



OPEN

## Subtle variation in sepsis-III definitions markedly influences predictive performance within and across methods

Samuel N. Cohen<sup>1,4</sup>, James Foster<sup>7</sup>, Peter Foster<sup>4</sup>, Hang Lou<sup>2</sup>, Terry Lyons<sup>1</sup>, Sam Morley<sup>1</sup>, James Morrill<sup>1</sup>, Hao Ni<sup>2✉</sup>, Edward Palmer<sup>3</sup>, Bo Wang<sup>4,5</sup>, Yue Wu<sup>6</sup>, Lingyi Yang<sup>1,4</sup> & Weixin Yang<sup>1</sup>

Early detection of sepsis is key to ensure timely clinical intervention. Since very few end-to-end pipelines are publicly available, fair comparisons between methodologies are difficult if not impossible. Progress is further limited by discrepancies in the reconstruction of sepsis onset time. This retrospective cohort study highlights the variation in performance of predictive models under three subtly different interpretations of sepsis onset from the sepsis-III definition and compares this against inter-model differences. The models are chosen to cover tree-based, deep learning, and survival analysis methods. Using the MIMIC-III database, between 867 and 2178 intensive care unit admissions with sepsis were identified, depending on the onset definition. We show that model performance can be more sensitive to differences in the definition of sepsis onset than to the model itself. Given a fixed sepsis definition, the best performing method had a gain of 1–5% in the area under the receiver operating characteristic (AUROC). However, the choice of onset time can cause a greater effect, with variation of 0–6% in AUROC. We illustrate that misleading conclusions can be drawn if models are compared without consideration of the sepsis definition used which emphasizes the need for a standardized definition for sepsis onset.

The rise in comprehensive electronic health record systems (EHRs) has enabled the application of state-of-the-art machine learning (ML) models in predictive diagnostics<sup>1–3</sup>. In this context, machine learning typically focuses on developing models to maximise predictive accuracy against a given target clinical definition. The aim for the research community is to develop and critique models to find the best solutions for prediction. However, little attention is paid to small variations in these definitions used to define a predictive target when operationalised within the EHRs. Consequently, variations in clinical definitions, and their interpretation within an EHRs, lead to severe difficulty in fair comparisons of predictive methods across the literature. An interesting research question that arises from this current practice is how robust are the conclusions drawn from a comparison of different studies if there exist subtle differences in target definition? We seek to explore this, in the context of predicting sepsis onset time, by quantifying the impact that variations in onset time definitions have on a suite of methods (which are chosen from tree-based methods, deep learning, and survival analysis methods to cover popular and performant methods used for sepsis onset prediction).

Sepsis—a heterogeneous syndrome characterized by infection-induced organ dysfunction<sup>4</sup>—is a global health concern. In 2017 alone, sepsis was estimated to affect nearly 50 million people worldwide, resulting in 11 million deaths<sup>5</sup>. EHRs do not contain a structured “ground truth” identifying sepsis. The sepsis-III definition operationalizes sepsis within an EHRs as an increase in the sequential organ failure assessment (SOFA) score<sup>6</sup> by two or more, in the presence of suspected or confirmed infection. In lieu of a gold standard label, it is a requirement for studies in this domain that clinical data are labelled with a “sepsis phenotype”<sup>7</sup>. In the sepsis-III study<sup>8,9</sup>,

<sup>1</sup>Mathematical Institute, University of Oxford, Oxford, UK. <sup>2</sup>Department of Mathematics, University College London, Room 603, 25 Gordon St, London WC1H 0AY, UK. <sup>3</sup>Bloomsbury Institute of Intensive Care Medicine, University College London, London, UK. <sup>4</sup>The Alan Turing Institute, London, UK. <sup>5</sup>Center for Precision Psychiatry, Massachusetts General Hospital, Boston, MA, USA. <sup>6</sup>Department of Mathematics and Statistics, University of Strathclyde, Glasgow, UK. <sup>7</sup>Department of Mathematical Sciences, University of Bath, Bath, UK. ✉email: h.ni@ucl.ac.uk

this phenotype was developed using the coincident administration of antibiotics and taking of blood cultures as a proxy for “suspicion of infection.” This allowed the labelling of patients who are likely to have sepsis, and so created a target definition for statistical models. Building on this definition, machine learning-based early warning systems have been studied for numerous clinical tasks, such as detection of sepsis<sup>10–13</sup> and septic shock<sup>14–16</sup>.

The sepsis-III definition does not explicitly define the sepsis onset time, leading to different interpretations when this is of primary interest<sup>8,11,17–22</sup>. A recent systematic review<sup>23</sup> of machine learning methods for sepsis prediction highlighted the inconsistencies in sepsis definition across different papers and that most of the associated code is not publicly available. See also<sup>24</sup> for a review and discussion on the controversy relating to sepsis definition from the clinical perspective. This lack of a precise onset time, together with the absence of verifiable implementations of machine learning models, prevents effective comparison between performance of different methods.

From the sepsis-III definition, we can see that it is ambiguous whether the onset should be defined as the time of organ failure, the time of the suspicion of infection, or the earlier of these two events. All three are used across the literature. We consider these three competing interpretations of sepsis definition by evaluating them with a suite of advanced models on the MIMIC-III dataset<sup>25</sup>. Each definition, depending on when organ dysfunction or suspected infection is identified, has a distinct clinical interpretation<sup>11,17,18</sup>. MIMIC-III contains high frequency data for patients admitted to ICU. We note that although this has some limitations, this is a dataset frequently utilized for sepsis prediction tasks and can highlight the issues relating to subtle variations in sepsis label construction. We found significant differences in predictive performance between definitions, confirming the need for precise and verifiable methods, to allow comparisons to be made.

This study is accompanied by code implementing a pipeline for the development of early warning scores and their evaluation, including the use of signature methods for time series analysis. All code used in this study has been made publicly available<sup>26</sup>, to allow further development and comparisons. This manuscript is prepared in accordance with the RECORD statement; the extension to STROBE for research based on routinely collected data<sup>27</sup>.

## Materials and methods

### Determining the onset of sepsis: $t_{\text{sepsis}}$

Determining the onset time of sepsis ( $t_{\text{sepsis}}$ ), requires identification of the onset of suspected infection ( $t_{\text{suspicion}}$ ) and the time at which the SOFA score deteriorates by at least two points ( $t_{\text{sofa}}$ ). We align to the sepsis-III definitions<sup>9</sup> as closely as possible and describe any necessary deviations.

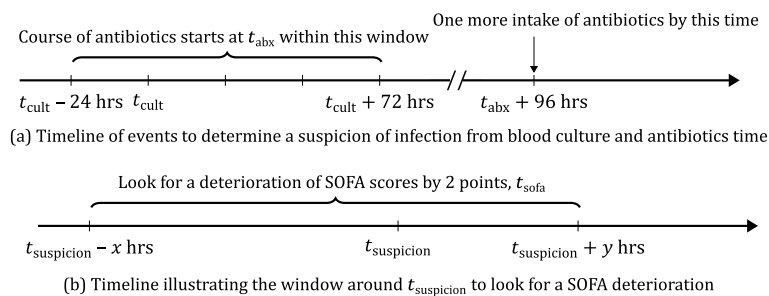
We used the drawing of blood cultures and the administration of antibiotics, within proximity to one another, as surrogates for clinical suspicion of infection. We identified the times that blood cultures were taken ( $t_{\text{cult}}$ ) and that a course of antibiotics was first administered ( $t_{\text{abx}}$ ). As patients typically took multiple doses of the same antibiotic, subsequent doses of the same antibiotic were classed as being in the same course, as long as consecutive doses were administered within two days of each other. Only the initial time of each course of antibiotics was used to look for clinical suspicion. We required at least two doses of any type of antibiotics to be administered within a 96-h period for a  $t_{\text{abx}}$  to count towards a suspicion of infection. One-off prophylactic antibiotics were not included.

A valid suspicion of infection required that if the antibiotic was given first, the culture must be obtained within 24 h, or that if the culture was taken first, the antibiotic must have been administered within 72 h. Once  $t_{\text{cult}}$  and  $t_{\text{abx}}$  had been identified with appropriate proximity, the onset of suspected infection ( $t_{\text{suspicion}}$ ) was defined as the first occurrence in this pair (Fig. 1a).

To identify organ dysfunction attributable to infection, we required an increase in the SOFA score of least two points (at time  $t_{\text{sofa}}$ ) in an interval around  $t_{\text{suspicion}}$  (Fig. 1). We call this interval the “SOFA window” described by the quantities  $\{x, y\}$  as in Fig. 1b. We calculated  $t_{\text{sofa}}$  as the first time that the SOFA score was two above its initial value. This is a popular approach for the early detection problem<sup>11,18</sup>.

For any given SOFA window, three competing onset times for sepsis ( $t_{\text{sepsis}}$ ) exist. These are labelled as **H1**, **H2** and **H3** and defined as follows:

- **H1:**  $t_{\text{sepsis}} = t_{\text{sofa}}$ , sepsis onset occurs at the onset of salient organ dysfunction.
- **H2:**  $t_{\text{sepsis}} = t_{\text{suspicion}}$ , sepsis onset occurs at the onset of suspicion of infection.



**Figure 1.** Timelines to determine the time of suspected infection (panel a) and the sepsis onset time (panel b). We name the interval around  $t_{\text{suspicion}}$  as the “SOFA window” where  $\{x, y\}$  denotes the time in hours before and after  $t_{\text{suspicion}}$  that is used for the detection of a change in SOFA score.

- **H3:**  $t_{\text{sepsis}} = \min(t_{\text{sofa}}, t_{\text{suspicion}})$ , sepsis onset occurs at the earlier of these two events.

### Data resource

We used the MIMIC-III dataset<sup>25</sup> to assess the impact of these definitions on the prediction of sepsis onset. MIMIC-III comprises de-identified patient-level data from over 40,000 patients that stayed in intensive care units (ICUs) between 2001 and 2012 at the Beth Israel Deaconess Medical Center, Boston, Massachusetts (USA). The MIMIC-III dataset was approved by the Institutional Review Boards of Beth Israel Deaconess Medical Center and the Massachusetts Institute of Technology.

Information on antibiotics was recorded infrequently prior to 2004. Since the year of hospitalization was removed during the de-identification process, we excluded all patients recorded under the “CareVue” system, which was known to be in operation only until 2008. Only those patients recorded under the “MetaVision” system is retained. Further details can be found at<sup>25</sup>.

Data were split into training (85%) and test (15%) sets by stratified sampling based on sex, age, intensive care length of stay, and whether the patient ever received invasive ventilation. This split was performed prior to applying further exclusion criteria, and only one hospital visit was retained per patient. We chose this split to balance between having enough data in the training set to calibrate the models (with cross validation) and having a reasonable size of test set (to reduce variance in results).

Granular information on the time of antibiotic administration and organ dysfunction was not available prior to ICU admission (for example, prescriptions only have data for date prescribed) hence precise sepsis onset time cannot be identified. Following<sup>28</sup>, we excluded patients who were prescribed antibiotics before their entry to ICU. Patients attending the cardiothoracic surgical ICU were excluded since their requirements for organ support are likely to produce false positive labels in the sepsis phenotype. Elective surgical patients were excluded as they are often prescribed prolonged antibiotic prophylaxis pre-arrival to the ICU and have, in general, a low risk of developing sepsis in their index ICU admission. We excluded patients whose length of stay in the ICU was shorter than four hours or longer than twenty days, to avoid the development of sepsis prior to ICU admission, or bias from atypical visits. Patients missing all vital sign data were removed since these likely represent cases with data quality issues.

Finally, we excluded patients who developed sepsis within four hours following entry to the ICU. Given the different interpretations of sepsis onset defined above (**H1–H3**) and SOFA windows of varying size, excluding patients who developed sepsis within four hours of entry to the ICU excluded a different number of patients depending on these choices. A study flow diagram is given in Fig. 2. See Appendix A for further information on the training and test set after exclusions.

### Modelling

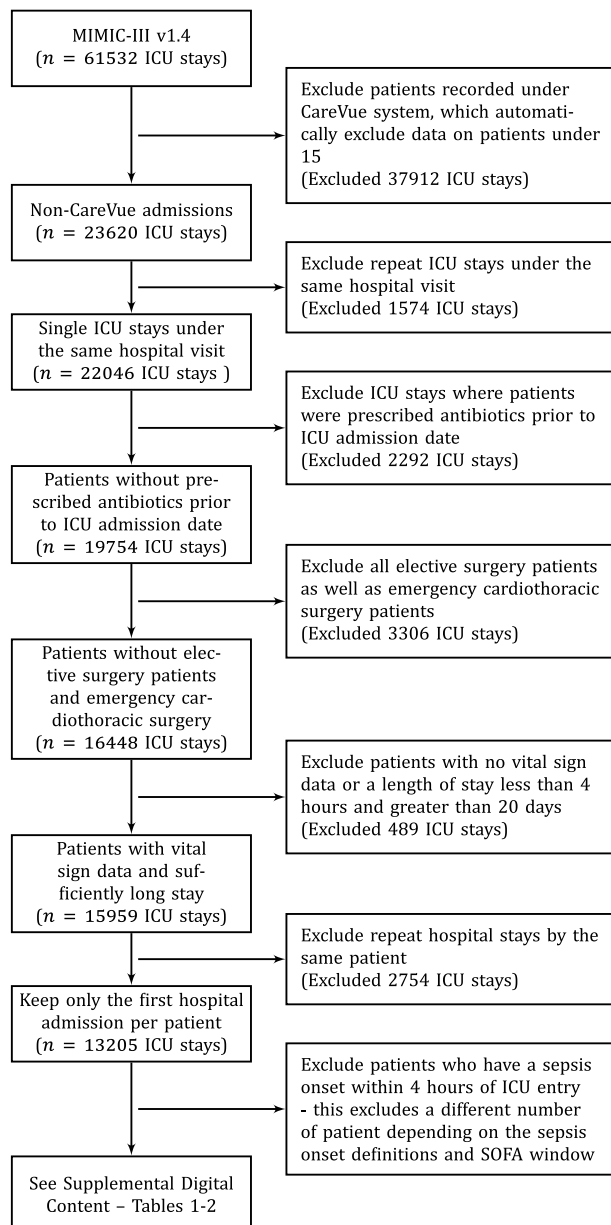
Using MIMIC-III data from the described cohort, we fitted several models to assign an hourly risk score, given all salient patient information up to the present, which indicates the likelihood a patient will develop sepsis over some pre-defined prediction horizon ( $T$  hours). For each definition of sepsis onset time (**H1–H3**) we evaluated three models which have been previously used in sepsis prediction tasks: a light gradient boosting machine (LGBM)<sup>19</sup>, a modified Cox proportional hazards model (CoxPHM)<sup>29</sup> and the long short-term memory (LSTM) neural network<sup>16</sup>. These are representative models from tree-based models<sup>30–32</sup>, survival analysis<sup>33–36</sup>, and deep-learning based models<sup>37–39</sup> respectively and further, these three classes cover a wide range of models used for this application. In particular, an LGBM model won the 2019 PhysioNet Computing in Cardiology Challenge on early prediction of sepsis<sup>21</sup>. Demonstrating how the performance of this winning model compares against other popular models on different definitions emphasizes the need for transparency in model comparison. See Appendix D and E for further details on the models and how we selected the hyperparameters for our study. We investigated the impact of subtle variations in sepsis definition by looking at the changes in performance and the relative ranking of these models.

Risk scores predicted by each model were converted into binary predictions by selecting a threshold risk score above which a patient would be classified as septic (Fig. 3). This threshold was set to achieve a sensitivity of 85% on the training set, as was used in a prior study<sup>11</sup>. We refer to this task as the real-time prediction task (See Appendix B for the precise problem formulation).

Models were trained on observations prior to the sepsis onset time ( $t_{\text{sepsis}}$ ), to ensure that the models were predictive, rather than replicating the chosen sepsis definition. Each model was trained against each definition of sepsis onset time (**H1–H3**) providing nine models for each set of  $\{x, y, T\}$ .

Models were fitted with 38 raw and 75 derived predictor variables, similar to those used in Morrill et al.<sup>13,14,16</sup>, though we did not use end tidal carbon dioxide, as it is typically unavailable in MIMIC-III. Raw variables were in four broad categories: demographics (e.g. age, sex), vital signs (e.g. heart rate, pulse oximetry), laboratory results (e.g. bicarbonate, pH), and admission information (e.g. time since admission to hospital/ICU). A full list of raw variables used is found in Supplemental Digital Content—Table 6. In addition to these raw variables, we derived time-dependent features to use in our models (Supplemental Digital Content—Table 7). Time varying features were processed using a rolling window and signature transformation; these capture key geometric and temporal properties of timeseries data<sup>21,40,41</sup>. Further discussion of feature extraction is found in Appendix C.

As our goal is real-time sepsis prediction, models were trained to optimize prediction of patient labels averaged over all patients and observation times. The hyperparameters of each model were tuned with 5-fold cross-validation performed against training data. Models with the best performing hyperparameters were subsequently refitted against the full training set to obtain final model parameters before evaluation on the test set. All methods were performed in accordance with the relevant guidelines and regulations.



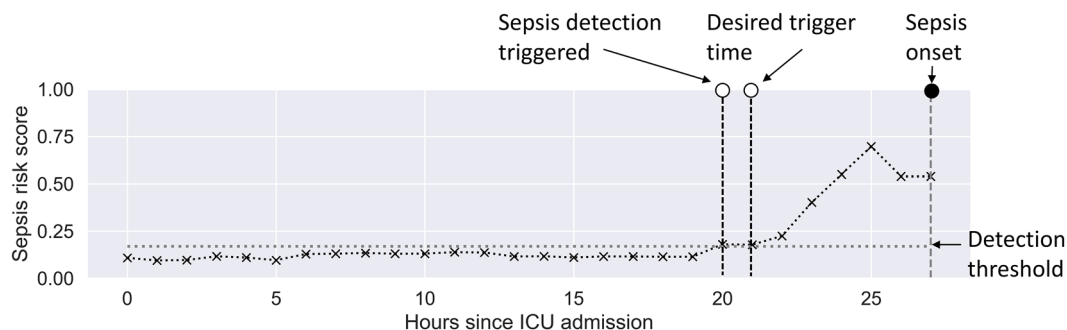
**Figure 2.** Flow Study flow diagram. Different analysis cohorts are produced depending upon the sepsis onset definitions and SOFA windows applied. These are detailed in Supplemental Digital Content—Table 1 and Supplemental Digital Content-Table 2.

## Evaluation

To evaluate model performance, we calculated the area under the receiver operator characteristic curve (AUROC, or c-statistic) of each model. This describes the predictive performance of models, averaged over all patients and observation times.

## Sensitivity analysis

Three main sensitivity analyses were considered to evaluate the sensitivity of the models and sepsis onset definitions. First, different sizes of SOFA window were considered. Second, we retrained all models for the **H1** and **H2** sepsis onset definitions, but with ICU admissions excluded using the strictest exclusion criteria from **H3**. This allowed us to examine the differences between **H1** and **H2** with the same cohort (as **H3**), and so isolate any discrepancies to the sepsis definitions and models, and not to differences in cohorts (input data). Last, microbiological samples other than blood cultures were included in the definition of  $t_{\text{sofa}}$ .



**Figure 3.** An illustrative example of converting sepsis risk scores to binary sepsis. The dotted black line with “x” markers indicates the risk score, while the horizontal dotted line indicates the chosen cutoff threshold to determine binary labels. In this case, the model predicted a positive sepsis label for the first time at 20 h after ICU admission. Here the prediction horizon  $T$  is set at 6. This indicates that the model is designed to predict 6 h ahead of the true sepsis onset time, which in this case is 21 h into the stay.

## Results

Following our study exclusions, between 737 and 1861 ICU admissions were available for model training, with 140 and 317 admissions in the test sets, depending upon the sepsis onset definition and SOFA window under investigation (Supplemental Digital Content—Table 1 and Table 2). Baseline characteristics for the cohort with SOFA window {24, 12} are shown in Supplemental Digital Content—Table 3.

### Prediction task

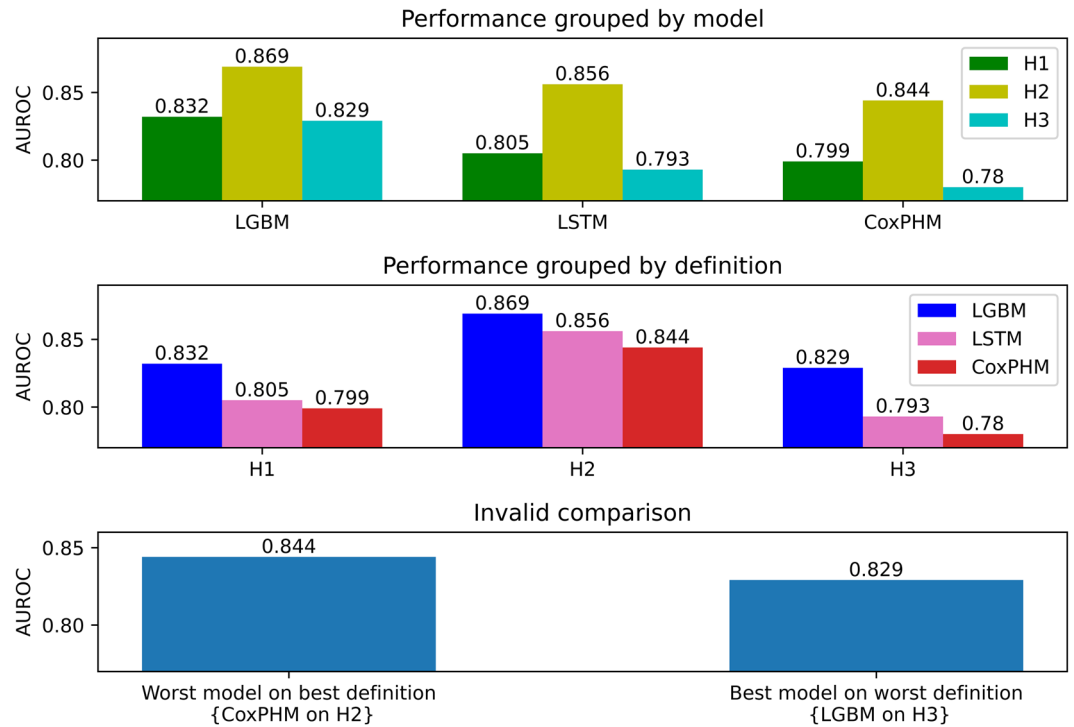
We compared the real-time prediction performance of LGBM, LSTM and CoxPHM models for each sepsis onset definition. Table 1 shows the performance of definitions in terms of test metrics (AUROC, specificity and accuracy), with **H2** yielding consistently higher scores than **H1** or **H3**, independent of the model chosen. The earliest sepsis onset time (**H3**) consistently gives the lowest performance scores. It is interesting to note that the variation in performance across sepsis criteria (0–6% AUROC) exceeds the variation in performance across models (1–5% AUROC).

The results are represented visually in the two top subplots of Fig. 4. We see that LGBM outperforms the others in that it has the highest AUROC for each fixed sepsis definition. However, in the bottom subplot, it can be seen that the best model on the worst performing definition (LGBM on **H3**) performs worse than the worst model on the best definition (CoxPHM on **H2**). This emphasizes the point that without full transparency on the design choice, fair comparison across models is not possible. Naive comparison across different definitions may result in misleading conclusions.

Full details of the evaluation, including the effects of changes in SOFA window and time horizon are detailed in Appendix F of Supplemental Digital Content—Tables 9 and 10. The evaluation of the real-time prediction

	AUROC	Specificity	Sensitivity	Accuracy
<b>H1</b>				
LGBM	0.832 [0.823,0.842]	0.784	0.723	0.725
LSTM	0.805 [0.796,0.815]	0.596	0.811	0.601
CoxPHM	0.799 [0.789,0.809]	0.546	0.850	0.553
<b>H2</b>				
LGBM	0.869 [0.862,0.876]	0.761	0.808	0.806
LSTM	0.856 [0.848,0.863]	0.718	0.813	0.721
CoxPHM	0.844 [0.836,0.852]	0.672	0.831	0.677
<b>H3</b>				
LGBM	0.829 [0.818,0.840]	0.768	0.713	0.714
LSTM	0.793 [0.781,0.805]	0.524	0.831	0.529
CoxPHM	0.780 [0.769,0.793]	0.490	0.840	0.496

**Table 1.** Summary of AUROC, specificity, accuracy of LGBM, LSTM and CoxPHM for the real-time prediction on the test set. Here the SOFA window is {24, 12} and prediction horizon  $T = 6$ . The mean [lower 95% confidence, upper 95% confidence] of AUROC with non-parametric bootstrapped confidence intervals drawn from 100 resamples are provided. Specificity and accuracy are calculated based on that the sensitivity on the training set was chosen to be 85%.



**Figure 4.** Highlights in the differences of AUROC for the SOFA window {24,12}. (Top panel) AUROC when models are evaluated on the test set, grouped by model. (Middle panel) AUROC when models are evaluated on the test set, grouped by sepsis onset definition. (Bottom) An illustration of an invalid comparison between models when the underlying sepsis definition is different. Here the best performing model on the worst definition, that is, LGBM on H3 has a lower AUROC than the worst performing model on the best definition, namely CoxPHM on H2.

method is consistent with model evaluation in the literature. However, we can infer other interesting outcomes by looking at predictions made on each patient. Further discussion can be found in Appendix G.

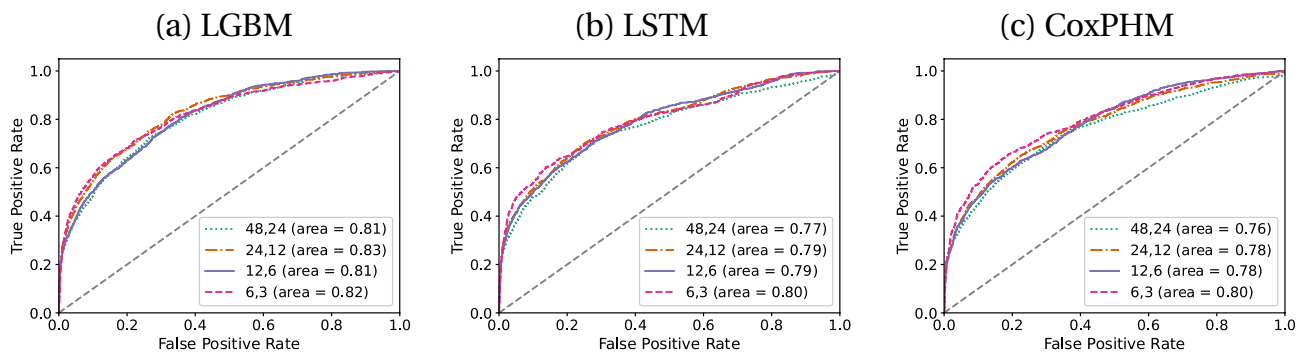
### Sensitivity analyses

#### Impact of SOFA deterioration observation window

All models were robust to variations of the SOFA window. Figure 5 demonstrates that, despite optimizing the models' hyperparameters using the SOFA window {24,12}, the receiver operating characteristic (ROC) curves of the prediction from each model are very close for different SOFA windows under H3. For similar plots of the other two definitions see Supplemental Digital Content—Fig. 3.

#### Impact of strict exclusion criteria

Table 2 shows that the performance of each model under H1 and H2 decreases with the use of the strict exclusion criteria (that is, excluding patients who develop sepsis, according to H3, within the first four hours). While the



**Figure 5.** ROC plots for real-time prediction at different SOFA windows for T = 6 and definition H3 on the test set for (a) LGBM, (b) LSTM and (c) CoxPHM model.



	H1	H2	H3
	LGBM/LSTM/CoxPHM	LGBM/LSTM/CoxPHM	LGBM/LSTM/CoxPHM
Main extraction method	0.832/0.805/0.798	0.869/0.856/0.844	0.829/0.793/0.780
Strict exclusion	0.832/0.796/0.785	0.851/0.805/0.795	0.829/0.793/0.780
Other cultures	0.816/0.792/0.778	0.847/0.837/0.816	0.801/0.756/0.751

**Table 2.** AUROC scores of the three models (LGBM/LSTM/CoxPHM) on the test data using main extraction method and the other data exclusion choices for the real-time prediction with SOFA window {24, 12} and  $T = 6$ .

models continue to be ranked similarly under these exclusion criteria, the performance variation of each model across definitions become smaller, especially LSTM and CoxPHM. This suggests that the superior performance observed earlier under **H1** and **H2** is largely due to the differences in patient exclusion—it appears the task of predicting sepsis is more difficult under **H3** than **H1** and **H2** principally due to the exclusion of more patients who develop sepsis soon after admission.

#### Impact of culture selection

The definition of suspected infection was broadened to include specimens other than blood culture. We note that in this case, the performance across all definitions and models shown in Table 2 decreased.

## Discussion

Using representative models from both classical statistics and machine learning, we have demonstrated that model performance in predicting sepsis onset was markedly sensitive to subtle variations in onset definitions. This performance impact was at times more pronounced than the gains from the different models themselves. Taking the LGBM model (which was consistently the best model for fixed definitions) to predict sepsis under **H3** gives a lower AUROC than using the CoxPHM (which performed worst for each fixed definition) to predict sepsis with **H2** definition. If these results were reported across different papers and compared without consideration of the underlying sepsis definitions used, we would get misleading conclusions.

There was a consistent ranking of onset definitions across all models, with the order **H2**, **H1** and **H3**. This characterization of performance is consistent with one prior large database study<sup>17</sup>. In the context of predictive modelling, this does not imply the models targeting the **H2** definition are any better at predicting the true sepsis condition of the patient. Rather, the models perform better at predicting the onset time defined using this definition and, in particular, using data after the corresponding exclusion criteria have been applied. This highlights the significance that a change in onset definition can have on model performance with everything else fixed. We note that these results are not evidence for whether **H2** should be chosen as the standard definition.

Significant variation in the application of sepsis onset definitions exists in the prediction literature, including examples of **H1**<sup>11</sup>, **H2**<sup>9</sup> and **H3**<sup>10,12,14,15</sup>. As a definition applied to retrospective data, **H1** appears to be the most clinically meaningful, given that sepsis is defined as a “life-threatening organ dysfunction caused by a dysregulated host response to infection”<sup>9</sup>. Thus, conditional on there being a confirmed suspicion of infection, the point of deterioration in organ function would best reflect the moment a patient becomes septic. However, we have demonstrated the sensitivity of a wide range of high-fidelity models to these target definitions.

There is no current consensus on the best SOFA window around  $t_{\text{suspicion}}$  to look for an increase in SOFA. Seymour et al.<sup>8</sup> used several SOFA windows, ranging from 3 to 48 h prior to  $t_{\text{suspicion}}$  and 3–24 h following  $t_{\text{suspicion}}$ . Others have used a SOFA window from 24 h prior to 12 h post  $t_{\text{suspicion}}$ <sup>11,17</sup>. Our results are reassuring, as this aspect of the definition has limited impact on predictive performance, even though the cohort size changes significantly.

Finally we note that modern data science advancements, for example, extracting signature features, can provide improved techniques for signaling the likelihood of an impending event from patient observations. However, it is impossible to compare mechanisms across the literature and develop quality tools without developing standard benchmark criteria. Our study highlights the pressing need for a gold-standard sepsis phenotype for machine learning research on early sepsis detection. The current definitions, based on SOFA, have been optimized for ease of use of application at the bedside. This imposes natural limitations for advanced statistical models. For example, the SOFA score comprises laboratory results, which are typically measured only once per day. In contrast, other constituents of SOFA comprise underlying biological data which may be continuously monitored; the juxtaposition of these may reduce statistical power<sup>42</sup>. Improvements to sepsis criteria, both in the precision of definition and in connection with clinical practice, will allow greater contributions to flow from machine learning research.

## Conclusion

To summarize, with the availability of electronic healthcare datasets, we can conduct numerous retrospective studies to design predictive models. However when comparing the vast amount of research, the precise construction of the target label can be overlooked. Our work demonstrates that a subtle difference in sepsis criteria leads to significant variation in model performance. Under any fixed sepsis definition, our implementation of LGBM was consistently the best model whilst CoxPHM was consistently the worst. However, if we compared

the performance of LGBM under the **H3** definition against CoxPHM under the **H2** definition, then we would find that the AUROC of CoxPHM is higher. If these were models presented by different papers (using different sepsis definitions), then we may inaccurately conclude that CoxPHM is the better model for sepsis prediction. The difference in the interpretation of the clinical criteria for sepsis created an issue which has not been addressed in literature.

In general, considering different methods, while ignoring variations in the criteria used to evaluate them, may lead to comparisons that draw invalid conclusions and hinder the progress to find the best predictive model. Therefore, we make the following recommendations for retrospective studies on EHRs data:

- The data management protocol should include the full details of extracting the clinical target of interest from EHRs, to allow full reproducibility. Publication of results should typically include a link to the code used.
- A high-quality early warning system should consistently outperform other methods and be robust to the variations in data.
- A gold-standard for the clinical outcome of interest, in a format which can be extracted from EHRs, should be established by the joint effort of the clinical community and machine learning community.

### Limitations

Owing to the availability of data assets like MIMIC, the majority of research in sepsis prediction in the ICU has been performed on cohorts that commence at the time of ICU admission, as is the case here. However, it is important to note that most sepsis cases admitted to an ICU display evidence of organ dysfunction prior to arrival in the ICU<sup>8</sup>. Extending this field of research to pre-ICU observational data comes with its own inherent challenges, since patients outside ICU are monitored far less frequently and will be subject to informative sampling based on their acute physiology. This highlights the importance of developing high-quality pre-ICU data as a target for future research.

Patients who received antibiotics prior to ICU admission were necessarily excluded from our study, since the MIMIC-III database does not contain granular detail on organ dysfunction prior to arrival in the ICU. It is probable that the sepsis onset time in these cases was prior to admission to the ICU, and including these patients would have unreasonably biased model fit.

Our primary aim is to investigate the discrepancy in model performance and potential risks in misleading model comparisons caused by a subtle variation on the sepsis definition. To conduct a systematic and comprehensive comparative study, we have narrowed our scope to three representative models, employing only forward-filling as our data imputation technique. This means that we explore neither the full range of predictive methods used for sepsis prediction nor the large literature of data imputation methods available. Therefore, although we have found that LGBM performs best in this analysis, we do not claim that it is the state-of-the-art model for early detection of sepsis. Moreover, we acknowledge the limitation of the forward-filling, which can be further improved by alternative imputation method for improving the predictive performance. These are important considerations when finding the best predictive model, but does not affect our conclusion that transparency and a standard sepsis definition is vital to identify such models.

### Data availability

The data we used in this paper is extracted from the MIMIC-III database. Once the required training and credentials are obtained, this dataset is accessible from PhysioNet at <https://physionet.org/content/mimiciii/1.4/>. The MIMIC-III project was approved by the Institutional Review Boards of Beth Israel Deaconess Medical Center (Boston, MA) and the Massachusetts Institute of Technology (Cambridge, MA). Requirement for individual patient consent was waived because the project did not impact clinical care and all protected health information was deidentified.

### Code availability

Our code is publicly available at <https://zenodo.org/record/5168789>. We have included detailed instructions on how to obtain the exact dataset used starting from the MIMIC-III database, as well as precise instructions on running the method so that we have an end-to-end pipeline to fully replicate the results.

Received: 4 April 2023; Accepted: 11 January 2024

Published online: 22 January 2024

### References

1. Neves, A. L. *et al.* Using electronic health records to develop and validate a machine-learning tool to predict type 2 diabetes outcomes: A study protocol. *BMJ Open* **11**, 461 (2021).
2. Zale, A. D., Abusamaan, M. S., McGready, J. & Mathioudakis, N. Development and validation of a machine learning model for classification of next glucose measurement in hospitalized patients. *EClinicalMedicine* **44**, 101290 (2022).
3. Tomašev, N. *et al.* Use of deep learning to develop continuous-risk models for adverse event prediction from electronic health records. *Nat. Protoc.* **16**, 2765–2787 (2021).
4. Pool, R., Gomez, H. & Kellum, J. A. Mechanisms of organ dysfunction in sepsis. *Crit. Care Clin.* **34**(1), 63–80 (2018).
5. Rudd, K. E. *et al.* Global, regional, and national sepsis incidence and mortality, 1990–2017: analysis for the Global Burden of Disease Study. *The Lancet* **395**(10219), 200–211 (2020).
6. 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. *Intens. Care Med.* **22**, 707–710 (1996).
7. Shah, A. D. *et al.* Descriptors of sepsis using the sepsis-3 criteria: A cohort study in critical care units within the UK National Institute for Health Research critical care health informatics collaborative. *Crit. Care Med.* **49**, 1883 (2021).



8. Seymour, C. W. *et al.* Assessment of clinical criteria for sepsis: For the Third International Consensus Definitions for Sepsis and Septic Shock (Sepsis-3). *JAMA* **315**, 762–774 (2016).
9. Singer, M. *et al.* The third international consensus definitions for sepsis and septic shock (Sepsis-3). *JAMA* **315**, 801–810 (2016).
10. Komorowski, M., Celi, L. A., Badawi, O., Gordon, A. C. & Faisal, A. A. The artificial intelligence clinician learns optimal treatment strategies for sepsis in intensive care. *Nat. Med.* **24**, 1716–1720 (2018).
11. Nemati, S., Holder, A. R. F., Stanley, M. C. G. & Timothy, B. An interpretable machine learning model for accurate prediction of sepsis in the ICU. *Crit. Care Med.* **46**(4), 547–553 (2018).
12. Futoma, J., Hariharan, S. & Heller, K. Learning to detect sepsis with a multitask Gaussian process RNN classifier. *Int. Conf. Mach. Learn.* **2019**, 1174–1182 (2017).
13. Goh, K. H. *et al.* Artificial intelligence in sepsis early prediction and diagnosis using unstructured data in healthcare. *Nat. Commun.* **12**, 711 (2021).
14. Fernandez-Gutierrez, F. *et al.* Mining primary care electronic health records for automatic disease phenotyping a transparent machine learning framework. *Diagnostics* **11**, 1908 (2021).
15. Giannini, H. M. *et al.* A machine learning algorithm to predict severe sepsis and septic shock: Development, implementation, and impact on clinical practice. *Crit. Care Med.* **2019**, 47 (2019).
16. Fagerström, J., Bång, M., Wilhelms, D. & Chew, M. S. Liseplstm: A machine learning algorithm for early detection of septic shock. *Sci. Rep.* **9**, 1–8 (2019).
17. Reyna, M. A. *et al.* Early prediction of sepsis from clinical data: The PhysioNet/Computing in Cardiology Challenge 2019. *Comput. Cardiol. (CinC)* **2019**, 1 (2019).
18. Desautels, T. *et al.* Prediction of sepsis in the intensive care unit with minimal electronic health record data: A machine learning approach. *JMIR Med. Inform.* **4**, 3 (2016).
19. Morrill, J. *et al.* The signature-based model for early detection of sepsis from electronic health records in the intensive care unit. *Comput. Cardiol. CinC* **2019**, 141 (2019).
20. Tran, L., Nguyen, M. & Shahabi, C. Representation learning for early sepsis prediction. *Comput. Cardiol. CinC* **2019**, 1–4 (2019).
21. Morrill, J. H. *et al.* Utilization of the signature method to identify the early onset of sepsis from multivariate physiological time series in critical care monitoring. *Crit. Care Med.* **48**, e976–e981 (2020).
22. Kam, H. J. K. H. Y. Learning representations for the early detection of sepsis with deep neural networks. *Comput. Biol. Med.* **89**, 248–255 (2017).
23. Moor, M., Rieck, B., Horn, M., Jutzeler, C. R. & Borgwardt, K. Early prediction of sepsis in the ICU using machine learning: A systematic review. *Front. Med. (Lausanne)* **8**, 348 (2021).
24. Giacobbe, D. R. *et al.* Early detection of sepsis with machine learning techniques: A brief clinical perspective. *Front. Med. (Lausanne)* **8**, 617486 (2021).
25. Johnson, A. E. W. *et al.* MIMIC-III, a freely accessible critical care database. *Sci. Data* **3**, 1–9 (2016).
26. The supplementary code of “Subtle variation in sepsis-III definitions influences the predictive performance of machine learning.” (2022). <https://zenodo.org/records/5168789>.
27. Benchimol, E. I. *et al.* The REporting of studies Conducted using Observational Routinely-collected health Data (RECORD) statement. *PLoS Med.* **12**, e1001885 (2015).
28. Pieroni, M., Olier, I., Ortega-Martorell, S., Johnston, B. W. & Welters, I. D. In-hospital mortality of sepsis differs depending on the origin of infection: An investigation of predisposing factors. *Front. Med. (Lausanne)* **9**, 915224 (2022).
29. Henry, K. E., Hager, D. N., Pronovost, P. J. & Saria, S. A targeted real-time early warning score (TREWScore) for septic shock. *Sci. Transl. Med.* **7**, 299–322 (2015).
30. Breiman, L., Friedman, J. H., Olshen, R. A. & Stone, C. J. *Classification and Regression Trees* (Routledge, 2017).
31. Ke, G. *et al.* Lightgbm: A highly efficient gradient boosting decision tree. *Adv. Neural Inf. Process. Syst.* **2017**, 30 (2017).
32. Biau, G. & Scornet, E. A random forest guided tour. *Test* **25**, 197–227 (2016).
33. Kleinbaum, D. G. *et al.* *Survival Analysis: A Self-Learning Text* (Springer, 2012).
34. Therneau, T. M. & Grambsch, P. M. *The cox Model. Modeling Survival Data: Extending the Cox Model* 39–77 (Springer, 2000).
35. Machin, D., Cheung, Y. B. & Parmar, M. *Survival Analysis: A Practical Approach* (Wiley, 2006).
36. Clark, T. G., Bradburn, M. J., Love, S. B. & Altman, D. G. Survival analysis part I: Basic concepts and first analyses. *Br. J. Cancer* **89**, 232–238 (2003).
37. Goodfellow, I., Bengio, Y. & Courville, A. *Deep Learning* (MIT Press, 2016).
38. Deng, L. *et al.* Deep learning: Methods and applications. *Found. Trends Signal Process.* **7**, 197–387 (2014).
39. Pouyanfar, S. *et al.* A survey on deep learning: Algorithms, techniques, and applications. *ACM Comput. Surv. (CSUR)* **51**, 1–36 (2018).
40. Kidger, P., Bonnier, P., Arribas, I. P., Salvi, C. & Lyons, T. Deep signature transforms. *Adv. Neural Inf. Process. Syst.* **2019**, 3099–3109 (2019).
41. Chevyrev, I. & Oberhauser, H. Signature moments to characterize laws of stochastic processes. *J. Mach. Learn. Res.* **23**(1), 7928–7969 (2022).
42. Altman, D. G. & Royston, P. The cost of dichotomising continuous variables. *BMJ* **332**, 1080 (2006).

## Author contributions

All the authors have played a key background role in setting up the framework and the experimental analysis related to the paper. The authors with \* after their names are primary authors, who have made significant contribution to this work. More specifically, S.C., H.L., T.L., J.M., H.N., E.P., Y.W., and L.Y. designed the study; P.F., H.L., J.M., H.N., Y.W., and L.Y. conducted the code implementation of the study and analyzed the data; S.C., H.L., T.L., J.M., H.N., E.P., Y.W., and L.Y. wrote the manuscript.

## Funding

S.C., J.F., P.F., H.L., T.L., S.M., J.M., H.N., E.P., Y.W., and L.Y. are supported by the Alan Turing Institute under the EPSRC grant EP/N510129/1. J.F., T.L., S.M., H.N., and Y.W. are supported by the EPSRC under the program grant EP/S026347/1. T. L. is supported in part by the Data Centric Engineering Programme (under the Lloyd’s Register Foundation grant G0095) and the Office of National Statistics Programme (funded by the UK Government) and in part by the Hong Kong Innovation and Technology Commission (InnoHK Project CIMDA). T.L. and P.F. are supported by the Defence and Security Programme at the Alan Turing Institute, funded by the UK Government. H.L. is supported by UCL-CSC scholarship by University College London and the China Scholarship Council (CSC) from the Ministry of Education of P.R. China. L.Y. and J.M. are supported by EPSRC grant EP/L015803/1 and L.Y. is also supported by the Clarendon Fund. E.P. is supported by an NIHR clinical

lectureship. B. W. is supported by “The Harvard Program in Precision Psychiatry” under the funding of Harvard Medical School and the Sang Foundation.

### Competing interests

The authors declare no competing interests.

### Additional information

**Supplementary Information** The online version contains supplementary material available at <https://doi.org/10.1038/s41598-024-51989-6>.

**Correspondence** and requests for materials should be addressed to H.N.

**Reprints and permissions information** is available at [www.nature.com/reprints](http://www.nature.com/reprints).

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



**Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article’s Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article’s Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

© The Author(s) 2024