JAMIA Open, 2025, **8(2)**, ooaf015
https://doi.org/10.1093/jamiaopen/ooaf015
Research and Applications

Research and Applications

Aligning prediction models with clinical information needs: infant sepsis case study

Lusha Cao, PhD[1,*], Aaron J. Masino, PhD[2,3,4], Mary Catherine Harris, MD[5,6], Lyle H. Ungar, PhD[7], Gerald Shaeffer, MHI[1], Alexander Fidel, MSI[1], Elease McLaurin, PhD[8], Lakshmi Srinivasan, MD[5,6], Dean J. Karavite, MSI[1], Robert W. Grundmeier, MD[1,6]

[1]Department of Biomedical and Health Informatics, Children's Hospital of Philadelphia, Philadelphia, PA 19146, United States, [2]School of Computing, Clemson University, Clemson, SC 29634, United States, [3]Center for Human Genetics, Clemson University, Clemson, SC 29634, United States, [4]School of Health Research, Clemson University, Clemson, SC 29634, United States, [5]Division of Neonatology, Children's Hospital of Philadelphia, Philadelphia, PA 19104, United States, [6]Perelman School of Medicine at the University of Pennsylvania, Philadelphia, PA 19104, United States, [7]Department of Computer and Information Science, University of Pennsylvania, Philadelphia, PA 19104, United States, [8]School of Medicine, Emory University, Atlanta, GA 30322, United States

*Corresponding author: Lusha Cao, PhD, Department of Biomedical and Health Informatics, Children's Hospital of Philadelphia, 2716 South Street, Philadelphia, PA 19146, United States (caol1@chop.edu).

Abstract

Objective: Sepsis recognition among infants in the Neonatal Intensive Care Unit (NICU) is challenging and delays in recognition can result in devastating consequences. Although predictive models may improve sepsis outcomes, clinical adoption has been limited. Our focus was to align model behavior with clinician information needs by developing a machine learning (ML) pipeline with two components: (1) a model to predict baseline sepsis risk and (2) a model to detect evolving (dynamic) sepsis risk due to physiologic changes. We then compared the performance of this two-component pipeline to a single model that combines all features reflecting both baseline risk and evolving risk.

Materials and Methods: We developed prediction models (two-stage pipeline and a single model) using logistic regression and XGBoost trained on electronic healthcare record data of an NICU cohort (1706 observations from 1094 patients, with a 1:1 ratio of cases to controls). We used nested 10-fold cross-validation to evaluate model performance on predictions made 1 h (T_{-1}) before actual clinical recognition.

Results: The single model (XGBoost) achieved the best performance with a sensitivity of 0.77 (0.74, 0.80), specificity of 0.83 (0.80, 0.85), and positive predictive value (PPV) of 0.82 (0.79, 0.84), at 1 h prior to clinical sepsis recognition (T_{-1}). The pipeline model (XGBoost) achieved a sensitivity of 0.72 (0.69, 0.75), specificity of 0.84 (0.82, 0.87), and PPV of 0.82 (0.80, 0.85) at T_{-1}.

Discussion: Our findings highlight the challenges of aligning machine learning with NICU clinical decision-making processes. The two-stage pipeline, designed to mirror clinicians' reasoning, underperformed compared to the single model. Future work should explore integrating continuous physiological data to enhance real-time risk assessment.

Conclusion: Although a pipeline model that separately estimates baseline and dynamic sepsis risk aligns with clinical information needs, at similar levels of specificity the observed sensitivity of the pipeline is inferior to that of a single model. Additional research is needed to better align model outputs with clinician information needs.

Lay Summary

Recognizing sepsis in infants in the Neonatal Intensive Care Unit (NICU) is a critical challenge, as delays in diagnosis can have devastating consequences. Predictive machine learning models can potentially improve early detection, but their use in clinical settings remains limited. Our research aimed to develop and evaluate a machine learning pipeline designed to align better with the needs of clinicians. This pipeline consists of 2 models: one to estimate baseline sepsis risk and another to detect dynamic risk changes based on infants' physiological data. We compared this 2-stage approach to a single, comprehensive model that integrates all relevant data. Our results show that the single model outperformed the 2-stage pipeline, achieving higher sensitivity while maintaining similar specificity, suggesting that the single model is more effective at identifying infants with sepsis 1 h before clinical recognition. However, the 2-stage pipeline may better address clinicians' information needs by separating baseline and evolving risk. Future efforts should focus on refining predictive tools to align with clinicians' needs, ultimately improving sepsis detection and outcomes for vulnerable infants in the NICU.

Key words: sepsis; Newborn Intensive Care Units; machine learning; clinical decision support.

Received: December 20, 2024; **Revised:** February 6, 2025; **Editorial Decision:** February 10, 2025; **Accepted:** February 11, 2025

Introduction

Sepsis among infants in the Neonatal Intensive Care Unit (NICU) is challenging to recognize, and delays in diagnosis can result in devastating consequences such as high mortality, chronic lung disease, and neurodevelopmental impairment.[1–5] The presentation of infant sepsis is heterogenous, sometimes presenting as fulminant disease and other times in an indolent manner making timely detection and potentially life-saving treatment difficult.[6,7] While delays in sepsis recognition increase mortality across all age groups, our recent work demonstrates that each 30-min delay in administration of antibiotics is associated with an alarming 1.4-fold increase in the odds of mortality in the NICU.[7,8]

Given the poor outcomes and financial burdens caused by delays in sepsis recognition, our team has utilized machine learning (ML) methods to support earlier recognition of sepsis. In our prior work, we developed machine learning models using readily available electronic health record (EHR) data to predict sepsis 4 h prior to clinical recognition.[9] Recently, additional ML models have been derived to predict pediatric sepsis in general wards,[10] emergency department (ED),[11–13] pediatric intensive care unit (PICU),[14,15] and NICU[16] settings. In the context of sepsis among very low birth weight infants, vital sign data from bedside monitors such as heart rate and pulse oximetry characteristics have demonstrated excellent prediction accuracy.[17,18] Implementation of a prediction rule based on heart rate characteristics has been shown to reduce mortality among very low birth weight infants.[19] However, model adoption has been hindered by the tension between clinical interpretability and algorithm performance challenges such as high false alarm rates at acceptable levels of sensitivity.[20–22] Models that achieve the highest accuracy tend to be more complex and less interpretable, whereas models more aligned with clinician expectations may be simpler to interpret but less accurate.[23,24]

To better support clinical information needs and motivated by interviews with clinician and nurse participants,[25] we designed a 2-step ML pipeline combining baseline and diagnostic assessments. The baseline risk model is intended to answer the question, "might this infant become infected?" while the diagnostic model is intended to answer the question, "is this infant currently infected?" For comparison, we also derived a single model from data available at the time of sepsis evaluation.

Our goal is to support future clinical decision support (CDS) efforts by developing a sepsis recognition model that best balances predictive performance with clinician expectations and utility. Our long-term aim is to enhance the detection and treatment of infant sepsis, ultimately improving clinical outcomes.

Methods

Data sources and setting

Our study leveraged data from the neonatal sepsis registry in the NICU at the Children's Hospital of Philadelphia (CHOP) from January 2013 to September 2019.[9] The CHOP NICU is a 100-bed quaternary referral unit treating outborn and inborn infants with complex medical and surgical conditions. Our neonatal sepsis registry contains detailed data from the electronic health record for all infants admitted to the CHOP NICU. The data are stored in an adapted version of the PCORnet common data model[26] augmented with manually adjudicated sepsis evaluation data. Consistent with our prior work, we defined sepsis evaluations as culture positive if the evaluation yielded a positive bacterial blood culture and clinically positive if cultures were negative, but antibiotics were administered for at least 5 days.[9] We selected the 5 days cutoff point as it is a treatment standard that has been used to define culture negative sepsis in prior multicenter neonatal studies.[27–30] Episodes of viral or fungal disease among our cohort were not analyzed further (Table 1).

The sepsis registry documents the precise time at which clinical evaluation for potential infection occurs, which is subsequently referred to as "clinical recognition" of a potential infection. However, the actual infectious process may have begun hours earlier, prior to clinical recognition, depending on the virulence of the infectious pathogen and the clinical signs demonstrated by the infant.

Our goal was to develop models that can support ongoing monitoring for the onset of sepsis among infants that arises after the time of NICU admission. We initiated our observation period on the third day post-admission, which excludes sepsis that was present on arrival and guarantees the inclusion of at least 2 days of baseline data for each infant. We excluded sepsis episodes that started on or after the infant's first birthday. If an infant was transferred out of the unit and then readmitted after an absence exceeding 4 h, this was regarded as a new admission. There is often some "change in status" related to lengthy departures from the NICU (eg, major surgical procedure or a transfer to a regular floor due to improving health and subsequent deterioration). Thus, a fresh 2-day baseline data collection period was mandated before the resumption of the observation period and corresponding inclusion of sepsis episodes in our evaluation.

All analyses were performed using Python version 3.9.18 and Scikit-learn version 1.4.1.[31] Our analyses were approved the CHOP's Institutional Review Board and a waiver of consent was granted.

Prediction features and feature engineering

Our models (baseline, diagnostic, and single) utilize features identified as important by clinicians and nurses based on our prior research.[25] These features include demographic characteristics, clinical signs and symptoms, the presence of indwelling hardware and artificial airways, vital signs, and comorbid conditions. Note, our 3 models each use distinct subsets of these variables (Table S1).

To ensure that the models can be easily integrated into clinical practice, we limited the features to those routinely collected and available in typical EHR workflows, which resulted in a final set of 28 candidate prediction features. We excluded lab results to avoid potential bias introduced by the variability in lab ordering patterns across sites and over time. For our diagnostic model (described further below in the section on model training), we selected 14 features out of our set of 28 features derived from vital signs and measurements. These dynamic features were selected based on (1) their potential to reflect short-term changes in clinical status and (2) their availability in the EHR in a timely fashion regardless of clinical workflow or variability in documentation practices.

Among these dynamic features and consistent with our original model derivation efforts, we included physiologically meaningful binary thresholds and vital sign difference

Table 1. Patient characteristics.

	Total	Infants evaluated but found to be uninfected	Infants never evaluated for sepsis and had no clinical suspicion	Infants with at least one sepsis episode[a]	Infants evaluated for infection and found to have viral or fungal etiology[b]
Patient, no.	1094	192	266	617	19
Sex, no. (%)					
Female	452 (41.3)	68 (35.4)	110 (41.4)	266 (43.1)	8 (42.1)
Male	642 (58.7)	124 (64.6)	156 (58.6)	351 (56.9)	11 (57.9)
Race, no. (%)					
American Indian or Alaska native	3 (0.3)	0 (0.0)	1 (0.4)	2 (0.3)	0 (0.0)
Asian	34 (3.1)	6 (3.1)	6 (2.3)	22 (3.6)	0 (0.0)
Black or African American	261 (23.9)	43 (22.4)	63 (23.7)	146 (23.7)	9 (47.4)
Native Hawaiian or other Pacific Islander	2 (0.2)	0 (0.0)	1 (0.4)	1 (0.2)	0 (0.0)
White	478 (43.7)	90 (46.9)	116 (43.6)	268 (43.4)	4 (21.1)
Multiple race	51 (4.7)	6 (3.1)	14 (5.3)	30 (4.9)	1 (5.3)
No information	10 (0.9)	3 (1.6)	3 (1.1)	4 (0.6)	0 (0.0)
Unknown	255 (23.3)	44 (22.9)	62 (23.3)	144 (23.3)	5 (26.3)
Ethnicity, no. (%)					
Hispanic or Latino	153 (14.0)	26 (13.5)	30 (11.3)	95 (15.4)	2 (10.5)
Not Hispanic or Latino	931 (85.1)	165 (85.9)	234 (88.0)	515 (83.5)	17 (89.5)
Unknown	10 (0.9)	1 (0.5)	2 (0.8)	7 (1.1)	0 (0.0)
Gestational age, median (IQR), weeks	33 (27-37)	33 (27-37)	35 (29-38)	32 (26-37)	30 (24-33)
Age at first admission, median (IQR), days	2 (0-37)	1 (0-43)	5 (0-51)	2 (0-30)	17 (2-44)
Age at NICU first discharge, median (IQR), days	113 (63-188)	135 (74-195)	69 (29-152)	120 (71-194)	214 (147-223)
Length of stay, median (IQR), days	88 (48-145)	97 (65-147)	46 (23-89)	104 (60-154)	160 (120-197)
All-cause mortality, no. (%)	119 (10.9)	6 (3.1)	3 (1.1)	109 (17.7)	1 (5.3)
Comorbidity, no. (%)					
Chronic lung disease	445 (40.7)	85 (44.3)	72 (27.1)	276 (44.7)	12 (63.2)
History of necrotizing enterocolitis	328 (30.0)	25 (13.0)	26 (9.8)	270 (43.8)	7 (36.8)
Congenital surgical malformations	380 (34.7)	63 (32.8)	80 (30.1)	234 (37.9)	3 (15.8)
Congenital cardiac disease	167 (15.3)	36 (18.8)	32 (12.0)	97 (15.7)	2 (10.5)
Central venous line, no. (%)[c]	951 (86.9)	164 (85.4)	173 (65.0)	596 (96.6)	18 (94.7)
Mechanical ventilator, no. (%)[c]	931 (85.1)	181 (94.3)	163 (61.3)	569 (92.2)	18 (94.7)
Age at first sepsis evaluation, median (IQR), days	22 (4-61)	27 (1-87)	—	21 (5-54)	54 (25-101)
Total evaluations, no.	2317	309	—	1964	44
Total positive sepsis evaluations, no.	972	0	—	972	0
Infants having positive sepsis evaluation, no. (%)	617 (56.4)	0	—	617 (100)	0
Infants having multiple positive sepsis evaluations, no. (%)	224 (20.5)	0	—	224 (36.3)	0

[a] Our eligibility criteria included infants hospitalized in the NICU >72 h. It is possible that infants may have been infected on admission to our NICU and as such, actual sepsis cases may be greater than our original numbers. Out of 617 infected, 587 were included as cases, the remainder were included as controls because the infection was occurred in the first 48 h within admission. Sepsis evaluations were defined as culture positive if the evaluation yielded a positive bacterial blood culture and clinically positive if cultures were negative, but antibiotics were administered for at least 5 days.

[b] Three infants have fungal infection, the rest of them have viral infection.

[c] Infants who ever had a CVL or mechanical ventilator.

("delta") features for selected vital signs.[9] The delta features for temperature, heart rate (HR), respiratory rate (RR), systolic blood pressure (SBP), and diastolic blood pressure (DBP) were calculated as the difference between the current value and the mean over the previous 24 h. We normalized infant weight to better reflect volume status rather than age. This normalization was completed by converting infant weight to *Z*-scores using either the Fenton standard[32] or WHO standard[33] reference data, depending on each infant's postmenstrual age at the time of weight measurement.

In our baseline and single model, we additionally included the remaining 14 features not included in the diagnostic model. Where necessary to address missing data, we carried forward central access (venous, arterial catheters, and presence of ECMO cannula) up to 48 h after last documentation,[34] as well as nurse assessments of poor perfusion, or lethargy up to 12 h which reflects the cadence of nursing assessment. We included all apnea, bradycardia, and desaturation (ABD) events, regardless of their severity. We included ABD events as a count of events to represent their intensity over the past 12 h. We included infant comorbid conditions (chronic lung disease [CLD], prior necrotizing enterocolitis [NEC], congenital surgical anomalies, congenital heart disease, and intraventricular hemorrhage) due to their potential associations with poor outcomes. CLD required some special handling in our feature engineering because lung disease among infants is only considered chronic if it remains present at least through day of life 30. Therefore, any mentions of CLD prior to this time were treated as if the condition started on day 30.

Model derivation

Selection of controls

A total of 1706 observations from 1094 neonatal patients were included in the training set, with a 1:1 ratio of cases to controls. Although sepsis among infants is a rare event and does not occur in a 1:1 ratio, we chose this ratio to reduce issues due to data imbalance. This approach allows us to achieve our primary goal of comparing the algorithm performance between our two modeling approaches (pipeline vs single model) but limits our ability to estimate important performance characteristics such as false alarm rates in real world populations.

Controls were selected from 3 cohorts: (1) infants who had at least one sepsis episode, but were uninfected at other time periods, (2) infants who were evaluated for sepsis but determined to be uninfected ("ruled-out" sepsis), and (3) infants who were never evaluated for sepsis. We stratified sampling from these cohorts in proportions reflective of those observed among our NICU infants (28% observation time among infants who were infected or presumed infected at least once, 45% observation time among infants who were evaluated but never infected, and 27% observation time from infants never evaluated). We did not match on any clinical characteristics to ensure all potentially relevant characteristics could be included in the model as predictors (eg, gestational age and clinical co-morbidities). To reduce the effect of data collection artifacts related to time of day or shift changes, we counterfactually assigned the time of day of sepsis evaluation for cases to the matched controls.

Model selection

Numerous ML algorithms were considered in our prior work and offered generally similar levels of performance on held out samples in a cross-validation study design.[9] For this work, regularized (lasso) logistic regression (LR) and XGBoost were selected as the candidate models, with LR selected for its explainability and generally good performance on datasets with appropriately engineered features, and XGBoost for its ability to learn non-linear relationships and variable interactions from the data.

Model training

To align our pipeline approach with clinical decision-making, we created a 2-stage set of models to support: (1) recognition of which infants may become septic in the future (sometimes referred to as "watchers")[35] and (2) identification of actively infected infants among those "watchers." We therefore created a baseline model to recognize infants that may become septic optimized for recognition 12 h prior to a potential sepsis event to mimic the first step in clinician decision making and align with forecasting that might be needed for the length of typical work shifts in the NICU. Additionally, we developed a diagnostic model to recognize currently septic infants optimized for recognition up to 1 h prior to clinician recognition. The diagnostic model is intended to mimic the second step in clinician decision making that may occur in response to a change in an infant's clinical status. Our pipeline model combines the baseline and diagnostic models into a single system of these 2 separately trained models to align with the overall clinical decision process. For purposes of comparison, we also trained a single model using all available features.

To begin our modeling development process, we binned the data into 1-h windows such that the final bin ends at the time of sepsis evaluation (T_0) for cases with data bin assignment based on its EHR timestamp. For controls, training data were taken from randomly selected portions of the sepsis-free time periods available for a given sample, and also binned into 1-h windows starting from the selected sepsis-free time. To optimize hyperparameters, we utilized 10-fold nested cross-validation. A high-level overview of data binning and time points related to model derivation and evaluation is shown in Figure 1.

Baseline risk model

The baseline risk model was trained on data curated 12 h prior to sepsis recognition (T_{-12}) for the case samples, and 12 h prior to the randomly selected sepsis-free time for the control samples, using all 28 features (Table S1).

Diagnostic model

The diagnostic model was trained on data at the time of sepsis evaluation (T_0) for case samples, and at the randomly selected sepsis-free time for the control samples. This model only used the 14 dynamic features that were reliably and frequently updated in the EHR system, such as vital signs, which were updated hourly (Table S1 and described previously in the feature engineering section).

Single model

The single model was trained on data at the time of sepsis evaluation (T_0) for case samples, and at the randomly selected sepsis-free time for the control samples, using all 28 features (Table S1).

Model evaluation

All models were evaluated using held out samples in the outer loop of the nested cross validation. Motivated by a desire to evaluate a system that can improve on the current state of clinical care, our model evaluation time point differs from the model training time points. Given even small delays in treatment for sepsis significantly impacts mortality and adverse outcome in NICU,[8] we evaluated model performance at 1 h prior to sepsis recognition (T_{-1}). In real-world implementations infants would be continuously monitored for possible sepsis, and even a 1-h earlier detection of sepsis could make a significant difference in reducing mortality and morbidity. Additionally, as retrospective analysis of infants in our NICU sometimes shows signs of infection several hours prior to clinical recognition,[36] we also assessed model performance 12 h prior to sepsis evaluation (T_{-12}), as this represents the earliest meaningful evaluation window. We selected a model predicted probability of sepsis decision threshold to achieve 80% sensitivity at the training hour of the model being evaluated (T_{-12} for the baseline risk model and T_0 for the diagnostic and single models). Based on the decision threshold, sensitivity, positive predictive value (PPV), area under the curve of the receiver operating characteristic (AUROC), and F1 score were calculated at our two evaluation times (T_{-1} for our primary analysis and T_{-12} for our secondary analysis).

Evaluation of the pipeline model

The 2-stage screening pipeline integrates both the baseline risk model and the diagnostic model, aiming to utilize their

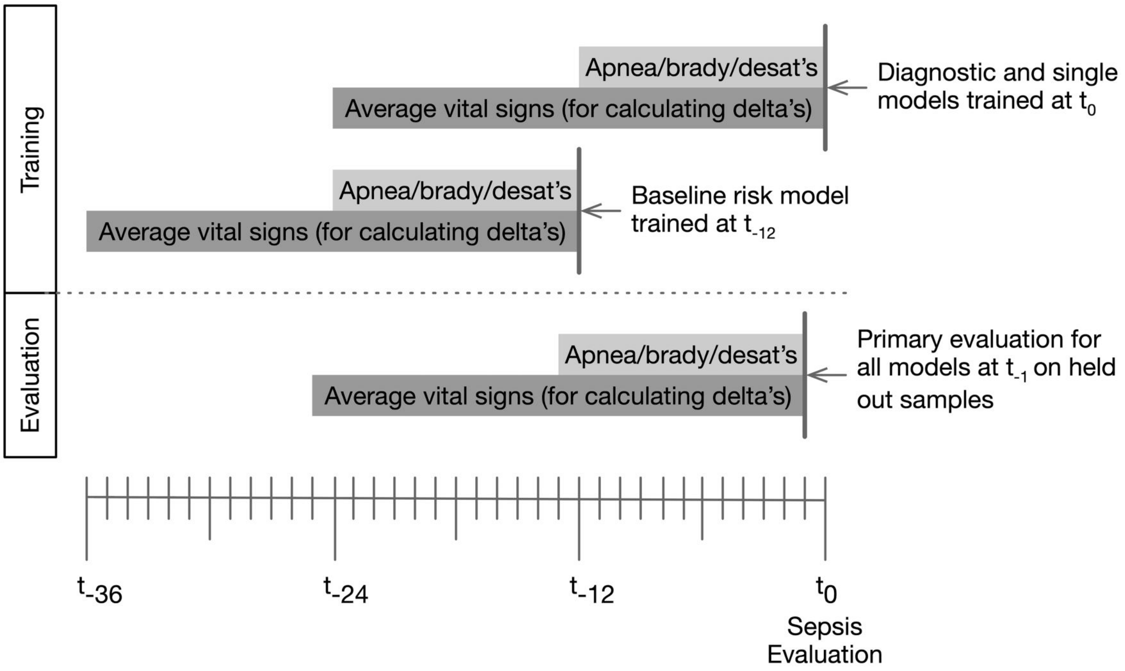
Training
Evaluation
Apnea/brady/desat's
Average vital signs (for calculating delta's)
Diagnostic and single models trained at t0
Baseline risk model trained at t-12
Primary evaluation for all models at t-1 on held out samples
t-36
t-24
t-12
t0
Sepsis Evaluation

Figure 1. High level overview of data binning for model training and evaluation. Note the primary evaluation time point for all models is 1 h prior to sepsis recognition in actual clinical care (T_{-1}). In a secondary analysis (not illustrated in this figure), we also evaluate model performance at T_{-12}. Model evaluation was performed on held out samples in the out loop of the nested cross validation.

predictions in sequence. If the baseline risk model flags a potential high risk, then those infants are monitored using the diagnostic model.

We did not do additional model training for the pipeline model (eg, to weigh the contributions of each model in the pipeline), rather we used the baseline risk and diagnostic models in tandem similar to how they might be used clinically. We defined the pipeline model predictions as follows: a positive prediction by the pipeline model occurs when both the baseline risk model and the diagnostic model flag a case as positive. A negative prediction by the pipeline occurs when either (1) the baseline risk model flags the case as negative or (2) the baseline model flags the case as positive, but the diagnostic model flags it as negative. True positive predictions are all cases of sepsis, for which both baseline risk and diagnostic model predict positive. True negative predictions are all cases that did not have a sepsis episode, for which either (1) the baseline risk model predicts negative (so case is not seen by diagnostic model) or (2) baseline risk model predicts positive, but the diagnostic model predicts negative.

Assessment of feature importance

To determine the feature importance for individual predictions, we used the Shapley Additive exPlanations (SHAP) method[37] and calculated the conditional SHAP values for each model.[38]

Results

Cohort demographic and clinical characteristics

The derivation set included 1094 patients < 1 year old, of which 617 (56.4%) had at least one sepsis episode (Table 1). Of those, 109 (17.7%) died (all-cause mortality), while among the 266 (24.3%) infants who were never evaluated for sepsis and had no clinical suspicion only 3 (1.1%) died. Infants with at least one sepsis episode had a median length of stay (LOS) of 104 days, while the median LOS of infants who had no clinical suspicion for sepsis was 46 days. Of note, 34.7% of total infants in our cohort presented with congenital surgical malformations, and 15.3% had congenital cardiac diseases, emphasizing the complexity and level of critical illness in our population. In addition, the prevalence of infants who ever required a CVL or mechanical ventilation was at 86.9% and 85.1%, indicating a high level of intervention within this population.

Model performance

Baseline and diagnostic models individually

Results for our primary evaluation at T_{-1} and secondary analysis at T_{-12} are shown in Table 2 and Table S2, respectively. Averaged over 10 outer folds of the nested cross-validation, both the LR and XGBoost baseline risk models had high sensitivity at T_{-1}. With a target sensitivity of 0.8 at T_{-12}, both achieved > 0.85 sensitivity at T_{-1}, thereby capturing a large proportion of patients at risk, although at modest levels of specificity (0.68 and 0.67 for LR and XGBoost, respectively). Evaluating the diagnostic models by themselves, the XGBoost diagnostic model achieved close to our target sensitivity (0.77 at T_{-1}) with a specificity of 0.74. In secondary analyses, the diagnostic models had low sensitivity at T_{-12} (LR: 0.59, XGBoost: 0.66).

Pipeline model and comparison to single model approach

When integrating the 2 models into a pipeline, the observed sensitivity was lower than our target at T_{-1} (LR: 0.65, XGBoost: 0.72), although specificity was somewhat higher (LR: 0.82, XGBoost: 0.84) (Table 2). Notably, the sensitivity of the pipeline model approach was lower than that of a

Table 2. Performance (95% confidence interval) of baseline risk model, diagnostic model, pipeline, and single model at 1 h prior (T_{-1}) to sepsis recognition.

	Baseline model[a] T_{-1}	Diagnostic model[b] T_{-1}	Pipeline T_{-1}	Single model[c] T_{-1}
Logistic regression				
Sensitivity	0.87 (0.85, 0.90)	0.69 (0.66, 0.72)	0.65 (0.62, 0.68)	0.76 (0.74, 0.79)
Specificity	0.68 (0.64, 0.71)	0.69 (0.65, 0.72)	0.82 (0.79, 0.85)	0.80 (0.78, 0.83)
PPV	0.73 (0.70, 0.76)	0.69 (0.66, 0.72)	0.78 (0.75, 0.81)	0.79 (0.77, 0.82)
NPV	0.84 (0.81, 0.87)	0.69 (0.66, 0.72)	0.70 (0.67, 0.73)	0.77 (0.75, 0.88)
AUC	0.86 (0.84, 0.88)	0.73 (0.71, 0.76)	—	0.87 (0.85, 0.88)
F1	0.79 (0.77, 0.82)	0.69 (0.66, 0.72)	0.71 (0.68, 0.74)	0.77 (0.77, 0.80)
XGBoost				
Sensitivity	0.85 (0.83, 0.88)	0.77 (0.75, 0.80)	0.72 (0.69, 0.75)	0.77 (0.74, 0.80)
Specificity	0.67 (0.64, 0.70)	0.74 (0.71, 0.77)	0.84 (0.82, 0.87)	0.83 (0.80, 0.85)
PPV	0.72 (0.69, 0.75)	0.75 (0.72, 0.78)	0.82 (0.80, 0.85)	0.82 (0.79, 0.84)
NPV	0.82 (0.79, 0.85)	0.77 (0.74, 0.80)	0.75 (0.73, 0.78)	0.78 (0.76, 0.81)
AUC	0.86 (0.84, 0.88)	0.83 (0.80, 0.85)	—	0.87 (0.86, 0.89)
F1	0.78 (0.75, 0.81)	0.76 (0.73, 0.79)	0.77 (0.74, 0.80)	0.79 (0.76, 0.82)

[a] Derived from 12 h prior to sepsis recognition (T_{-12}) using all 28 features.
[b] Derived from the time at sepsis recognition (T_0) using 14 dynamic features only.
[c] Dervied from T_0 using all 28 features.

single model trained on all available features at a similar level of specificity (the single model's sensitivities were 0.76 and 0.77 and specificities were 0.80 and 0.83 for LR and XGBoost, respectively). This observed difference in prediction outcome distribution at T_{-1} was statistically significant (McNemar's test $P < .0001$ for both LR and XGBoost models).

Feature importance assessment

We conducted SHAP analysis to evaluate feature importance and their impact on the XGBoost models (Figure 2). In the SHAP violin plots, each dot is a patient where red indicates higher risk; for example, higher temperatures predict increased risk while increased age predicts decreased risk. Indwelling central venous line (CVL), high fraction of inspired oxygen (FiO2) and mechanical ventilation were the top 3 predictors in the baseline risk model. For the diagnostic model, elevated temperature, high FiO2, and low systolic blood pressure (SBP) were predictive of sepsis. For the single model, the highest ranked 5 features are CVL, temperature, FiO2, chronological age and mechanical ventilation, which represents a blend of the important features from the baseline risk and diagnostic models.

Discussion

As evidenced by the clinical pathways developed at our institution,[39,40] clinicians often think about infant sepsis in the NICU using a 2-stage process: (1) who are the "watcher" infants who might become infected and (2) among those "watchers" which infants show signs of deterioration that suggests infection. Our main objective was to determine if a 2-step pipeline model, which aligns better with current clinical decision-making processes in the NICU, has superior performance compared to a single model for diagnosing infant sepsis. Unfortunately, in our analysis, a 2-stage pipeline approach had inferior test characteristics—notably unacceptable sensitivity—compared to a model that combines both baseline and dynamic input features into a single model. The sensitivity of the 2-stage pipeline is inadequate for clinical use given the risks of not treating potentially infected infants. These results underscore the importance and ongoing challenges of thoughtful efforts to implement machine learning algorithms in clinical care in ways that support the needs of clinical teams.[41]

We anticipated the baseline model would have high sensitivity with potentially poor precision as it is based on relatively static or otherwise slowly changing variables. Simultaneously, we anticipated the diagnostic model would have high precision. We did not combine these into a single model because we expected that the baseline risk factors were likely to dominate the diagnostic variables. Our hypothesis was that the pipeline model could combine the baseline and diagnostic models in such a way that overall false positives could be reduced (ie, increased precision) without sacrificing sensitivity. Emergence of that outcome requires that one of the individual models provides a negative prediction in cases where the other provides a false positive or where an overall model using all features provides a false positive. In our exploration, the pipeline model was not able to achieve this balance. However, performance could potentially be improved by incorporating data from bedside monitors or other physiological monitoring systems.[18,19,42,43] For instance, the Pulse Oximetry Warning System (POWS) models utilize statistical analyses of beat-to-beat heart rate variability and abnormal oxygen saturation patterns collected from bedside monitors. These models have demonstrated robust dynamic risk prediction for late-onset sepsis in very low birth weight infants.[17] Moreover, spikes in POWS have been clinically correlated with infection and respiratory deterioration.[44] This represents an intriguing direction for further investigation. In our future work, we plan to evaluate and explore the potential benefits of integrating bedside and other physiological monitoring data to improve predictive modeling.

It is worth noting that the SHAP values show that FiO2 is consistently ranked among the top 3 most important features for the baseline risk, diagnostic, and the single models (Figure 2). Although FiO2 theoretically varies significantly, in practice, it tends to remain relatively stable, making it a

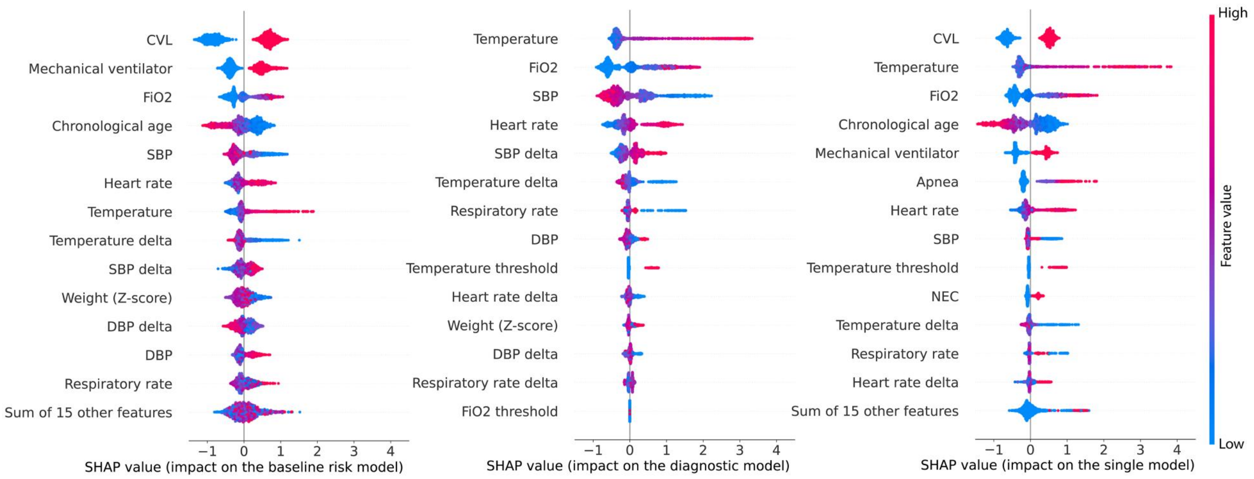

Figure 2. SHAP feature importance plot displaying patient-level conditional SHAP value for the XGBoost algorithm. Left: baseline risk model. Middle: Diagnostic model. Right: single model. Abbreviations: CVL, central venous line; FiO2, high fraction of inspired oxygen; NEC, prior necrotizing enterocolitis; SBP, systolic blood pressure; DBP, diastolic blood pressure).

key factor in all 3 models. In O'Sullivan et al.'s scoping review of previous machine learning studies on neonatal sepsis diagnosis, the gestational age, C-reactive protein levels, and white blood cell count are the strongest predictors to diagnose sepsis.[45] However, gestational age is a static feature, meaning its value remains unchanged across all possible prediction windows prior to sepsis. Although the C-reactive protein and white blood cell count were used in our original model development as laboratory data,[9] all lab-based features have been excluded from our current models for 2 key reasons: (1) to reduce bias, as the ordering of lab tests could indicate extant clinician concern for patient deterioration and (2) to support generalization of the models, given that ordering patterns may vary over time and across sites. Temperature and systolic blood pressure were among the most important dynamic features in our models (Figure 2). In the NICU setting these dynamic features are typically updated hourly in the EHR. Although vital signs are updated much more frequently on bedside monitors, our goal was to balance ease of implementation with the ability to help clinicians better understand the evolving state of the infant over time using readily available data.

Our work relates to numerous studies that utilize traditional ML and deep learning models to support sepsis recognition and management. In the setting of adult sepsis, several models have been shown to accurately predict sepsis onset and have been effectively adopted into clinical practice.[46,47] Fewer models are available for neonatal patients, given age-specific differences in physiology, difficulty in data acquisition for neonates and the complexity of EHR data for infants in the NICU.[45]

In current clinical practice, an ideal implementation approach for infant sepsis predictive models in the NICU should support both the need to identify infant "watchers" who might become infected and support a more active monitoring process using models tuned to detect changes in clinical status that may be due to sepsis among those "watchers."[35] It is possible due to advances in machine learning and artificial intelligence that this clinical paradigm may evolve, and new clinical information needs may arise. For example, the incorporation of sepsis biomarkers and multimodal data available continuously from bedside monitors and video cameras may result in highly accurate sepsis prediction models where greater degrees of automation in sepsis monitoring might be reasonable. However, in the present state where tradeoffs must be made between sensitivity and false alarm rates, clinician judgment remains an essential aspect of the decision-making process. Consequently, near-term efforts should continue to seek ways to support clinician information needs in alignment with the current clinical best practices for ensuring the timely recognition of sepsis among infants in NICU settings.

Limitations

Our evaluation was performed in a single health system and did not use an independent held-out test set to externally evaluate algorithm performance. Also, due to our sampling strategy (1:1 ratio of cases and controls), important real world performance characteristics such as false alarm rate cannot be assessed. Our future efforts will involve evaluating the potential impact of our ML models under real-world conditions in partnership with clinical teams. These efforts will involve evaluating the impact of false alarms, and determining appropriate risk thresholds to trigger clinical actions related to sepsis monitoring, evaluation and treatment. We will also design interfaces in partnership with these clinical teams to present model outputs in a manner that is easily interpretable and actionable in alignment with their information needs.

Conclusion

In this study, we demonstrated that a single model trained on both baseline risk and dynamic physiologic features available at the time of sepsis evaluation has better inference performance compared to a two-stage pipeline model comprising a separate baseline risk model and a diagnostic model trained on data available 12 h and 1 h prior to sepsis evaluation, respectively. Although aspects of the 2-stage screening pipeline may better support clinical information needs, our evaluation suggests that future efforts should consider ways that outputs from a single optimized model can be used to support the two distinct information needs of current clinical practice. Specifically, clinical teams need model outputs or visualizations that help them both identify infants at risk of sepsis (ie, "watchers") and support the timely recognition of clinical deterioration among those at-risk infants.

Our over-arching goal is to improve the recognition of infant sepsis and the clinical use of predictive models to help to reduce morbidity and mortality in this vulnerable population. Due to imperfect sensitivity of all models we evaluated, it is important to note that while these models can help identify infants who may have sepsis, they should not replace clinical judgment.

Acknowledgments

We thank Dr Tanis Fenton for her generosity in sharing the Fenton Preterm Growth Chart reference data with us.

Author contributions

Lusha Cao (Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Project administration, Software, Validation, Visualization), Aaron J. Masino (Conceptualization, Data curation, Formal analysis, Methodology, Software), Mary Catherine Harris (Conceptualization, Funding acquisition, Investigation, Project administration, Resources, Supervision), Lyle Ungar (Investigation, Methodology), Gerald Shaeffer (Investigation, Software), Alexander Fidel (Investigation, Project administration), Elease McLaurin (Methodology), Lakshmi Srinivasan (Methodology), Dean J. Karavite (Investigation, Methodology, Project administration, Software, Validation), and Robert W. Grundmeier (Conceptualization, Data curation, Formal analysis, Funding acquisition, Investigation, Methodology, Project administration, Resources, Software, Supervision, Validation, Visualization)

Supplementary material

Supplementary material is available at *JAMIA Open* online.

Funding

Funding for this study was provided by National Library of Medicine Award R01LM013526 (MPI R.W.G. and M.C.H.), Situation Awareness to Improve Infant Sepsis Recognition in the Presence of Clinical Uncertainty.

Conflicts of interest

The authors confirm that the research presented in this article met the ethical guidelines, including adherence to the legal requirements, of the United States of America and received approval from the Children's Hospital of Philadelphia. A.F. owns equity in Together AI. His spouse is employed by Together AI. Together AI had no involvment in this study. Other authors have no conflicts of interest to report.

Data availability

The data underlying this article cannot be shared publicly because it contains potentially identifiable health information. The data will be shared on reasonable request to the corresponding author.

References

1. Hartman ME, Linde-Zwirble WT, Angus DC, Watson RS. Trends in the epidemiology of pediatric severe sepsis. *Pediatr Crit Care Med*. 2013;14:686-693. https://doi.org/10.1097/PCC.0b013e3182917fad
2. Stoll BJ, Hansen N, Fanaroff AA, et al. Late-onset sepsis in very low birth weight neonates: the experience of the NICHD neonatal research network. *Pediatrics*. 2002;110:285-291. https://doi.org/10.1542/peds.110.2.285
3. Watson RS, Carcillo JA, Linde-Zwirble WT, Clermont G, Lidicker J, Angus DC. The epidemiology of severe sepsis in children in the United States. *Am J Respir Crit Care Med*. 2003;167:695-701. https://doi.org/10.1164/rccm.200207-682OC.
4. Li J, Xiang L, Chen X, et al. Global, regional, and national burden of neonatal sepsis and other neonatal infections, 1990-2019: findings from the global burden of disease study 2019. *Eur J Pediatr*. 2023;182:2335-2343. https://doi.org/10.1007/s00431-023-04911-7.
5. Stoll BJ, Hansen NI, Sanchez PJ, et al.; Eunice Kennedy Shriver National Institute of Child Health and Human Development Neonatal Research Network. Early onset neonatal sepsis: the burden of group B streptococcal and E. coli disease continues. *Pediatrics*. 2011;127:817-826. https://doi.org/10.1542/peds.2010-2217.
6. Srinivasan L, Harris MC. New technologies for the rapid diagnosis of neonatal sepsis. *Curr Opin Pediatr*. 2012;24:165-171. https://doi.org/10.1097/MOP.0b013e3283504df3
7. Weiss SL, Fitzgerald JC, Balamuth F, et al. Delayed antimicrobial therapy increases mortality and organ dysfunction duration in pediatric sepsis. *Crit Care Med*. 2014;42:2409-2417. https://doi.org/10.1097/CCM.0000000000000509
8. Schmatz M, Srinivasan L, Grundmeier RW, et al. Surviving sepsis in a referral neonatal intensive care unit: association between time to antibiotic administration and In-Hospital outcomes. *J Pediatr*. 2020;217:59-65 e1. https://doi.org/10.1016/j.jpeds.2019.08.023.
9. Masino AJ, Harris MC, Forsyth D, et al. Machine learning models for early sepsis recognition in the neonatal intensive care unit using readily available electronic health record data. *PLoS One*. 2019;14:e0212665. https://doi.org/10.1371/journal.pone.0212665.
10. Lee JW, Lee B, Park JD. Pediatric septic shock estimation using deep learning and electronic medical records. *Acute Crit Care*. 2024;39:400-407. https://doi.org/10.4266/acc.2024.00031
11. Scott HF, Colborn KL, Sevick CJ, et al. Development and validation of a predictive model of the risk of pediatric septic shock using data known at the time of hospital arrival. *J Pediatr*. 2020;217:145-151 e6. https://doi.org/10.1016/j.jpeds.2019.09.079.
12. Goto T, Camargo CA, Jr., Faridi MK, Freishtat RJ, Hasegawa K. Machine learning-based prediction of clinical outcomes for children during emergency department triage. *JAMA Netw Open*. 2019;2:e186937. https://doi.org/10.1001/jamanetworkopen.2018.6937.
13. Mercurio L, Pou S, Duffy S, Eickhoff C. Risk factors for pediatric sepsis in the emergency department: a machine learning pilot study. *Pediatr Emerg Care*. 2023;39:e48-e56. https://doi.org/10.1097/PEC.0000000000002893.
14. Qin Y, Kernan KF, Fan Z, et al. Machine learning derivation of four computable 24-h pediatric sepsis phenotypes to facilitate enrollment in early personalized anti-inflammatory clinical trials. *Crit Care*. 2022;26:128. https://doi.org/10.1186/s13054-022-03977-3.
15. Liu R, Greenstein JL, Fackler JC, Bergmann J, Bembea MM, Winslow RL. Prediction of impending septic shock in children with sepsis. *Crit Care Explor*. 2021;3:e0442. https://doi.org/10.1097/CCE.0000000000000442.
16. Meeus M, Beirnaert C, Mahieu L, et al. Clinical decision support for improved neonatal care: the development of a machine learning model for the prediction of late-onset sepsis and necrotizing enterocolitis. *J Pediatr*. 2024;266:113869. https://doi.org/10.1016/j.jpeds.2023.113869.
17. Kausch SL, Brandberg JG, Qiu J, et al. Cardiorespiratory signature of neonatal sepsis: development and validation of prediction

models in 3 NICUs. *Pediatr Res*. 2023;93:1913-1921. https://doi.org/10.1038/s41390-022-02444-7.

18. Griffin MP, Lake DE, Bissonette EA, Harrell FE, Jr., O'Shea TM, Moorman JR. Heart rate characteristics: novel physiomarkers to predict neonatal infection and death. *Pediatrics*. 2005;116:1070-1074. https://doi.org/10.1542/peds.2004-2461
19. Fairchild KD, Schelonka RL, Kaufman DA, et al. Septicemia mortality reduction in neonates in a heart rate characteristics monitoring trial. *Pediatr Res*. 2013;74:570-575. https://doi.org/10.1038/pr.2013.136.
20. Bonafide CP, Lin R, Zander M, et al. Association between exposure to nonactionable physiologic monitor alarms and response time in a children's hospital. *J Hosp Med*. 2015;10:345-351. https://doi.org/10.1002/jhm.2331.
21. Albanowski K, Burdick KJ, Bonafide CP, Kleinpell R, Schlesinger JJ. Ten years later, alarm fatigue is still a safety concern. *AACN Adv Crit Care*. 2023;34:189-197. https://doi.org/10.4037/aacnacc2023662
22. Ruppel H, Makeneni S, Rasooly IR, Ferro DF, Bonafide CP. Pediatric characteristics associated with higher rates of monitor alarms. *Biomed Instrum Technol*. 2023;57:171-179. https://doi.org/10.2345/0899-8205-57.4.171.
23. Linardatos P, Papastefanopoulos V, Kotsiantis S. Explainable AI: a review of machine learning interpretability methods. *Entropy (Basel)*. 2020;23:18. https://doi.org/10.3390/e23010018.
24. Bohlen L, Rosenberger J, Zschech P, Kraus M. Leveraging interpretable machine learning in intensive care. *Ann Oper Res*. 2024;1-40. https://doi.org/10.1007/s10479-024-06226-8
25. Karavite DJ, Harris MC, Grundmeier RW, Srinivasan L, Shaeffer GP, Muthu N. Using a sociotechnical model to understand challenges with sepsis recognition among critically ill infants. *ACI Open*. 2022;06:e57-e65. https://doi.org/10.1055/s-0042-1749318
26. PCORnet Common Data Model. Secondary PCORnet Common Data Model 2024. Accessed December 10, 2024. https://pcornet.org/news/resources-pcornet-common-data-model/.
27. Flannery DD, Puopolo KM, Hansen NI, et al.; Eunice Kennedy Shriver National Institute of Child Health and Human Development Neonatal Research Network. Neonatal infections: insights from a multicenter longitudinal research collaborative. *Semin Perinatol*. 2022;46:151637. https://doi.org/10.1016/j.semperi.2022.151637.
28. Greenberg RG, Chowdhury D, Hansen NI, et al.; Eunice Kennedy Shriver National Institute of Child Health and Human Development Neonatal Research Network. Prolonged duration of early antibiotic therapy in extremely premature infants. *Pediatr Res*. 2019;85:994-1000. https://doi.org/10.1038/s41390-019-0300-4.
29. Mukhopadhyay S, Puopolo KM, Hansen NI, et al.; NICHD Neonatal Research Network. Neurodevelopmental outcomes following neonatal late-onset sepsis and blood culture-negative conditions. *Arch Dis Child Fetal Neonatal Ed*. 2021;106:467-473. https://doi.org/10.1136/archdischild-2020-320664.
30. Dimopoulou V, Klingenberg C, Naver L, et al.; AENEAS Study Group. Antibiotic exposure for culture-negative early-onset sepsis in late-preterm and term newborns: an international study. *Pediatr Res*. 2024; https://doi.org/10.1038/s41390-024-03532-6.
31. Pedregosa F, Varoquaux G, Gramfort A, et al. Scikit-learn: machine learning in python. *J Mach Learn Res*. 2011;12:2825-2830.
32. Fenton TR, Sauve RS. Using the LMS method to calculate Z-scores for the Fenton preterm infant growth chart. *Eur J Clin Nutr*. 2007;61:1380-1385. https://doi.org/10.1038/sj.ejcn.1602667.
33. Group WHOMGRS. WHO child growth standards based on length/height, weight and age. *Acta Paediatr Suppl*. 2006;450:76-85. https://doi.org/10.1111/j.1651-2227.2006.tb02378.x
34. Ji R, He Z, Zhou J, Fang S, Ge L. Antibiotic use at planned central line removal in reducing neonatal post-catheter removal sepsis: a systematic review and meta-analysis. *Front Pediatr*. 2023;11:1324242. https://doi.org/10.3389/fped.2023.1324242
35. Eisenberg MA, Balamuth F. Pediatric sepsis screening in US hospitals. *Pediatr Res*. 2022;91:351-358. https://doi.org/10.1038/s41390-021-01708-y.
36. Coggins S, Harris MC, Grundmeier R, Kalb E, Nawab U, Srinivasan L. Performance of pediatric systemic inflammatory response syndrome and organ dysfunction criteria in late-onset sepsis in a quaternary neonatal intensive care unit: a case-control study. *J Pediatr*. 2020;219:133-139 e1. https://doi.org/10.1016/j.jpeds.2019.12.064.
37. Lundberg S, Lee S-I. *A unified approach to interpreting model predictions*. 2017; https://doi.org/10.48550/arxiv.1705.07874
38. Ltu, L. Towards cotenable and causal Shapley feature explanations *AAAI 2021*. Workshop: Trustworthy AI for Healthcare; 2021.
39. Yellow Zone—Sepsis Watcher. Initial Laboratory Testing Considerations for Suspected Bacterial Infection without Organ Dysfunction. Secondary Yellow Zone—Sepsis Watcher. Initial Laboratory Testing Considerations for Suspected Bacterial Infection without Organ Dysfunction; 2024. https://pathways.chop.edu/clinical-pathway/sepsis-yellow-zone-sepsis-watcher-initial-laboratory-testing-considerations-ed-inpatient-picu.
40. Coggins S, Srinivasan L, Gattoline S, et al. N/IICU Clinical Pathway for Evaluation and Treatment of Suspected Sepsis in Newborns and Infants. Secondary N/IICU Clinical Pathway for Evaluation and Treatment of Suspected Sepsis in Newborns and Infants; 2018. https://pathways.chop.edu/clinical-pathway/sepsis-niicu-clinical-pathway.
41. Li RC, Asch SM, Shah NH. Developing a delivery science for artificial intelligence in healthcare. *NPJ Digit Med*. 2020;3:107. https://doi.org/10.1038/s41746-020-00318-y.
42. Sullivan BA, Nagraj VP, Berry KL, et al. Clinical and vital sign changes associated with late-onset sepsis in very low birth weight infants at 3 NICUs. *J Neonatal Perinatal Med*. 2021;14:553-561. https://doi.org/10.3233/NPM-200578
43. Moorman JR, Carlo WA, Kattwinkel J, et al. Mortality reduction by heart rate characteristic monitoring in very low birth weight neonates: a randomized trial. *J Pediatr*. 2011;159:900-906 e1. https://doi.org/10.1016/j.jpeds.2011.06.044.
44. Kausch SL, Slevin CC, Duncan A, et al. Clinical correlates of a high cardiorespiratory risk score for very low birth weight infants. *Pediatr Res*. 2024; https://doi.org/10.1038/s41390-024-03580-y.
45. O'Sullivan C, Tsai DH, Wu IC, et al. Machine learning applications on neonatal sepsis treatment: a scoping review. *BMC Infect Dis*. 2023;23:441. https://doi.org/10.1186/s12879-023-08409-3.
46. Fleuren LM, Klausch TLT, Zwager CL, et al. Machine learning for the prediction of sepsis: a systematic review and meta-analysis of diagnostic test accuracy. *Intensive Care Med*. 2020;46:383-400. https://doi.org/10.1007/s00134-019-05872-y.
47. Adams R, Henry KE, Sridharan A, et al. Prospective, multi-site study of patient outcomes after implementation of the TREWS machine learning-based early warning system for sepsis. *Nat Med*. 2022;28:1455-1460. https://doi.org/10.1038/s41591-022-01894-0.

JAMIA Open, 2025, 8, 1–9
https://doi.org/10.1093/jamiaopen/ooaf015
Research and Applications