# Pediatric Severe Sepsis Prediction Using Machine Learning

Thomas Desautels PhD[1], Jana Hoffman PhD[1], Christopher Barton MD[2], Qingqing Mao PhD[1], Melissa Jay BA[1], Jacob Calvert MSc[1], Ritankar Das MSc[1]

## Introduction

Early detection of pediatric severe sepsis is necessary in order to administer effective treatment. In this study, we assessed the efficacy of a machine-learning-based prediction algorithm applied to pediatric electronic healthcare record (EHR) data for the early detection and prediction of severe sepsis onset.

## Modeling

This work used boosted ensembles of decision trees with carry-forward imputation. Our classifiers were trained on a set of features from the EHR that included patient age, diastolic and systolic blood pressures, heart rate, temperature, respiration rate, and peripheral oxygen saturation ($SpO_2$). Additionally, the values of Glasgow Coma Score, white blood cell count, and platelet count were used if available. We used de-identified chart data from pediatric (ages 2 to 17 years) inpatient and emergency encounters at the University of California San Francisco (UCSF) Medical Center, from June 2011 to March 2016, resulting in 11,127 included encounters.

## Results

We evaluated predictive performance of the machine learning-based predictor by training and testing at the time of pediatric severe sepsis onset and four hours before onset. Figures 1 and 2 show the area under the receiver operator characteristic (AUROC) curves, compared with the Pediatric Logistic Organ Dysfunction (PELOD-2) and pediatric Systemic Inflammatory Response Syndrome (SIRS).

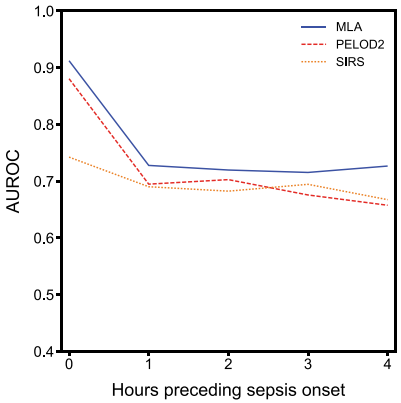


**Figure 3:** Average AUROC for pediatric severe sepsis detection and prediction hours preceding onset.

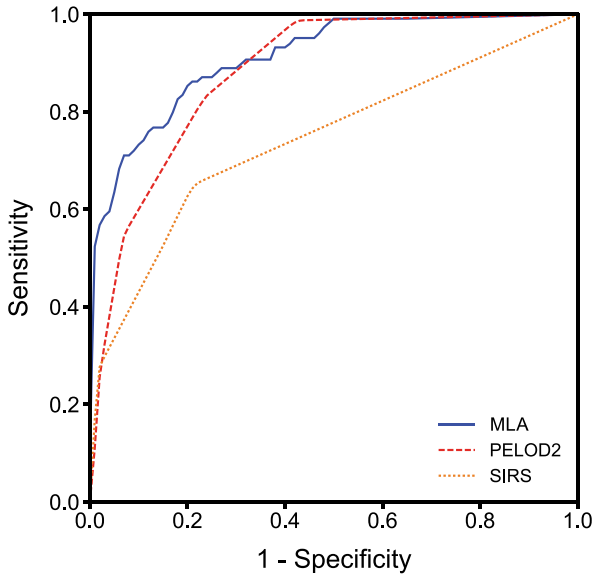


**Figure 1:** ROC curves (averaged across the four test folds) for the machine learning algorithm (MLA), PELOD-2, and SIRS at time of onset.

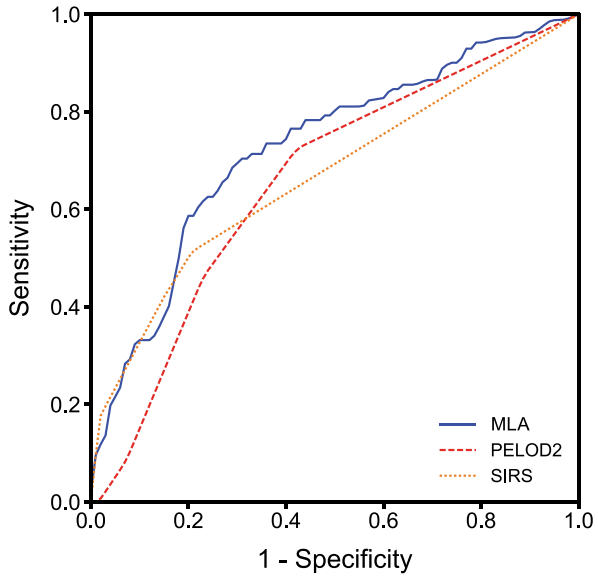


**Figure 2:** ROC curves (averaged across the four test folds) for the machine learning algorithm (MLA), PELOD-2, and SIRS at 4 hours pre-onset.

| | Predictive Algorithm (Onset) Mean (SE) | PELOD-2 (Onset) Mean (SE) | SIRS (Onset) Mean (SE) | Predictive Algorithm (4 hours pre-onset) Mean (SE) | PELOD-2 (4 hours pre-onset) Mean (SE) | SIRS (4 hours pre-onset) Mean (SE) |
|---|---|---|---|---|---|---|
| **AUROC** | 0.912 ± (0.023) | 0.881 ± (0.007) | 0.727 ± (0.052) | 0.727 ± (0.052) | 0.658 ± (0.053) | 0.667 ± (0.044) |
| **Sensitivity** | 0.797 ± (0.008) | 0.828 ± (0.041) | 0.651 ± (0.094) | 0.797 ± (0.008) | 0.724 ± (0.069) | 0.514 ± (0.089) |
| **Specificity** | 0.868 ± (0.058) | 0.767 ± (0.008) | 0.787 ± (0.007) | 0.505 ± (0.182) | 0.581 ± (0.008) | 0.792 ± (0.005) |
| **DOR** | 34.211 ± (19.494) | 17.152 ± (5.097) | 7.586 ± (2.575) | 5.161 ± (3.372) | 3.98 ± (1.397) | 4.379 ± (1.812) |

**Table 1:** Performance metrics for the machine learning algorithm and pediatric scoring systems. This procedure chose an operating point from the ROC curve where the sensitivity was the largest possible value ≤ 0.80; the selected PELOD-2 and SIRS sensitivity values for 4-hour pre-onset prediction were considerably below this value. SE is the standard error and DOR is the diagnostic odds ratio.

## Discussion

These experiments demonstrate that the machine learning-based sepsis prediction system can detect and predict pediatric severe sepsis onset with high accuracy and AUROC performance.

The Machine Learning-based system can be computed automatically in the absence of biomarkers, nursing reports, or access to clinical notes. The input consists of vital sign and clinical data extracted directly from the patients' electronic health records.

The high AUROC performance offers clinicians and hospitals a variety of useful alert operating points to suit their sepsis alerting needs. Using these tools, clinicians will be better able to allocate finite clinical resources, identify patients before their condition deteriorates, and avoid adverse outcomes.

1 Dascena Inc., Hayward, CA, USA
2 Department of Emergency Medicine, University of California San Francisco, San Francisco, CA, USA

# Multicenter validation of a machine learning algorithm for 48 hour all-cause mortality prediction

Hamid Mohamadlou PhD[1], Anna Lynn-Palevsky[1]*, Christopher Barton MD[2], Grant Fletcher MD[3], Lisa Shieh MD PhD[4], Philip B Stark PhD[5], Uli Chettipally MD[2,6], David Shimabukuro MD[7], Mitchell Feldman MD[8], Ritankar Das MSc[1]

1 Dascena, Inc., Hayward, CA, United States 2 Department of Emergency Medicine, University of California San Francisco, San Francisco, CA, United States 3 Division of Internal Medicine, University of Washington School of Medicine, Seattle, WA, United States 4 Department of Medicine, Stanford University School of Medicine, Stanford, CA, United States 5 Department of Statistics, University of California Berkeley, Berkeley, CA, United States 6 Kaiser Permanente South San Francisco Medical Center, South San Francisco, CA, United States 7 Department of Anesthesia and Perioperative Care, University of California San Francisco, San Francisco, CA, United States 8 Division of General Internal Medicine, Department of Medicine, University of California San Francisco, San Francisco, CA, United States

## Introduction

Timely identification of patients with elevated risk for in-hospital mortality is necessary to best allocate limited and costly hospital resources, focus care to prevent patients from deteriorating, and anticipate probable patient outcomes.

## Hypothesis

Machine learning algorithms can improve patient mortality prediction by calculating a score that depends not only on linear combinations of patient input variables, but also on trends in those variables.

## Methods

Retrospective validation of a machine learning mortality risk prediction tool with a 48 hour prediction horizon.

- Six vital signs (heart rate, respiratory rate, peripheral oxygen saturation ($SpO_2$), temperature, systolic blood pressure, diastolic blood pressure, and Glasgow Coma Scale values) were used as inputs. At least one recorded observation of each measurement was required for patient inclusion.
- Data was extracted from Electronic Health Record datasets for patients over 18 years of age between June 2011 and March 2016 at UCSF (94,222 encounters); between December 2008 and May 2017 at Stanford Medical Center (77,142 encounters); and between January 2014 and March 2017 at UW (32,936 encounters).
- The classifier was created using gradient boosted trees and carry-forward imputation.

| Tested On: | UCSF 12 hrs (95% CI) | UCSF 24 hrs (95% CI) | UCSF 48 hrs (95% CI) | Stanford 12 hrs (95% CI) | Stanford 24 hrs (95% CI) | Stanford 48 hrs (95% CI) | UW 12 hrs (95% CI) | UW 24 hrs (95% CI) | UW 48 hrs (95% CI) |
|---|---|---|---|---|---|---|---|---|---|
| **Trained on UCSF** | **0.955** (0.951, 0.958) | **0.933** (0.929, 0.936) | **0.899** (0.894, 0.904) | **0.894** (0.862, 0.926) | **0.869** (0.837, 0.900) | **0.839** (0.810, 0.868) | **0.948** (0.946, 0.950) | **0.919** (0.916, 0.922) | **0.896** (0.893, 0.899) |
| **Trained on Stanford** | **0.862** (0.838, 0.886) | **0.763** (0.734, 0.791) | **0.751** (0.725, 0.774) | **0.930** (0.920, 0.940) | **0.915** (0.902, 0.927) | **0.887** (0.865, 0.908) | **0.819** (0.801, 0.838) | **0.764** (0.729, 0.798) | **0.770** (0.747, 0.791) |
| **Trained on UW** | **0.927** (0.925, 0.929) | **0.914** (0.911, 0.917) | **0.870** (0.866, 0.874) | **0.908** (0.898, 0.918) | **0.878** (0.860, 0.895) | **0.836** (0.819, 0.853) | **0.918** (0.912, 0.925) | **0.914** (0.905, 0.923) | **0.845** (0.830, 0.860) |

**Table 1:** AUROC and 95% confidence interval values for each data set for cross-population training experiments. Testing was performed 12, 24, and 48 hours in advance of patient death. Risk scores were computed using systolic blood pressure, diastolic blood pressure, heart rate, temperature, and respiratory rate.

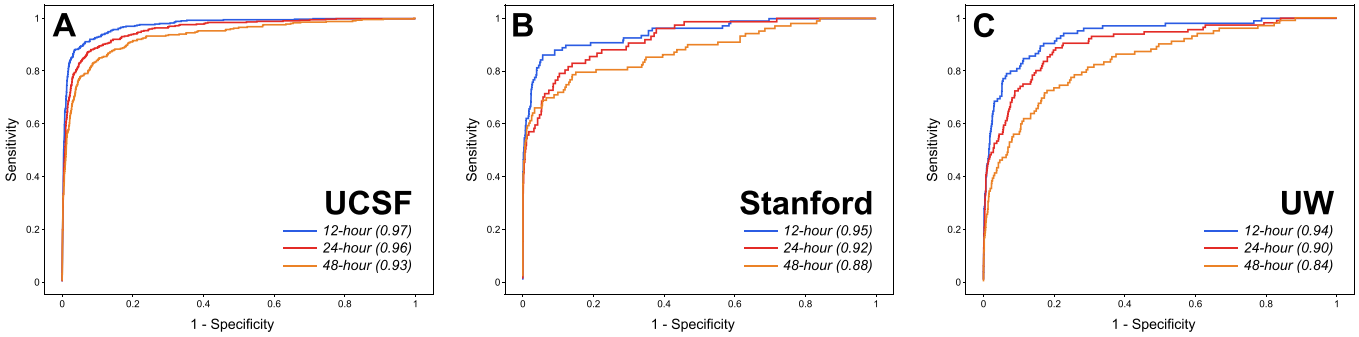


**Figure 1:** Comparison of the receiver operating characteristic curve (ROC). ROC and area under the ROC (AUROC) for the machine learning algorithm are presented for 12, 24, and 48 hour mortality prediction with training and testing performed on a) University of California, San Francisco (UCSF) patient data, b) Stanford patient data, c) University of Washington (UW) patient data.

| | | Machine Learning Algorithm | MEWS Score | qSOFA Score |
|---|---|---|---|---|
| **12 hours before death** | **AUROC (95% CI)** | 0.972 (0.970, 0.975) | 0.929 | 0.909 |
| | Partial AUROC | 0.178 (0.177, 0.180) | 0.161 | 0.140 |
| | DOR | 180.93 | 39.87 | 49.02 |
| | Sensitivity | 0.782 | 0.709 | 0.806 |
| | Specificity | 0.980 | 0.942 | 0.922 |
| **24 hours before death** | **AUROC (95% CI)** | 0.960 (0.957, 0.963) | 0.907 | 0.890 |
| | Partial AUROC | 0.171 (0.170, 0.172) | 0.150 | 0.131 |
| | DOR | 64.24 | 27.17 | 35.24 |
| | Sensitivity | 0.817 | 0.685 | 0.783 |
| | Specificity | 0.935 | 0.926 | 0.907 |
| **48 hours before death** | **AUROC (95% CI)** | 0.935 (0.931, 0.939) | 0.877 | 0.851 |
| | Partial AUROC | 0.157 (0.156, 0.159) | 0.137 | 0.117 |
| | DOR | 40.99 | 15.96 | 19.85 |
| | Sensitivity | 0.796 | 0.649 | 0.745 |
| | Specificity | 0.913 | 0.896 | 0.872 |

**Table 2:** Comparison of the area under the receiver operating characteristic curve (AUROC), Diagnostic Odds Ratio (DOR), sensitivity, and specificity for the machine learning algorithm and qSOFA and MEWS scoring systems for mortality prediction. Predictions were performed 12, 24, and 48 hours in advance patient death. Partial AUROC corresponds to specificity above 0.80. 95% Confidence Intervals are provided only for the machine learning algorithm.

## Results

- The machine learning algorithm predicted patient mortality with a high area under the receiver operating characteristic curve (AUROC) for all prediction windows on all datasets (Figure 1).
- For all prediction windows, the MLA had a significantly higher Diagnostic Odds Ratio (DOR) than either MEWS or qSOFA (Table 2).
- Additional statistics and partial AUROC, corresponding to specificity above 0.80, in the clinical operating range for the predictor, for the machine learning algorithm trained on UCSF data, MEWS, and qSOFA scores are presented in Table 2.
- When trained and tested on patient datasets from distinct institutions, the boosted tree predictor maintained high levels of accuracy as demonstrated by AUROC and DOR values, up to two days in advance of patient death (Table 1).

## Discussion

- The boosted tree predictor accurately predicted patient mortality 48 hours in advance of death using only routinely collected patient vital signs. The algorithm demonstrated significantly improved accuracy over commonly used mortality risk stratification tools.
- Even when trained and tested on separate patient datasets, the boosted tree predictor maintained high levels of accuracy as demonstrated by AUROC and DOR values, up to two days in advance of patient death.
- The boosted tree predictor's combination of high sensitivity and specificity means that it can identify more at-risk patients while also reducing the number of false alarms.

In a clinical setting, this algorithm may help clinicians identify those patients for whom more intensive care would prevent deterioration. In future studies, we intend to test the algorithm prospectively using real-time clinical data.