***Improving prediction models’ propriety in intensive care units by enforcing an advance notice period***

*Tomer Hermelin1, Pierre Singer2 & Nadav Rappoport1*

*1Department of Software and Information Systems Engineering, Faculty of Engineering, Ben-Gurion University of the Negev, ISRAEL*

*2Beilinson, ISRAEL*

*hermelit@post.bgu.ac.il*

*psinger@clalit.org.il*

*nadavrap@bgu.ac.il*

*Key phrases:*

Forecasting (D005544); Intensive Care Units (D007362); Hospital Mortality (D017052); Sepsis (D018805); Deep Learning (D000077321)

*Word Count: 3,923*

Table of Contents

[Abstract 3](#_Toc68961960)

[Objective 3](#_Toc68961961)

[Materials and Methods 3](#_Toc68961962)

[Results 3](#_Toc68961963)

[Conclusions 3](#_Toc68961964)

[Introduction 4](#_Toc68961965)

[Background 4](#_Toc68961966)

[Objective 6](#_Toc68961967)

[Materials and methods 6](#_Toc68961968)

[Formal problem definition 7](#_Toc68961969)

[Evaluation 8](#_Toc68961970)

[Clinical data 9](#_Toc68961971)

[Cohort development and feature selection 9](#_Toc68961972)

[Data preprocessing and feature engineering 10](#_Toc68961973)

[Model development 10](#_Toc68961974)

[Results 12](#_Toc68961975)

[Discussion 12](#_Toc68961976)

[Conclusions 13](#_Toc68961977)

[References 14](#_Toc68961978)

# Abstract

## Objective

Intensive care units (ICUs) are time-critical environments, and sufficient reaction time is crucial. There are existing systems for alerting staff to life-threatening events in an ICU, although these models suffer from “immediate event” bias. Here, we present a new approach for outcome prediction in ICU admissions, which takes into consideration the advance notice of a predicted outcome. We provide examples of the approach using mortality and sepsis-3 prediction. Further, we examine whether models need to be trained for a specific notice period, or whether the approach could be incorporated at the evaluation level.

## Materials and Methods

We have created a set of neural network models that implement and evaluate the suggested approach using MIMIC-III data. We trained and evaluated the models with and without adding a constraint representing an “alert interval” between the prediction time and the prediction window.

## Results

We show that enforcing a notice period can significantly affect performance, but not for all outcome predictions. Additionally, we showed that the alert interval could be defined post-model training, with no significant loss in performance, within the bounds of the trained lookahead.

## Conclusions

When evaluating the applicability of predictive models for ICUs, incorporating an advance-notice constraint to the model for some scenarios can be crucial and, in some cases, can significantly change the results. This could be done for pre-existing, previously trained models. The concept of adding an alert interval could be applied to other clinical scenarios where having advance notice is essential.

# Introduction

An intensive care unit (ICU) is a specialized department of a hospital or healthcare facility that provides intensive treatment and care. Patients admitted to the ICU usually have severe or even life-threatening illnesses and injuries, and therefore are at high risk of mortality. The admitted patients are provided with constant care and close supervision. The goal of the ICU is to nurse patients to a vigorous and stable condition, so they can be released from the ICU and continue to receive the care they need in a step-down unit or at home. However, not all admissions have successful outcomes. Statistics show that around 11.5% of patients admitted to the ICU die during admission.[1] Close monitoring and the adoption of electronic medical records (EMRs) has enabled patient data from ICUs to be abundant and frequently sampled, to leverage data science solutions. As response time in ICUs is critical, leveraging these data to provide risk alerts for patients’ future events (including death, onset of sepsis, cardiac arrest, organ failure, etc.) can improve the care given in an ICU and reduce the mortality rate.

## Background

Predicting mortality, sepsis onset, or other types of events among ICU admissions is not a new subject for research. Good literature reviews have been conducted that cover mortality outcome prediction [2] and sepsis onset prediction.[3] Prediction models used in this clinical scenario can be categorized into two main groups based on their approach: Cut-off and intervallic.

A cut-off model is a model that uses information from the first X hours to predict the outcome (for example, death or discharge) of an ICU admission (or, in some cases, the outcome after a given time period, e.g., a patient’s status 24 hours after admission or 30 days following discharge). In this type of model, there is a single prediction per ICU admission. Common values for X are 24 and 48 hours.[3–13] A well-known clinical score that matches this profile is APACHE-II (Acute Physiology And Chronic Health Evaluation II),[14] which is applied within 24 hours of admission and assigns a risk of death score, according to various measurements. In contrast, an intervallic model is a model that provides multiple predictions during an ICU admission. Each prediction refers to a “prediction window”, a slice of time from the patient’s admission. The prediction is based on the patient’s data up to the prediction window, and it predicts whether the patient will have an event within the prediction window’s timeframe. Common settings used for a prediction window are 1, 6, and 24 hours.[15–23] Some models use multiple prediction window sizes, evaluating the forecasting ability of different “horizons”.[24]

**Figure 1.**

****

*A graphical illustration of two common prediction types followed by the new proposed notice approach. In the cut-off model, each admission is assigned a single prediction at a pre-defined time, and the prediction is typically about the admission outcome. In the intervallic model, each admission is assigned multiple predictions, depending on the length of admission and the prediction’s window-size. In the notice approach, there are also multiple predictions per admission. However, the prediction window is distanced from the prediction time.*

A recent study addressed and implemented both of these types of models.[25] There are additional variations of these two types of models, which are less common or have fewer applications, that we have not included for comparison with the model we propose. One variation is the “rolling cut-off” model, which is a hybrid of the two methods. Rolling cut-off models have a sliding prediction point, similar to intervallic models, but the prediction is made with regard to the rest of the admission.[26,27]

Each type of model has its drawbacks. Cut-off models are not scoped in time, making it difficult to focus efforts when read alerts are needed most. For example, if the cut-off time is after 24 hours and the duration of admission is 168 hours, the prediction does not tell us when, in the remaining 144 hours of the admission, the risk of the event is greatest. Therefore, it may not be useful as a real-time alert system. Additionally, it only leverages the data available up to the cut-off point, regardless of the patient’s duration of admission and when a prediction is needed or asked for. Looking at the above example, if 100 hours into admission a prediction is required, the model cannot leverage any data between hours 24 and 100, including the most recent data.

In contrast, the intervallic approach does provide a scoped prediction (for a specific prediction window) and leverages all data up to the required prediction point. However, by definition, such models have a prediction window that immediately follows the time of prediction. This results in two disadvantages: (i) Application-wise, this does not ensure a minimum advance notice period for intervention. For example, consider the case where the model predicts a patient’s status every 6 hours. A patient who undergoes an event 31 hours after admission will receive a negative prediction for the prediction window within 24 to 30 hours of admission. For a prediction window of 30 to 36 hours, if a prediction is correct, the patient will receive a positive prediction that gives just a 1-hour alert prior to the event time. (ii) Performance evaluation-wise: it can be easier for prediction models to predict events that occur close to the prediction time rather than events that occur further from the prediction time. Clinical events are often gradual, progressive events. Predictions that occur adjacent to a predicted event can rely on signals that indicate that event in a straightforward manner. Therefore, this can be considered to be a type of data leaking. These disadvantages are also relevant to the cut-off approach, although they are less important.

One way to demonstrate the effect of the distance of the event from the prediction point on performance evaluation is by breaking the receiver operating characteristic (ROC) area under the curve (AUC) performance of a cut-off model by the time of the event. As shown in Figure 2 for our example outcomes, when predicting mortality from data from the first 24 hours of admission, the ROC AUC decreases as admission times get longer. However, for the prediction of sepsis, the drop is mild.

**Figure 2.**

******

***Prediction performance of a cut-off model as measured by ROC AUC per admission length bin, using 10-fold cross-validation.******(A) Mortality prediction and (B) Sepsis onset prediction.*** *Performance drops for more prolonged admissions in mortality, while sepsis onset prediction is more stable. The bins increase in width to avoid bins with low numbers of samples. The ticks mark the bin’s start time, where the last bin has no upper limit. The error bars represent the confidence intervals for each averaged data point over the x-axis.*

# Objective

Here, we are introducing and implementing a new approach for outcome prediction, and we demonstrate the use of this approach for mortality and sepsis onset prediction during ICU admissions. This new approach is designed to take into account a minimal advance notice for alerting, while maintaining the prediction scoped in a similar way to the intervallic approach. We achieve this by adding an “alert interval” between the prediction time and the prediction window. We call this new type of model a “notice model”.

# Materials and methods

A schematic overview of the study is shown in Figure 3. The new approach we introduce, the notice model and its predictions, is depicted in Figure 1 and contrasted with the cut-off and intervallic models, also illustrated in Figure 1.

**Figure 3.**



***Study overview****. The numbers in the flow vary a little between configurations and the predicted outcome, as we filter admissions that are shorter than “start time” + ”alert interval” (defined in the “Formal problem definition”, below) to prevent the prediction of events that are within the model’s input signals. The numbers in the above flow-chart are for the example notice configuration of a 6-hour alert interval, predicting mortality, as detailed in Table 1.*

## Formal problem definition

Formally, we wish to generate a predictor, , where:

* – Single ICU admission data, limited to events in time window , where , in hours
* – Label. Indicator for whether the predicted outcome occurred for an ICU admission within time window , where , in hours
* – Start time. Defines the time from admission of the first prediction point, in hours
* – Prediction step. Defines the time interval between each two prediction points, in hours
* – Prediction window size. Defines the length of each prediction window, in hours
* – Alert interval. Sets the minimal notice in advance time for the prediction, in hours

Then, for a given ICU admission, our target function is:

* =
* ICU admission has not concluded until time
* is the lookback (or observation window) chosen for the model. It is addressed as a hyperparameter to tune.

We refer to the set of values as a “configuration” when examining different models in this paper.

## Evaluation

We evaluated our models using area under the receiver operating characteristics (AUC), which is conventionally used to evaluate such risk-prediction models. The drawback of evaluating AUC in intervallic/notice models is that longer admissions are counted more times than shorter admissions, as these admissions appear in more prediction windows. However, there is sense in evaluating the model in a way that gives each admission the same weight. For this we used weighted AUC (WAUC), where every sample is weighted inversely proportionally to the number of samples (prediction windows) of that admission. The weights of all the predictions that belong to the same admission sum up to 1.

**AUC: Evaluate all predicted time windows evenly:**

measure for predictions made at time for the outcomes at time

**WAUC: Evaluate all predictions, ICU normalized:**

## Clinical data and cohort

For this study, we used the MIMIC-III (Medical Information Mart for Intensive Care III) dataset from the Beth Israel Deaconess Medical Center (BIDMC), Boston, Massachusetts. The MIMIC-III database contains clinical data from 53,423 adult ICU stays from 38,597 adult patients.[29] All admissions of patients under the age of 18 years and any admissions shorter than the time of the first prediction window were removed. As a result, we excluded 8656 admissions, resulting in 44,767 admissions remaining. For sepsis-3 outcome prediction, we also filtered 21,208 admissions of patients whose first occurrence of sepsis occurred prior to their first admission prediction window (as per the definition given in the “Acquiring labels” section), resulting in 23,559 admissions. For both outcomes, the split to train-validation-test was performed at patient level to avoid information leakage across ICU stays of the same patient, such that all time windows of the same patients were not split between train, validation, and test set. The training, validation, and testing sets comprised 80%, 10%, and 10% of the patients, respectively.

## Acquiring labels

Patient mortality and its time is logged in the MIMIC-III dataset’s “Admissions” table. However, a parallel label does not exist for sepsis-3 events. Therefore, we followed the approach for identifying ICU-acquired sepsis described by Goldstein et al.[25] A sepsis-3 case was defined by a patient having a culture sampling and antibiotic administration within 24 hours of each other, while also having a sequential organ failure assessment (SOFA)[30] score of 2 or above within 24 hours. Sepsis time was defined as the first time of culture sampling or when antibiotics were administered. To avoid interpreting ongoing cases as several instances, we kept only the first case of sepsis in each patient and added a constraint that there was no additional administration of antibiotics in the 24 hours preceding the antibiotics administered for the diagnosed case.

## Data preprocessing

The dataset is constructed so that each prediction point and its associated data and prediction window is an independent entry. The prediction point is defined as the time during an admission when a prediction is taking place. The prediction window is defined as the time span about which is being predicted. A prediction window’s label (i.e., case or control) is determined by whether the predicted outcome’s time falls within that time window.

The data used for prediction include admissions data from both Chart-Event and Lab-Event tables (containing laboratory test results, vital signs, diagnosis, provided procedures, medication administered, etc.), acquired prior to the prediction point and during the defined lookback window by the configuration. In the notice model, we drop out the data accumulated between the prediction point and the prediction window (the alert-interval timeframe) (Figure 1). For each entry, we used the following features:

* Demographic features, including age, sex, admission type, and an indicator for whether the patient had a recorded previous ICU admission in the database.
* For each numerical type of measurement taken from the data (such as heart rate, oxygen saturation, etc.), we computed over its values within the data that is available to the prediction point:
	+ Count of measures
	+ Minimal value
	+ Maximal value
	+ Average value
	+ Variance of the values
	+ First value
	+ Last value
	+ Warning/flag count for abnormal measurements.
* For categorical measures (e.g., an oxygen delivery device), we used only the count/existence-indicator of measure.

The dataset is sparse and contains different scales because, as would be expected, most admissions did not have values for most types of measures, and different measures have different value ranges. However, the missing data are not missing at random,[31] as missing data are often a result of caregivers’ decisions. For example, a caregiver may decide to order some specific laboratory tests and not others, as the latter may not currently be relevant for the patient. As a result, the missing values have an inherent bias, compared with the general population, as they are the result of medical considerations. When handling missing values and normalizing, we wanted to make sure that the information about not having a value for a certain feature did not get lost (the missing values are assigned a designated, separable value from all non-missing values). We chose the following approach for normalizing each feature:

where is a set of all original values of the feature in the training set, and is set to 0.1. In this way, all missing values are replaced with 0, and non-missing values are scaled such that the range of the training set is [0.1,1.1]. In this way, missing values are represented with a unique value, 0, which does not appear in non-missing results. For non-missing results, we get a min-max scaling translated by a constant .

## Model development

We trained a fully connected, deep neural network model for each configuration examined. Each model type had its own hyperparameters tuned, using the validation set and the ROC AUC metric (rather than the WAUC). Then, the models were evaluated using the testing set for both AUC and WAUC.

For simplicity, we focus here on a small set of configurations. However, other configurations can be easily set up, changing the alert interval, prediction-window size, etc. For each predicted outcome, we implemented the three models: cut-off, intervallic, and notice. We created the intervallic and notice configurations as a pair so that the prediction horizon for the models was the same. However, an alert interval was added for the notice model, making the notice configuration a more challenging task, as the events close to the prediction point are excluded.

For each configuration, we tuned the hyperparameters independently to try and maximize its potential. The hyperparameters included conventional deep learning hyperparameters along with a parameter of the lookback size the model used to compute the features for each prediction entry.

The example configurations and models are detailed in Table 1.

**Table 1**

|  |  |  |
| --- | --- | --- |
| **Configuration** | **Configuration specifications** | **Hyperparameters** |
| **ST** | **PS** | **PWS** | **AI** |
| Mortality Cut-Off | 24 | Inf | Inf | 0 | Layer dims: 1200, 600, 250Weight decay: 0.0005Lookback: 24 hours |
| Mortality Intervallic | 24 | 18 | 18 | 0 | Layer dims: 1000, 500, 200 Weight decay: 0.00075Lookback: 24 hours |
| Mortality Notice | 24 | 12 | 12 | 6 | Layer dims: 900, 400, 200 Weight decay: 0.0004Lookback: 24 hours |
| Sepsis Cut-Off | 24 | Inf | Inf | 0 | Layer dims: 800, 400, 150Weight decay: 0.00075Lookback: 24 hours |
| Sepsis Intervallic | 24 | 18 | 18 | 0 | Layer dims: 900, 400, 150 Weight decay: 0.0004Lookback: 24 hours |
| Sepsis Notice | 24 | 12 | 12 | 6 | Layer dims: 600, 300, 100 Weight decay: 0.0005Lookback: 36 hours |

*Configuration parameters: ST, start time; PS, prediction step; PWS, prediction window size; AI, alert interval. Dropout was set to 0.25 for all configurations.*

# Results

Table 2 details the results of each model configuration that was trained and evaluated with 10 different random model initializations. Results are presented as mean AUC and mean WAUC, with 95% confidence intervals. In Table 3, we detail the results of the intervallic configuration models evaluated on their paired counterpart notice configuration testing sets.

**Table 2. Performances of the three model architectures for both outcomes.**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Outcome** | **Configuration** | **Mean AUC**  | **AUC CI** | **Mean WAUC**  | **WAUC CI** |
| Mortality | Cut-Off | 0.869 | 0.0026 | 0.869 | 0.0026 |
| Mortality | Intervallic | 0.891 | 0.0015 | 0.933 | 0.002 |
| Mortality | Notice | 0.866 | 0.0017 | 0.91 | 0.0021 |
| Sepsis | Cut-Off | 0.783 | 0.0046 | 0.783 | 0.0046 |
| Sepsis | Intervallic | 0.76 | 0.0034 | 0.739 | 0.0068 |
| Sepsis | Notice | 0.76 | 0.0065 | 0.722 | 0.0093 |

**Table 3. Evaluation of the notice model on the intervallic testing set for both outcomes.**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Outcome** | **Configuration** | **Test Set** | **AUC**  | **AUC CI** | **WAUC** | **WAUC CI** |
| Mortality | Intervallic  | Notice | 0.864 | 0.0016 | 0.906 | 0.003 |
| Sepsis | Intervallic  | Notice | 0.758 | 0.0048 | 0.727 | 0.0087 |

When examining the results in Table 2, we can see that the WAUC scores are higher than their AUC counterparts for mortality prediction but lower for sepsis onset prediction. When comparing performance across different configurations, the ranking could change, depending on the metric used, AUC or WAUC. For mortality prediction, the cut-off model outscored the notice model with respect to AUC, but the notice model outperformed the cut-off, with a higher WAUC.

For mortality prediction, the intervallic model outperformed the notice model, when evaluating each model with their configuration’s testing set. We anticipated this behavior, as the prediction horizons for both tasks are the same, but the notice model was not evaluated using close events, which are “easier” to predict, as illustrated in Figure 2A. When evaluating the intervallic model using its paired notice testing sets, which incorporated alert intervals (the size of the injected alert-interval to the intervallic testing set was equivalent to the parallel notice model), the intervallic model’s results decreased, becoming closer to those of the notice model.

For sepsis onset prediction, the intervallic and notice models performed quite similarly according to the AUC metric, while there was a slight difference in favor of the intervallic model in the WAUC. Once again, when evaluating the intervallic models on an adjusted testing set, the intervallic model’s results decreased close to those of the notice model.

# Discussion

We suggest a more appropriate approach for evaluating alert systems in the clinical setting, incorporating a constraint for advance notice into the model’s evaluation. This type of evaluation may be more suitable for such alert systems, as alerting for an event that is going to happen within a short time period may not be helpful for staff, as they may already know about it, or they may not be able to do anything to change it. The concept of creating an alert system with a prediction model is not limited to mortality or sepsis onset prediction and can be used for other clinical and even non-clinical settings.

Mortality prediction demonstrated a scenario where “immediate” signals gave a strong indication of the upcoming outcome and, when enforcing advance notice, the results changed significantly. In sepsis onset prediction, the immediate signals did not greatly affect the prediction. We see this sits well with the observation illustrated in Figure 2. The slope of the mortality prediction is steep and the slope for sepsis is relatively stable, with a similar performance for both faraway events and close ones.

Comparing variant solutions and different architectures could result in having the intervallic and notice approaches rank models differently (due to different performance gaps between intervallic and notice models in each solution), changing the selected “best” model, depending on which approach is taken. In future work, we plan to examine comparisons where this is the case. Additionally, we argue there is still work to be done in incorporating this concept into a system suitable for application in ICUs. Having a confident short-notice prediction is also valuable and should be considered when planning a holistic solution. It has been shown that alert systems integrated into an ICU have a much lower AUC than expected.[32] We scope out from this paper the topic of generating a production alert-system from the models.

The fact that the intervallic model performs similarly to the notice model on the notice testing set (rather than the notice model outperforming the intervallic on the notice) shows there is nothing to be gained from “focusing” on this specific subset of events in the prediction window. This means that the alert interval could be defined independently from the model development process, configurable in size after the model is trained. Although one can argue that for mortality prediction there was a statistically significant gain, we believe that this gain is not sufficient and that the fact the models were tuned independently could also contribute to differences in performance.

While the cut-off model does not have an alert interval, it generally predicts on further events than the notice model. On the other hand, the bound that the notice model provides on a predicted event is much tighter and more informative than the cut-off model. These are aspects way against each other, thus it’s hard to rank the tasks’ difficultness.

Naturally, different scenarios require different alert-intervals and different configurations in general. Therefore, we have kept the formal problem definition in general form. When shortening the alert interval of a configuration, the results catch up to the intervallic results, until the tasks unite when the alert interval is 0. We think there are other scenarios that would benefit from incorporating notice models, using different configurations. For example, seizure prediction might benefit from shorter alert intervals, while discharge-readiness might require longer alert intervals. Moreover, the problem definition defined above could be further generalized by transitioning to be:

This would enable a problem’s fixed size parameters to be dynamically defined per prediction. It could serve applications like having smaller prediction windows or prediction steps at the start of an admission and expanding them as the admission duration increases or tuning the lookback according to the time during admission the prediction takes place. In this work, we have not focused on these generalizations; however, we believe they could be useful in future research. The fact that the alert interval can be defined and applied after training the model could be an advantage here.

# Conclusion

There are currently two main types of approaches for predicting outcomes in ICU admissions, cut-off and intervallic models. The intervallic approach is more applicable. We suggest our new notice approach further improves the applicability of the intervallic approach, in scenarios that benefit a heads-up on the predicted event of at least a pre-defined time. This is achieved by adding an alert interval constraint over the model’s data. Empirical experiments show that adding this constraint can significantly affect model performance in some outcome predictions, resulting in better model evaluation (and better model selection, when comparing several models). Adding the alert interval could be carried out at inference time alone (and not necessarily during training). This allows the alert interval to be configured post-training and to be applied to pre-existing, trained models. The concept of adding an alert interval could be applied to other clinical scenarios, where having advance notice is important. We also saw that there are scenarios where there is a significant difference between measuring this task using WAUC rather than with AUC.

# References

1 MIMIC-III, a freely accessible critical care database | Scientific Data. https://www.nature.com/articles/sdata201635 (accessed 14 Jul 2020).

2 Fu L-H, Schwartz J, Moy A, *et al.* Development and validation of early warning score system: A systematic literature review. *J Biomed Inform* 2020;**105**:103410. doi:10.1016/j.jbi.2020.103410

3 Islam MdM, Nasrin T, Walther BA, *et al.* Prediction of sepsis patients using machine learning approach: A meta-analysis. *Comput Methods Programs Biomed* 2019;**170**:1–9. doi:10.1016/j.cmpb.2018.12.027

4 Ge W, Huh J-W, Park YR, *et al.* An Interpretable ICU Mortality Prediction Model Based on Logistic Regression and Recurrent Neural Networks with LSTM units. *AMIA Annu Symp Proc* 2018;**2018**:460–9.

5 Awad A, Bader-El-Den M, McNicholas J, *et al.* Early hospital mortality prediction of intensive care unit patients using an ensemble learning approach. *Int J Med Inf* 2017;**108**:185–95. doi:10.1016/j.ijmedinf.2017.10.002

6 Marafino BJ, John Boscardin W, Adams Dudley R. Efficient and sparse feature selection for biomedical text classification via the elastic net: Application to ICU risk stratification from nursing notes. *J Biomed Inform* 2015;**54**:114–20. doi:10.1016/j.jbi.2015.02.003

7 McMillan S, Chia C-C, Esbroeck AV, *et al.* ICU Mortality Prediction using Time Series Motifs. ;:4.

8 Taori RN, Lahiri KR, Tullu MS. Performance of PRISM (Pediatric Risk of Mortality) Score and PIM (Pediatric Index of Mortality) Score in a Tertiary Care Pediatric ICU. *Indian J Pediatr* 2010;**77**:5.

9 Zhu Y, Fan X, Wu J, *et al.* Predicting ICU Mortality by Supervised Bidirectional LSTM Networks. ;:12.

10 Luo Y, Xin Y, Joshi R, *et al.* Predicting ICU Mortality Risk by Grouping Temporal Trends from a Multivariate Panel of Physiologic Measurements. ;:9.

11 Yu R, Zheng Y, Zhang R, *et al.* Using a Multi-Task Recurrent Neural Network With Attention Mechanisms to Predict Hospital Mortality of Patients. *IEEE J Biomed Health Inform* 2020;**24**:486–92. doi:10.1109/JBHI.2019.2916667

12 Marafino BJ, Park M, Davies JM, *et al.* Validation of Prediction Models for Critical Care Outcomes Using Natural Language Processing of Electronic Health Record Data. *JAMA Netw Open* 2018;**1**:e185097–e185097. doi:10.1001/jamanetworkopen.2018.5097

13 Nachimuthu SK, Haug PJ. Early Detection of Sepsis in the Emergency Department using Dynamic Bayesian Networks. *AMIA Annu Symp Proc* 2012;**2012**:653–62.

14 Barton C, Chettipally U, Zhou Y, *et al.* Evaluation of a machine learning algorithm for up to 48-hour advance prediction of sepsis using six vital signs. *Comput Biol Med* 2019;**109**:79–84. doi:10.1016/j.compbiomed.2019.04.027

15 Knaus WA, Draper EA, Wagner DP, *et al.* APACHE II: A severity of disease classification system. *Crit Care Med* 1985;**13**:818–29.

16 Aczon M, Ledbetter D, Ho L, *et al.* Dynamic Mortality Risk Predictions in Pediatric Critical Care Using Recurrent Neural Networks. *ArXiv170106675 Cs Math Q-Bio Stat* Published Online First: 23 January 2017.http://arxiv.org/abs/1701.06675 (accessed 4 Sep 2020).

17 Alves T, Laender A, Veloso A, *et al.* Dynamic Prediction of ICU Mortality Risk Using Domain Adaptation. *ArXiv191210080 Cs Stat* Published Online First: 20 December 2019. doi:10.1109/BigData.2018.8621927

18 HOURLY PREDICTION OF ORGAN FAILURE AND OUTCOME IN INTENSIVE CARE BASED ON DATA MINING TECHNIQUES: In: *Proceedings of the 12th International Conference on Enterprise Information Systems*. Funchal, Madeira, Portugal: SciTePress - Science and Technology Publications 2010. 270–7. doi:10.5220/0002903802700277

19 Yu K, Zhang M, Cui T, *et al.* Monitoring ICU Mortality Risk with A Long Short-Term Memory Recurrent Neural Network. *Pac Symp Biocomput Pac Symp Biocomput* 2020;**25**:103–14.

20 Desautels T, Calvert J, Hoffman J, *et al.* Prediction of Sepsis in the Intensive Care Unit With Minimal Electronic Health Record Data: A Machine Learning Approach. *JMIR Med Inform* 2016;**4**:e28. doi:10.2196/medinform.5909

21 Johnson AEW, Mark RG. Real-time mortality prediction in the Intensive Care Unit. *AMIA Annu Symp Proc* 2018;**2017**:994–1003.

22 An Interpretable Machine Learning Model for Accurate Prediction of Sepsis in the ICU. - Abstract - Europe PMC. https://europepmc.org/article/PMC/5851825 (accessed 3 Sep 2020).

23 Mao Q, Jay M, Hoffman JL, *et al.* Multicentre validation of a sepsis prediction algorithm using only vital sign data in the emergency department, general ward and ICU. *BMJ Open* 2018;**8**:e017833. doi:10.1136/bmjopen-2017-017833

24 Shashikumar SP, Li Q, Clifford GD, *et al.* Multiscale network representation of physiological time series for early prediction of sepsis. *Physiol Meas* 2017;**38**:2235–48. doi:10.1088/1361-6579/aa9772

25 Goldstein BA, Pencina MJ, Montez-Rath ME, *et al.* Predicting mortality over different time horizons: which data elements are needed? *J Am Med Inform Assoc* 2017;**24**:176–81. doi:10.1093/jamia/ocw057

26 Schvetz M, Fuchs L, Novack V, *et al.* Outcomes prediction in longitudinal data: Study designs evaluation, use case in ICU acquired sepsis. *J Biomed Inform* 2021;**117**:103734. doi:10.1016/j.jbi.2021.103734

27 van Wyk F, Khojandi A, Mohammed A, *et al.* A minimal set of physiomarkers in continuous high frequency data streams predict adult sepsis onset earlier. *Int J Med Inf* 2019;**122**:55–62. doi:10.1016/j.ijmedinf.2018.12.002

28 van Wyk F, Khojandi A, Kamaleswaran R. Improving Prediction Performance Using Hierarchical Analysis of Real-Time Data: A Sepsis Case Study. *IEEE J Biomed Health Inform* 2019;**23**:978–86. doi:10.1109/JBHI.2019.2894570

29 Johnson AEW, Pollard TJ, Shen L, *et al.* MIMIC-III, a freely accessible critical care database. *Sci Data* 2016;**3**:160035. doi:10.1038/sdata.2016.35

30 The SOFA (Sepsis-related Organ Failure Assessment) score to describe organ dysfunction/failure. ;:4.

31 Rubin DB. *Multiple Imputation for Nonresponse in Surveys*. John Wiley & Sons 2004.

32 Wong A, Otles E, Donnelly JP, *et al.* External Validation of a Widely Implemented Proprietary Sepsis Prediction Model in Hospitalized Patients. *JAMA Intern Med* 2021;**181**:1065–70. doi:10.1001/jamainternmed.2021.2626