Supplementary Digital Content for: An Interpretable Machine Learning Model for Accurate Prediction of Sepsis in the ICU

Authors: Shamim Nemati PhD[1*], Andre Holder MD MSc[2], Fereshteh Razmi[1], Matthew D. Stanley[3], Gari D. Clifford[1,4] , Timothy Buchman MD PhD[3,5]

APPENDIX A

TABLE A1 Summary of development Cohort (Emory dataset) Characteristics

	Overall Cohort			Development Cohort		
Demographics	**All Patients**	**Non-Septic**	**Septic**	**All Patients**	**Non-Septic**	**Septic**
Patients (#)	31179	23720	7459 (23.9%)	27527	25152	2375 (8.6%)
Male (%)	52.8	52.4	54.4 *	52.7	52.4	56.2 *
Age (year)	61 [49 – 71]	61 [49 – 71]	61 [50 – 71]	61 [49 – 71]	61 [49 – 71]	61 [50 – 71]
Race (%) **Caucasian** **Black** **Asian** **Hispanic**	 47.3 44.6 1.4 0.04	 48.8 43.2 1.3 0.02	 42.3 * 48.9 * 1.4 0.09 *	 48.6 † 43.3 † 1.3 0.03	 48.9 43.1 1.3 0.02	 45.0 * 45.4 * 1.3 0.08
ICU LOS (hours)	52 [30 – 104]	46 [27 – 79]	109 * [54 – 238]	48 † [28 – 90]	46 [27 – 77]	141 * [77 – 258]
Inpatient Mortality (%) **Inpatient Hospice (%)**	6.2% 5.9%	3.2% 3.5%	15.6 * 13.7 *	3.9 † 4.2 †	2.9 3.5	14.5 * 12.5 *
ICD-9 (%): 995.92 or 785.52	13.2	4.3	41.4 *	6.1 †	4.2	26.7 *
SOFA	2.5 [0.8 – 4.7]	1.7 [0.5 – 3.6]	5.0 * [3.1 – 7.4]	1.9 † [0.6 – 4.0]	1.7 [0.5 – 3.6]	5.0 * [3.1 – 7.4]
CCI	3 [1 – 5]	2 [1 – 4]	4 * [2 – 6]	2 † [1 – 4]	2 [1 – 4]	4 * [2 – 6]
ICU Admission to Sepsis (hours)	—	—	1.3 [-3.1 – 13.8]	—	—	23.9 [9.8 – 55.8]

* Statistically significant difference between septic and non-septic patients within each cohort

† Statistically significant difference between overall cohort and the development cohort

APPENDIX B

Features: All relevant static (demographic, historical, and contextual) data such as age, gender, and ethnicity, along with dynamic clinical and laboratory features commonly recorded by bedside nurses were included for analysis as stored in the clinical data warehouse. Dynamic clinical features included Mean Arterial Pressure (MAP), Heart Rate (HR), Peripheral capillary Oxygen Saturation (SpO2), Systolic Blood Pressure (SBP), Diastolic Blood Pressure (DBP), Respiration Rate (RESP), Glasgow Coma Score (GCS), and Temperature (Temp). Some of the dynamic laboratory data included white blood cell count (WBC), and serum lactate among others. We also extracted a number of features that captured history, comorbidity, and the clinical context of the patients, including Charlson Comorbidity Index (CCI), Mechanical Ventilation, care unit (medical, surgical, cardiac care, or neuro-intensive care), as well as the surgical specialty (cardiovascular, neurosurgery, urology, etc.) and wound type (clean, contaminated, dirty, or infected) if the patient had a surgery in past 12 hours.

All dynamic features were organized into 1-hour non-overlapping time series bins to accommodate for different sampling frequencies of available data. The 1-hour time bin interval was selected as a balance between having short windows with too many missing data points (low-frequency clinical data) and having time windows too long to make any meaningful prediction. Non-overlapping bins simplified the modeling schema by minimizing autocorrelation. EMR features with sampling frequencies higher than once every hour were uniformly resampled into 1-hour time bins, by taking the median values if multiple measurements were available. Features were updated hourly when new data became available; otherwise, the old values were kept (sample-and-hold interpolation). The renal component of the SOFA score was slightly modified to account for poor data quality of urine output, and only used serum creatinine. Otherwise, the SOFA score was calculated as outlined in the original manuscript. [1] Mean imputation was used to replace all remaining missing values (mainly at the start of each record).

The bedside monitor data (HR and MAP with 0.5 Hz resolution) was matched and time synchronized to each patient's EMR data. The following features from the HR and MAP time series were derived from the bedside monitor's proprietary software using the ECG and blood pressure waveforms: standard Deviation of HR (HR_{STD}), Standard Deviation of MAP (MAP_{STD}), Multiscale Entropy [2] and Multiscale Conditional Entropy [3] of R-R intervals (60/HR) and MAP (HRV1, HRV2 and BPV1, BPV2, respectively). The time series-related features were updated every hour, using a 6-hour sliding window with five hours overlap. For each window, 17 different scales (scales 1, 4, 7, … 49) were considered for all variability measurements of heart rate and blood pressure, and the average value of multiscale entropy and conditional entropy over all

scales were included as features in our machine learning prediction model. **A complete list of these features (total of 65) are provided here**:

High-resolution dynamical features (calculated using 6 hours sliding windows, with 5 hours overlap; **6** features): standard deviation of RR intervals and MAP (RRSTD and MAPSTD), average multiscale entropy[1] of RR and MAP (HRV1 and BPV1) and average multiscale conditional entropy of RR and MAP (HRV2 and BPV2).

Clinical features (**10** features): Mean Arterial Blood Pressure (MAP), Heart Rate (HR), Oxygen Saturation (O_2Sat), Systolic Blood Pressure (SBP), Diastolic Blood Pressure (DBP), Respiratory Rate (RESP), Temperature (Temp), Glasgow Coma Scale (GCS), Partial Pressure of Arterial Oxygen (PaO_2), Fraction of Inspired O_2 (FIO_2).

Laboratory (General; **25** features): White Blood Count (WBC), Hemoglobin, Hematocrit, Creatinine, Bilirubin and Bilirubin direct, Platelets, International Normalized Ratio (INR), Partial Prothrombin Time (PTT), Aspartate Aminotransferase (AST), Alkaline Phosphatase, Lactate, Glucose, Potassium, Calcium, blood urea nitrogen (BUN), Phosphorus, Magnesium, Chloride, B-type Natriuretic Peptide (BNP), Troponin, Fibrinogen, CRP, Sedimentation Rate, Ammonia.

Laboratory (Arterial Blood Gas or ABG; **5** features): pH, pCO_2, HCO_3, Base Excess, SaO_2.

Demographics/History/Context (**19** features): Care Unit (Surgical, Cardiac Care, or Neuro-intensive care), Surgery in the past 12 hours, Wound Class (clean, contaminated, dirty, or infected), Surgical Specialty (Cardiovascular, Neuro, Ortho-Spine, Oncology, Urology, etc.), Number of antibiotics in the past 12, 24, and 48 hours, Age, Charleston Comorbidity Index (CCI), Mechanical Ventilation, maximum change in SOFA score over the past 6 hours.

The **reduced model** excluded the less commonly measured clinical variables, such as ABG laboratory features, as well as Bilirubin direct, Glucose, B-type Natriuretic Peptide (BNP), Fibrinogen, CRP, Sedimentation Rate, Ammonia, resulting in a total of **53** features. As shown in Fig. B1, performance of the reduced model was comparable to the full model.

[1] https://www.physionet.org/physiotools/mse/tutorial/tutorial.pdf

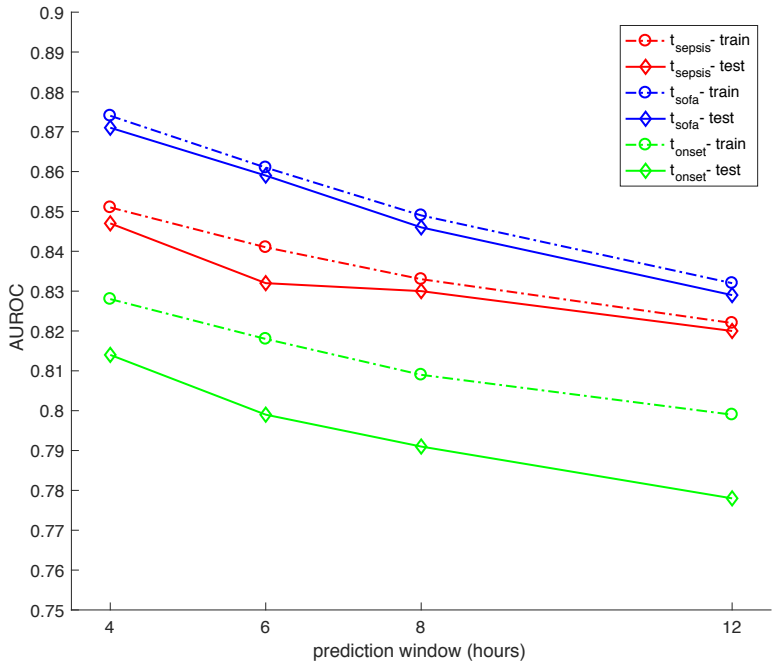

FIGURE B1. Summary of the training set (dashed lines) and testing set (solid lines) prediction performance of AISE on the Emory cohort, Area under the ROC curve (AUROC) as a function of prediction window shows a decreasing pattern, as expected. The reduced model shows similar performance to the full model (Fig. 2) across all windows and all prediction tasks, indicating robustness to uncommonly measured clinical laboratory values.

APPENDIX C

Machine Learning: The proposed Artificial Intelligence Sepsis Expert (AISE) algorithm is based on a modified Weibull-Cox proportional hazards model, designed to predict onset of sepsis in the proceeding T hours (where T = 12, 8, 6 or 4 hours). The Weilbull proportional hazards model is a robust parametric counterpart of the more familiar Cox time-to-event analysis. The Weibull-Cox model assumes a traditional Cox proportional hazards hazard rate but with a Weibull base hazard rate.

We assume we have observed sepsis-related data $D = \{(x_1, \tau_1, s_1), (x_2, \tau_2, s_2), \ldots, (x_N, \tau_N, s_N)\}$ for N observation windows across the entire patient population, where x_i is a set of features, $\tau_i>0$ is the time until a sepsis event, and $s_i=0$ indicates a sepsis event that occurred within the i-th observation window, which $s_i=1$ indicates right censoring (sepsis event did not occur within the observation window; but may have occurred outside the observation window). A rigorous mathematical treatment of the Weibull-COX proportional hazard model is beyond the scope of this article. Instead we make an attempt to provide an intuitive explanation of the model. First, let us define a few terms:

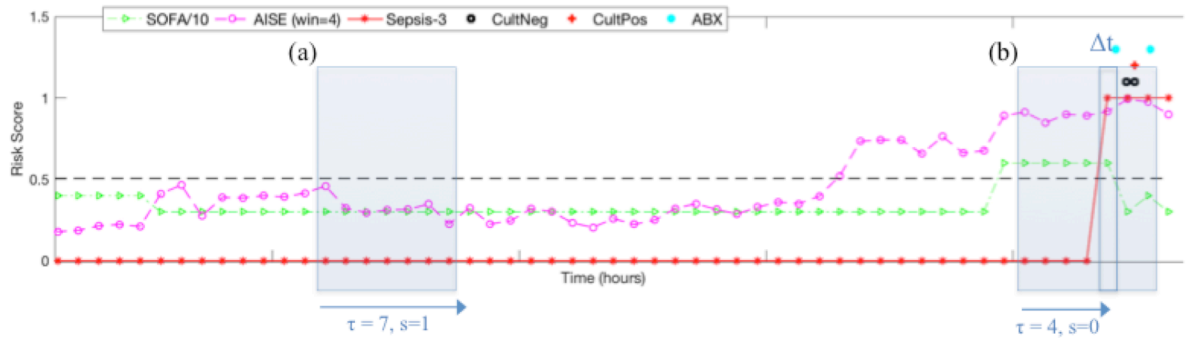

Figure C1. Depiction of the proposed sliding window Weibull-COX proportional hazard based prediction. A sliding window (e.g., 6 hours duration) moves through the time series, hour-by-hour. If no sepsis event falls within the window (case (a)), we set s_i=1 indicating right censoring and τ>6. If a sepsis event falls within the window time-to-sepsis is recorded (e.g., τ=4 in case (b)) and we set s_i=0 indicating a sepsis event. Finally, if the patient record was terminated within a given prediction window (due to death, transfer or discharge from the ICU) that window was also marked as censored.

- Survival function S(τ): Gives the probability of not getting sepsis up to time τ. S(τ)=1-F(τ), where F(τ) is the probability that a sepsis event occurred by time τ.
- Hazard function h(τ): Gives the conditional probability that the sepsis event will occur in the time bin (1hour width) τ, given that it did not occur up to time τ and will occur sometime t >= τ. In other words, h(τ) is the instantaneous risk of sepsis at time τ. This risk is often characterized in terms of a baseline risk (or population level risk; denoted by h_0) and a patient specific risk, which is a function of patient's features (vitals, labs, etc.).

The Weibull-Cox model chooses a Weibull base hazard function: $h_0(\tau|\,\lambda, k) = (k/\lambda)(\tau/\lambda)^{k-1}$,

where λ>0 is a scale parameter and k>0 is a shape parameter. The hazard rate for the *i*-th window is given by: $h_i(\tau|x_i, \lambda, k, \beta) = h_0(\tau|\,\lambda, k)e^{\beta x_i}$, which is the product of the baseline hazard and the patient specific risk as related to features x_i . And the survival curve (probability of not getting sepsis up to time τ) is given by $S(\tau|x_i, \lambda, k,) = e^{-\Lambda_0(\tau_i)e^{\beta x_i}}$, where $\Lambda_0(\tau) = (\tau/\lambda)^k$ is the cumulative base hazard rate.

The data likelihood (how well the observed data is explained by the model) is given by:

$$P(D|\lambda, k, \beta) = \prod_{i=1}^{N} [h_0(\tau_i|\,\lambda, k)e^{\beta x_i}]^{1-s_i} e^{-\Lambda_0(\tau_i)e^{\beta x_i}}$$

$$= \prod_{i=1}^{N} [h_0(\tau_i|\,\lambda, k)e^{\beta x_i}]^{1-s_i} S(\tau_i|x_i, \lambda, k,). \quad (1)$$

Intuitively, maximizing the data likelihood corresponds to maximizing the probability that an event did not occur before time τ_i (i.e., survival term $S(\tau_i)$) and maximizing the probability of actual sepsis events, when events are not censored (or 1-s_i is equal to1), i.e., the term inside the bracket.

Parameters of this model (λ, k, β) are learned through a maximum likelihood approach, i.e., the model parameters are tweaked in an iterative fashion (using a mini-batch stochastic gradient descent approach) in order to maximize the log likelihood of the data (logarithm of Eq. (1)). In practice, a regularization term (we used L1-L2 regularization) is added to the log likelihood to minimize overfitting and optimize generalizability of the learned model.

For a given prediction horizon T, the sepsis risk score is defined as:

$$\mathrm{Prob}(t_{\mathrm{sepsis}} <= \mathrm{T}) = 1 - S(\mathrm{T}|x_i, \lambda, \mathrm{k},) \quad (2)$$

This approach allowed us to make meaningful predictions, as opposed to predicting sepsis many days in advance. [5] However, using a *T*-hour sliding-window prediction approach resulted in over 1 million prediction windows within the development cohort training set, and roughly ¼ million prediction windows within the testing set, with only roughly 2% of these windows corresponding to a positive outcome (class imbalance). We used mini-batch stochastic gradient descent [6] with backpropagation to fit the model parameters. This approach made the learning algorithm scalable when dealing with millions of training examples, and provided a systematic way of handling class imbalance via oversampling the underrepresented class within each mini-batch. Moreover, backpropagation allowed us to quantify the time-varying contribution of each input feature to changes in the risk score, [7] thus making the AISE algorithm transparent and interpretable.

Visualization and Interpretability: Importance of each feature can be calculated using an approach similar to saliency maps in Deep Learning. [25] We simply take the derivative (or gradient) of the risk score in Eq. (2) with respect to all input features and multiply it by the input features. This is also known as the *relevance score,* which simply says that an input feature is relevant if it is both present in the data and if the model reacts to it (the derivative term). We calculate the Z-scored relevance score of all features and report any feature with a Z-score of larger than 1.96 (corresponding to 95% confidence interval). A similar technique has been previously applied in the context of Bayesian classification, [8] but to the best of our knowledge this is the first time such an approach has been applied in the context of survival analysis and its medical applications. Figures C2 and C3 show examples of such relevance scores for two subjects.

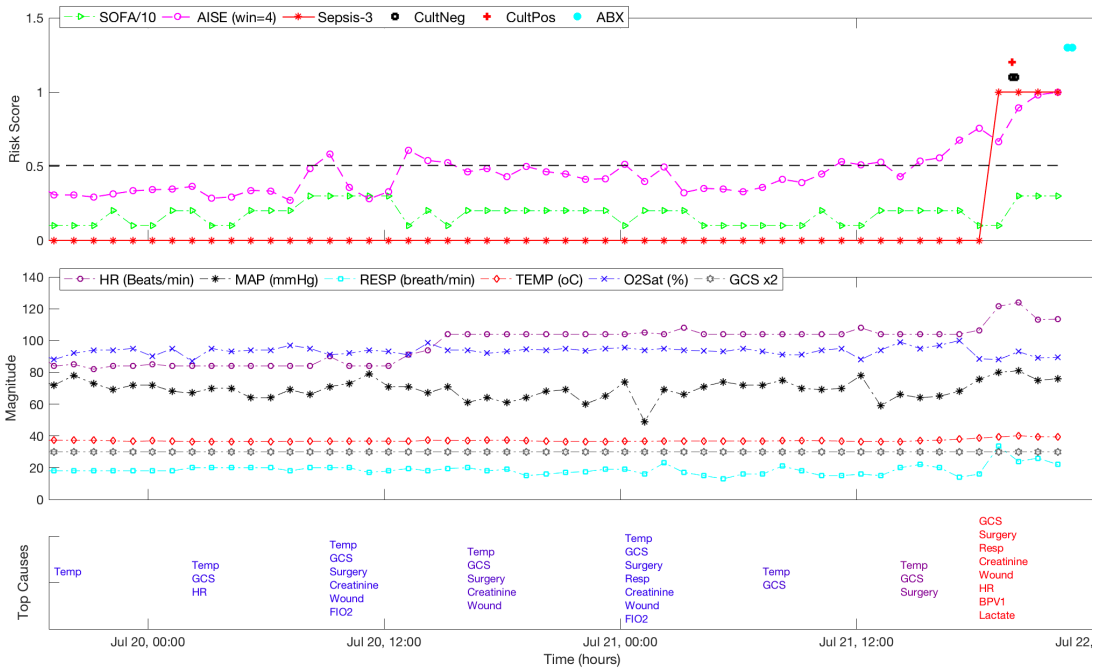

FIGURE C2. An illustrative example of the prediction performance of AISE. Hourly calculated Sequential Organ Failure Assessment (SOFA) Score, Sepsis-3 definition, and the AISE score are shown for one patient in Panel (A). Superimposed on the figure is the order-time of three blood cultures, and the administration-time of two antibiotics. In Panel (B), commonly recorded hourly vital signs of the patient, including heart rate (HR), Mean Arterial Blood Pressure (MAP), Respiratory Rate (RESP), Temperature (TEMP), Oxygen Saturation (O_2Sat) and the Glascow Coma Score (GCS) are shown. Panel (C) shows the most significant features contributing to the AISE score (for clarity of presentation only selected time-points are shown).

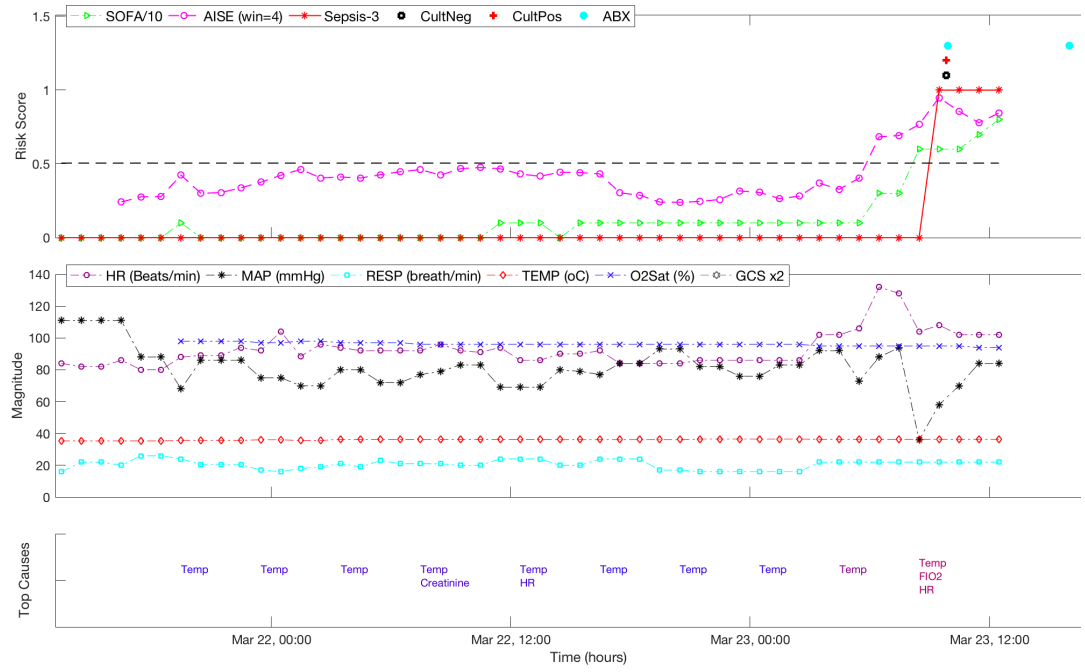

FIGURE C3. Another illustrative example of the prediction performance of AISE.

APPENDIX D

Distribution of Event Times: As shown in Fig. D1 (panel D), approximately 22% of the time clinical recognition of sepsis occurs before physiological manifestations of organ failure (as captured by 2 points change in SOFA), although the median difference between clinical suspicion of sepsis and a two-point change in SOFA was 10.3 [2.7, 19.4] hours.

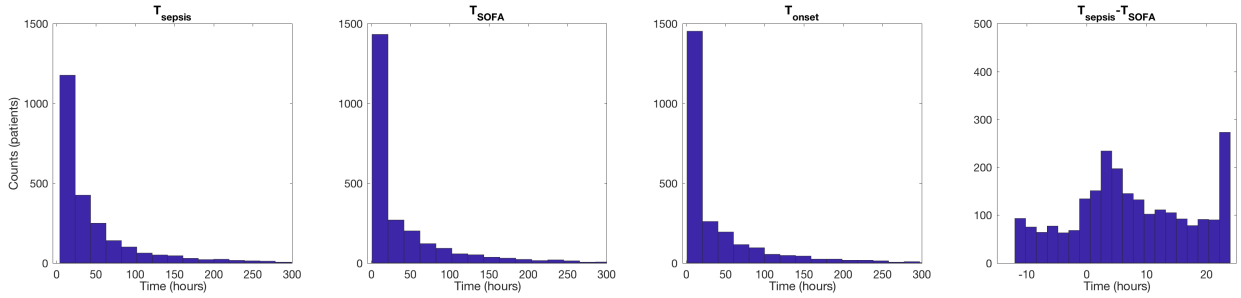

FIGURE D1. Distribution of elapsed times (in hours) from ICU admission to t_{sepsis} (Panel A), t_{SOFA} (Panel B), t_{onset} (Panel C), and t_{sepsis} -T_{SOFA} (Panel D).

APPENDIX E

Validation Cohort Results

Table E1. Summary of Validation Cohort (MIMIC-III) Characteristics

	Overall Cohort			Development Cohort		
Demographics	**All Patients**	**Non Septic**	**Septic**	**All Patients**	**Non Septic**	**Septic**
Patients (#)	52098	39224	12874 (25.0%)	42411	38566	3845 (9.1%)
Male (%)	56.4	56.3	56.6	56.5	56.3	58.8 *
Age (year)	66 [53 – 78]	65 [52 – 77]	67 * [54 – 79]	66 [52 – 77]	65 [52 – 77]	66 [50 – 71]
Race (%) **Caucasian** **Black** **Asian** **Hispanic**	71.8 9.5 2.3 3.4	71.5 9.2 2.2 3.4	72.7 * 10.3 * 2.6 * 3.4	71.6 9.2 2.2 3.4	71.5 9.2 2.2 3.4	72.4 8.6 2.4 3.4
ICU LOS (hours)	50 [28 – 100]	45 [26 – 75]	107 * [50 – 240]	47 † [27 – 88]	45 [26 – 74]	158 * [83 – 266]

Mortality (%)	12.2	8.2	24.3 *	9.3 †	7.8	25.0 *
ICD-9 (%): 995.92 or 785.52	8.9	2.6	28.3 *	3.9 †	2.6	17.2 *
SOFA	1.9 [0.8 – 3.5]	1.5 [0.6 – 2.9]	3.3 * [2.0 – 5.1]	1.6 † [0.65 – 3.1]	1.5 [0.6 – 2.9]	3.3 * [2.0 – 5.1]
CCI	2 [1 – 3]	2 [1 – 3]	2 * [1 – 4]	2 † [1 – 3]	2 [1 – 3]	2 * [1 – 4]
ICU Admission to Sepsis (hours)	—	—	-0.9 [-9.3 – 18.9]	—	—	31.2 [13.3 – 70.2]

* Statistically significant difference between septic and non-septic patients within each cohort

† Statistically significant difference between overall cohort and the development cohort

Table E2. Summary of algorithm performance on the MIMIC-III cohort

Performance metric	**4 hours**	**6 hours**	**8 hours**	**12 hours**
t_{sepsis} Prediction Testing set (Training set)				
AUROC	0.84 (0.84)	0.82 (0.82)	0.82 (0.81)	0.79 (0.80)
Specificity*	0.64 (0.66)	0.62 (0.63)	0.62 (0.60)	0.57 (0.58)
Accuracy	0.64 (0.66)	0.62 (0.64)	0.62 (0.61)	0.58 (0.59)
t_{sofa} Prediction Testing set (Training set)				
AUROC	0.85 (0.86)	0.83 (0.84)	0.82 (0.83)	0.80 (0.82)
Specificity*	0.66 (0.69)	0.61 (0.65)	0.60 (0.62)	0.56 (0.59)
Accuracy	0.67 (0.69)	0.62 (0.65)	0.60 (0.63)	0.57 (0.60)
t_{onset} Prediction Testing set (Training set)				
AUROC	0.80 (0.81)	0.78 (0.80)	0.77 (0.79)	0.76 (0.78)
Specificity*	0.57 (0.61)	0.54 (0.58)	0.52 (0.55)	0.51 (0.54)
Accuracy	0.57 (0.61)	0.55 (0.58)	0.53 (0.56)	0.52 (0.55)

* Sensitivity was fixed at 0.85 (catching 85% of sepsis cases)

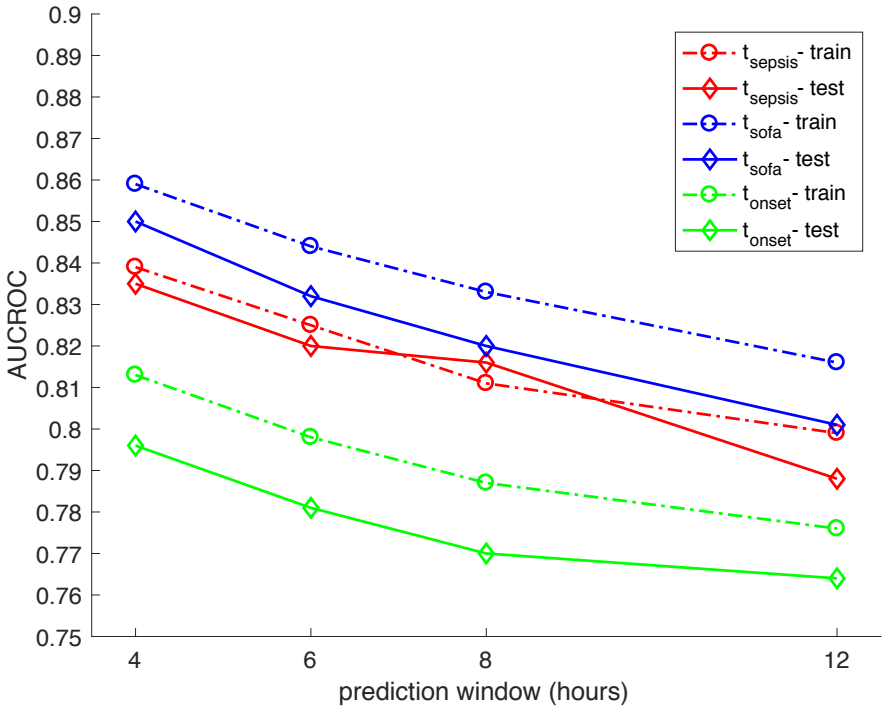

FIGURE E1. Summary of training set (dashed lines) and testing set (solid lines) prediction performance of AISE on the MIMIC cohort. Area under the ROC curve (AUROC) as a function of prediction window shows a decreasing pattern. Across all windows, the best performance is achieved for predicting t_{SOFA}, followed by t_{Sepsis}, and finally t_{onset}. A close agreement between the training set and testing set performance indicates good generalizability.

APPENDIX F

GLOSSARY OF MACHINE LEARNING TERMINOLOGY:

- *Backpropagation* – An approach used in certain types of differentiable machine learning models to calculate the sensitivity of model output with respect to model parameters and model input (i.e., how model reacts to an input). Backpropagation utilizes the *Chain rule* from calculus to accomplish this.
- *Overfitting* (aka, lack of generalizability) – When a model performs well on the training data (seen patients) and performs poorly on the testing data (unseen patients). Regularization is often used to minimize overfitting and optimize generalizability of machine leaning algorithms.
- *Bin* – The interval of time over which data of different sampling frequencies are collected to facilitate time series analysis.
- *Blood pressure variability (BPV)* – A group of metrics that measure variability in blood pressure over a period of time. They are calculated in short time intervals, or scales (e.g. 20 seconds). Some examples include the average standard deviation between two arterial wave complexes, and entropy measurements. These changes are believed to represent the interaction between the cardiovascular and other organ systems (e.g. neuro-cardiac). Physiologic health correlates to more variability within a scale (more organ-organ interactions/cross-talk).
- *Class imbalance* – The outcome, or "class" of interest (ICU sepsis) is not equally as likely as other outcomes (no ICU sepsis). It can introduce bias in machine learning, since the algorithm will preferentially choose the class/outcome which is more prevalent in the dataset from which it learns.
- *Entropy* – A group of metrics (e.g. multiscale entropy, multiscale conditional entropy, sample entropy) used to measure the complexity of high-resolution physiologic dynamics

over time. Higher entropy values correspond to more complex dynamics in a system or organism.

- *Features* – input variables used by machine learning algorithms.
- *Heart rate variability (HRV)* – A group of metrics that measure changes in time intervals between successive heart beats (i.e. time between two QRS complexes). They are calculated in short time intervals, or scales (e.g. 20 seconds). Some examples include the average standard deviation between two QRS complexes, and entropy measurements. These changes are believed to represent the interaction between the cardiovascular and other organ systems (e.g. neuro-cardiac). Physiologic health correlates to more variability within a scale (more organ-organ interactions/cross-talk).
- *High-resolution/high-frequency* – Data collection at frequencies over 1 Hz (once per second). Used to describe data derived directly from physiologic waveforms using automated methods.
- *Low-resolution/low-frequency* – Data collection at frequencies of less than once per hour. Used to describe data collected from the electronic medical record (laboratory data, manually-derived vitals).
- *Non-overlapping (time series) bins* – Consecutive intervals of time with no common time segments or datapoints.
- *Overfitting* – A statistical model incorporates noise or random error into prediction, usually because there are too many features compared to the number of observations
- *Overlapping (time series) bins* - Consecutive intervals of time which share some common time segments or datapoints.
- *Prediction horizon/time window* – The time interval (lead time) over which prediction occurs.
- *Sampling frequency* – Number times a measurement (e.g., HR) is measured per second (in units of 1/sec or Hz). A sampling frequency of 2Hz means the measurement was done twice every second. Conversely, a sampling frequency of 0.5Hz mean the measurement was done once every 2 seconds.
- *Scale* – The interval of time from which variability measurements such as sample entropy are derived.
- *Stochastic* – When values in a time series include a random component, not completely predetermined from prior values.

REFERENCES

[1] Vincent JL, Moreno R, Takala J, et al: The SOFA (Sepsis-related Organ Failure Assessment) score to describe organ dysfunction/failure. On behalf of the Working Group on Sepsis-Related Problems of the European Society of Intensive Care Medicine. *Intensive Care Med* 1996. **22**(7): p. 707-10.

[2] Costa M, Goldberger AL, Peng CK: Multiscale entropy analysis of complex physiologic time series. Physical review letters. 2002 Jul 19;**89**(6):068102.

[3] Nemati S, Edwards BA, Lee J, et al: Respiration and heart rate complexity: effects of age and gender assessed by band-limited transfer entropy. Respiratory physiology & neurobiology. 2013 Oct 1;**189**(1):27-33.

[4] Carroll KJ. On the use and utility of the Weibull model in the analysis of survival data. Controlled clinical trials. 2003 Dec 31;24(6):682-701.

[5] Henry KE, Hager DN, Pronovost PJ, et al: A targeted real-time early warning score (TREWScore) for septic shock. *Sci Transl Med* 2015. **7**(299): p. 299ra122.

[6] Hinton G, Srivastava N, and Sewrsky K. Overview of mini-batch gradient descent, in Neural Networks for Machine Learning. 2012.

[7] Simonyan K, Vedaldi A, and Zisserman: A. Deep inside convolutional networks: Visualising image classification models and saliency maps., U.o.O. Visual Geometry Group, Editor. 2014.

[8] Baehrens D, et al: How to explain individual classification decisions. JMLR, 11:1803–1831, 2010.