AI-Driven Predicton of ICU Mortality Through Digital Processing of Vital Signs

[1]Paul Rogers, [1]Dong Wang, [2]Youjin Wang, [3]Khalid Puthawala, [4]Tariq Fahmi, [4]Beverly Lyn-Cook, [1]Wen Zou, [1]Weida Tong

[1]Division of Bioinformatics and Biostatistics, National Center for Toxicological Research,U.S. Food & Drug Administration, Jefferson, AR 72079, USA
[2]Office of Surveillance and Epidemiology (OSE), Center for Drug Evaluation and Research; U.S. Food & Drug Administration, Silver Springs, MD 20993, USA
[3]Division of Pulmonology, Allergy, and Critical Care, Center for Drug Evaluation and Research; U.S. Food & Drug Administration, Silver Springs, MD 20993, USA
[4]Division of Biochemical Toxicology, National Center for Toxicological Research,U.S. Food & Drug Administration, Jefferson, AR 72079, USA

Introduction

Patient vital signs in the intensive care unit (ICU) are traditionally monitored and recorded on an hourly basis. Several factors can influence patient vital signs, including medications, treatments, and the condition or injury from which the patient suffers. Variation in vital signs is believed to be key in predicting impending patient death or recovery. Although some patients exhibit significant changes in vital signs as death approaches, others do not.

Providing advance warning of mortality risk to health-care providers allows the opportunity for interventions to improve the patient's chances of survival. The increasing number of ICU patients within the US makes this topic a significant public health issue.

The ICU data for this study was obtained from the Medical Information Mart for Intensive Care version IV (MIMIC-IV), generated from the Beth Israel Deaconess Medical Center (BIDMC) in Boston from 2008 through 2019. We selected subjects who were 20 years and older and in one of the 9 ICUs for at least 24 hours but no more than 7 days.

The vital signs data, in addition to the use of mechanical ventilation, were restructured into an hourly longitudinal format for each eligible patient over the entirety of their ICU stay. In other words, we denormalized the MIMIC-IV dataset to present the data in its most granular form (Fig1).

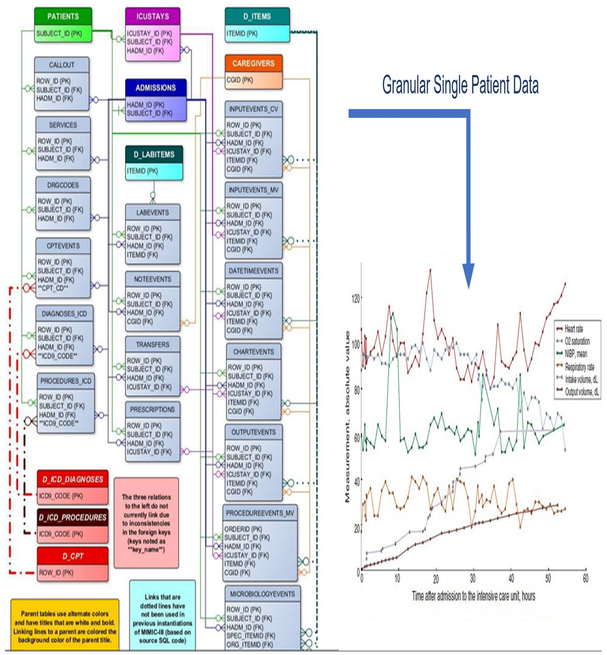

Figure 1. Conceptual diagram of how the MIMIC-IV vital signs are denormalized and restructured into a longitudinal hour-by-hour format to expose the most granular level of individual patient data.

Materials and methods

Denormalizing MIMIC-IV and reorganizing the data into a longitudinal database on a per-patient basis requires an algorithm that can predict whether mortality occurs and the specific hour it happens. When the data is structured in this way, the occurrence of mortality becomes a rare event, leading to a highly imbalanced dataset. When dealing with a rare outcome in a classification algorithm, metrics like Accuracy can be misleading.

A Bidirectional Long Short-Term Memory (LSTM) AI algorithm processed these vital signs to make a mortality prediction for each patient. We chose an LSTM model as they handle sequential data, capturing temporal dependencies and patterns vital to understanding patient health trajectories. LSTM models are good at remembering long-term dependencies that are key in identifying improving or deteriorating patient conditions. In addition to their ability to filter statistical noise and focus on underlying patterns, they are adept at handling missing values. LSTMs are able to identify both linear and non-linear relationships between different physiological variables. An optimal LSTM model enables the possibility for early detection of critical events, personalized care, and, ultimately, improving patient outcomes in the ICU.

The LSTM algorithm examines the patterns of each patient's vital signs, including heart rate, systolic blood pressure, diastolic blood pressure, mean arterial pressure, percent of oxygen in blood (SpO2), fraction of inspired oxygen (FiO2), respiratory rate, temperature, mechanical ventilation, and the Glasgow coma scores when available. There were no other laboratory tests or patient demographics outside of Age and Sex in predicting patient outcomes. The algorithm examines the patterns in two directions, as described in Fig 2.

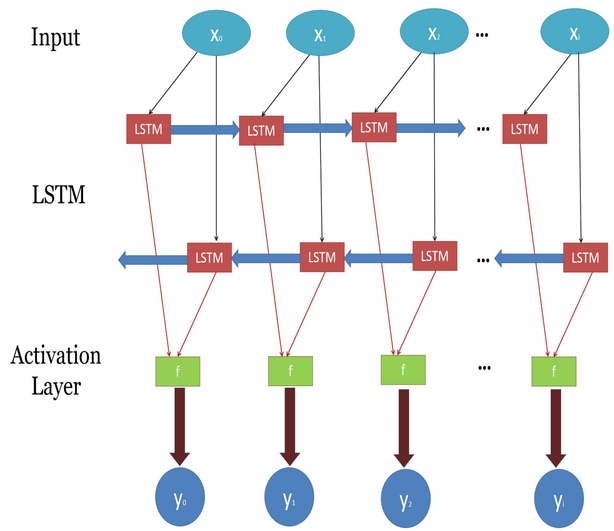

Figure 2. Conceptual diagram of a bidirectional LSTM algorithm.

The algorithm is, in effect, a screening test designed to identify patients at risk of mortality based on their last 24 hours of vital signs. Sensitivity and the Positive Predictive Value (PPV) are metrics for assessing the performance of screening tests (Fig3). Alternate labels for Sensitivity and PPV are Recall and Precision, respectively.

	Gold Standard		
Test Results	True Positive	True Negative	Row Sums
Test Positive	TP	FP	TP + FP
Test Negative	FN	TN	FN + TN
Column Sums	TP + FN	FP + TN	TP + FN + FP + TN

$$PPV = \frac{TP}{TP + FP} \qquad NPV = \frac{TN}{TN + FN}$$

$$Sensitivity = \frac{TP}{TP + FN} \qquad Specificity = \frac{TN}{TN + FP}$$

Figure 3. The traditional 2x2 screening test table with formulas for Sensitivity, Positive Predictive Value (PPV), Negative Predictive Value (NPV), and Specificity written in terms of True Positives (TP), False Positives (FP), True Negatives (TN), and False Negatives (FN).

In the context of our problem, Sensitivity is the proportion of patients who are going to die that are correctly identified as positive by the algorithm. We want to reduce the number of False Negatives (FN) to maximize Sensitivity. Reducing the number of FNs also maximizes the Negative Predictive Value (NPV). In the clinical setting, however, a different question may be important for the physician: If the test results are positive in the patient, what is the probability that this patient will die? This is the PPV. We want to reduce the number of False Positives (FP) to maximize PPV. Reducing the number of FPs also maximizes the Specificity.

The PPV is highly susceptible to the prevalence. As the outcome becomes less common, the PPV's performance declines significantly. The performance of Sensitivity is based on the number of FNs. The cutoff threshold can be adjusted to balance the number of FNs and FPs generated by the test. This is, in essence, a tradeoff in performance between Recall and Precision, as depicted in Fig 4. Since we are screening patients for risk of mortality, a FN is much more damaging than a FP in this scenario. Therefore, we tuned our model for high Recall at the expense of Precision. The F beta score described in Equation 1 is a performance metric that considers this tradeoff.

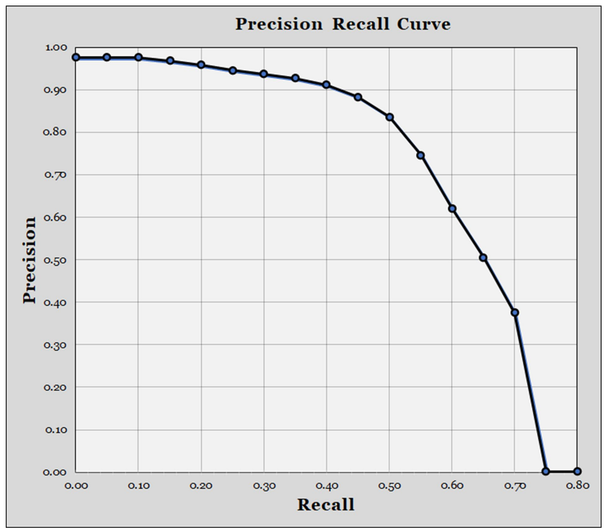

Figure 4. The Precision-Recall curve for patients who spent 1 to 7 days in the ICU in a model tuned for high Precision (PPV).

$$F_{beta} = (1 + \beta^2)\frac{precision * recall}{\beta^2 * precision + recall}$$

Equation 1. The F beta score combines Precision and Recall into a single metric. The more you care about Recall over Precision, then beta should be greater than 1. When beta is less than 1 (0 < beta < 1), the more we favor Precision.

Results and discussion

Descriptive statistics for the ICU patient population over the age of 20 are given in Table 1.

Descriptive Statistics for ICU Patients 20 Years and Older						
Sex	N	Avg Age	ICU stay from 1 to 3 days	ICU stay from 1 to 5 days	ICU stay from 1 to 7 days	Died
Male	40,664	62.5	68%	82%	88%	4,528
Female	32,237	64.5	69%	83%	89%	3,790

Table 1. Summary statistics for BIDMC ICU patients over the age of 20 with stays varying from 1 to 7 days.

Our implementation of a bidirectional LSTM AI algorithm coupled with highly granular ICU patient data was predictive of mortality risk. The model could be tuned for a high Sensitivity or high PPV, as seen in Table 2. Since the algorithm functions as a population screening tool, higher Sensitivity is preferred over a higher PPV. A Sensitivity of 99% was achieved for ICUpatients with stays of 1 to 7 days.

Six Models Tuned for High Precision or Recall			
Models Tuned for High Precision			
	1-3 Day ICU	1-5 Day ICU	1-7 Day ICU
Binary Accuracy	0.9987	0.9990	0.9990
Recall (Sensitivity)	0.3262	0.4153	0.4089
Precision (PPV)	0.9302	0.8916	0.9080
$F_{(1/100)}$	0.9300	0.8915	0.9079
Models Tuned for High Recall			
	1-3 Day ICU	1-5 Day ICU	1-7 Day ICU
Binary Accuracy	0.9790	0.9682	0.9781
Recall (Sensitivity)	0.9757	0.9935	0.9901
Precision (PPV)	0.0776	0.0491	0.0671
F_{100}	0.9746	0.9916	0.9887

Table 2. Performance metrics for six different models tuned for high Precision and high Recall involving BIDMC patients with varying ICU stays.

Conclusion

The results confirm that there is potential in predicting mortality risk within ICU patients utilizing routinely collected vital signs. **An AI algorithm tuned for high Sensitivity offers the opportunity of an early warning to health care providers concerning patient mortality risk.**

Disclaimer

This poster reflects the views of the authors and does not necessarily reflect those of the U.S. Food and Drug Administration.