SCIENTIFIC REPORTS

OPEN

DeepSOFA: A Continuous Acuity Score for Critically Ill Patients using Clinically Interpretable Deep Learning

Received: 17 August 2018
Accepted: 28 December 2018
Published online: 12 February 2019

Benjamin Shickel[1], Tyler J. Loftus[2], Lasith Adhikari[3,5], Tezcan Ozrazgat-Baslanti[3,5], Azra Bihorac[3,5] & Parisa Rashidi[1,4,5]

Traditional methods for assessing illness severity and predicting in-hospital mortality among critically ill patients require time-consuming, error-prone calculations using static variable thresholds. These methods do not capitalize on the emerging availability of streaming electronic health record data or capture time-sensitive individual physiological patterns, a critical task in the intensive care unit. We propose a novel acuity score framework (DeepSOFA) that leverages temporal measurements and interpretable deep learning models to assess illness severity at any point during an ICU stay. We compare DeepSOFA with SOFA (Sequential Organ Failure Assessment) baseline models using the same model inputs and find that at any point during an ICU admission, DeepSOFA yields significantly more accurate predictions of in-hospital mortality. A DeepSOFA model developed in a public database and validated in a single institutional cohort had a mean AUC for the entire ICU stay of 0.90 (95% CI 0.90–0.91) compared with baseline SOFA models with mean AUC 0.79 (95% CI 0.79–0.80) and 0.85 (95% CI 0.85–0.86). Deep models are well-suited to identify ICU patients in need of life-saving interventions prior to the occurrence of an unexpected adverse event and inform shared decision-making processes among patients, providers, and families regarding goals of care and optimal resource utilization.

Critically ill patients in the intensive care unit (ICU) have a life-threatening condition or the propensity to develop one at any moment. Early recognition of evolving illness severity in the ICU is invaluable. Timely and accurate illness severity assessments may identify patients in need of life-saving interventions prior to the occurrence of an unexpected adverse event and may inform shared decision-making processes among patients, providers, and families regarding goals of care and optimal resource utilization.

One of the most commonly used tools for assessing ICU patient acuity is the Sequential Organ Failure Assessment (SOFA) score[1]. SOFA considers 13 variables representing six different organ systems (cardiovascular, respiratory, nervous, liver, coagulation, and renal) and uses their worst measurements over a given interval (typically 24 hours) in conjunction with static value thresholds to assign numerical scores for each component. The sum of these component scores yields the overall SOFA score, which can be used to assess illness severity and predict mortality[2–4]. Although SOFA provides a reasonably accurate assessment of a patient's overall condition and mortality risk, its accuracy is hindered by fixed cutoff points for each component score, and SOFA variables are often infrequent or missing in electronic health records. In particular, Glasgow Coma Scale scores and measurements of serum bilirubin and partial pressure of arterial oxygen are often sparse. Badawi *et al.*[5] performed retrospective hourly recalculations of several acuity scores for ICU patients at 208 hospitals in the Philips eICU Research Institute database, reporting that hourly SOFA scores predicted ICU mortality with mean area under the receiver operating characteristic curve (AUC) of 0.86. Although hourly acuity score calculations may provide

[1]Department of Computer and Information Science and Engineering, University of Florida, Gainesville, FL, 32611, USA. [2]Department of Surgery, University of Florida, Gainesville, FL, 32611, USA. [3]Department of Medicine, University of Florida, Gainesville, FL, 32611, USA. [4]Department of Biomedical Engineering, University of Florida, Gainesville, FL, 32611, USA. [5]Precision and Intelligent Systems in Medicine (PRISMAP), University of Florida, Gainesville, FL, 32611, USA. Azra Bihorac and Parisa Rashidi contributed equally. Correspondence and requests for materials should be addressed to A.B. (email: abihorac@ufl.edu)

advantages despite the potentially confounding impact of transient and self-limited fluctuations in real-time data[6], they are only feasible if implemented as an autonomous real-time process.

The availability of temporal trends and high-fidelity physiologic measurements in the ICU offers the opportunity to apply computational approaches beyond existing conventional models[7–9]. Our primary aim was to develop an acuity score framework that encompasses the full scope of a patient's physiologic measurements over time to generate dynamic in-hospital mortality predictions. Our solution uses deep learning, a branch of machine learning that encompasses models and architectures that learn optimal features from the data itself, capturing increasingly complex representations of raw data by combining layers of nonlinear data transformations[10,11]. Deep learning models automatically discover latent patterns and form high-level representations from large amounts of raw data without the need for manual feature extraction based on *a priori* domain knowledge or practitioner intuition, which is time-consuming and error-prone. Deep learning has revolutionized natural language processing, speech recognition, and computer vision, and is gaining momentum within healthcare[12]. Computer vision has been used to identify diabetic retinopathy[13] and recognize skin cancer with accuracy similar to that of a board-certified dermatologist[14]. Deep models have also been used to predict pain responses[15], the onset of heart failure[16], and ICU mortality[17].

Here we report the development and external validation of DeepSOFA, a deep learning model that employs a clinician-interpretable variant of recurrent neural network (RNN) to analyze multivariate temporal clinical data in the ICU. Experiments were performed with two independent hospital populations and were designed to be cross-institutional; we report internal and externally validated results for both hospital cohorts. Cohorts were derived from ICU admissions at the University of Florida Health Hospital and the publicly available Medical Information Mart for Intensive Care (*MIMIC*-III) dataset that contains records for ICU patients from the Beth Israel Deaconess Medical Center in Boston, Massachusetts[18]. We compared deep learning mortality prediction models trained on hourly measurements with baseline models using traditional SOFA score definitions and the same hourly measurements using the entirety of a patient's data stream over the same time period. Two baseline SOFA models were tested: a Bedside SOFA model using published mortality rates correlating with any given total SOFA score[2], and a Traditional SOFA model in which hourly SOFA scores are correlated with in-hospital mortality for individual patients[5]. Because deep models automatically learn the complex, nonlinear associations among input variables, we hypothesized that DeepSOFA would yield greater accuracy in predicting in-hospital mortality among ICU patients compared with traditional SOFA techniques.

Results

Development of DeepSOFA model.

Two datasets (*UFHealth* and *MIMIC*) derived from two distinct cohorts of ICU patients from two academic medical centers, University of Florida Health (Gainesville, FL) and Beth Israel Deaconess Medical Center (Boston, MA), respectively, were used for model development and external cross validation (Table 1). The *UFHealth* cohort included 36,216 ICU admissions for 27,660 patients, and the *MIMIC* cohort included 48,948 ICU admissions for 35,993 patients. To ensure that results would be generalizable to all patients entering an ICU at any phase of a hospital admission, all ICU admissions and readmissions for all patients were analyzed. Cohorts were comparable in terms of patient characteristics and outcomes, with slightly shorter ICU stays with a median of 2.1 days (25th–75th percentiles 1.2–4.1) vs. 2.9 days (25th–75th percentiles 1.5–5.9), shorter time between hospital and ICU admission with a median of 0.1 hours (25th–75th percentiles 0.0–24.6) vs. 7.0 hours (25th–75th percentiles 1.8–21.9), longer time between ICU and hospital discharge with a median of 73.1 hours (25th–75th percentiles 27.0–143.0) vs. 48.4 hours (25th–75th percentiles 0.0–122.6), and greater proportion of ICU stays requiring mechanical ventilation (47.8% vs. 30.4%) for the *MIMIC* cohort compared to *UFHealth* cohort. The *MIMIC* cohort also included a greater proportion of Medical ICU admissions (38.9% vs. 25.1%) and Cardiac ICU admissions (32.4% vs. 18.3%), with fewer Surgical ICU admissions (28.7% vs. 33.0%). The DeepSOFA model was trained and internally validated with 5-fold cross validation in each cohort separately.

The internal validation demonstrated excellent AUC performance for models in each cohort (Supplementary Figs S2 and S3). DeepSOFA developed and validated in the *UFHealth* cohort had AUC ranging from 0.72, 95% CI 0.71–0.72 ($p < 0.05$ compared with Bedside SOFA AUC of 0.61, 95% CI 0.60–0.62 and $p < 0.05$ compared with Traditional SOFA AUC of 0.63, 95% CI 0.62–0.63) at one hour after ICU admission to 0.93, 95% CI 0.93–0.94 ($p < 0.05$ compared with Bedside SOFA AUC of 0.82, 95% CI 0.81–0.83 and $p < 0.05$ compared with Traditional SOFA AUC of 0.88, 95% CI 0.88–0.89) at time of ICU discharge. DeepSOFA developed and validated in the *MIMIC* cohort had AUC ranging from 0.67, 95% CI 0.66–0.67 ($p < 0.05$ compared with Bedside SOFA AUC of 0.59, 95% CI 0.58–0.60 and $p < 0.05$ compared with Traditional SOFA AUC of 0.61, 95% CI 0.61–0.62) at one hour after ICU admission to 0.94, 95% CI 0.93–0.94 ($p < 0.05$ compared with Bedside SOFA AUC of 0.81, 95% CI 0.80–0.82 and $p < 0.05$ compared with Traditional SOFA AUC of 0.85, 95% CI 0.84–0.86) at time of ICU discharge.

External validation of DeepSOFA model.

For predicting in-hospital mortality, DeepSOFA significantly outperformed traditional SOFA models in external validation cohorts regardless of which cohort was used for model development. The DeepSOFA model developed in the *MIMIC* cohort and validated in the *UFHealth* cohort had a mean AUC for the entire ICU stay of 0.90, 95% CI 0.90–0.91 ($p < 0.05$ compared with Bedside SOFA AUC of 0.79, 95% CI 0.79–0.80 and $p < 0.05$ compared with Traditional SOFA AUC of 0.85, 95% CI 0.85–0.86, Fig. 1A), while the DeepSOFA model developed in the *UFHealth* cohort and validated in the *MIMIC* cohort had mean AUC of 0.90, 95% CI 0.90–0.90 ($p < 0.05$ compared to Bedside SOFA AUC of 0.78, 95% CI 0.77–0.79 and $p < 0.05$ compared to Traditional SOFA AUC of 0.82, 95% CI 0.81–0.82, Fig. 1B).

DeepSOFA models had significantly higher AUC across all hours of ICU stays, starting at the second hour of ICU admission with AUC of 0.74 (95% CI 0.73–0.75) in the *UFHealth* cohort (compared to Bedside SOFA AUC of 0.65, 95% CI 0.64–0.66, $p < 0.05$ and the Traditional SOFA AUC of 0.67, 95% CI 0.66–0.67, $p < 0.05$) and AUC of 0.68 (95% CI 0.67–0.68) in the *MIMIC* cohort (compared to Bedside SOFA AUC of 0.62, 95% CI 0.62–0.63,

	UFHealth (n = 36216)	*MIMIC* (n = 48948)
Hospital admissions, n	**33953**	**45748**
Patients, n	27660	35993
Female gender, n (%)	15170 (44.7%)	19558 (42.8%)
Age, median (25th, 75th)	61.5 (49.2, 71.5)	64.5 (52.0, 76.1)
Body mass index, median (25th, 75th)	27.0 (23.1, 31.9)	27.3 (23.7, 31.7)
Charlson comorbidity index, median (25th, 75th)	2 (0, 3)	2 (2, 2)
Race, n (%)		
White	25837 (76.1%)	32462 (71.0%)
African-American	5450 (16.1%)	4471 (9.8%)
Hispanic	1081 (3.2%)	1714 (3.7%)
Asian	238 (0.7%)	1079 (2.4%)
Other/Missing	1347 (4.0%)	6022 (13.2%)
Hospital LOS, days, median (25th, 75th)	7.1 (3.9, 13.0)	6.9 (4.1, 11.9)
In-hospital mortality rate, %	10.40%	10.80%
Discharged to hospice, n (%)	1011 (3.0%)	488 (1.1%)
Hospital LOS for non-survivors, days, median (25th, 75th)	6.2 (2.4, 13.4)	6.1 (2.2, 13.2)
ICU admissions, n	**36216**	**48948**
ICU stays per hospital admission, median (min, max)	1.0 (1.0, 6.0)	1.0 (1.0, 7.0)
ICU LOS, days, median (25th, 75th)	2.9 (1.5, 5.9)	2.1 (1.2, 4.1)
ICU type, n (%)		
Medical ICU	9102 (25.1%)	19054 (38.9%)
Cardiac ICU	6621 (18.3%)	15844 (32.4%)
Surgical ICU	11941 (33.0%)	14050 (28.7%)
Neurological ICU	7633 (21.1%)	0 (0.0%)
Burn ICU	642 (1.8%)	0 (0.0%)
Pediatric ICU	277 (0.8%)	0 (0.0%)
Deaths occurring in ICU, n (% all deaths)	2768 (71.3%)	3873 (68.9%)
Hours between hospital admission and ICU admission, median (25th, 75th)	7.0 (1.8, 21.9)	0.1 (0.0, 24.6)
Hours between ICU discharge and hospital discharge/death, median (25th, 75th)	48.4 (0.0, 122.6)	73.1 (27.0, 143.0)
Variables, n	**14**	**14**
ICU stays requiring any vasopressor, n (%)	6216 (17.2%)	9418 (19.2%)
ICU stays requiring mechanical ventilation, n (%)	11004 (30.4%)	23410 (47.8%)
Urine, mL (min, max)	(0.0, 1090.0)	(0.0, 1090.0)
Frequency, hours, median (25th, 75th)	1.0 (1.0, 1.9)	1.0 (1.0, 1.5)
Measured value, median (25th, 75th)	100.0 (50.0, 200.0)	85.0 (45.0, 160.0)
ICU stays missing any measurement, n (%)	1213 (3.3%)	2080 (4.2%)
Bilirubin, mg/dL (min, max)	(0.1, 49.2)	(0.1, 50.0)
Frequency, hours, median (25th, 75th)	24.2 (19.0, 48.0)	23.0 (10.8, 25.1)
Measured value, median (25th, 75th)	0.6 (0.3, 1.5)	1.1 (0.5, 3.7)
ICU stays missing any measurement, n (%)	17278 (47.7%)	27884 (57.0%)
Creatinine, mg/dL (min, max)	(0.1, 26.5)	(0.1, 29.1)
Frequency, hours, median (25th, 75th)	15.2 (6.3, 24.0)	13.3 (8.5, 23.5)
Measured value, median (25th, 75th)	0.9 (0.7, 1.4)	1.0 (0.7, 1.7)
ICU stays missing any measurement, n (%)	1714 (4.7%)	942 (1.9%)
FiO2, % (min, max)	(21.0, 100.0)	(21.0, 100.0)
Frequency, hours, median (25th, 75th)	1.0 (1.0, 2.0)	3.0 (2.0, 7.0)
Measured value, median (25th, 75th)	40.0 (30.0, 40.0)	37.0 (29.0, 50.0)
ICU stays missing any measurement, n (%)	0 (0.0%)	0 (0.0%)
PaO2, mmHg (min, max)	(9.0, 720.0)	(1.0, 775.0)
Frequency, hours, median (25th, 75th)	4.2 (2.6, 8.3)	3.9 (1.8, 7.6)
Measured value, median (25th, 75th)	119.0 (87.9, 155.0)	109.0 (83.0, 151.0)
ICU stays missing any measurement, n (%)	17257 (47.7%)	19739 (40.3%)
Mean arterial pressure, mmHg (min, max)	(1.0, 300.0)	(0.4, 300.0)
Frequency, hours, median (25th, 75th)	1.0 (0.2, 1.0)	1.0 (0.5, 1.0)
Measured value, median (25th, 75th)	79.0 (69.0, 91.0)	77.7 (68.0, 89.0)
ICU stays missing any measurement, n (%)	0 (0.0%)	0 (0.0%)

Continued

	UFHealth (n = 36216)	*MIMIC* (n = 48948)
Platelet count, $\times 10^3$/mm^3 (min, max)	(1.0, 832.0)	(5.0, 832.0)
Frequency, hours, median (25th, 75th)	17.7 (6.5, 24.0)	18.4 (8.0, 24.0)
Measured value, median (25th, 75th)	175.0 (106.0, 257.0)	186.0 (118.0, 273.0)
ICU stays missing any measurement, n (%)	1949 (5.4%)	1187 (2.4%)
GCS, score (min, max)	(3.0, 15.0)	(3.0, 15.0)
Frequency, hours, median (25th, 75th)	1.0 (1.0, 4.0)	4.0 (2.0, 4.0)
Measured value, median (25th, 75th)	14.0 (10.0, 15.0)	14.0 (9.0, 15.0)
ICU stays missing any measurement, n (%)	1595 (4.4%)	83 (0.2%)

Table 1. Demographics and summary of included variables for *UFHealth* and *MIMIC* cohorts. Variable summary statistics were obtained after imputing FiO2 and removing outliers as detailed in Supplementary Table S2.

$p < 0.05$ and the Traditional SOFA AUC of 0.64, 95% CI 0.64–0.65, $p < 0.05$). This advantage remained statistically significant for the remaining duration of ICU admissions. Although all models gained accuracy over time as more input data became available, DeepSOFA accuracy increased at a greater rate during the first 24 hours following ICU admission (Fig. 1A,B).

In addition to using all 14 variables in our primary DeepSOFA model, we also assessed the accuracy of six separate DeepSOFA models using only individual subsets of variables defined in each of the SOFA organ systems (Fig. 1C,D). The individual DeepSOFA component systems most predictive of mortality were central nervous system (Glasgow Coma Scale (GCS) score), respiratory (partial pressure of arterial oxygen, fraction of inspired oxygen, and mechanical ventilation status), and cardiovascular (mean arterial pressure and vasopressor administration), all relying on more frequent time series (GCS assessed every three hours, oxygen saturation and mean arterial blood pressure assessed every minute and averaged per hour).

We also examined model accuracy as a function of the predictive window being further away from the time of death or hospital discharge. As expected, all models achieved maximum AUC in the last hour of the predictive window when using data available from the entire ICU stay in both the *UFHealth* cohort (DeepSOFA AUC of 0.93, 95% CI 0.93–0.94, $p < 0.05$ compared to Bedside SOFA AUC of 0.82, 95% CI 0.81–0.83 and $p < 0.05$ compared to Traditional SOFA AUC of 0.88, 95% CI 0.88–0.89) and the *MIMIC* cohort (DeepSOFA AUC of 0.93, 95% CI 0.92–0.93, $p < 0.05$ compared to Bedside SOFA AUC of 0.81, 95% CI 0.80–0.82 and $p < 0.05$ compared to Traditional SOFA AUC of 0.85, 95% CI 0.84–0.86). Although model performance decreased slightly when prediction occurred over a longer time window, DeepSOFA retained excellent AUC above 0.87, 95% CI 0.87–0.88 in the *UFHealth* cohort and above 0.83, 95% CI 0.82–0.83 in the *MIMIC* cohort up to 100 hours away from discharge, regardless of the mortality time point of interest (Fig. 2). These findings were consistent across all development and validation cohorts.

Usability of DeepSOFA.

To demonstrate the feasibility of clinical application, DeepSOFA and Bedside SOFA scores were applied to a single patient encounter from the *UFHealth* cohort. The patient was a 25-year-old female with cystic fibrosis who was admitted to a Medical ICU following angioembolization of the blood supply to a lung abscess and remained in the ICU for 112 hours prior to death following cardiac arrest. Figure 3A illustrates the predicted probability of death according to DeepSOFA and Bedside SOFA scores. During the second day of ICU admission, despite increased supplemental oxygen requirements and worsening chest pain (Fig. 3D,E), the patient's vital signs remained relatively stable over time (Fig. 3B), and the Bedside SOFA model continued to estimate a low probability of death (<5%). However, predicted mortality according to DeepSOFA increased during these events, and continued to increase significantly as the patient developed increased work of breathing and required procedures to decompress the stomach and place a breathing tube, estimating a 50–80% probability of death, while the Bedside SOFA model continued to estimate a 5% probability of death. The Bedside SOFA score did not reflect clinical decompensation until the time of cardiac arrest. In the final five hours before death, the Bedside SOFA model estimated a 51.5% probability of mortality, while DeepSOFA estimated a 99.6% probability of mortality.

Translating our mortality prediction task into a real-time continuous acuity score is possible by examining the predicted probability of death at each hour of a patient's ICU stay (Fig. 4, Supplementary Fig. S4). Given mean mortality probabilities stratified by survival status, the traditional SOFA score tended to underestimate the severity of illness, predicting relatively low chances of death for both survivors (<5%) and non-survivors (20–30%). In contrast, DeepSOFA is better equipped to quantify illness severity for non-survivors, estimating mortality probability of 60–90% among non-survivors compared with 20–40% for survivors. DeepSOFA overestimated the probability of death for survivors, but Bedside SOFA underestimated the probability of death for non-survivors by a greater margin.

DeepSOFA Interpretability.

DeepSOFA includes added mechanisms designed to improve the human interpretability of mortality predictions (see Supplemental Section *Model Details*). Our self-attention approach is designed to highlight particular time steps of the input time series that the model believes to be most important in formulating its final mortality prediction. Since DeepSOFA is focused on real-time prediction, at each new hour after ICU admission, the model learns to distribute its internal "attention" in such a way to assign more weight to time steps it deems more influential for overall prediction.

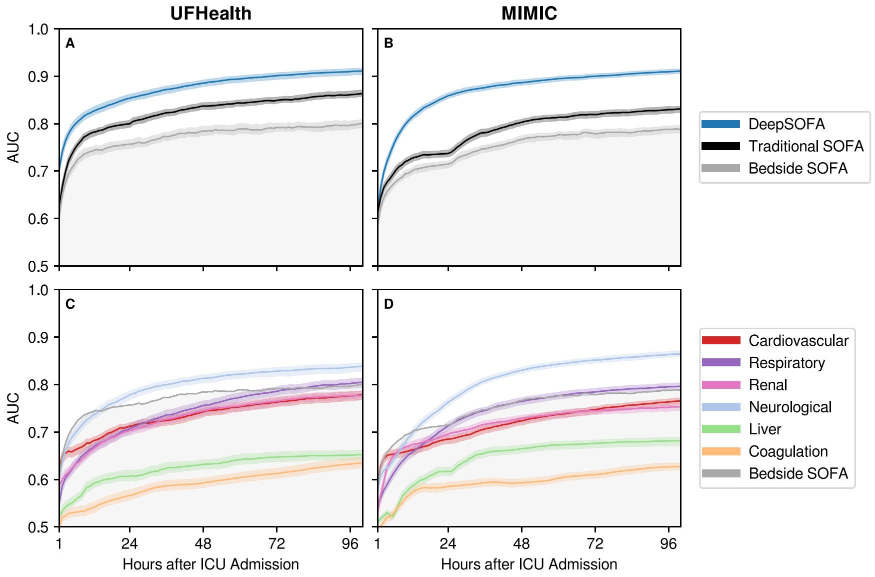

Figure 1. DeepSOFA performance in two external validation cohorts. (**A**,**B**) Externally validated DeepSOFA, Bedside SOFA, and Traditional SOFA score accuracy in predicting in-hospital mortality, expressed as area under the receiver operating characteristic curve (AUC) for the first 100 hours following ICU admission. (**C**,**D**) Externally validated DeepSOFA accuracy for individual models corresponding to variable sets derived from SOFA organ system classification for the first 100 hours following ICU admission. Shaded regions represent 95% confidence intervals based on 100 bootstrapped iterations. Columns specify the validation cohort. DeepSOFA model validated in *UFHealth* (**A**,**C**) was trained using *MIMIC*, and DeepSOFA model validated in *MIMIC* (**B**,**D**) was trained using *UFHealth*. SOFA: Sequential Organ Failure Assessment.

Self-attention can be visualized as a two-dimensional matrix. At each time step after ICU admission (columns), the model assigns weights to all preceding time steps (rows) in such a way that the column weights sum to 1. Figure 5 shows examples of self-attention matrices for one survivor and one non-survivor, along with raw time series aligned by hours after ICU admission. For the example survivor, the model focused on what occurred five hours after ICU admission and continued to focus on that hour for the remaining seven hours of the encounter. By consulting the raw time series, it appears that a clinically significant decrease in creatinine and clinically significant increases in urine output and GCS contributed to DeepSOFA's overall survival prediction.

Figure 3C features a modified version of self-attention, where we visualize only the diagonal of the two-dimensional matrix. This answers the question, "how important was each time step of data at the moment it was received by the model?" For the example non-survivor, we see several attention updates in the beginning and end of the ICU stay, with changes corresponding to salient changes in the clinical time series.

Discussion

In large, heterogeneous populations of ICU patients, we have developed and externally validated a dynamic deep learning model (DeepSOFA) that uses a time-honored illness severity score framework to predict in-hospital mortality with significantly greater accuracy than traditional methods. We also demonstrate that the deep model may be used to generate real-time prognostic data for a single patient with visual representation of model attention, indicating time periods during which model inputs made a significant impact on predictions, improving model interpretability and application. When used to predict the likelihood of death for a single patient, DeepSOFA exhibited a consistent and proportionate response to clinical events. Because DeepSOFA may be automated, it is well suited to capitalize on the emerging availability of streaming EHR data. In this regard, deep models may augment clinical decision-making by serving as an early warning system to identify patients in need of therapeutic interventions and by informing the shared decision-making processes among patients, providers, and families regarding goals of care and resource utilization by instantaneously assessing large volumes of data over time, a task which is difficult and time-consuming for clinicians.

The superior accuracy of deep models is partially attributable to their ability to learn latent structure and complex relationships from low-level data, including temporal trends in the case of recurrent neural networks. Due to their internal memory mechanisms, recurrent mortality prediction models based on sequential time series learn temporal patterns from potentially long-term dependencies in time series variables. These complex relationships are lost in traditional models, especially when applying worst-value thresholds like SOFA score calculations.

Previous work has often employed multivariable regression models in predicting mortality for ICU patients. The Simplified Acute Physiology Score (SAPS)[19,20] and Mortality Probability Model (MPM)[21] have each been used to predict in-hospital mortality using data available within one hour of ICU admission. Afessa *et al.*[22] evaluated the accuracy of SAPS III and MPM III in predicting mortality among large cohorts (SAPS III: 16,784 patients from 281 hospitals across five continents; MPM III: 124,855 patients from 98 hospitals in the United States), and found that each had strong accuracy (SAPS III: AUC 0.85, MPM III: AUC 0.82). In the same study, the Acute

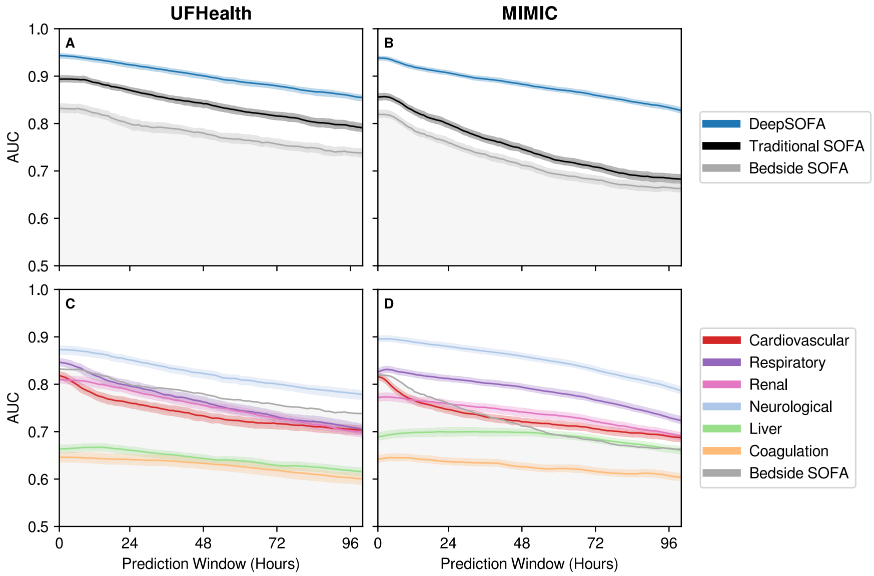

Figure 2. Model accuracy for prediction windows of increasing time from hospital discharge or death. (**A**,**B**) Externally validated DeepSOFA, Bedside SOFA, and Traditional SOFA score accuracy in predicting in-hospital mortality, expressed as area under the receiver operating characteristic curve (AUC) for 100 hours preceding death or hospital discharge. (**C**,**D**) Externally validated DeepSOFA accuracy for individual models corresponding to variable sets derived from SOFA organ system classification for the 100 hours preceding death or hospital discharge. Shaded regions represent 95% confidence intervals based on 100 bootstrapped iterations. Columns specify the validation cohort. DeepSOFA model validated in *UFHealth* (**A**,**C**) was trained using *MIMIC*, and DeepSOFA model validated in *MIMIC* (**B**,**D**) was trained using *UFHealth*. SOFA: Sequential Organ Failure Assessment.

Physiology and Chronic Health Evaluation (APACHE) IV[23] score was used with data from the first 24 hours of ICU admission for 110,558 patients from 45 hospitals in the United States, and achieved AUC 0.88. Although these methods have produced reasonably accurate predictions of in-hospital mortality, their accuracy is inferior to that of deep models, and their clinical application is cumbersome compared with automated models that have the capacity for integration of streaming electronic health record data.

This study was limited by using data from hospitals within a single country. Patient populations and practice patterns from *UFHealth* and *MIMIC*-III may differ from that of other ICU settings, limiting the generalizability of these findings. This study is also limited by restricting the deep learning input data to SOFA components rather than the full spectrum of variables in electronic health records. Future studies should apply DeepSOFA to live streaming electronic health record data and investigate the efficacy of expanding the input variables beyond SOFA score components to include the full spectrum of variables in electronic health records.

To our knowledge, DeepSOFA is the first application of deep learning toward generating real-time patient acuity scores. Our interpretability mechanism is also a novel application of recent advances in deep learning self-attention, where sequence elements involved in the self-attention calculation are distinct hours of a patient's ICU trajectory. We utilized these attention scores to determine and visualize the severity of fundamental time series patterns and their overall effect on the resulting acuity scores, an important contribution toward the interpretability of deep learning techniques in clinical setting.

DeepSOFA models trained on time series data were more accurate than baseline SOFA models for predicting in-hospital mortality among ICU patients. Baseline SOFA models significantly underestimated the probability of death, especially among non-survivors; DeepSOFA overestimated the probability of death among survivors, albeit to a lesser degree. Magnitude of error aside, the latter is less likely to contribute to a scenario in which clinicians fail to rescue a decompensating patient, a primary concern in ICUs. DeepSOFA may be applied to individual patients, exhibiting consistent and proportionate responses to clinical events, with visual representation of the probability of death and time periods during which model inputs disproportionately contributed to predictions. These findings suggest that the SOFA score can be augmented with more nuanced and intelligent mechanisms for assessing patient acuity. Deep learning technology may be used to augment clinician decision-making by generating accurate real-time prognostic data to identify patients in need of therapeutic interventions and inform shared decision-making processes among patients, providers, and families.

Methods

Study Design.

Using the University of Florida Health Integrated Data Repository as Honest Broker, we created a single-center longitudinal dataset (referred to as *UFHealth*) that was extracted directly from the electronic medical records derived from 84,350 patients 18 years or older at University of Florida Health during their admissions between January 1, 2012 and April 1, 2016 as well as all encounters within one-year history and one-year follow-up. All electronic health records were de-identified, except that dates of service were maintained.

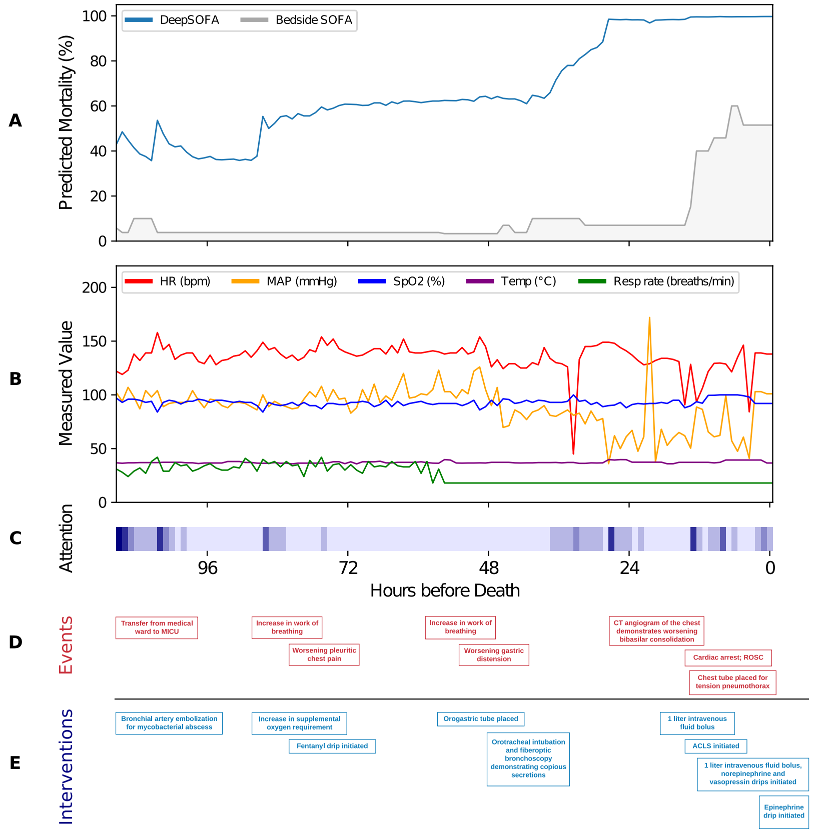

Figure 3. Externally validated DeepSOFA and Bedside SOFA score predicted mortality (**A**) for a single patient from the *UFHealth* cohort, correlated with vital signs (**B**), clinical events (**D**), and clinical interventions (**E**). Shown also are model self-attention weights (**C**) for each hour after ICU admission, where darker bars indicate increased model focus. Attention weights taken from the diagonal of full self-attention matrix and indicate how important the model believes each hour's data is, as it is encountered in real-time. SOFA: Sequential Organ Failure Assessment, SpO2: oxygen saturation, MICU: Medical Intensive Care Unit, CT: computed tomography, ACLS: Advanced Cardiac Life Support, ROSC: return of spontaneous circulation.

The dataset includes structured and unstructured clinical data, demographic information, vital signs, laboratory values, medications, diagnoses, and procedures. Among these hospital encounters, there were 33,953 distinct encounters related to 27,660 unique patients and 36,216 ICU stays in which the patient was at least 18 years old, had their ICU stay last between 4 hours and 30 days, and had at least one measurement of mean arterial pressure and either PaO2 or SpO2 (Supplementary Fig. S5). Identical selection criteria were applied to the publicly available *MIMIC*-III[18] database of 45,748 hospital encounters and 48,948 ICU stays from 35,993 patients admitted to the Beth Israel Deaconess Medical Center in Boston, Massachusetts between 2001 and 2012. The study was approved by the University of Florida Institutional Review Board 589–2016 with waiver of informed consent.

This was a retrospective study. To predict in-hospital mortality, we made predictions every hour starting when a subject first entered the ICU, with the first mortality predictions generated one hour after ICU admission and ending at the time of ICU discharge or death. Prediction modeling was limited to data accrued during ICU admission. For patients transferred out of the ICU to an intermediate care unit or hospital ward, the end-point of hospital discharge or death was assessed at the conclusion of that hospital admission. For every prediction we used all information for our selected 14 variables available in the EHR up to the time at which the prediction was made.

Data Processing. For both cohorts, all raw time series data were extracted for the 14 variables in electronic health records (mean arterial pressure, fraction of inspired oxygen, partial pressure of oxygen, mechanical ventilation status, Glasgow Coma Scale, urine output, platelet count, serum bilirubin, serum creatinine and dosing for dopamine, dobutamine, epinephrine and norepinephrine) used in the original SOFA score, as well as for blood oxygen saturation, a commonly used respiratory measurement when partial pressure of oxygen is unavailable (Table 1, Supplementary Table S2). Although additional variables would have likely improved mortality prediction accuracy, the deep learning models were limited to the use of SOFA input variables to facilitate direct comparison with baseline SOFA models and as a starting point for real-time continuous acuity assessment. Variable time series began at ICU admission and ended at ICU discharge or death.

Following variable extraction, measurement outliers were removed from both cohorts according to rules in Supplementary Table S2, which come from both expert-defined ranges and modified Z-scores. We also employed

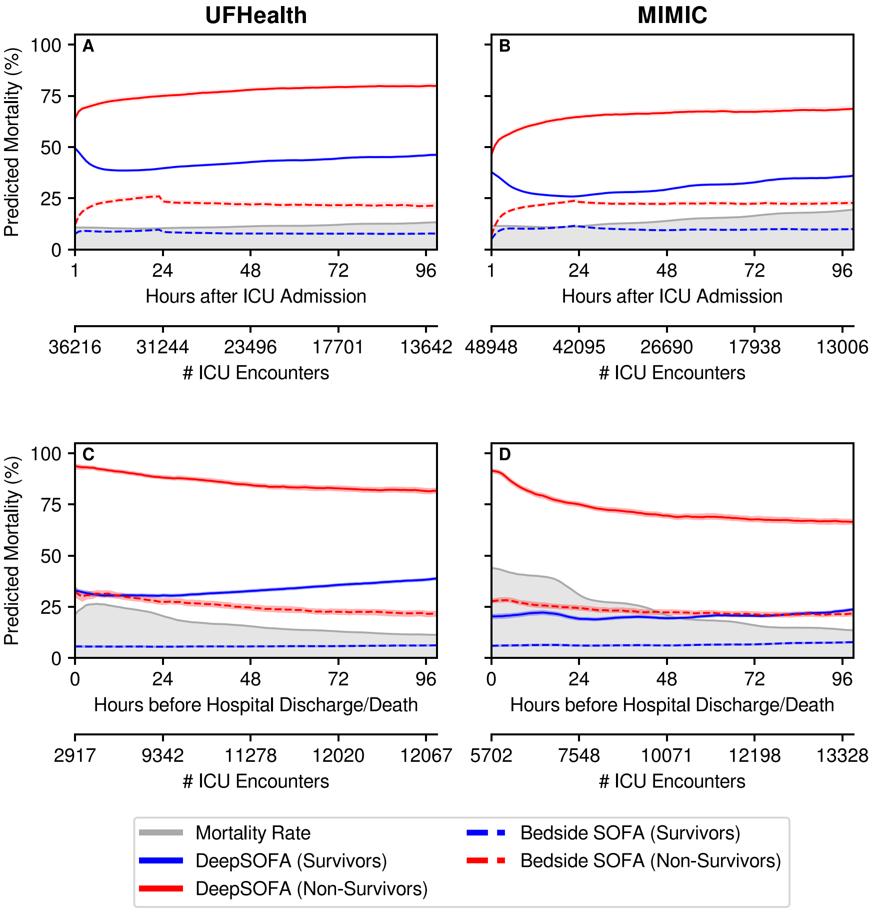

Figure 4. Mean predicted mortality probabilities for externally validated DeepSOFA and Bedside SOFA models stratified by outcome. Probabilities shown both for first 100 hours after ICU admission (**A**,**B**) and final 100 hours before hospital discharge or death (**C**,**D**). Number of ICU encounters at each time point shown below each panel. Shaded regions around each line represent 95% confidence intervals based on 100 bootstrapped iterations. Gray shared area denotes hourly mortality rate for active ICU encounters. Columns specify the validation cohort. DeepSOFA model validated in *UFHealth* (**A**,**C**) was trained using *MIMIC*, and DeepSOFA model validated in *MIMIC* (**B**,**D**) was trained using *UFHealth*. SOFA: Sequential Organ Failure Assessment.

an FiO2 imputation strategy outlined in Supplementary Table S1 and Supplementary Fig. S6 for calculating FiO2 based on respiratory device and oxygen flow rates.

Raw time series were then resampled to an hourly frequency, taking the mean value when multiple measurements existed for the same encounter during the same one-hour window. Following resampling, gaps in the resulting time series were filled by forward-propagating previous values for vital signs and laboratory tests and substituting 0 for vasopressor rates and the use of mechanical ventilation. For all remaining missing values, including instances in which a variable was missing entirely from an admission or before the first measurement became available, clinically normal ranges defined by experts were imputed (Supplementary Table S2).

The primary outcome was in-hospital mortality. Discharges to hospice in which death occurred within 7 days of hospital discharge (3% of encounters in *UFHealth* and 1.1% in *MIMIC*) were treated as mortalities.

Model Development and Analysis. For predicting in-hospital mortality, we used a modification of a recurrent neural network (RNN) with gated recurrent units (GRU)[24], a deep learning model ideal for working with sequentially ordered temporal data. Figure 6 shows a high-level overview of our model at three increasing levels of abstraction. Background, motivation, and detailed technical specification for our model can be found in the Supplementary Section *Model Details*. Briefly, the RNN internally and continuously updates its parameters based on multivariate inputs from both the current time step and previous time steps. As such, a mortality prediction incorporates patterns detected across the entirety of an ICU admission, with recognition of longer-range temporal relationships aided by the addition of GRUs.

One of the weaknesses of deep learning techniques is the inherent difficulty in understanding the relative importance of model inputs in generating the output. In the case of mortality prediction, clinicians are interested not only in the likelihood of death, but also in knowing which factors are primarily responsible for the risk of death. If such factors are modifiable, then they represent therapeutic targets. If such factors are not modifiable, then the sustained provision of life-prolonging interventions may reach futility. To improve clinical interpretability, inspired by state-of-the-art results in other deep learning domains, we modified the traditional GRU-RNN

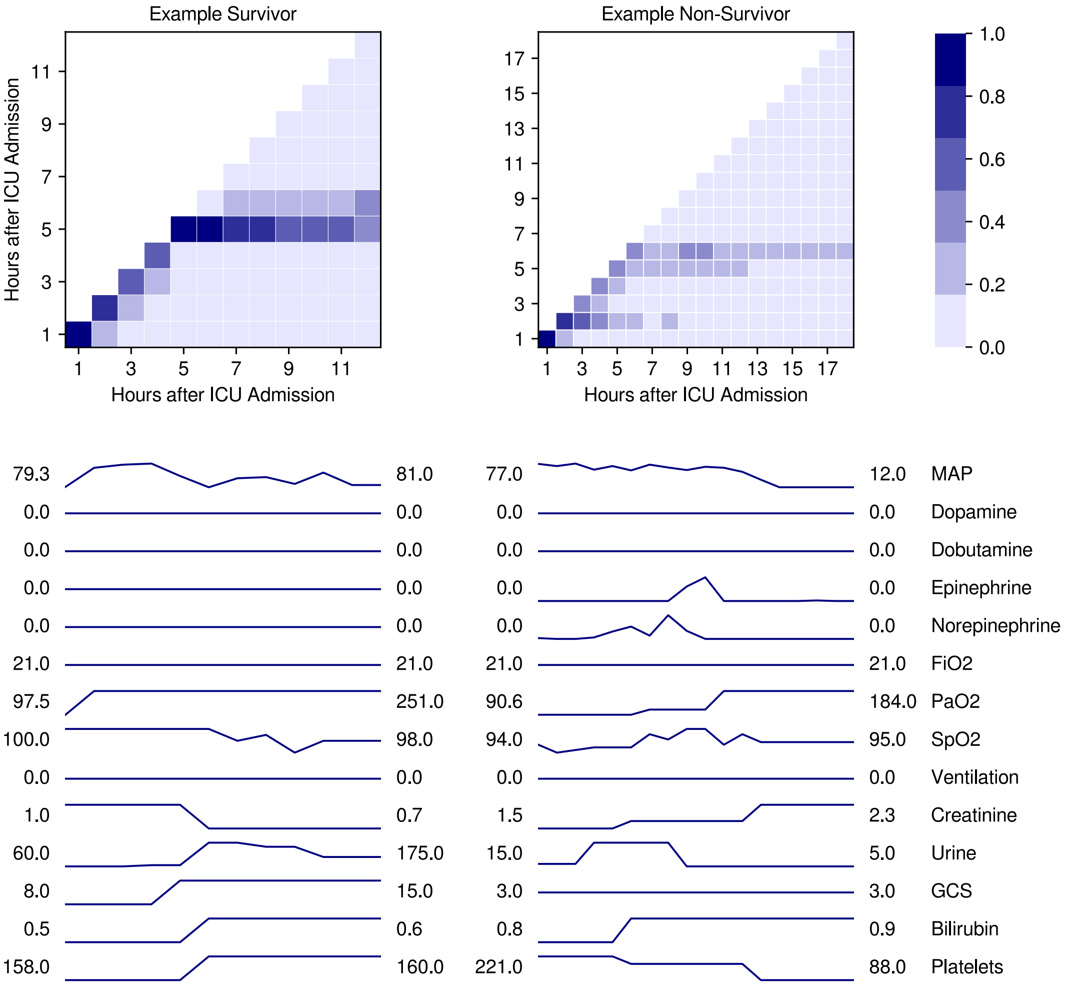

Figure 5. Visualized self-attention distributions for an example survivor and non-survivor from the *UFHealth* cohort, using DeepSOFA trained on the *MIMIC* cohort. Darker squares indicate increased model focus as a function of the passage of time (x-axis). Shown also are variable time series at each hour of the ICU stay, with initial and final measurement values shown on the left and right, respectively. MAP: mean arterial pressure, FiO2: fraction of inspired oxygen, PaO2: partial pressure of oxygen, SpO2: oxygen saturation, GCS: Glasgow Coma Scale.

network to include a final self-attention mechanism to allow clinicians to understand why the deep network is making its predictions. At each hour during a real-time ICU stay, the model's attention mechanism focuses on salient deep representations of all previous time points, assigning relevance scores to every preceding hour that determine the magnitude of each hour's contribution to the model's overall mortality prediction. Subject to the constraint that each hour's relevance scores must sum to 1, we are able to see exactly which hours of the multivariate time series the model thinks are most important, and how sudden the shift in attention happens. An example of this interpretable attention mechanism is shown in Fig. 5 where along with a mapping back to the original input time series, the model is able to justify its mortality predictions by changes in each of the input variables.

DeepSOFA mortality predictions were compared with two baseline models using traditional SOFA scores, which were calculated at each hour using the previous 24 hours of EHR data. The mortality predictions associated with calculated SOFA scores were derived from both published mortality rate correlations with any given score[2], which we refer to as "Bedside SOFA", and to overall AUC derived from raw SOFA scores, which we refer to as "Traditional SOFA". At any hour during an ICU admission, the Bedside SOFA baseline model associated the current SOFA score with a predicted probability of mortality, as would be performed using an online calculator, in which total SOFA scores correlate with mortality ranges. The Traditional SOFA model is based on retrospective analysis that derives AUC from raw SOFA scores and outcomes, and while not suitable for real-time prediction in practice, is a reasonable and contemporary baseline and an appropriate challenger to compare with DeepSOFA. A high-level comparison between the prediction and AUC calculation for all three models used in our experiments can be found in Supplementary Table S3.

Our baselines are based on both current practice (Bedside SOFA) and recent retrospective methods (Traditional SOFA). Both of these baselines utilize a single feature (current SOFA score) from patient time series for making hourly predictions. As a sensitivity analysis, we also trained two additional conventional machine

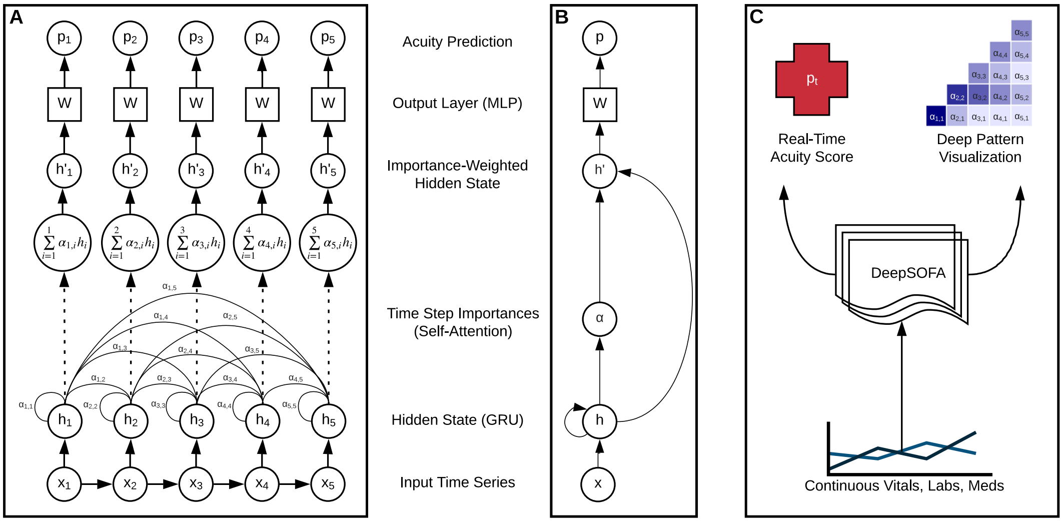
A
p1 p2 p3 p4 p5
W W W W W
h'1 h'2 h'3 h'4 h'5
$\sum_{i=1}^{1}\alpha_{1,i}h_i$ $\sum_{i=1}^{2}\alpha_{2,i}h_i$ $\sum_{i=1}^{3}\alpha_{3,i}h_i$ $\sum_{i=1}^{4}\alpha_{4,i}h_i$ $\sum_{i=1}^{5}\alpha_{5,i}h_i$
α1,5 α1,4 α2,5 α1,3 α2,4 α3,5 α1,2 α2,3 α3,4 α4,5
α1,1 α2,2 α3,3 α4,4 α5,5
h1 h2 h3 h4 h5
x1 x2 x3 x4 x5
Acuity Prediction
Output Layer (MLP)
Importance-Weighted Hidden State
Time Step Importances (Self-Attention)
Hidden State (GRU)
Input Time Series
B
p
W
h'
α
h
x
C
pt
Real-Time Acuity Score
α5,5 α4,4 α5,4 α3,3 α4,3 α5,3 α2,2 α3,2 α4,2 α5,2 α1,1 α2,1 α3,1 α4,1 α5,1
Deep Pattern Visualization
DeepSOFA
Continuous Vitals, Labs, Meds

Figure 6. Three identical views of our DeepSOFA model at increasing levels of abstraction, from most technical (**A**) to highest level operation (**C**). Panel A illustrates mortality prediction calculation for an example data segment *x* of five hours in the ICU and corresponding hourly acuity assessments *p*, where at each time point only the current and previous time steps are used in attention calculations and mortality predictions. Panel B displays a more general and compact form for the same stages of data transformation. Panel C describes the high-level inputs and outputs for DeepSOFA, where along with overall acuity assessment, interpretable prediction rationale by way of salient sequence patterns are visualized by a heatmap of self-attention weights. A more technical description of each stage of DeepSOFA can be found in the Supplemental Section *Model Details*.

learning models (logistic regression, random forest) using 84 aggregate features recalculated at every hour after ICU admission, including the following for each of the 14 SOFA variables: minimum value, maximum value, mean value, standard deviation, first value, and last value. A summary of these results can be found in Supplementary Tables S4 and S5 for internal and external validation, respectively. The SOFA baselines included in our study outperformed these additional machine learning models.

Model Evaluations and Statistical Analysis. Using each of *UFHealth* and *MIMIC* as primary cohorts, we performed several internal and externally-validated experiments and baseline comparisons. For each cohort, internal validation experiments consisted of training and testing the DeepSOFA model on the same data source. In this setting, we performed 5-fold cross-validation in which a model was trained on a random 80% of ICU admissions and tested on the remaining 20%, repeated for 5 non-overlapping iterations to yield a prediction trajectory for every ICU stay in the cohort. In the external validation experiments, a DeepSOFA model was trained on the entirety of one cohort and tested on the entirety of the other cohort. Baseline models did not require training and were applied to the same testing cohorts as DeepSOFA. All reported results are from the external validation experiments. Internal cross-validation results, which are more optimistic than their external counterparts, can be found in the online supplement.

Predictions involved in our experiments were performed on individual ICU encounters; as such, a single patient could have multiple ICU stays that appear as distinct prediction units. We performed a sensitivity analysis involving two variations of adjusting for patients with multiple ICU encounters in their EHR, including (1) only keeping their first ICU encounter, and (2) removing such patients entirely from the dataset. All models were retrained and tested using these modified datasets, and a summary of these prediction results can be found in Supplementary Tables S6 and S7 for internal and external validation, respectively. Both modifications resulted in increased performance for all models.

For all models, a prediction was obtained at every hour, beginning one hour after ICU admission and ending at the time of ICU discharge or death. We assessed model discrimination by calculating area under the receiver operating characteristic curve (AUC), and calculated 95% confidence intervals using 100 bootstrapped iterations of sampling mortality prediction probabilities with replacement. At each hour, all ICU stays were included in reported results; for encounters with duration less than the current hour, the final prediction was used in AUC calculations. Figures 4 and S4 show the number of active ICU encounters and corresponding mortality rates by hour.

Data Availability

The *MIMIC* cohort is derived from the publicly available *MIMIC*-III database[18]. *UFHealth* cohort data are available from the University of Florida Institutional Data Access/Ethics Committee for researchers who meet the criteria for access to confidential data and may require additional IRB approval.

References

1. Vincent, J. L. *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* **22**, 707–710 (1996).
2. Ferreira, F., Bota, D., Bross, A., Mélot, C. & Vincent, J. Serial evaluation of the sofa score to predict outcome in critically ill patients. *J. Am. Med. Assoc.* **286**, 1754–1758 (2001).
3. Vincent, J.-L. *et al*. Use of the SOFA score to assess the incidence of organ dysfunction/failure in intensive care units. *Crit. Care Med.* **26**, 1793–1800 (1998).
4. Minne, L., Abu-Hanna, A. & de Jonge, E. Evaluation of SOFA-based models for predicting mortality in the ICU: A systematic review. *Crit. Care* **12**, R161 (2008).
5. Badawi, O., Liu, X., Hassan, E., Amelung, P. J. & Swami, S. Evaluation of ICU Risk Models Adapted for Use as Continuous Markers of Severity of Illness Throughout the ICU Stay. *Crit. Care Med.* **46**, 361–367 (2018).
6. Maslove, D. M. With Severity Scores Updated on the Hour, Data Science Inches Closer to the Bedside. *Crit. Care Med.* **46**, 480–481 (2018).
7. Kim, S., Kim, W. & Park, R. W. A Comparison of Intensive Care Unit Mortality Prediction Models through the Use of Data Mining Techniques. *Healthc. Inform. Res.* **17**, 232–243 (2011).
8. Meyfroidt, G., Güiza, F., Ramon, J. & Bruynooghe, M. Machine learning techniques to examine large patient databases. *Best Pract. Res. Clin. Anaesthesiol.* **23**, 127–143 (2009).
9. Clermont, G., Angus, D. C., DiRusso, S. M., Griffin, M. & Linde-Zwirble, W. T. Predicting hospital mortality for patients in the intensive care unit: A comparison of artificial neural networks with logistic regression models. *Crit. Care Med.* **29** (2001).
10. LeCun, Y., Bengio, Y. & Hinton, G. Deep learning. *Nature* **521**, 436–444 (2015).
11. Goodfellow, I., Bengio, Y. & Courville, A. *Deep Learning*. (MIT Press, 2016).
12. Shickel, B., Tighe, P. J., Bihorac, A. & Rashidi, P. Deep EHR: A Survey of Recent Advances in Deep Learning Techniques for Electronic Health Record (EHR)Analysis. *IEEE J. Biomed. Heal. Informatics* **22**, 1589–1604 (2018).
13. Gulshan, V. *et al*. Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. *J. Am. Med. Assoc.* **316**, 2402–2410 (2016).
14. Esteva, A. *et al*. Dermatologist-level classification of skin cancer with deep neural networks. *Nature* **542**, 115–118 (2017).
15. Nickerson, P., Tighe, P., Shickel, B. & Rashidi, P. Deep neural network architectures for forecasting analgesic response. In *Proceedings of the Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, 2966–2969 (2016).
16. Choi, E., Schuetz, A., Stewart, W. F. & Sun, J. Using recurrent neural network models for early detection of heart failure onset. *J. Am. Med. Inform. Assoc.* **292**, 344–350 (2016).
17. Du, H., Ghassemi, M. M. & Feng, M. The effects of deep network topology on mortality prediction. In *Proceedings of the Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, 2602–2605 (2016).
18. Johnson, A. E. W. *et al. MIMIC*-III, a freely accessible critical care database. *Sci. Data* **3**, 1–9 (2016).
19. Metnitz, P. G. H. *et al*. SAPS 3-From evaluation of the patient to evaluation of the intensive care unit. Part 1: Objectives, methods and cohort description. *Intensive Care Med.* **31**, 1336–1344 (2005).
20. Moreno, R. P. *et al*. SAPS 3 - From evaluation of the patient to evaluation of the intensive care unit. Part 2: Development of a prognostic model for hospital mortality at ICU admission. *Intensive Care Med.* **31**, 1345–1355 (2005).
21. Higgins, T. L. *et al*. Assessing contemporary intensive care unit outcome: an updated Mortality Probability Admission Model (MPM0-III). *Crit. Care Med.* **35**, 827–35 (2007).
22. Afessa, B., Gajic, O. & Keegan, M. T. Severity of Illness and Organ Failure Assessment in Adult Intensive Care Units. *Crit. Care Clin.* **23**, 639–658 (2007).
23. Zimmerman, J. E., Kramer, A. A., McNair, D. S. & Malila, F. M. Acute Physiology and Chronic Health Evaluation (APACHE) IV: hospital mortality assessment for today's critically ill patients. *Crit. Care Med.* **34**, 1297–1310 (2006).
24. Cho, K. *et al*. Learning Phrase Representations using RNN Encoder-Decoder for Statistical Machine Translation. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*. 1724–1734 (2014).

Acknowledgements

The authors acknowledge Gigi Lipori, MBA and Paul Nickerson, MS for assistance with data retrieval, and the University of Florida Integrated Data Repository (IDR) and the UF Health Office of the Chief Data Officer for providing the analytic data set for this project. A.B., T.O.B. and P.R. were supported by R01 GM110240 from the National Institute of General Medical Sciences. P.R. was supported by CAREER award, NSF-IIS 1750192, from the National Science Foundation (NSF), Division of Information and Intelligent Systems (IIS). A.B. and T.O.B. were supported by Sepsis and Critical Illness Research Center Award P50 GM-111152 from the National Institute of General Medical Sciences. T.O.B. has received grant (97071) from Clinical and Translational Science Institute, University of Florida. This work was supported in part by the NIH/NCATS Clinical and Translational Sciences Award to the University of Florida UL1 TR000064 and UL1 TR001427. T.L. was supported by a post-graduate training grant (T32 GM-008721) in burns, trauma and perioperative injury awarded by the National Institute of General Medical Sciences (NIGMS). B.S. was supported by internal funds at J. Crayton Pruitt Family Department of Biomedical Engineering, University of Florida. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript. A.B., B.S. and T.O.B. had full access to all the data in the study and take responsibility for the integrity of the data and the accuracy of the data analysis. The Titan X Pascal partially used for this research was donated by the NVIDIA Corporation.

Author Contributions

B.S., T.L., A.B., L.A. and T.O.B. had full access to the data in the study and take responsibility for the integrity of the data and the accuracy of the data analysis. Study design performed by B.S., T.L., A.B. and P.R. B.S. and T.L. drafted the manuscript. Analysis was performed by B.S., T.L., T.O.B., L.A., A.B. and P.R. Funding was obtained by A.B. and P.R. Administrative, technical, material support was provided by A.B. and P.R. Study supervision was performed by A.B. and P.R. All authors contributed to the acquisition, analysis, and interpretation of data. All authors contributed to critical revision of the manuscript for important intellectual content.

Additional Information

Supplementary information accompanies this paper at https://doi.org/10.1038/s41598-019-38491-0.

Competing Interests: The authors declare no competing interests.

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.