record review checklist.	Four sampled clusters; 27 cord-sepsis records(13 before and 14 during implementation).	Cord sepsis	Protocol-based mHealth decision-support rules	Mean cord-sepsis protocol adherence: intervention 52.0% to 75.0% (p=0.24); control 53.8% to 60.8% (p=0.21).	A: Concurrent improvement activities; incomplete mortality data; unmeasured provider-related factors. R: Four purposively selected clusters; limited cord-sepsis sample (n=27); potential selection bias.
Pokhylko et al. Ukraine 2020 (43)	Identify EOS risk factors and develop a prognostic model for preterm neonates	Predictive/ model-development study	Retrospective cohort study	STATA v11 software	152 preterm neonates (121 EOS cases and 31 comparison infants)	EOS	Stepwise multivariable logistic regression	Final EOS model: AUROC 0.9377, sensitivity 82.2%, specificity 93.55%, PPV 97.98%, and NPV 58%.	A: Not explicitly reported. R: Small retrospective cohort (n=152); stepwise predictor selection; no independent validation.
Helguera-Repetto et al. Mexico ,2020 (44)	Develop an ANN model for neonatal sepsis diagnosis using maternal and neonatal variables.	AI/ML diagnostic model-development study	Single-center retrospective ANN development and internal evaluation.	Graphpad Prism 5 software, MatLab software	238 analyzed neonates: 106 culture-confirmed sepsis (22 EOS, 84 LOS) and 132 non-septic neonates.	EOS and LOS	Artificial neural network (ANN; 25 input variables)	ANN: AUROC 0.944, sensitivity 93.3%, specificity 80.0%, PPV 82.35%, and NPV 92.3%.	A: Small single-center dataset; dichotomized clinical variables; potential unnecessary antibiotic treatment following false-positive predictions. R: Selected study population; no independent external validation.

Notes: *A = author-reported limitations; R = review-identified methodological or applicability concerns. Key strengths summarize documented study-design or methodological features and, where available, strengths reported by the original authors; they do not necessarily indicate clinical effectiveness. Formal methodological quality, risk-of-bias, and applicability assessments are presented separately in Table 3 and Supplementary Tables S4–S7. Sample sizes are reported according to the original units of analysis (e.g., infants, records, monitoring samples, or admissions). Adib et al. reported inconsistent sensitivity and specificity values between the Abstract and the Results section; values from Table 1 follow the main Results and Table 3. Helguera-Repetto et al. reported 236 neonates in the Methods section and 238 in the Results section; Table 1 follows the latter. Shaw et al. reported 1,946 records, although the stated training and test sample sizes total 1,945; Table 1 retains the reported overall total. For Alvi et al., the EOS-specific applicability of Dataset 2, derived from Masino et al., remains uncertain. The performance estimates reported by Blue et al. and Shaw et al. concern composite perinatal morbidity and infant mortality, respectively, rather than sepsis-specific diagnosis.* ***Abbreviations****: AI, artificial intelligence; ANN, artificial neural network; AR-HMM, autoregressive hidden Markov model; AUROC, area under the receiver operating characteristic curve; CDSS, clinical decision support system; CNN, convolutional neural network; CPOE, computerized provider order entry; DNN, deep neural network; EHR, electronic health record; EOS, early-onset sepsis; LSTM-RNN, long short-term memory recurrent neural network; LOS, late-onset sepsis; MDTA, multidimensional temporal abstraction; ML, machine learning; NICU, neonatal intensive care unit; NPV, negative predictive value; PPV, positive predictive value; SGA, small for gestational age.*

Stages of Decision-Support Development and Clinical Implementation

The 13 studies show different points along the development pathway [Table 2]. Seven built or tested prediction models. Two looked at diagnostic biomarkers or decision rules. Two dealt with computerized monitoring systems or technical architectures. Two examined decision-support systems already running in practice. The prediction-model studies mainly reported apparent performance, internal cross-validation, or split-sample evaluation. None confirmed independent external validation of a fixed sepsis-prediction model. The technical studies showed that prototypes and system integration were possible, but they did not show routine clinical use or effectiveness of a validated sepsis-specific CDSS. Of the two operational systems, Sheehan et al. described antibiotic-prescribing alerts in a NICU, while Amoakoh et al. assessed adherence to neonatal care protocols during an mHealth implementation. Neither showed improvements in sepsis-related patient outcomes. On the whole, the evidence supports diagnostic, predictive, and technical feasibility rather than demonstrated clinical effectiveness.

Table 2. Classification of included studies by stage of decision-support development, validation, implementation, and outcome evaluation.

Study	Primary study category	Model development	Validation / technical evaluation	Technical / CDSS integration	Clinical implementation	System / process outcomes	Intervention-related patient outcomes	Key evidence limitation
Adib et al. (2006)	Diagnostic biomarker	NA	Diagnostic accuracy	NA	NA	NE	NE	CD64 evaluation; no computerized CDSS
Stacey et al. (2007)	Technical architecture	NA	Prototype testing	Prototype	NE	NE	NE	No clinical validation or routine deployment
Sheehan et al. (2009)	Operational alert-based CDSS	NA	NA	Yes	Yes	Alert frequency/ delivery only	NE	Prescribing behavior and patient effects not evaluated
Stanculescu et al. (2014)	LOS prediction model	Yes	Internal CV	NE	NE	NE	NE	Retrospective evaluation; no clinical deployment
Ebenebe et al. (2021)	Diagnostic biomarker / rule	NA	Diagnostic accuracy	NA	NA	NE	NE	IL-6 and perinatal factors; no computerized CDSS
Singh et al. (2019)	Integrated monitoring system	Technical scoring component described	Technical evaluation	Yes – monitoring system	Sepsis-specific routine use NR	NE	NE	No validated sepsis-specific clinical intervention

Amoakoh et al. (2020)	Operational mHealth CDSS	NA	NA	Yes	Yes	Protocol adherence evaluated	NE	No statistically significant adherence improvement; causal attribution limited
Pokhylko et al. (2020)	EOS prediction model	Yes	Apparent performance	NE	NE	NE	NE	No independent model validation
Helguera-Repetto et al. (2020)	ANN diagnostic model	Yes	Internal CV + hold-out	NE	NE	NE	NE	No independent external validation
Alvi et al. (2021)	CNN/ LSTM prediction models	Yes	Internal split-sample evaluation; two datasets	NE	NE	NE	NE	Fixed-model external validation not established
Blue et al. (2022)	Composite morbidity models	Yes	Apparent discrimination and calibration	NE	NE	NE	NE	Sepsis was one component of a composite outcome
Shaw et al. (2023)	Mortality prediction / XAI	Yes	Internal split-sample	NE	NE	NE	NE	Mortality, rather than sepsis, was the prediction target
Honoré et al. (2023)	LOS detection model	Yes	Patient-level CV	NE	NE	NE	NE	Offline model evaluation; no clinical deployment

Notes: *NA = not applicable; NE = not evaluated; NR = not reported or insufficiently documented; CV = cross-validation; CDSS = clinical decision support system; EOS = early-onset sepsis; LOS = late-onset sepsis; ANN = artificial neural network; CNN = convolutional neural network; LSTM = long short-term memory; XAI = explainable artificial intelligence. Model-performance evaluation was distinguished from clinical effectiveness. Technical integration or prototype testing was not considered equivalent to routine implementation of a validated sepsis-specific CDSS. Alert frequency and delivery were classified as system activity rather than evidence of changed clinical practice. Intervention-related patient outcomes refer to observed clinical effects of decision-support interventions, not outcomes used as prediction targets. No included study established improvements in sepsis-related patient outcomes attributable to a CDSS intervention.*

Validation and clinical utility

Validation strategies varied substantially across the included studies. Diagnostic biomarker studies generally assessed performance within their study populations, whereas prediction-model studies relied mainly on apparent performance, cross-validation, or internal train–test evaluation. Stanculescu et al. used leave-one-out cross-validation, with nested cross-validation for model tuning. Helguera-Repetto et al. performed five-fold internal cross-validation and evaluated their final model using a held-out test set from the same data source. Honoré et al. used patient-level cross-validation. Alvi et al. evaluated models using two datasets; however, independent external validation of a fixed model was not clearly established. Reported AUROCs for neonatal sepsis prediction included 0.9377 for Pokhylko et al., 0.944 for Helguera-Repetto et al., and approximately 0.82 for Honoré et al. Alvi et al. reported AUROCs of 0.9741 and 0.9915 for CNN and LSTM-RNN models, respectively, in Dataset 1. Ebenebe et al. reported an AUROC of 0.804 for IL-6-based EOS diagnosis. In contrast, Blue et al. reported AUROCs of 0.72 and 0.84 for composite perinatal morbidity, whereas Shaw et al. reported 98.54% accuracy for infant mortality prediction. These estimates reflected different prediction targets and evaluation designs and were therefore not directly comparable. Measures relevant to clinical utility were inconsistently reported. Pokhylko et al. reported an NPV of 58%, whereas Honoré et al. reported low prediction-frame-level PPVs (0.009–0.022), despite an AUROC of approximately 0.82. Calibration assessment was uncommon. Blue et al. compared observed and predicted morbidity rates across risk deciles within the development dataset, while Honoré et al. applied isotonic regression to the training data. Decision thresholds were frequently data-derived or lacked independent validation, and missing-data handling was inconsistently described. No confirmed independent temporal or external validation of a fixed prediction model, or prospective clinical validation of its effectiveness, was identified. Further details are presented in Supplementary Table S2. Overall, evidence regarding model transportability and prospective clinical utility remained limited.

Clinical Implementation and Process-Related Outcomes

Two studies examined decision-support systems that were already in use in clinical practice. Both focused on care processes rather than on what happened to patients. Sheehan et al. documented 6,863 alerts in the NICU, of which 1,660 (24%) were antibiotic-related. Within that group, 54% of alerts concerned laboratory history, 37% renal-impairment dosing, and 8.8% gestational age. The study did not examine whether clinicians accepted or overrode the alerts, whether antibiotic prescribing changed, whether medication errors occurred, or whether patients did better or worse. Amoakoh et al. reported that mean adherence to cord-sepsis protocols increased from 52.0% to 75.0% in the intervention clusters and from 53.8% to 60.8% in the control clusters. Neither change was statistically significant within its own group ($p = 0.24$ for the intervention clusters; $p = 0.21$ for the controls). Because other neonatal health-improvement activities were taking place at the same time, and because of other methodological constraints, the observed changes could not be reliably attributed to the mHealth CDSS on its own. Neither study established that the decision-support systems reduced antibiotic exposure, treatment duration, or sepsis-related patient outcomes. Alert fatigue was also not formally assessed. Evidence from clinical implementation therefore remained limited largely to descriptions of prescribing alerts and tracking of protocol adherence; none of the systems had been shown to change clinical outcomes.

Methodological Quality and Risk of Bias

The methodological appraisals raised concerns in several domains across the included studies [Supplementary Tables S4–S7]. Eight studies with a prediction-model component were assessed using PROBAST+AI; one of these, Singh et al., was also assessed using QuADS. All eight studies received a High overall concern rating for model development quality. In most cases, this reflected limitations in the Analysis domain, including sample size, predictor selection, overfitting, and validation procedures. High concern ratings in the Participants and Data Sources domain were recorded for Helguera-Repetto et al., Stanculescu et al., Honoré et al., and Alvi et al. (Dataset 2). In the Outcome domain, High concern ratings were recorded for Helguera-Repetto et al., Pokhylko et al., and Alvi et al. (Dataset 2). Shaw et al. also received a High concern rating in the Predictors domain. Seven studies had assessable model-performance evaluations, and all seven received a High overall risk-of-bias rating. For Singh et al., performance-evaluation risk of bias was not assessed because the reporting was insufficient. Applicability ratings also varied across the prediction-model studies. High outcome-applicability concern ratings were recorded for Blue et al., Shaw et al., and Singh et al. Their prediction targets were composite perinatal morbidity, infant mortality, and general neonatal morbidity, respectively, rather than sepsis-specific outcomes. Alvi et al. (Dataset 2) also received a High outcome-applicability concern rating because the classification of sepsis observations was uncertain. These applicability ratings were kept separate from the eligibility criteria used in the broader review. Both diagnostic

accuracy studies assessed using QUADAS-2, Adib et al. and Ebenebe et al., received a High overall risk-of-bias rating. Adib et al. also received a High patient-selection applicability concern rating. In Ebenebe et al., the combined IL-6 and CRP rules raised additional concerns about incorporation bias, because CRP formed part of the diagnostic reference definition. The implementation study by Amoakoh et al., assessed using EPHPP, received a Weak global rating, with Weak ratings for selection bias, confounders, and data collection methods. QuADS assessments of Stacey et al. and Singh et al. identified limitations in methodological reporting and sampling. Stacey et al. scored 0 for recruitment information, whereas Singh et al. scored 1 for stakeholder involvement. In the MMAT assessment of Sheehan et al., three quantitative descriptive criteria were rated Can't tell because information on representativeness, measurement appropriateness, and nonresponse or data completeness was insufficient. The appraisal tools addressed different methodological dimensions, so their findings were reported separately and no combined quality score was calculated. Summary findings are presented in Table 3, with detailed assessments in Supplementary Tables S4–S7.

Table 3. Summary of methodological appraisal and risk-of-bias assessment of included studies

Study	Appraisal tool	Overall finding
Adib et al.(36)	QUADAS-2	High RoB
Stacey et al.(37)	QuADS	Criterion-level assessment
Sheehan et al.(38)	MMAT 2018	Criterion-level assessment
Stanculescu et al. (39)	PROBAST+AI	High development concerns; High performance-evaluation RoB
Ebenebe et al.(40)	QUADAS-2	High RoB
Singh et al.(41)	PROBAST+AI; QuADS	High development concerns; performance evaluation not assessed; QuADS criterion-level assessment
Amoakoh et al.(42)	EPHPP	Weak global rating
Pokhylko et al.(43)	PROBAST+AI	High development concerns; High performance-evaluation RoB
Helguera-Repetto et al.(44)	PROBAST+AI	High development concerns; High performance-evaluation RoB
Alvi et al.(45)	PROBAST+AI	High overall concerns across assessed model–dataset combinations
Blue et al.(46)	PROBAST+AI	High overall concerns for both models
Shaw et al.(47)	PROBAST+AI	High development concerns; High performance-evaluation RoB
Honoré et al.(28)	PROBAST+AI	High development concerns; High performance-evaluation RoB

***Notes**: For PROBAST+AI, concerns about model development quality and risk of bias in model performance evaluation were assessed separately. Applicability was evaluated in relation to the intended clinical use of neonatal sepsis prediction or diagnosis. For Singh et al., performance-evaluation risk of bias was not assessed because of insufficient reporting. QuADS and MMAT results were reported at the criterion level without assigning overall quality categories. Detailed assessments are provided in Supplementary Tables S4–S7. **Abbreviations**: EPHPP, Effective Public Health Practice Project; MMAT, Mixed Methods Appraisal Tool; PROBAST+AI, Prediction Model Risk of Bias Assessment Tool plus Artificial Intelligence; QuADS, Quality Assessment with Diverse Studies; QUADAS-2, Quality Assessment of Diagnostic Accuracy Studies-2; RoB, risk of bias psis-related patient outcomes attributable to a CDSS intervention.*

Discussion

This review identified, classified, and critically synthesized decision-support approaches for neonatal sepsis. Thirteen studies were reviewed, and the results showed considerable heterogeneity across clinical populations, sepsis phenotypes, computational approaches, validation strategies, and stages of clinical implementation. Seven

developed or evaluated prediction models, two examined diagnostic biomarkers or decision rules, two investigated monitoring systems or technical architectures, and two evaluated systems already deployed in clinical practice. Diagnostic accuracy, predictive performance, and technical feasibility are not the same as demonstrated clinical benefit, and the included evidence should be read with that distinction in mind.

The studies sit at different points along a decision-support pathway: model development, validation, CDSS integration, clinical implementation, and patient-outcome evaluation. This is a framework for organizing the evidence. Adib et al. (36) and Ebenebe et al. (40), for example, evaluated diagnostic biomarkers without testing a computerized CDSS. Most prediction-model studies assessed algorithm development or performance without showing how the model would function in routine workflows. Stacey et al. (37) and Singh et al. (41) concentrated on technical architecture, physiological monitoring, or system development rather than sepsis-specific clinical effectiveness. The two studies that evaluated clinically implemented systems did not test sepsis-specific prediction models integrated into routine care: Sheehan et al. (38) examined antibiotic-prescribing alerts in a general NICU, while Amoakoh et al. (42) assessed adherence to neonatal care protocols, including cord-sepsis management. Their results speak to implementation and process, not to the clinical effectiveness of a sepsis-specific predictive CDSS.

The reported findings diverge for reasons that are partly clinical and partly methodological. EOS and LOS differ in timing, risk factors, clinical presentation, and relevant predictors. Pokhylko et al. (43) and Ebenebe et al. (40) used maternal, perinatal, and laboratory variables in EOS populations. Stanculescu et al. (39) and Honoré et al. (28) relied heavily on physiological monitoring data for LOS prediction. Gestational-age eligibility ranged from very preterm populations to broader neonatal cohorts. Sepsis definitions, diagnostic reference standards, and outcome-assessment timing also varied, and these differences can shape reported model performance. When reference standards differ, models may be developed against clinically distinct outcome definitions, which limits comparability and transferability.

Not all included studies targeted sepsis specifically. Blue et al. (46) developed models for composite perinatal morbidity, in which sepsis was only one component. Shaw et al. (47) examined infant mortality using sepsis-related clinical variables. The EOS-specific applicability of the second dataset used by Alvi et al. (45) remained unclear. These studies provide useful information about neonatal risk prediction, but their reported performance should not be interpreted as evidence of accurate neonatal sepsis diagnosis. Differences in populations, prediction targets, and reference standards need to be considered when interpreting performance estimates, rather than comparing models solely on numerical accuracy or discrimination.

Validation strategies differed across studies. Stanculescu et al. (39) used leave-one-out cross-validation. Helguera-Repetto et al. (44) evaluated their model with internal cross-validation and a held-out test set from the same source population. Honoré et al. (28) performed patient-level cross-validation. Alvi et al. (45) assessed performance using two datasets, but independent external validation of a fixed model without further development or tuning was not clearly established. Internal validation can offer preliminary evidence, but it cannot show whether performance will hold across different NICUs, patient populations, or periods of care.

Transportability is affected by differences in gestational-age distributions, disease prevalence, monitoring practices, laboratory testing, and clinical workflows. Using prospectively collected data is not the same as prospective clinical validation, because model development and evaluation may still be retrospective and may not influence clinical decisions. Even when multicenter data were available, as in Blue et al., the predicted outcome was not neonatal sepsis specifically. Confirmed independent external validation of sepsis-prediction models is largely absent from this literature.

Good statistical performance may suggest potential clinical usefulness, but it does not prove it. AUROC, sensitivity, and specificity describe certain aspects of model performance, yet they do not fully capture usefulness at specific decision thresholds. Predictive values, calibration, and the balance between false-positive and false-negative predictions also matter. Pokhylko et al. (43) reported an NPV of 58%. Honoré et al. (28) reported low prediction-frame-level PPVs (0.009–0.022) despite an AUROC of approximately 0.82. Predictive values also depend on disease prevalence and the unit of analysis, which limits direct comparisons across studies and populations.

False-positive predictions can lead to unnecessary diagnostic investigations or antibiotic administration. False-negative predictions can delay sepsis recognition and treatment. Helguera-Repetto et al. (44) acknowledged the

possibility of unnecessary antibiotic treatment following false-positive predictions, but the actual clinical consequences of these errors were not prospectively evaluated in the included prediction-model studies.

Calibration and decision-threshold validation were reported inconsistently. Blue et al. assessed agreement between predicted and observed composite morbidity rates, and Honoré et al. (28) applied isotonic regression during model training, but independent calibration performance was not established for most models. Missing-data handling was similarly inconsistent: some studies described approaches for incomplete observations, while others provided too little information. These limitations restrict reproducibility and make it difficult to determine how reliably the models would perform in routine clinical datasets.

Clinical implementation raises considerations that retrospective prediction-model evaluation cannot address. Sheehan et al. (38) examined antibiotic-related alerts generated by an operational NICU prescribing system. This demonstrated that computerized decision support was being delivered within a clinical workflow, but the study did not establish whether clinicians accepted or acted on the alerts, whether prescribing practices improved, or whether patient outcomes changed. The large number of generated alerts also highlights the importance of alert burden, although formal assessment of alert fatigue was not reported.

Amoakoh et al. (42) evaluated adherence to neonatal care protocols during implementation of a mHealth decision-support system. Adherence to cord-sepsis protocols increased numerically in the intervention clusters, but the changes were not statistically significant, and concurrent neonatal health-improvement activities limited causal attribution. Neither implementation study established reductions in antibiotic exposure, treatment duration, sepsis-related morbidity, or mortality. The absence of demonstrated clinical benefit should not be taken as proof that these interventions were ineffective; the available studies simply did not provide sufficiently rigorous outcome evaluations.

The methodological quality assessments support a cautious reading of the findings. PROBAST+AI identified High overall concerns about model development quality in all eight assessed studies and High overall risk of bias in the seven studies with assessable model-performance evaluations. Model-performance evaluation was not assessed for Singh et al. (41) because of insufficient reporting. The main methodological concerns involved analytical methods, sample size, predictor selection, outcome assessment, and validation procedures.

The diagnostic accuracy studies assessed with QUADAS-2 also received High overall risk-of-bias ratings, particularly because of concerns about patient selection and diagnostic assessment. The implementation study by Amoakoh et al. (42) received a Weak global EPHPP rating, reflecting limitations related to selection bias, confounding, and data collection methods. Reported performance should therefore be read alongside methodological quality rather than treated as evidence of reliable or transferable clinical benefit.

The included studies also span a broad publication period. Earlier studies investigated diagnostic biomarkers, computerized monitoring approaches, and technical decision-support architectures. More recent investigations increasingly incorporated statistical prediction and machine-learning methods. These approaches represent different stages of technological development and are not equivalent in computational capabilities, clinical integration, or evaluation methods. Including them provides historical context for the development of neonatal sepsis-related decision support. The findings are more informative when interpreted according to the clinical purpose and developmental stage of each approach, rather than through direct comparisons of technologies developed under different conditions.

Future research should move from retrospective model development to rigorous clinical evaluation. Clear definitions of EOS and LOS, appropriate gestational-age stratification, and consistent outcome definitions would improve interpretability and comparability across studies. Prediction models should undergo independent temporal and external validation, with transparent reporting of calibration, predictive values, clinically relevant decision thresholds, missing-data handling, and false-positive burden. Implementation studies should evaluate integration into clinical workflows, clinician acceptance, and responses to alerts, usability, and alert fatigue.

Prospective comparative evaluations should also investigate clinically meaningful outcomes, including time to sepsis recognition, antibiotic initiation and duration, unnecessary antibiotic exposure, morbidity, and mortality.

Reporting guidance such as TRIPOD+AI (48) and DECIDE-AI (49) may support more transparent evaluation of clinical prediction models and early-stage AI-driven decision-support interventions, respectively. Such evidence is needed to determine whether promising diagnostic and computational approaches actually translate into measurable benefits in neonatal care.

Strengths and Limitations of the Review

A key strength of this review was the explicit distinction between diagnostic biomarkers, prediction models, technical monitoring systems, and clinically implemented decision-support interventions. This classification allowed the evidence to be interpreted according to its stage of development, validation, integration, and clinical evaluation. The use of design-appropriate methodological appraisal tools also supported separate evaluation of prediction-model studies, diagnostic accuracy studies, technical systems, and clinical implementation research. Several limitations should also be noted. Heterogeneity in study populations, sepsis definitions, computational approaches, validation strategies, and reported outcomes precluded meaningful quantitative pooling and limited direct comparisons across studies. The broad eligibility criteria allowed inclusion of studies in which neonatal sepsis was an indirect prediction target, a component of a broader clinical outcome, or a condition addressed by a general neonatal decision-support intervention, which reduced the specificity of conclusions regarding sepsis-focused CDSS. Only two studies evaluated clinically implemented decision-support systems, limiting conclusions about clinical effectiveness. Methodological weaknesses and incomplete reporting in several primary studies restricted the assessment of model transportability, clinical utility, and patient outcomes. Finally, the review was restricted to English-language publications, and the search included studies available through 19 April 2025; relevant studies outside the searched sources or published after that date may not have been captured. These limitations should be considered when interpreting the findings and their applicability to contemporary neonatal care.

Conclusion

The decision-support approaches reviewed in this article, encompassing diagnostic, predictive, technical, and clinical applications, were highly heterogeneous. Some showed promising diagnostic or predictive performance, while technical studies demonstrated the feasibility of continuous neonatal monitoring and automated data processing in NICU settings. However, most prediction models underwent internal validation only. None of the included studies clearly demonstrated independent external validation of a fixed sepsis-prediction model, and evidence of the integration of such models into routine clinical workflows was limited. The two studies evaluating clinical decision-support systems in practice did not test sepsis-specific prediction models. One examined antibiotic-prescribing alerts, while the other assessed adherence to neonatal care protocols. Neither study established improvements in sepsis-related patient outcomes. Methodological limitations and high risk-of-bias ratings further limited confidence in the reported performance estimates. Future research should focus on independent temporal and external validation, clinical integration, and prospective evaluation of clinically relevant outcomes, including sepsis recognition, antibiotic use, and neonatal morbidity and mortality. Until such evidence is available, the clinical effectiveness of decision-support approaches for neonatal sepsis cannot be considered established.

Data availability statement

All data extracted and analyzed in this systematic review are presented in the manuscript and Supplementary Materials. The original studies are available through their respective published sources.

Ethics statements

Patient consent for publication

Not applicable.

Ethics approval

Not applicable.

Competing interests

The authors declare no competing interests.

Funding

This study did not receive funding from any sources.

Author's contributions

LSH, SP and UMK conceived the idea for the review. SP, ES, and UMK were involved in the study selection, quality assessment, and data extraction. ES, ABY, and UMK conducted the data analysis and synthesis. LSH, SP, ES, ABY, and UMK wrote the first draft of the manuscript. All authors reviewed the manuscript, contributed to critical changes, and approved the final version of the manuscript for submission.

Acknowledgements

The authors thank all the participants in the study.

Declaration of AI Use

During the preparation of this manuscript, the authors used Grammarly and ChatGPT-4 solely for language editing and grammar improvement. The tool was not used to generate scientific content, interpret data, or draw conclusions

Author's information

First corresponding author: Alireza Banaye Yazdipour, Email: yazdipoura@mums.ac.ir, alirezayazdipour20@gmail.com; Orcid ID: 0000-0001-9284-2640.

Second corresponding author: Usama Mohammed Kasem, Email: ukasem@moh.gov.ps

Supplementary Materials

The full search strategy is provided in Supplementary Table S1.
The validation methods and performance measures are provided in Supplementary Table S2.
The study characteristics, decision-support approaches, implementation status, and clinical outcomes are provided in Supplementary Table S3.
The PROBAST+AI assessments are provided in Supplementary Table S4.
The QUADAS-2 assessments are provided in Supplementary Table S5.
The EPHPP quality assessment is provided in Supplementary Table S6.
The QuADS assessments are provided in Supplementary Table S7-A.
The MMAT assessment is provided in Supplementary Table S7-B.
The geographical distribution of the included studies is shown in Supplementary Figure S1.
The keyword co-occurrence and density visualizations are shown in Supplementary Figure S2.
The completed PRISMA 2020 Checklist is provided as a separate supplementary file.

References

1. Hu YQ, Xue SF, Yang RZ, Bao YW, Zhang H, Hu J, et al. Risk factors for healthcare-associated infection and colonization of Serratia marcescens in neonates: a systematic review and meta-analysis. Antimicrob Resist Infect Control. 2025;15(1):4.
2. Celik IH, Hanna M, Canpolat FE, Pammi M. Diagnosis of neonatal sepsis: the past, present and future. Pediatric research. 2022;91(2):337-50.
3. Fleischmann C, Reichert F, Cassini A, Horner R, Harder T, Markwart R, et al. Global incidence and mortality of neonatal sepsis: a systematic review and meta-analysis. Archives of Disease in Childhood. 2021;106(8):745-52.
4. Wattal C, Kler N, Oberoi J, Fursule A, Kumar A, Thakur A. Neonatal sepsis: mortality and morbidity in neonatal sepsis due to multidrug-resistant (MDR) organisms: part 1. The Indian Journal of Pediatrics. 2020;87:117-21.
5. Kim F, Polin RA, Hooven TA. Neonatal sepsis. bmj. 2020;371.
6. Russell N, Barday M, Okomo U, Dramowski A, Sharland M, Bekker A. Early-versus late-onset sepsis in neonates - time to shift the paradigm? Clin Microbiol Infect. 2024;30(1):38-43.
7. Hayes R, Hartnett J, Semova G, Murray C, Murphy K, Carroll L, et al. Neonatal sepsis definitions from randomised clinical trials. Pediatric Research. 2023;93(5):1141-8.

8. Chan YTV, Lau SYF, Hui SYA, Ma T, Kong CW, Kwong LT, et al. Incidence of neonatal sepsis after universal antenatal culture-based screening of group B streptococcus and intrapartum antibiotics: A multicentre retrospective cohort study. BJOG: An International Journal of Obstetrics & Gynaecology. 2023;130(1):24-31.
9. Harrison ML, Dickson BFR, Sharland M, Williams PCM. Beyond Early- and Late-onset Neonatal Sepsis Definitions: What are the Current Causes of Neonatal Sepsis Globally? A Systematic Review and Meta-analysis of the Evidence. Pediatric Infectious Disease Journal. 2024;43(12):1182-90.
10. Flannery DD, Edwards EM, Coggins SA, Horbar JD, Puopolo KM. Late-Onset Sepsis Among Very Preterm Infants. Pediatrics. 2022;150(6):e2022058813.
11. Strunk T, Molloy EJ, Mishra A, Bhutta ZA. Neonatal bacterial sepsis. The Lancet. 2024;404(10449):277-93.
12. Fung A, Shafiq Y, Driker S, Rees CA, Mediratta RP, Rosenberg R, et al. Diagnostic Accuracy of Clinical Sign Algorithms to Identify Sepsis in Young Infants Aged 0 to 59 Days: A Systematic Review and Meta-analysis. Pediatrics. 2024;154(Supplement 1):e2024066588D.
13. Russell NJ, Stöhr W, Plakkal N, Cook A, Berkley JA, Adhisivam B, et al. Patterns of antibiotic use, pathogens, and prediction of mortality in hospitalized neonates and young infants with sepsis: A global neonatal sepsis observational cohort study (NeoOBS). PLOS Medicine. 2023;20(6):e1004179.
14. Kuzniewicz MW, Escobar GJ, Forquer H, Li S, Shu D, Kipnis P, et al. Update to the Neonatal Early-Onset Sepsis Calculator Utilizing a Contemporary Cohort. Pediatrics. 2024;154(4):e2023065267.
15. The Prematurity-Related Ventilatory Control I, Kausch SL, Lake DE, Di Fiore JM, Weese-Mayer DE, Claure N, et al. Apnea, Intermittent Hypoxemia, and Bradycardia Events Predict Late-Onset Sepsis in Infants Born Extremely Preterm. The Journal of Pediatrics. 2024;271.
16. Kainth D, Prakash S, Sankar MJ. Diagnostic Performance of Machine Learning-based Models in Neonatal Sepsis: A Systematic Review. Pediatric Infectious Disease Journal. 2024;43(9):889-901.
17. Dong Y, Basmaci R, Titomanlio L, Sun B, Mercier J-C. Neonatal sepsis: within and beyond China. Chinese medical journal. 2020;133(18):2219-28.
18. Organization WH. World health statistics 2018: monitoring health for the SDGs, sustainable development goals: World Health Organization; 2018.
19. Procianoy RS, Silveira RC. The challenges of neonatal sepsis management. Jornal de pediatria. 2020;96:80-6.
20. Persad E, Jost K, Honoré A, Forsberg D, Coste K, Olsson H, et al. Neonatal sepsis prediction through clinical decision support algorithms: A systematic review. Acta Paediatr. 2021;110(12):3201-26.
21. Ackermann K, Baker J, Festa M, McMullan B, Westbrook J, Li L. Computerized Clinical Decision Support Systems for the Early Detection of Sepsis Among Pediatric, Neonatal, and Maternal Inpatients: Scoping Review. JMIR Med Inform. 2022;10(5):e35061.
22. Sutton RT, Pincock D, Baumgart DC, Sadowski DC, Fedorak RN, Kroeker KI. An overview of clinical decision support systems: benefits, risks, and strategies for success. NPJ Digit Med. 2020;3:17.
23. White NM, Carter HE, Kularatna S, Borg DN, Brain DC, Tariq A, et al. Evaluating the costs and consequences of computerized clinical decision support systems in hospitals: a scoping review and recommendations for future practice. Journal of the American Medical Informatics Association. 2023;30(6):1205-18.
24. Schäfer H, Lajmi N, Valente P, Pedrioli A, Cigoianu D, Hoehne B, et al. The Value of Clinical Decision Support in Healthcare: A Focus on Screening and Early Detection. Diagnostics. 2025;15(5):648.
25. Joshi M, Ashrafian H, Arora S, Khan S, Cooke G, Darzi A. Digital alerting and outcomes in patients with sepsis: systematic review and meta-analysis. Journal of medical Internet research. 2019;21(12):e15166.
26. O'Sullivan C, Tsai DH, Wu IC, Boselli E, Hughes C, Padmanabhan D, et al. Machine learning applications on neonatal sepsis treatment: a scoping review. BMC Infect Dis. 2023;23(1):441.
27. Li L, Ackermann K, Baker J, Westbrook J. Use and Evaluation of Computerized Clinical Decision Support Systems for Early Detection of Sepsis in Hospitals: Protocol for a Scoping Review. JMIR Res Protoc. 2020;9(11):e24899.
28. Honoré A, Forsberg D, Adolphson K, Chatterjee S, Jost K, Herlenius E. Vital sign-based detection of sepsis in neonates using machine learning. Acta Paediatrica. 2023;112(4):686-96.
29. Page MJ, McKenzie JE, Bossuyt PM, Boutron I, Hoffmann TC, Mulrow CD, et al. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. BMJ. 2021;372:n71.
30. Ouzzani M, Hammady H, Fedorowicz Z, Elmagarmid A. Rayyan—a web and mobile app for systematic reviews. Systematic Reviews. 2016;5(1):210.
31. Moons KGM, Damen JAA, Kaul T, Hooft L, Andaur Navarro C, Dhiman P, et al. PROBAST+AI: an updated quality, risk of bias, and applicability assessment tool for prediction models using regression or artificial intelligence methods. BMJ. 2025;388:e082505.
32. Whiting PF, Rutjes AWS, Westwood ME, Mallett S, Deeks JJ, Reitsma JB, et al. QUADAS-2: A Revised Tool for the Quality Assessment of Diagnostic Accuracy Studies. Annals of Internal Medicine. 2011;155(8):529-36.
33. Thomas BH, Ciliska D, Dobbins M, Micucci S. A Process for Systematically Reviewing the Literature: Providing the Research Evidence for Public Health Nursing Interventions. Worldviews on Evidence-Based Nursing. 2004;1(3):176-84.
34. Harrison R, Jones B, Gardner P, Lawton R. Quality assessment with diverse studies (QuADS): an appraisal tool for methodological and reporting quality in systematic reviews of mixed- or multi-method studies. BMC Health Services Research. 2021;21(1):144.
35. Hong QN, Fàbregues S, Bartlett G, Boardman F, Cargo M, Dagenais P, et al. The Mixed Methods Appraisal Tool (MMAT) version 2018 for information professionals and researchers. Education for Information. 2018;34(4):285-91.
36. Adib M, Navaei F, Oreizi F, Saheb-Fosoul F, Ostadi V. Evaluation of CD64 expression on peripheral blood neutrophils for early detection of neonatal sepsis. Iranian Journal of Immunology. 2006;3(1):9-14.
37. Stacey M, McGregor C, Tracy M. An architecture for multi-dimensional temporal abstraction and its application to support neonatal intensive care. Annu Int Conf IEEE Eng Med Biol Soc. 2007;2007:3752-6.
38. Sheehan B, Chused A, Graham Iii PL, Stetson P, Currie L. Frequency and types of alerts for antibiotic prescribing in a neonatal ICU. IOS Press; 2009. p. 521-5.
39. Stanculescu I, Williams CK, Freer Y. Autoregressive hidden Markov models for the early detection of neonatal sepsis. IEEE J Biomed Health Inform. 2014;18(5):1560-70.
40. Ebenebe CU, Hesse F, Blohm ME, Jung R, Kunzmann S, Singer D. Diagnostic accuracy of interleukin-6 for early-onset sepsis in preterm neonates. Journal of Maternal-Fetal and Neonatal Medicine. 2021;34(2):253-8.
41. Singh H, Kaur R, Gangadharan A, Pandey AK, Manur A, Sun Y, et al. Neo-Bedside monitoring Device for Integrated Neonatal Intensive Care Unit (iNICU). Ieee Access. 2019;7:7803-13.

42. Amoakoh HB, Klipstein-Grobusch K, Agyepong IA, Amoakoh-Coleman M, Kayode GA, Reitsma JB, et al. Can an mHealth clinical decision-making support system improve adherence to neonatal healthcare protocols in a low-resource setting? BMC Pediatr. 2020;20(1):534.
43. Pokhylko V, Cherniavska Y, Adamchuk N, Tsvirenko S, Klimchuk Y. CLINICAL PREDICTION OF EARLY ONSET SEPSIS IN PRETERM NEONATES. Wiadomosci lekarskie (Warsaw, Poland : 1960). 2020;73(10):2219-23.
44. Helguera-Repetto AC, Soto-Ramírez MD, Villavicencio-Carrisoza O, Yong-Mendoza S, Yong-Mendoza A, León-Juárez M, et al. Neonatal Sepsis Diagnosis Decision-Making Based on Artificial Neural Networks. Frontiers in Pediatrics. 2020;8.
45. Alvi RH, Rahman MH, Khan AAS, Rahman RM. Deep learning approach on tabular data to predict early-onset neonatal sepsis. Journal of Information and Telecommunication. 2021;5(2):226-46.
46. Blue NR, Allshouse AA, Grobman WA, Day RC, Haas DM, Simhan HN, et al. Developing a predictive model for perinatal morbidity among small for gestational age infants. Journal of Maternal-Fetal & Neonatal Medicine. 2022;35(25):8462-71.
47. Shaw P, Pachpor K, Sankaranarayanan S. Explainable AI Enabled Infant Mortality Prediction Based on Neonatal Sepsis. Computer Systems Science and Engineering. 2023;44(1):311--25.
48. Collins GS, Moons KGM, Dhiman P, Riley RD, Beam AL, Van Calster B, et al. TRIPOD+AI statement: updated guidance for reporting clinical prediction models that use regression or machine learning methods. BMJ. 2024;385:e078378.
49. Vasey B, Nagendran M, Campbell B, Clifton DA, Collins GS, Denaxas S, et al. Reporting guideline for the early stage clinical evaluation of decision support systems driven by artificial intelligence: DECIDE-AI. BMJ. 2022;377:e070904.