# Supplementary Appendix

## Table of contents

## Section S1: Additional methods

**a) Data sources and further cohort information**

In this study, electronic health record data was harmonized from four large ICU databases (versions are provided wherever available): i) AUMC: 1.0.2, ii) eICU: 2.0, iii) MIMIC-III: 1.4, and iv) HiRID: 1.1.1. Hourly-resolved sepsis annotations were implemented using the Sepsis-3 definition[1]. Depending on the database, the data were mostly collected in the last decade. Specifically, for AUMC between 2003 and 2016, for eICU between 2014 and 2015, for MIMIC-III between 2001 and 2012, for HiRID between 2008 and 2016. Hence, all included ICU datasets have no temporal overlap with the COVID-19 pandemic. The used variables were modeled using the open-source software package ricu. Additionally, in our supplemental data, we provide SNOMED codes for the used variables. An additional data source, the Emory dataset of the Physionet 2019 Computing in Cardiology Challenge[2], was intended to be included into this study. However, due to low availability of variables (only 35 variables overlapped with our set of 63 variables), as well as a preprocessed sepsis label that did not follow our implementation of Sepsis-3, this dataset was removed from our set of cohorts.

**b) Sepsis label annotation**

The Sepsis-3 criterion defined sepsis as co-occurrence of suspected infection and a SOFA increase of two or more points. We followed the approach of the original authors as closely as possible[1 3]. Suspected infection was defined as co-occurrence of antibiotic treatment and body fluid sampling. If antibiotic treatment occurred first, it needed to be followed by fluid sampling within 24 hours. Alternatively, if fluid sampling occurred first, it needed to be followed by antibiotic treatment within 72 hours in order for it to be classified as suspected infection. The earlier of the two times is taken as the suspected infection (SI) time. After this, the SI window is defined to begin 48 hours before the SI time and end 24 hours after the SI time. Within the SI window, a SOFA increase of two or more points is defined as the time of sepsis onset. While to our understanding the original Sepsis-3 definition is not committal on how to define the time of sepsis onset, we use the moment of SOFA increase as this makes it likely that both organ dysfunction and infection are co-present (for instance, infection may be clinically suspected with delay). By contrast, if SI time were used for defining sepsis onset time, we expect that in many cases, time points of patients prior to sepsis (e.g., still suffering from a focal infection) would be mislabelled as septic which could dilute the phenotype. Fig. S2B provides a visual representation of our implementation of Sepsis-3. Table S2D and Table S2E list all concepts that were used for constructing the SOFA score. SOFA is a composite score to assess the damage of different organ systems which is calculated over a look-back window of 24 hours. Consistent with previous works, missing values of SOFA components resulted in a score of 0 for those components. For the SOFA components,

there are two remarks we make about the score construction. For all patients who were sedated, we set the Glasgow Coma Scale (GCS) score to 15. Furthermore, the 24-hour urine output is not evaluated before 12 hours into ICU stay and appropriately scaled during hours 12 through 24 into ICU stay.
The eICU dataset reports only a small number of body fluid samplings, while the HiRID dataset reports no body fluid samplings at all. For this reason, the original definition of suspected infection is hard to implement on these datasets. Therefore, on these two datasets, we used an alternative definition of suspected infection, which was defined as a co-occurrence of multiple antibiotic administrations. We then tested whether the alternative definition coincided with the original definition on the MIMIC-III and AUMC datasets, where both antibiotic treatment and fluid sampling were reported. The comparison of the two definitions, in terms of Venn diagrams, is given in Fig. S3, where we use data prior to filtering of early onsets. What can be observed is that on the MIMIC-III dataset, the two definitions overlap to a large extent (Jaccard similarity 0.69). On the AUMC dataset, the multiple antibiotics definition was broader than the original definition, but it included almost all patients in the original definitions (Jaccard similarity 0.42). The reason for this was that the majority of admissions in the AUMC database were surgical, and such patients are frequently prescribed antibiotics for prophylactic purposes. Finally, a comparison of the two definitions on the subset of non-surgical admissions in the AUMC database showed a very strong overlap (Jaccard similarity 0.78).

**c) Exclusion criteria**

This section defines the exact exclusion criteria that were used for patient filtering. The steps are also visually shown in the flow diagram in Fig. S1. Namely, the following filtering steps were applied:

a) non-adult patients (< 14 years) were excluded,
b) hospital centres in the eICU dataset that had low Sepsis-3 prevalence were removed (< 15%), as it is likely that such hospitals would contribute negative cases (controls), which might in fact correspond to Sepsis-3 cases, but are not labelled as such due to data missingness; the list of hospital IDs that were used can be found in the config/cohorts.json file in the code repository.

Furthermore, we excluded patients that satisfied at least one of the conditions outlined below:

i) an ICU length of stay shorter than 6 hours,
ii) recorded measurements at fewer than 4 different hourly time points,
iii) a missing data window longer than 12 hours,
iv) the onset of sepsis outside of the ICU stay,
v) the onset of sepsis before 4 hours into ICU stay, or after 168 hours into ICU stay.

To detect (and penalise) delayed alarms, we include 24 hours of data after sepsis onset, which means that for sepsis cases we consider the first time window of up to 168 + 24 hours, or 7 days (plus 1 day buffer). Accordingly, for controls we include all data up to 168 + 24 hours into ICU stay.

**d) Human subject data**

The data in MIMIC-III has been previously de-identified, and the institutional review boards (IRBs) of the Massachusetts Institute of Technology (No. 0403000206) and BIDMC (2001-P-001699/14) approved the use of the database for research. The eICU database is released under the Health Insurance Portability and Accountability Act (HIPAA) safe harbour provision. The re-identification risk was certified as meeting safe harbour standards by Privacert (Cambridge, MA) (HIPAA Certification no. 1031219-2). The IRB of the Canton of Bern approved the HiRID data collection and its anonymised use for research. The AUMC dataset was certified according to NEN7510 (ISO 27001), which ensures that strict information governance protocols are in place for the protection of source data. Furthermore, according to the original source[4], two independent teams auditing the project reported that the design, database management and governance were state-of-the-art, and therefore the data was considered as anonymous information in the context of the General Data Protection Regulation.

**e) Machine learning methods**

In this study, we investigated a comprehensive selection of supervised ML approaches. This includes i) deep self-attention models (attn)[5] ii) recurrent neural networks employing gated recurrent units (gru)[6] iii) LightGBM gradient boosting trees (lgbm)[7], and iv) LASSO-regularised[8] logistic regression (lr). All models were tuned to minimise the binary cross-entropy loss of the prediction score with the binary sepsis label on the time-point level. For the non-deep classifiers, we ran a randomised hyperparameter search with 50 iterations of a stratified 5-fold cross-validation of the first training split. For the deep learning models depth, width, learning rate, batch size, and weight decay were selected based on a dedicated online validation split of the training data. The performance in terms of cross-entropy loss on this split was monitored during training and the model state where the performance was best was used for the final evaluation. The hyperparameter selection process was split into two stages: first, a search over a coarse grid (see Table S3 for the exact values) was performed by

randomly selecting 25 of the possible parameter combinations and evaluating the models on them. Afterwards a second finer grid around the previously selected parameters was constructed and further 25 randomly selected parameter combinations were evaluated. The model architecture (depth and width) was not changed in the second stage. All models were trained with a positive weight dependent on the class imbalance. Training was stopped after 100 epochs (i.e., iterations through the dataset) and the hyperparameter search was always performed on the training split of the first repetition of train-validation splitting. Finally, using the best hyperparameters, for each method 5 repetition models were fitted on 5 different repetitions of train-validation splitting of the development data. This strategy was applied to each database independently.

Next, we list more details about the used ML architectures, starting with the deep self-attention model (attn). First, each position (pos) (the relative time step measured as hours since ICU admission) is embedded with a 10-dimensional positional encoding (PE) which is computed as follows: $PE(pos, 2i) = \sin(pos \cdot s_i)$ and $PE(pos, 2i + 1) = \cos(pos \cdot s_i)$, where $i$ enumerates 5 different time scales that are being used from 0 to 4, and $s_i$ indicates the actual time scaling factors that are computed as: $s_i = a \cdot \exp\left(\frac{-\log\left(\frac{b}{a}\right) \cdot i}{S-1}\right)$, where $a$ and $b$ represent the minimal and maximal time scales, with $a = 1$ and $b = 500$, and $S = 5$ denoting the number of time scales that are employed. Next, the attn model consists of a first linear layer that maps the 59-dimensional input time series (concatenated with 10 dimensions of the positional encoding) to a model dimension d which is treated as a hyperparameter. Then, two Transformer layers are sequentially applied that each consist of the following components (that are also sequentially applied):

1. an 8-head multi-head attention layer, which is followed by
2. a residual connection that combines the attention output (after applying dropout) with the Transformer layer's input
3. a feedforward Multilayer perceptron that has a single hidden dimension of size $4 \cdot$ d and that uses a rectified linear unit activation, followed by
4. a final residual connection combining the outputs of step 2 and 3.

Causal masking is applied to the multi-head attention layers to prevent the leakage of future information. After the Transformer layers, a final linear layer projects the d-dimensional hidden representation to a one-dimensional output (the logits). For the recurrent neural network employing Gated Recurrent Units (gru), we provided the time series of 59 input variables as sequential input to this recurrent architecture. Static variables were used as initial state of the recurrent model via the projection of a linear layer. For gru, we allowed a network depth of up to 3 layers (Table S3). The LightGBM model was trained by randomly sampling configurations from the following choice of hyperparameters. Number of estimators: 100, 300, 500, 1000, 2000; Boosting type: gbdt, dart; learning rate: 0.001, 0.01, 0.1, 0.5; number of leaves: 30, 50, 100; L1-regularisation strength: 0, 0.1, 0.5, 1, 3, 5. The logistic regression model was optimized by allowing for two solvers: liblinear and saga. Additionally, L1-regularisation was employed whereas for the regularization strength parameter, the range between $10^{-3}$ and $10^{2}$ was partitioned into 50 parameter choices that were equally spaced (in log space).

**f) Feature engineering**

For the 59 temporal variables (vital signs and lab tests), hourly intervals were considered and the median measurement value was taken as representative value of each interval. If no measurement occurred in an interval, it would be treated as a missing value (see below in this section).

For the included non-deep classifiers, which were not originally designed for sequential data, we extracted 1,269 features based on multi-scale look-back statistics (mean, median, variance, minimum and maximum over the last 4, 8, and 16 hours), measurement counts, missingness indicators, the raw measurement values of the current time step, and hand-crafted features derived from domain knowledge (for instance shock index, oxygenation index, or available lab and vital components of baseline scores like SOFA, SIRS, or MEWS). For the deep models, which are capable of automatically learning feature representations from sequential data, we refrained from employing manual multi-scale look-back statistics, resulting in 190 features including 59 observed measurements, 59 missingness indicators, 59 measurement counts, 9 derived domain-knowledge features (ratios and partial scores) and 4 static variables representing demographic information. All features were standardised on the respective training split by computing the mean and standard deviation from all measurements of a given variable (across time). Missing values were accounted for with a missingness indicator feature while the missing measurement was imputed with the feature mean (i.e., a value of zero after standardisation). Validation samples were standardised using the standardisation moments calculated from the given training split, respectively.

During preliminary experiments, we observed that including the 4 static variables had a slight detrimental effect on performance of the attention model (not on the other models); we therefore discarded this information specifically for the attention model.

**g) Variable importance**

We measured the importance and relevance of individual features by calculating Shapley values[9] via the integrated gradients method[10] for our attention-based deep learning model. This method "explains" a prediction of a trained model by assessing the contribution of each feature. This enables us to explain and interpret the predictions of a model in a unified manner, which even accounts for the directionality of an effect. We focused on (patho-)physiological signals by restricting

ourselves to "raw" measurements, i.e., we discarded count variables, missingness indicator variables, and derived features. All computations were performed on the dataset-specific held-out test split (with the model having no access to such test data during training, as described in Section S2h). For this, we consistently inspected the model as fitted on the first repetition split of the development split. We sampled 500 patient stays at random from each dataset and repeated this procedure five times, as the implementation of integrated gradients precludes calculations on the full dataset due to excessive memory requirements. Since we were specifically interested in contributions ahead of a potential sepsis onset, we chose the time point with the maximum model prediction as an anchor for each patient stay.
For each patient, we then focused on how the model utilised such changes in the period of up to 16 hours preceding the time step at which the prediction score was maximal (making the resulting time windows at most 16 hours long) and calculated Shapley values for each of those time points. The advantage of this procedure is that it restricts our view to *recent* time points, making the Shapley values more pertinent for a specific prediction, as opposed to incorporating a potentially large number of time points, which may in turn result in de-emphasising a potential signal.

**h) Experimental setup**

All datasets were split into a development set (90%) and a held-out test set (10%). For 5 independent repetitions, the development set was randomly divided into a training set and validation set such that they accounted for 80% and 10% of the total data. All splits were stratified to preserve the dataset-specific sepsis prevalence. We optimised all ML models on the first partition (out of 5 repetitions) of the development set, whereas the validation split was used for fine-tuning hyperparameters. Next, each model was refitted on the respective train split of each of the five repetitions, which allowed us to assess model robustness with respect to varying training data. Finally, all performance measures are reported on the respective held-out test split. To maximise comparability between internal and external evaluations, in both settings, identical test splits are evaluated. For the sake of performance metrics being comparable across datasets, upon testing time we harmonised the prevalence of sepsis cases to the across-dataset average of 17%. This was achieved by randomly subsampling either controls for increasing the prevalence, or cases for decreasing the prevalence. To ensure that valuable case information is not lost during this procedure, we repeated all subsamplings 10 times and confirmed that the coverage of sepsis cases is above 98.3%. The across-dataset average of 17% was computed at a stage where an additional dataset from the Emory hospital was still included in our cohort. It was removed during the revision process as it limited the joint feature set of our overall cohort. Nevertheless, we kept the 17% prevalence average, that was computed on a larger cohort, more closely reflecting the ground truth. For fine-tuning, a given pre-trained model was fine-tuned on a small split of the testing database (of size 10%) which was independent from the testing split used for final evaluation.

**i) Evaluation and statistical analyses**

We formulated the problem of detecting sepsis as an online prediction task, where a model is continuously provided with new data to make the next, hourly prediction for a given patient. This setup was chosen in order to simulate a realistic deployment setting already during the (retrospective) model development phase. By contrast, to devise a clinically sensible model evaluation strategy, we report the following performance metrics on the *patient* level: i) area under the ROC curve (AUC) and ii) positive predictive value (PPV) and earliness for a fixed sensitivity of 80%. In both cases, we considered the set of all observed unnormalised prediction scores (logits) and divided them into 100 evenly spaced thresholds. For each threshold, we swept through the prediction scores of an ICU stay and triggered an alarm the first (and only the first) time the prediction score surpassed the current threshold. The set of raised alarms was then used to fill the cells of a confusion matrix, where a sepsis case with an alarm was considered true positive, a control ICU stay without an alarm a true negative, and so on. ROC curves were then computed from these confusion matrices over all thresholds. For the second measure, we determined the threshold with 80% recall (or sensitivity) and reported the precision (or positive predictive value) as obtained from the confusion matrix corresponding to this threshold, as well as alarm *earliness*, computed as the median number of hours the alarm preceded sepsis onset. In case there was no threshold with exactly 80% recall, we linearly interpolated all measures based on the two closest thresholds. This evaluation strategy is conservative in that no repeated alarms (that would improve recall at the cost of alarm fatigue) are permissible. While this makes recognising sepsis cases more challenging, it also guarantees that at most only a single false alarm can be raised in a control stay.
Means and standard deviations (SDs) of these performance measures are estimated by first averaging the measures over 10 subsamplings for each repetition split and by then computing mean and SD over the 5 repetitions of the training/validtion splits of the development part of each database. 95% confidence intervals are computed as percentile intervals by treating all of these 50 iterations as bootstrap samples.
Hypothesis testing was conducted using the Wilcoxon signed-rank test to assess whether the pooling of predictors performs significantly different (two-sided) to the best performing predictor as derived on the separate datasets. Our bootstraps derived from the subsampling and repetition splits allowed for a natural pairing of the AUC values between both settings (pooled predictors versus best pair-wise predictor), which was used for computing the Wilcoxon test. Next, we tested whether pooling predictors is inferior to pooling the datasets already during model derivation (which is more effortful as retraining on large datasets is necessary and sensitive patient data needs to be shared across sites) with a one-sided Wilcoxon signed-rank test. In both cases, for four core datasets, multiple testing was accounted for using Bonferroni correction, resulting in $\alpha = 0.0125$. Additionally, to assess inferiority of prediction pooling versus dataset pooling across all core

datasets, we also report the one-sided Wilcoxon signed-rank test on the AUC mean over datasets. We assessed model calibration of our attention model on each core dataset by means of reliability diagrams.
To calibrate our models, we used Platt scaling which was tuned on the respective validation split of the cohort, whereas all reliability diagrams are computed on the held-out test split of each cohort. To align with the evaluation of model discrimination, which was performed on the patient level (as opposed to a time point level), the calibration analysis was also performed on the patient level. To create the reliability diagram, for a given ICU admission, we used the models average predicted score and the overall label, whether or not this admission was a sepsis case. Since calibration is strongly connected with the specific prevalence of a cohort, we used the raw cohorts (without repeated subsampling for harmonizing the prevalences) and assessed the attention model that was trained on the given dataset.
Finally, we conducted an ablation analysis to assess the informativeness of individual feature categories as well as the model's performance on different patient cohorts (Fig. S10).
All computing code used for this study will be made publically available under
https://github.com/BorgwardtLab/multicenter-sepsis.

## Section S2: Additional analyses

**a) Comparison of dataset pooling and predictor pooling**

As compared to our pooling strategy where *predictors* are combined that were developed on the different databases, in this auxiliary analysis, we assessed the value of pooling already on the *data* level when performing the external validation across databases. For each of the four core datasets, one dataset is chosen as the external testing site, whereas the remaining datasets are merged for developing the model (using our deep learning architecture), respectively. Averaged over the datasets, with data pooling we observed an AUC of 0·755 (95% CI, 0·743 to 0·763). The corresponding ROC curves are visualised in Fig. S10. In contrast, with our strategy to pool predictors, where no additional retraining on large datasets was necessary, we observed an average AUC of 0·761 (95% CI, 0·747 to 0·770). On the dataset level, we observed significant AUC improvements of dataset pooling over predictor pooling for MIMIC-III and eICU ($P < 0·001$ on each dataset), albeit no improvements on HiRID and AUMC ($P = 0·99$ in both datasets). When averaging the AUCs over the four core datasets, we find no improvement when employing dataset pooling ($P = 0·99$).

**b) Model calibration**

On each core dataset we plotted reliability diagrams of our attention model before and after calibration using platt scaling (Figures S24 and S25). We observe that the predicted risk of the respective calibrated models aligns well with the true risk of sepsis.

**c) Temporal analysis**

We assess performance over the course of the ICU stay. For this, in Figure S15 we plot four quantities: (i) true positive rate over the preceding 6-hour period (TPR); (ii) false positive rate over the preceding 6-hour period (FPR); (iii) proportion of patients still to be diagnosed with sepsis (PTBD); (iv) the inverse positive likelihood ratio $(LR_+)^{-1}$ which puts the FPR and TPR in direction relation: $LR_+ = TPR / FPR$ and $(LR_+)^{-1} = FPR/TPR$. We note that the FPR and TPR are highly correlated, which indicates that there is likely no deterioration in the performance of the alarm system. However, the figure suggests that it may be worthwhile to consider alarm thresholds which are *time-dependent*. Such an approach could be used to stabilise the TPR and FPR values throughout the period of ICU stay.

**d) Antibiotics analysis**

We calculated the fraction of true positive alarms (only at most one alarm per patient is raised) that occurred before any kind of antibiotic was given. Pooled across all datasets, this fraction is 42.5% (SD 7.4 std). This means that almost half of the true positive alarms were raised before any antibiotic was given – which is likely actionable. However, the reverse conclusion, i.e., that the remaining TP alarms would not be actionable, does not necessarily hold: Many patients receive multiple drugs (polypharmacy) and may enter intensive care due to an ongoing focal infection (like e.g. pneumonia or UTI). Just because these patients will receive antibiotics early on to treat a potentially focal infection, does not imply that a) the given antibiotic is the right choice for treating a later fulminant sepsis or that b) the clinician is at all aware of an imminent sepsis. Hence, to fully understand the clinicians' intentions would require a detailed case-by-case qualitative analysis (planned ahead of data collection) that cannot be readily applied to our large-scale retrospective study. For more conclusive answers on utility and actionability, clinical trials will be indispensable.

## Section S3: Figures

**Figure S1: Study flow chart of filtering steps**

Patient filtering steps applied to the data sets yielding the final study cohort of 156,309 patients. Parenthesised numbers refer to the individual data sets MIMIC-III (M), eICU (E), HiRID (H), and AUMC (A), respectively.

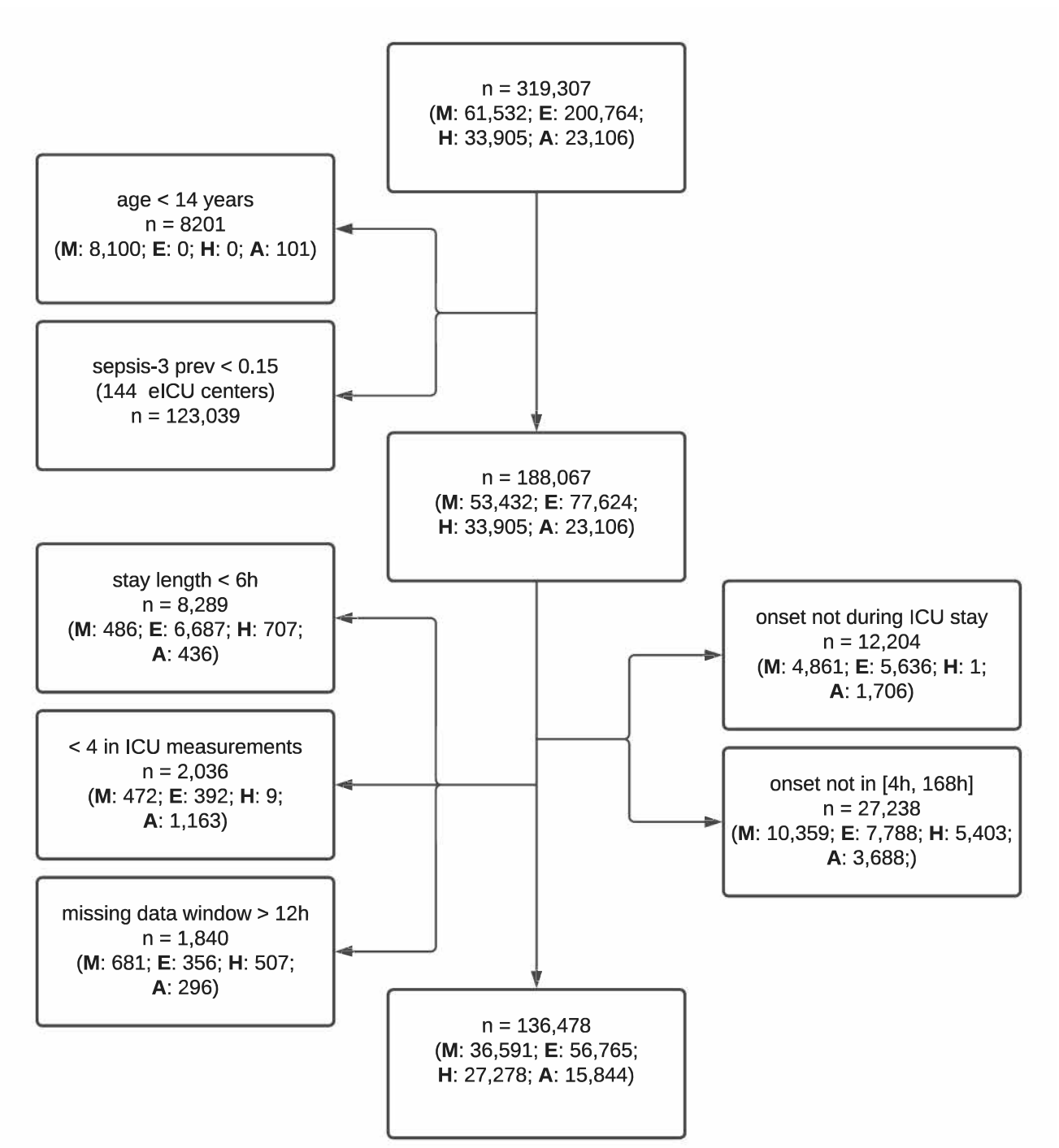


**Figure S2A: Data harmonisation**

Monitoring of distributions of measurements across all five datasets. These checks were applied to all variables to ensure that units were harmonised. Here, from left to right, a subset of variables is illustrated: Mean arterial pressure, temperature, creatinine and lactate.

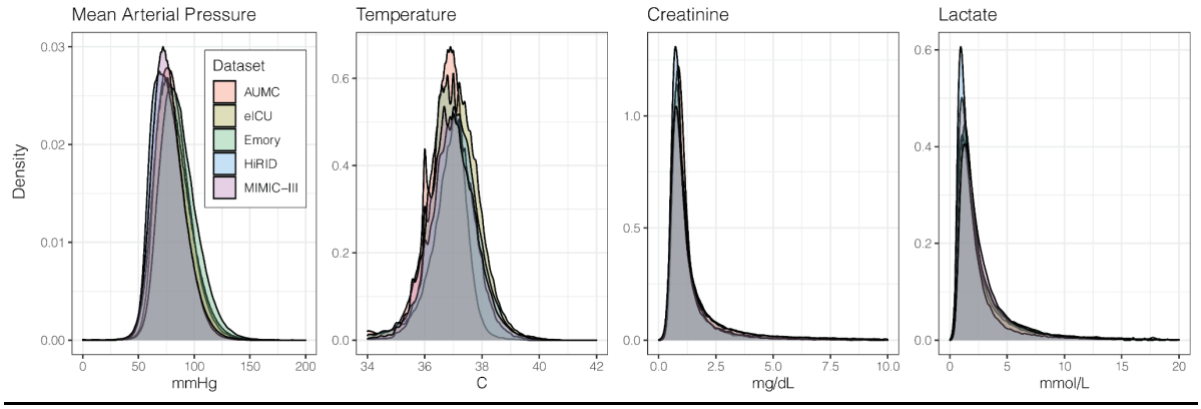



**Figure S2B: Sepsis-3 label**

Visual representation of how the onset time of sepsis was defined following Sepsis-3. The label is constructed by evaluating the SOFA score on hourly basis and determining suspicion of infection (SI) windows based on body fluid culture sampling and antibiotics administration. The criterion is fulfilled if (and at the time of) an acute increase in SOFA of at least 2 points coincides with an SI window

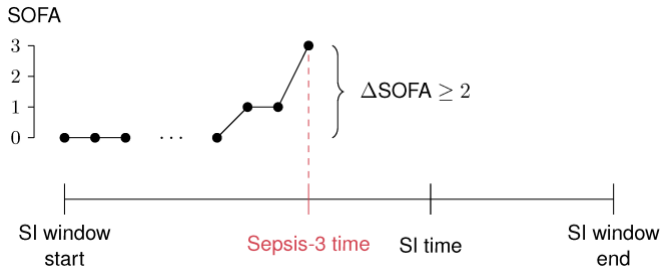



**Figure S3: Comparison of two alternative suspicion of infection tags**

Comparison of the original suspected infection (SI) definition (antibiotics + fluid sampling) against the alternative suspected infection definition based on multiple antibiotics (which was used when no culture sampling was available) displayed as Venn diagrams for MIMIC-III and AUMC, the two datasets that allowed for both implementations. In Panel A, the two variations of SI showed a Jaccard similarity of 0·69 for MIMIC-III. Panel B shows that in AUMC, the multiple antibiotics definition was broader than the original definition, but included most patients of the original definitions (Jaccard similarity 0·42). This is likely due to a foremost surgical cohort where antibiotics are frequently prescribed for prophylactic purposes. In Panel C, a comparison of the two definitions on the subset of non-surgical admissions in the AUMC database shows a very strong overlap (Jaccard similarity 0·78).

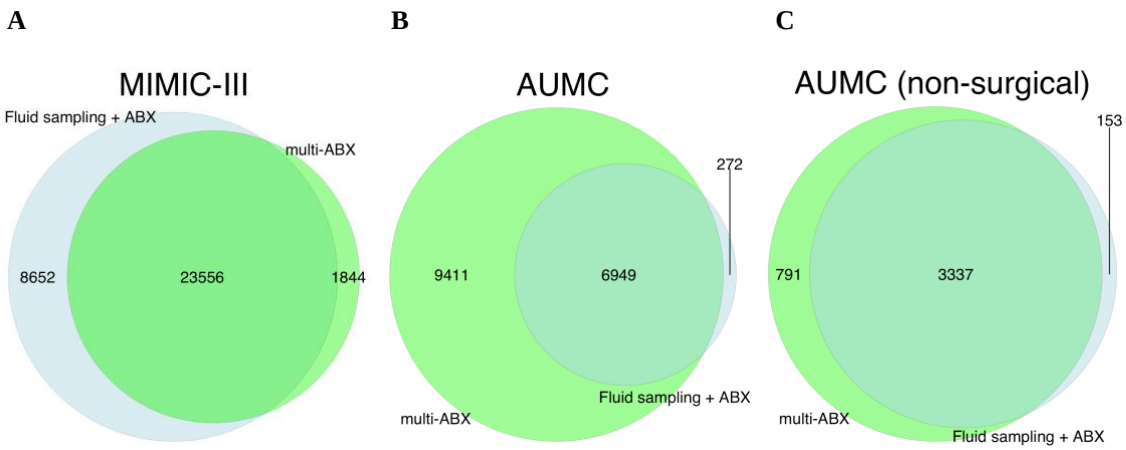


**Figure S4: Classification performances on the AUMC dataset.**
Predictive performance plots for the AUMC dataset and all methods. In Panel A (and Panel B), ROC curves of the internal (and external) validation performance are shown. In Panel C (and Panel D), PPV at 80% Sensitivity is plotted against median earliness for the internal (and external) validation. Our deep learning approach (attn) is visualised together with the comparison methods including clinical baselines (SOFA, MEWS, NEWS, qSOFA and SIRS), and machine learning methods: logistic regression model (lr), LightGBM (lgbm), and recurrent neural network employing Gated Recurrent Units (gru). Error bands (Panel A and B) and error bars (Panel C and D) indicate standard deviation of the reported metric over 5 repetitions of train-validation splitting.

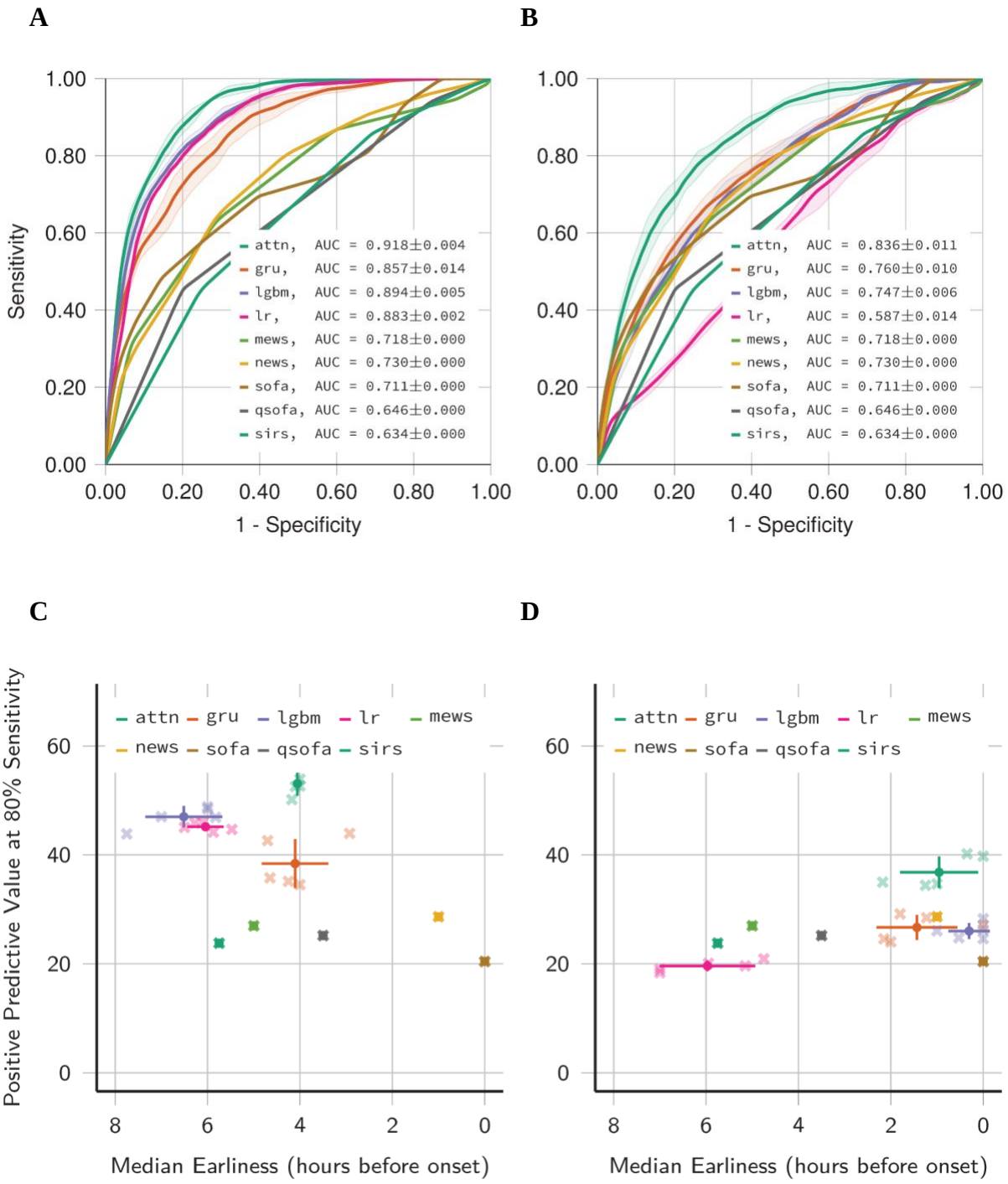


**Figure S5: Classification performances on the eICU dataset.**
Predictive performance plots for the eICU dataset and all methods. In Panel A (and Panel B), ROC curves of the internal (and external) validation performance are shown. In Panel C (and Panel D), PPV at 80% Sensitivity is plotted against median earliness for the internal (and external) validation. Our deep learning approach (attn) is visualised together with the comparison methods including clinical baselines (SOFA, MEWS, NEWS, qSOFA and SIRS), and machine learning methods: logistic regression model (lr), LightGBM (lgbm), and recurrent neural network employing Gated Recurrent Units (gru). Error bands (Panel A and B) and error bars (Panel C and D) indicate standard deviation deviation of the reported metric over 5 repetitions of train-validation splitting.

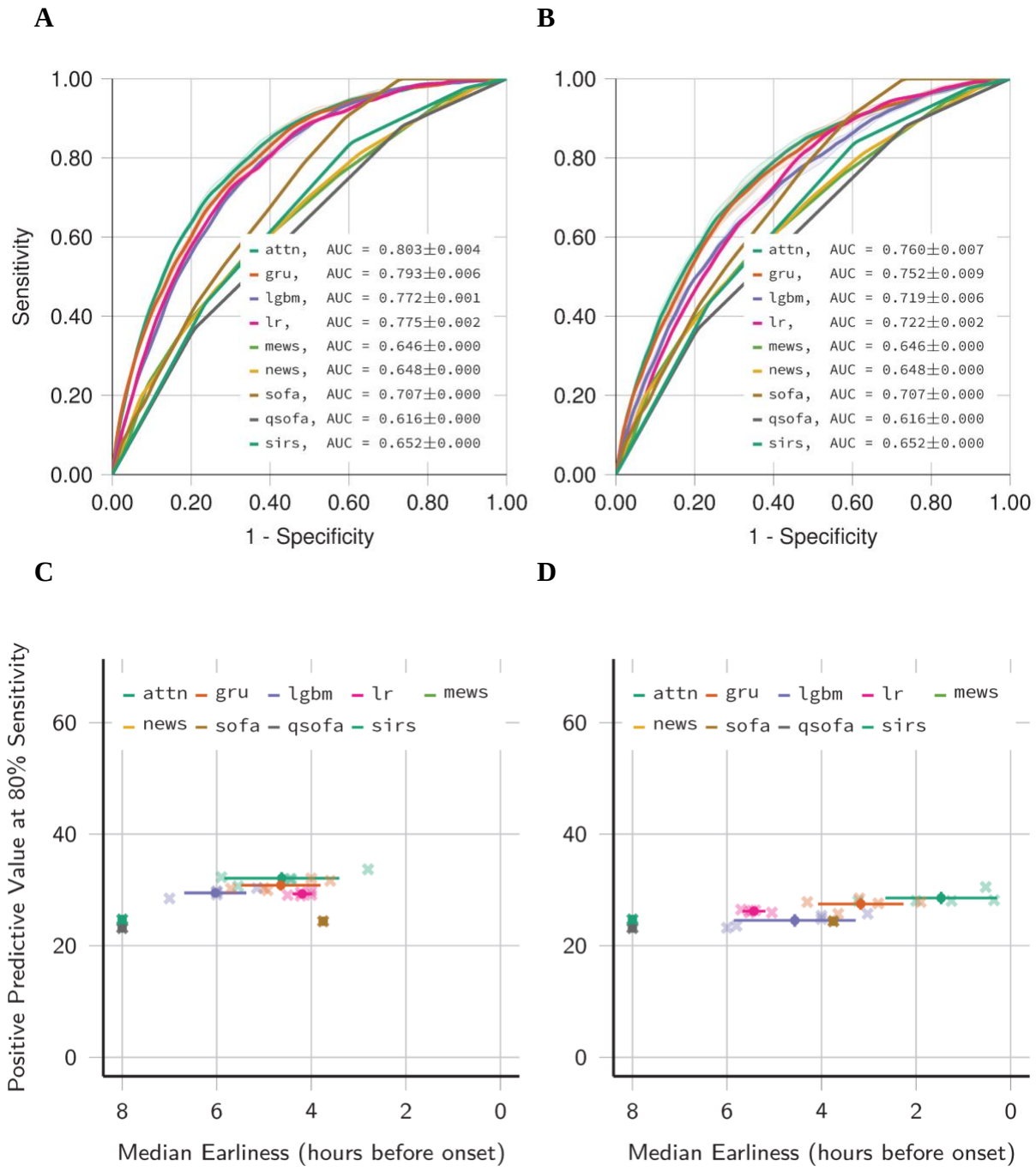


**Figure S6: Classification performances on the HiRID dataset.**
Predictive performance plots for the HiRID dataset and all methods. In Panel A (and Panel B), ROC curves of the internal (and external) validation performance are shown. In Panel C (and Panel D), PPV at 80% Sensitivity is plotted against median earliness for the internal (and external) validation. Our deep learning approach (attn) is visualised together with the comparison methods including clinical baselines (SOFA, MEWS, NEWS, qSOFA and SIRS), and machine learning methods: logistic regression model (lr), LightGBM (lgbm), and recurrent neural network employing Gated Recurrent Units (gru). Error bands (Panel A and B) and error bars (Panel C and D) indicate standard deviation deviation of the reported metric over 5 repetitions of train-validation splitting.

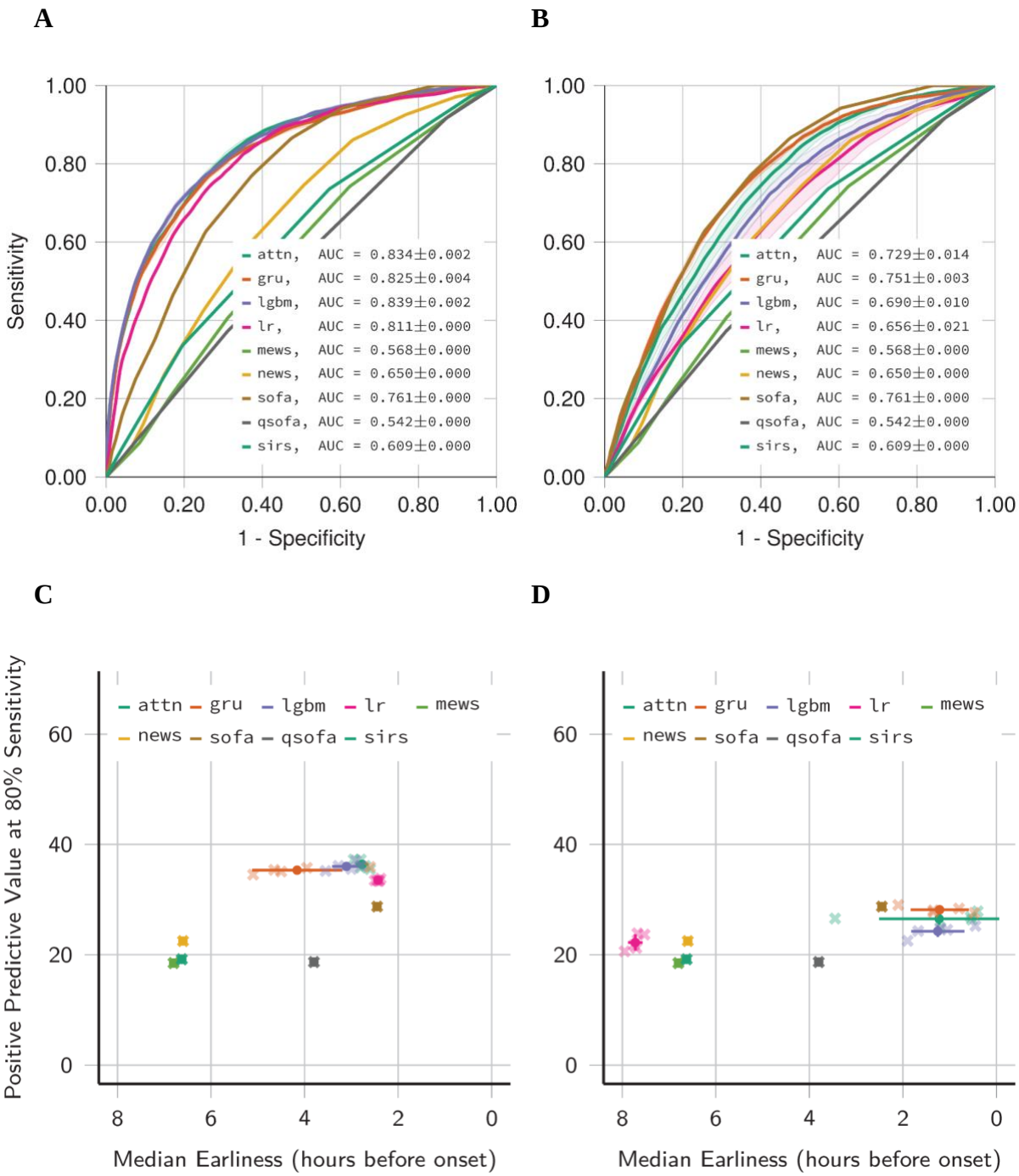


**Figure S7: Classification performances on the MIMIC-III dataset.**
Predictive performance plots for the MIMIC-III dataset and all methods. In Panel A (and Panel B), ROC curves of the internal (and external) validation performance are shown. In Panel C (and Panel D), PPV at 80% Sensitivity is plotted against median earliness for the internal (and external) validation. Our deep learning approach (attn) is visualised together with the comparison methods including clinical baselines (SOFA, MEWS, NEWS, qSOFA and SIRS), and machine learning methods: logistic regression model (lr), LightGBM (lgbm), and recurrent neural network employing Gated Recurrent Units (gru). Error bands (Panel A and B) and error bars (Panel C and D) indicate standard deviation deviation of the reported metric over 5 repetitions of train-validation splitting.

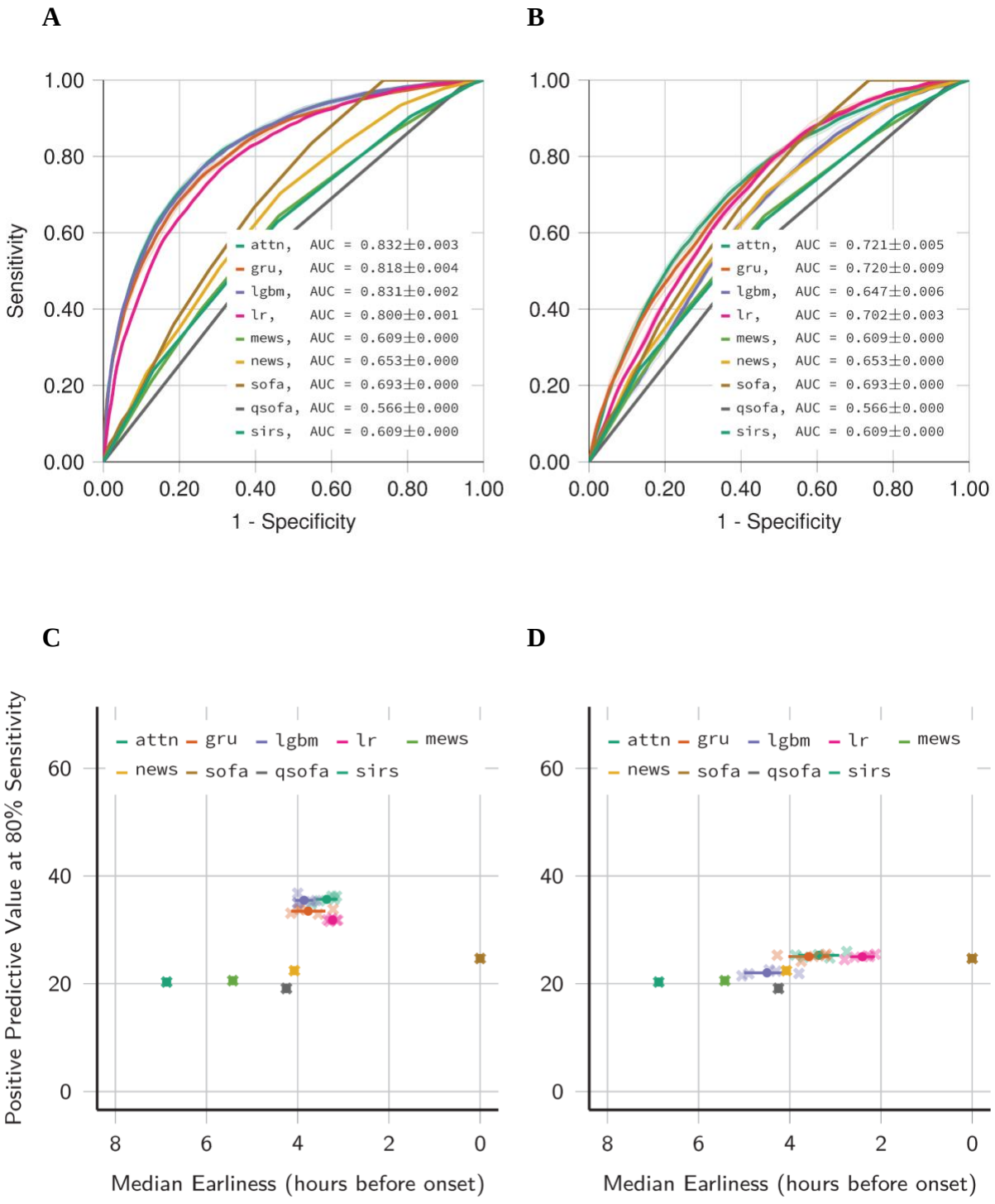
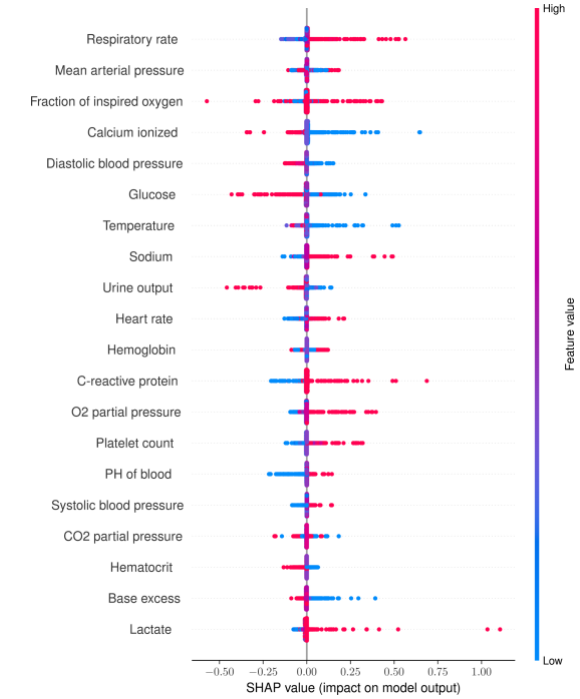


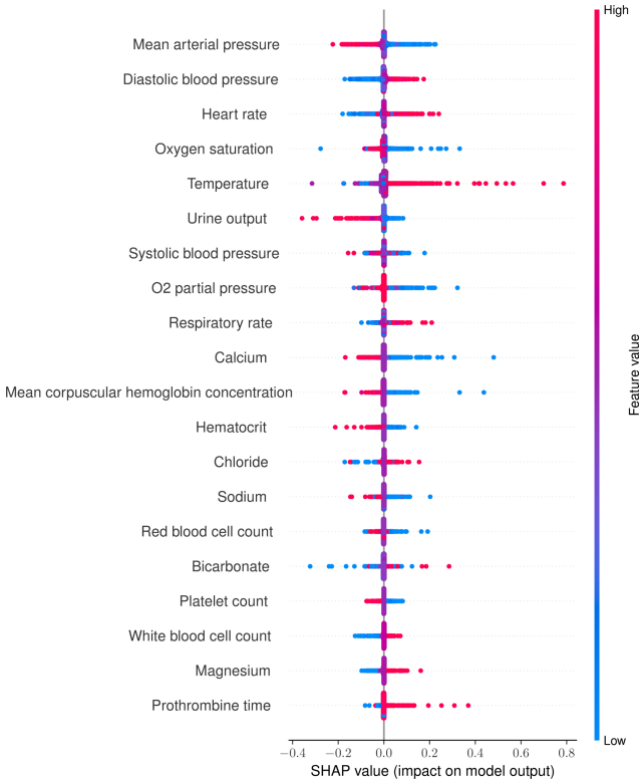
**Figure S8: Additional Shapley distributions of the raw measurements.**
Additional Shapley distributions are provided for all four core datasets: AUMC (Panel A), eICU (Panel B), MIMIC-III (Panel C) and HiRID (Panel D). The top 20 raw measurements are being shown in the visualisation. While the precise ranking differs between databases, mean arterial pressure and heart rate are consistently part of the top 10 features.

**A**




**B**


**C**




**D**

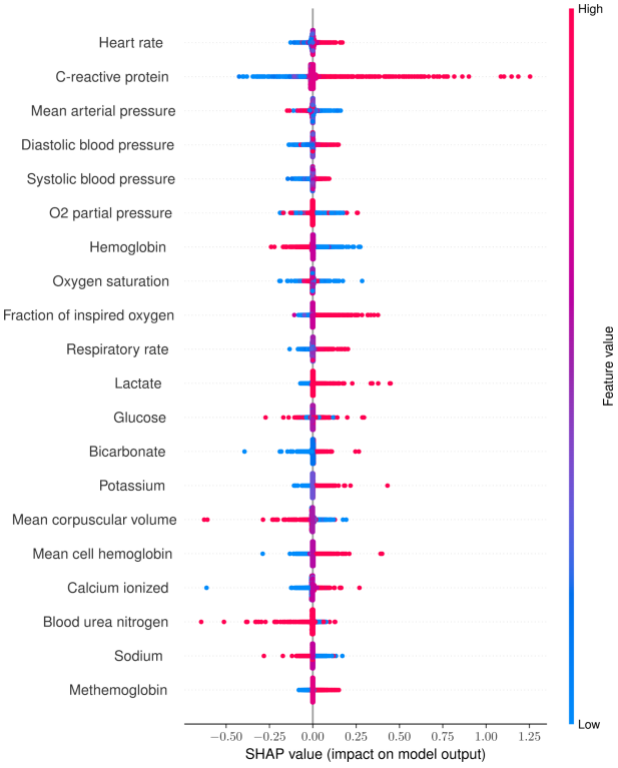
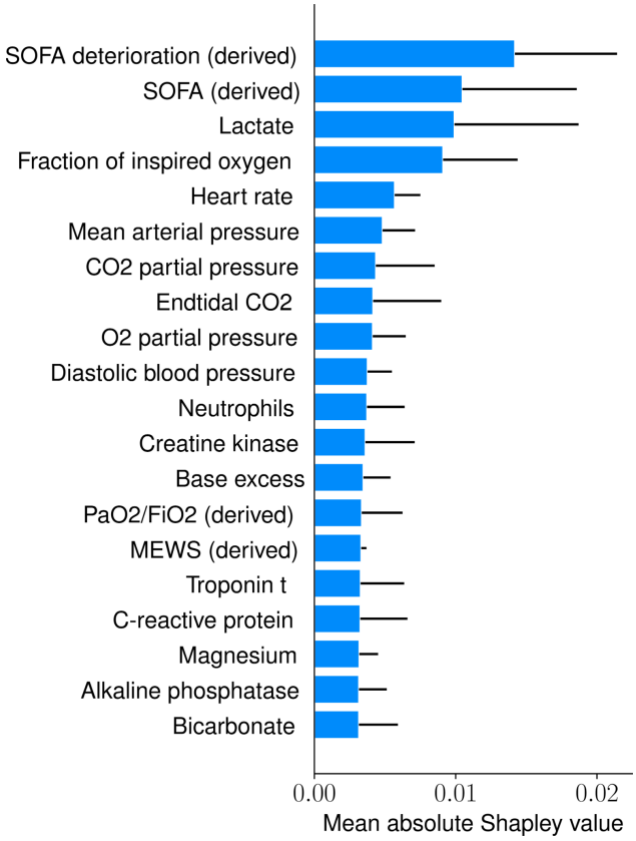


**Figure S9: Variable importances including non-physiological signals.**
We display a repetition of the Shapley value analysis presented in Fig. 4 that considers all features types used in the deep learning model, i.e., raw, count, indicator. Panel A shows the mean absolute Shapley (SHAP) values averaged over all datasets (error bars indicate standard deviation over datasets) and lists the top 20 variables for which the features were most informative in terms of increasing the predicted sepsis risk. Panel B illustrates the corresponding beeswarm plot of Shapley distributions for the eICU dataset.

**A**




**B**

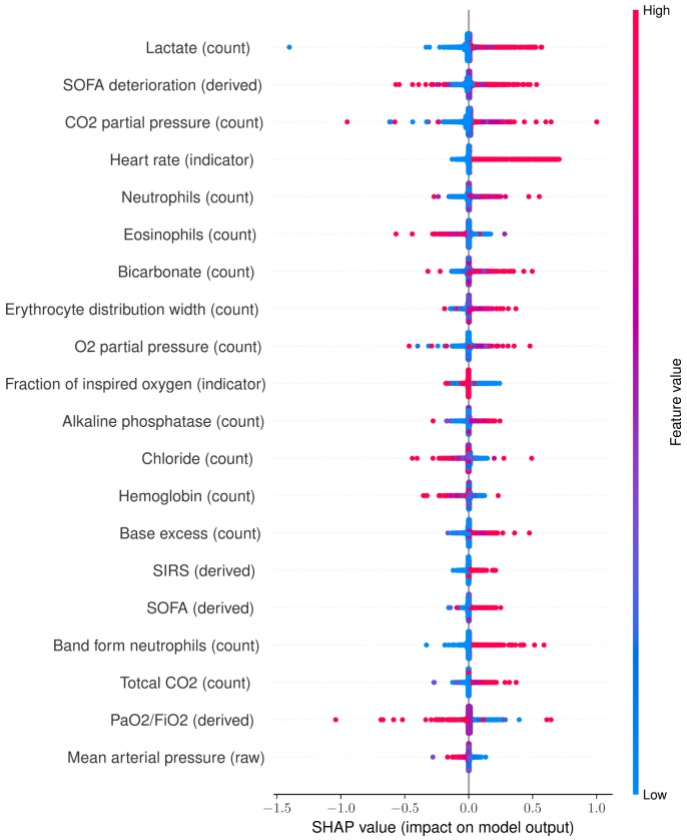


**Figure S10: Performance ablations.**
Ablations of cohorts and feature categories. We observed the best performance on AUMC, a dataset with a large proportion (80%) of surgical patients. Thus, in Panel A we applied a deep model that was trained on AUMC to the medical and surgical cohorts both in-distribution (AUMC) and out-of-distribution (MIMIC-III) and observed that the surgical cohort within AUMC is indeed easier to classify, whereas this does not necessarily generalise to other datasets.
Having observed that count features appeared frequently among the top 20 features explaining high prediction scores in Fig. S9, in Panel B we investigated how a model performed when trained solely on counts or raw measurements (here internally validated on MIMIC-III). We further subdivided the features into lab tests and vital signs and confirmed that lab tests carry sampling information (as a clinician requested them), whereas this was less the case for vital signs. The legend displays AUC means +/- SDs over the 5 repetition splits of the development split, respectively (variations due to subsampling are averaged out beforehand).

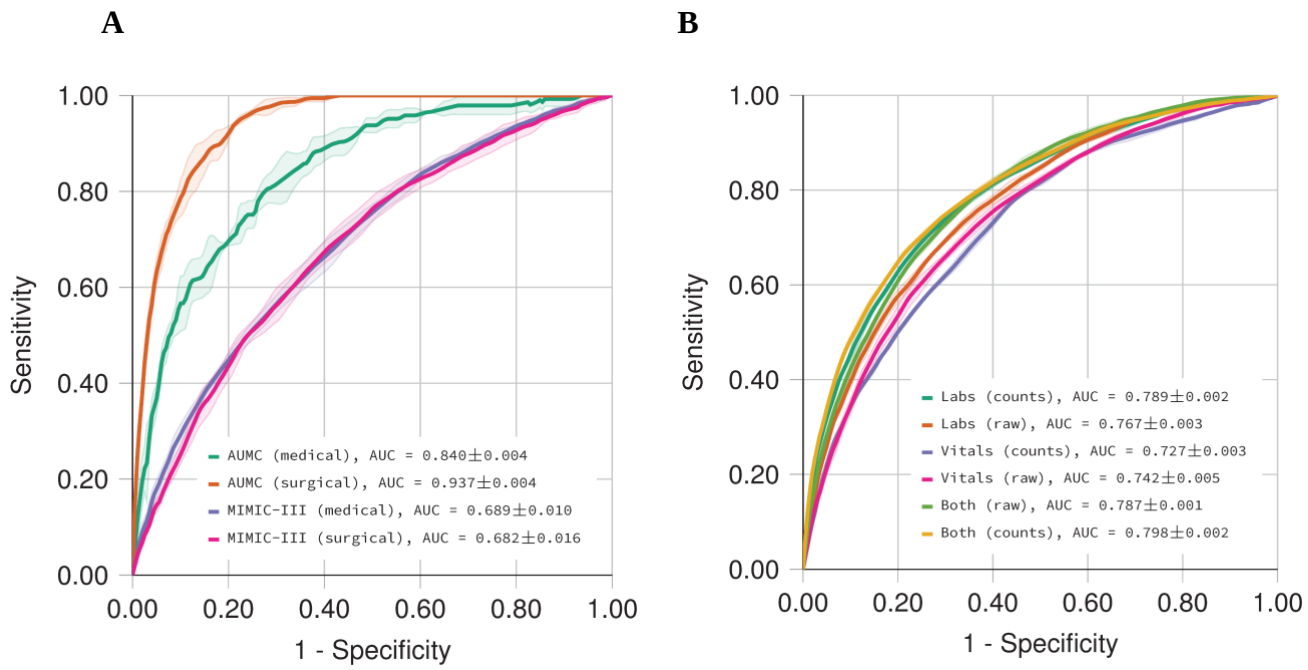


**Figure S11: Pooling datasets during training.**
In an auxiliary analysis of our deep model, we pooled datasets already during training, as opposed to pooling predictions of models trained on different datasets. This is computationally more costly as in each new setting (i.e., new dataset) the model needs to be retrained again, and this on increasingly bigger datasets. For each displayed dataset, the model was trained on a merged dataset comprising the remaining core datasets. Averaged over datasets, we observe an AUC of 0.754 (95% CI, 0.742 to 0.761). The legend displays AUC means +/- SDs over the 5 repetition splits of the development split, respectively (variations due to subsampling are averaged out beforehand).

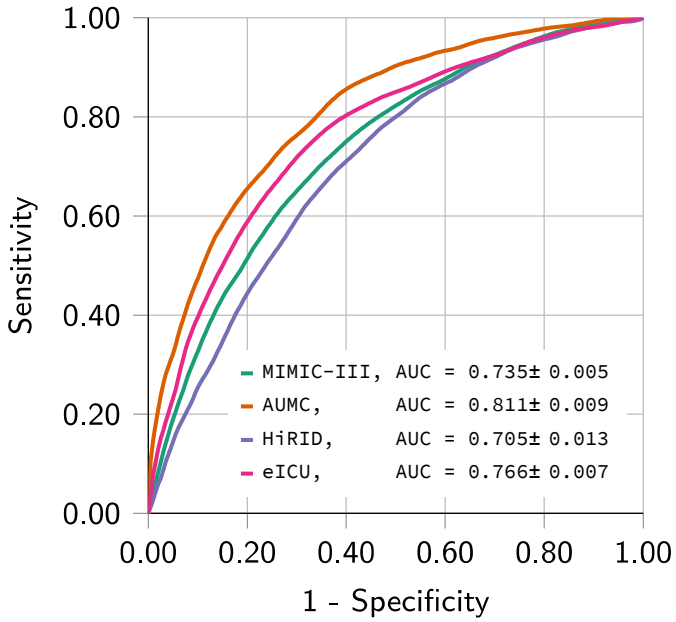



**Figure S12: Prediction task visualisation**
An illustration of our prediction task for a given sepsis case and control patient. Data is available for the whole ICU stay. We set the label (prediction target) of the case to be 1 starting from 6 hours before sepsis onset (red line) until 24 hours after sepsis onset. This allows for penalising late false negative predictions. If more than 24 hours after onset have passed, any further available data is not considered (grey region).

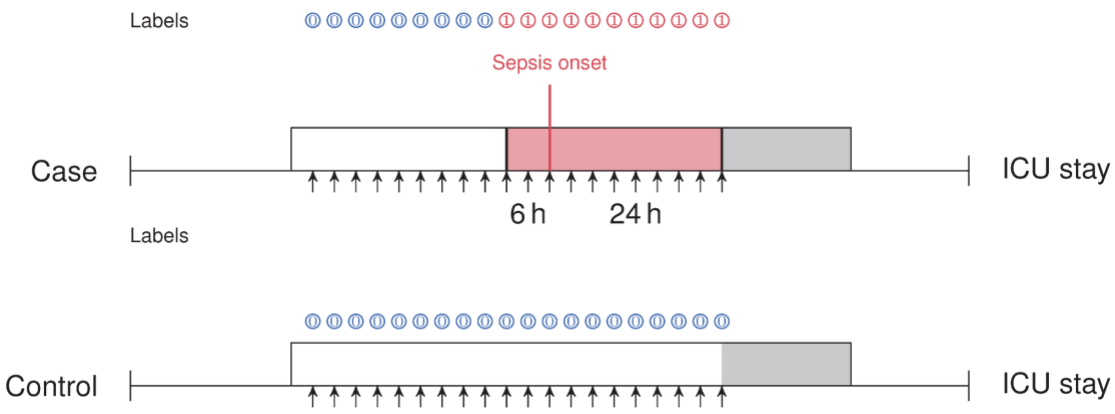


**Figure S13: Model reliability before calibration**

In Panel A, a reliability diagram is shown for the attention models that were trained on the four core datasets before applying calibration. First, 10 bins of patients are created such that they show an increasing mean predicted risk (ranging from 0 to 1) as averaged over the patients in each bin. Next, the true fraction of positives, i.e., of sepsis cases are shown on the y-axis. Here, we see that the uncalibrated models tend to be overconfident in that they predict higher risks as compared to the empirically observed sepsis risk. In Panel B, a histogram displays the patient counts in each bin for each dataset. Solid lines indicate the mean, error bands indicate the standard deviation over the five repetitions of the train/validation splitting.

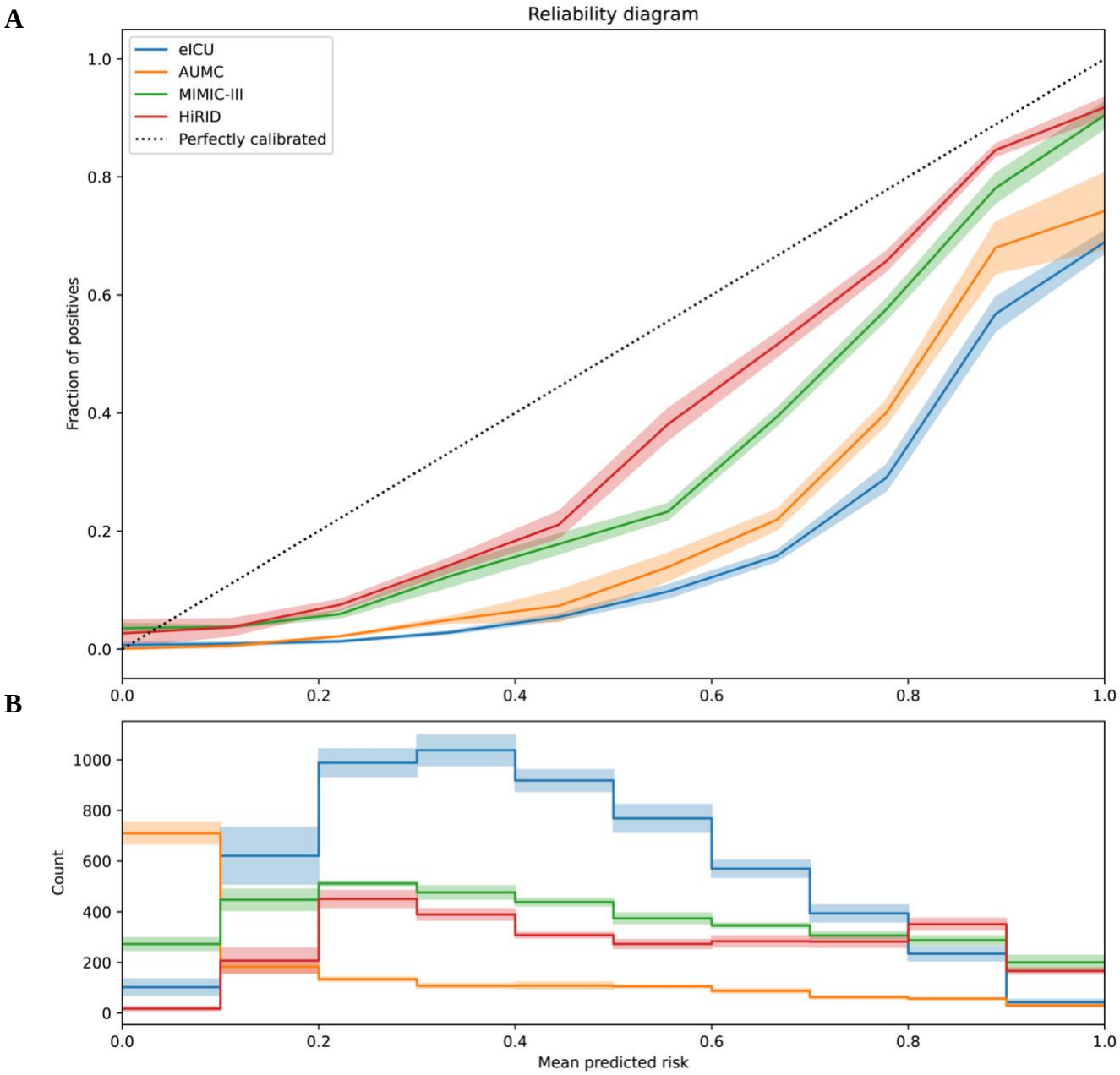


**Figure S14: Model reliability after calibration**

In Panel A, a reliability is shown for the attention models that were trained and calibrated using platt scaling on the four core datasets. First, 10 bins of patients are created such that they show an increasing mean predicted risk (ranging from 0 to 1) as averaged over the patients in each bin. Next, the true fraction of positives, i.e., of sepsis cases are shown on the y-axis. Here, we see that the predicted risk of the calibrated models aligns well with the empirical risk of sepsis. In Panel B, a histogram displays the patient counts in each bin for each dataset. Solid lines indicate the mean, error bands indicate the standard deviation over the five repetitions of the train/validation splitting.

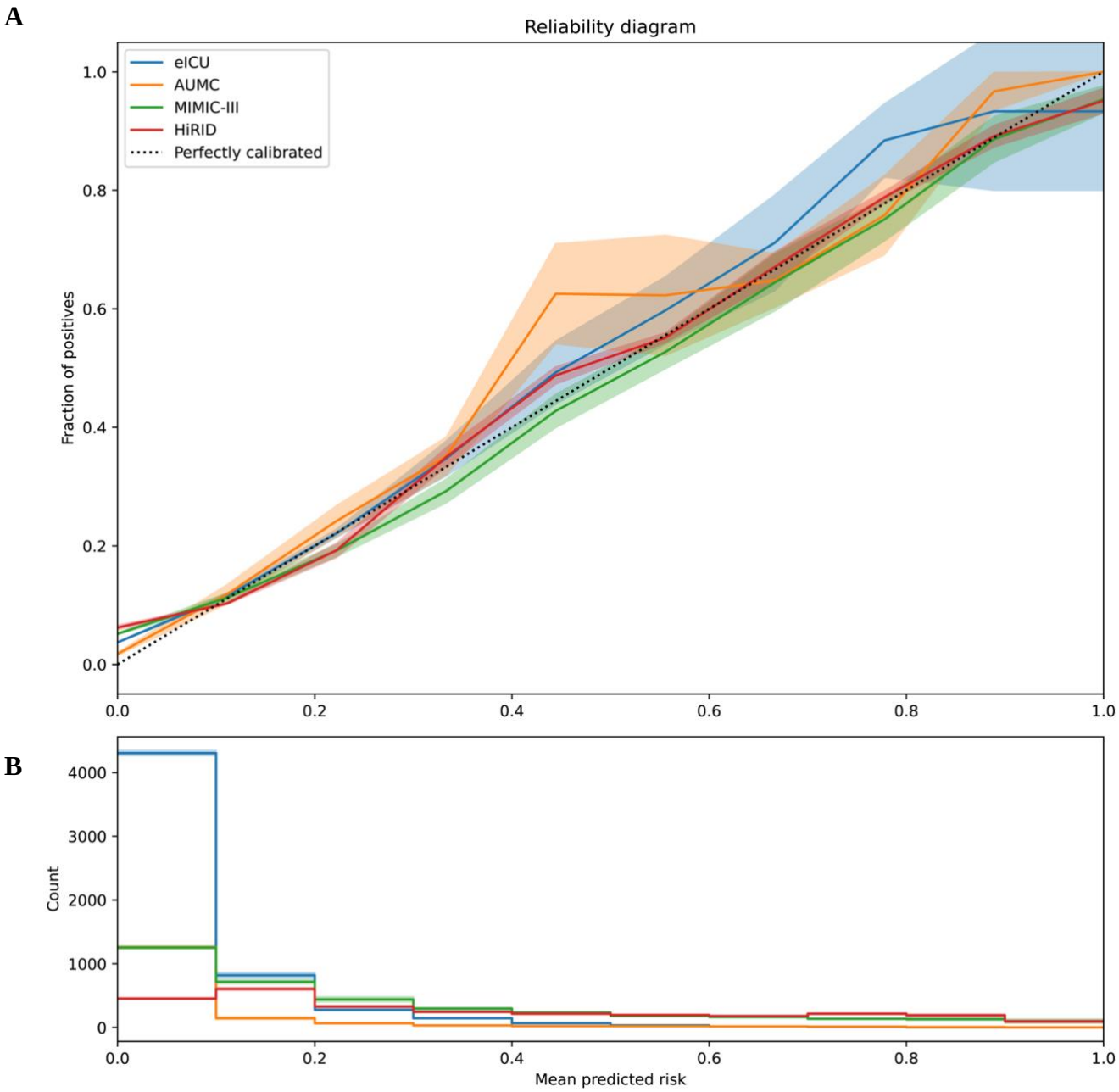
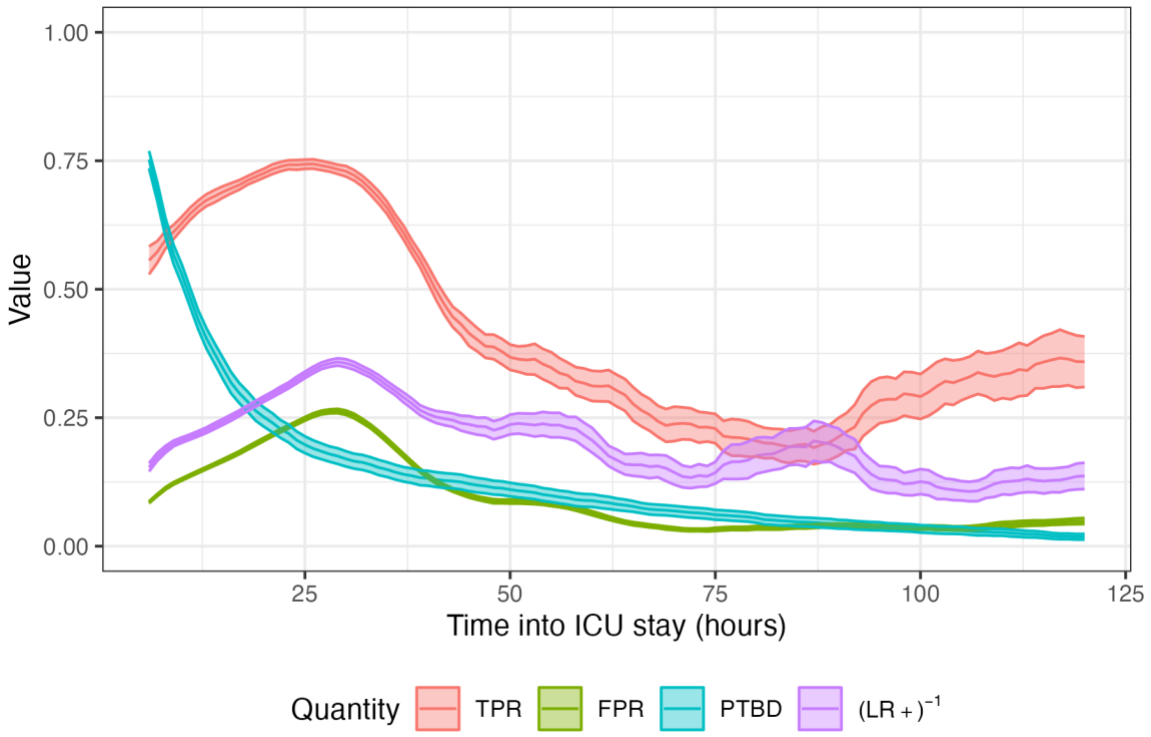


**Figure S15: Temporal analysis of performance**

Four quantities are shown over time into ICU stay: (i) true positive rate over the preceding 6-hour period (TPR); (ii) false positive rate over the preceding 6-hour period (FPR); (iii) proportion of patients still to be diagnosed with sepsis (PTBD) and (iv) the inverse positive likelihood ratio $(LR_+)^{-1}$ which puts the FPR and TPR in direction relation ($LR_+$ = TPR / FPR). We use the reciprocal value $(LR_+)^{-1}$ to fit the value into the same [0,1] range.



## Section S4: Tables

**Table S1A: Variables used for sepsis prediction**

Variables that were provided to the sepsis warning system. Availability per database is indicated.

| Name | Description | MIMIC-III | eICU | HiRID | AUMC |
|---|---|---|---|---|---|
| age | patient age | ✓ | ✓ | ✓ | ✓ |
| alb | albumin | ✓ | ✓ | ✓ | ✓ |
| alp | alkaline phosphatase | ✓ | ✓ | ✓ | ✓ |
| alt | alanine aminotransferase | ✓ | ✓ | ✓ | ✓ |
| ast | aspartate aminotransferase | ✓ | ✓ | ✓ | ✓ |
| basos | basophils | ✓ | ✓ | X | ✓ |
| be | base excess | ✓ | ✓ | ✓ | ✓ |
| bicar | bicarbonate | ✓ | ✓ | ✓ | ✓ |
| bili | total bilirubin | ✓ | ✓ | ✓ | ✓ |
| bili_dir | bilirubin direct | ✓ | ✓ | ✓ | ✓ |
| bnd | band form neutrophils | ✓ | ✓ | ✓ | ✓ |
| bun | blood urea nitrogen | ✓ | ✓ | ✓ | ✓ |
| ca | calcium | ✓ | ✓ | ✓ | ✓ |
| cai | calcium ionized | ✓ | ✓ | ✓ | ✓ |
| ck | creatine kinase | ✓ | ✓ | ✓ | ✓ |
| ckmb | creatine kinase MB | ✓ | ✓ | ✓ | ✓ |
| cl | chloride | ✓ | ✓ | ✓ | ✓ |
| crea | creatinine | ✓ | ✓ | ✓ | ✓ |
| crp | C-reactive protein | ✓ | ✓ | ✓ | ✓ |
| dbp | diastolic blood pressure | ✓ | ✓ | ✓ | ✓ |
| eos | eosinophils | ✓ | ✓ | X | ✓ |
| esr | erythrocyte sedimentation rate | ✓ | X | ✓ | ✓ |
| etco2 | endtidal CO2 | ✓ | X | ✓ | ✓ |
| fgn | fibrinogen | ✓ | ✓ | ✓ | ✓ |
| fio2 | fraction of inspired oxygen | ✓ | ✓ | ✓ | ✓ |
| glu | glucose | ✓ | ✓ | ✓ | ✓ |
| hbco | carboxyhemoglobin | X | ✓ | ✓ | ✓ |
| hct | hematocrit | ✓ | ✓ | X | ✓ |
| height | patient height | ✓ | ✓ | ✓ | ✓ |
| hgb | hemoglobin | ✓ | ✓ | ✓ | ✓ |
| hr | heart rate | ✓ | ✓ | ✓ | ✓ |
| inr_pt | prothrombin time / international normalized ratio | ✓ | ✓ | ✓ | ✓ |
| k | potassium | ✓ | ✓ | ✓ | ✓ |
| lact | lactate | ✓ | ✓ | ✓ | ✓ |
| lymph | lymphocytes | ✓ | ✓ | ✓ | ✓ |
| map | mean arterial pressure | ✓ | ✓ | ✓ | ✓ |
| mch | mean cell hemoglobin | ✓ | ✓ | ✓ | ✓ |
| mchc | mean corpuscular hemoglobin concentration | ✓ | ✓ | ✓ | ✓ |
| mcv | mean corpuscular volume | ✓ | ✓ | ✓ | ✓ |
| methb | methemoglobin | ✓ | ✓ | ✓ | ✓ |

**Table S1B: Variables used for sepsis prediction (continued)**
Variables that were provided to the sepsis warning system. Availability per database is indicated.

| Name | Description | MIMIC-III | eICU | HiRID | AUMC |
|---|---|---|---|---|---|
| mg | magnesium | ✓ | ✓ | ✓ | ✓ |
| na | sodium | ✓ | ✓ | ✓ | ✓ |
| neut | neutrophils | ✓ | ✓ | ✓ | ✓ |
| o2sat | oxygen saturation | ✓ | ✓ | ✓ | ✓ |
| pco2 | CO2 partial pressure | ✓ | ✓ | ✓ | ✓ |
| ph | pH of blood | ✓ | ✓ | ✓ | ✓ |
| phos | phosphate | ✓ | ✓ | ✓ | ✓ |
| plt | platelet count | ✓ | ✓ | ✓ | ✓ |
| po2 | O2 partial pressure | ✓ | ✓ | ✓ | ✓ |
| pt | prothrombine time | ✓ | ✓ | X | ✓ |
| ptt | partial thromboplastin time | ✓ | ✓ | ✓ | ✓ |
| rbc | red blood cell count | ✓ | ✓ | X | ✓ |
| rdw | erythrocyte distribution width | ✓ | ✓ | X | ✓ |
| resp | respiratory rate | ✓ | ✓ | ✓ | ✓ |
| sbp | systolic blood pressure | ✓ | ✓ | ✓ | ✓ |
| sex | patient sex | ✓ | ✓ | ✓ | ✓ |
| tco2 | totcal CO2 | ✓ | ✓ | X | X |
| temp | temperature | ✓ | ✓ | ✓ | ✓ |
| tnt | troponin t | ✓ | ✓ | ✓ | ✓ |
| tri | troponin I | ✓ | ✓ | X | X |
| urine | urine output | ✓ | ✓ | ✓ | ✓ |
| wbc | white blood cell count | ✓ | ✓ | ✓ | ✓ |
| weight | patient weight | ✓ | ✓ | ✓ | ✓ |

**Table S2A: Variables used for clinical baseline scores.**
Modified early warning score (MEWS)

| Name | Description | MIMIC-III | eICU | HiRID | AUMC |
|---|---|---|---|---|---|
| sbp | systolic blood pressure | ✓ | ✓ | ✓ | ✓ |
| hr | heart rate | ✓ | ✓ | ✓ | ✓ |
| resp | respiratory rate | ✓ | ✓ | ✓ | ✓ |
| temp | temperature | ✓ | ✓ | ✓ | ✓ |
| egcs | GCS eye | ✓ | ✓ | ✓ | ✓ |
| mgcs | GCS motor | ✓ | ✓ | ✓ | ✓ |
| vgcs | GCS verbal | ✓ | ✓ | ✓ | ✓ |
| tgcs | GCS total | ✓ | ✓ | X | X |
| trach | tracheostomy | ✓ | X | ✓ | ✓ |
| rass | Richmond agitation sedation scale | ✓ | ✓ | ✓ | ✓ |

**Table S2B: Variables used for clinical baseline scores**

National early warning score (NEWS)

| Name | Description | MIMIC-III | eICU | HiRID | AUMC |
|---|---|---|---|---|---|
| resp | respiratory rate | ✓ | ✓ | ✓ | ✓ |
| o2sat | oxygen saturation | ✓ | ✓ | ✓ | ✓ |
| vent_start | ventilation start | ✓ | ✓ | ✓ | ✓ |
| vent_end | ventilation end | ✓ | ✓ | ✓ | ✓ |
| fio2 | fraction of inspired oxygern | ✓ | ✓ | ✓ | ✓ |
| temp | temperature | ✓ | ✓ | ✓ | ✓ |
| sbp | systolic blood pressure | ✓ | ✓ | ✓ | ✓ |
| hr | heart rate | ✓ | ✓ | ✓ | ✓ |
| egcs | GCS eye | ✓ | ✓ | ✓ | ✓ |
| mgcs | GCS motor | ✓ | ✓ | ✓ | ✓ |
| vgcs | GCS verbal | ✓ | ✓ | ✓ | ✓ |
| tgcs | GCS total | ✓ | ✓ | X | X |
| trach | tracheostomy | ✓ | X | ✓ | ✓ |
| rass | Richmond agitation sedation scale | ✓ | ✓ | ✓ | ✓ |

**Table S2C: Variables used for clinical baseline scores**

Quick SOFA (qSOFA)

| Name | Description | MIMIC-III | eICU | HiRID | AUMC |
|---|---|---|---|---|---|
| egcs | GCS eye | ✓ | ✓ | ✓ | ✓ |
| mgcs | GCS motor | ✓ | ✓ | ✓ | ✓ |
| vgcs | GCS verbal | ✓ | ✓ | ✓ | ✓ |
| tgcs | GCS total | ✓ | ✓ | X | X |
| trach | tracheostomy | ✓ | X | ✓ | ✓ |
| rass | Richmond agitation sedation scale | ✓ | ✓ | ✓ | ✓ |
| sbp | systolic blood pressure | ✓ | ✓ | ✓ | ✓ |
| resp | respiratory rate | ✓ | ✓ | ✓ | ✓ |

**Table S2D: Variables used for clinical baseline scores**

Sequential organ failure assessment (SOFA). The SOFA subscores were not available in the data and needed to be computed from the raw data.

| Name | Description | MIMIC-III | eICU | HiRID | AUMC |
|---|---|---|---|---|---|
| sresp | SOFA respiratory component | - | - | - | - |
| scoag | SOFA coagulation component | - | - | - | - |
| sliver | SOFA liver component | - | - | - | - |
| scardio | SOFA cardiovascular component | - | - | - | - |
| scns | SOFA central nervous system component | - | - | - | - |
| srenal | SOFA renal component | - | - | - | - |

**Table S2E: Variables used for clinical baseline scores**

SOFA components

| Component | Name | Description | MIMIC-III | eICU | HiRID | AUMC |
|---|---|---|---|---|---|---|
| sresp | po2 | O2 partial pressure | ✓ | ✓ | ✓ | ✓ |
| | fio2 | fraction of inspired oxygen | ✓ | ✓ | ✓ | ✓ |
| | vent_start | | ✓ | ✓ | ✓ | ✓ |
| | vent_end | | ✓ | ✓ | ✓ | ✓ |
| scoag | plt | | ✓ | ✓ | ✓ | ✓ |
| sliver | bili | | ✓ | ✓ | ✓ | ✓ |
| scardio | map | | ✓ | ✓ | ✓ | ✓ |
| | dopa_rate | dopamine rate | ✓ | ✓ | X | ✓ |
| | dopa_dur | dopamine duration | ✓ | ✓ | X | ✓ |
| | norepi_rate | norepinephrine rate | ✓ | ✓ | ✓ | ✓ |
| | norepi_dur | norepinephrine duration | ✓ | ✓ | ✓ | ✓ |
| | dobu_rate | dobutamine rate | ✓ | ✓ | ✓ | ✓ |
| | dobu_dur | dobutamine duration | ✓ | ✓ | ✓ | ✓ |
| | epi_rate | epinephrine | ✓ | ✓ | ✓ | ✓ |
| | epi_dur | epinephrine duration | ✓ | ✓ | ✓ | ✓ |
| scns | egcs | GCS eye | ✓ | ✓ | ✓ | ✓ |
| | mgcs | GCS motor | ✓ | ✓ | ✓ | ✓ |
| | vgcs | GCS verbal | ✓ | ✓ | ✓ | ✓ |
| | tgcs | GCS total | ✓ | ✓ | X | X |
| | trach | tracheostomy | ✓ | X | ✓ | ✓ |
| | rass | Richmond agitation sedation scale | ✓ | ✓ | ✓ | ✓ |
| srenal | crea | creatinine | ✓ | ✓ | ✓ | ✓ |
| | urine | urine output | ✓ | ✓ | ✓ | ✓ |

**Table S2F: Variables used for clinical baseline scores**
Systemic inflammatory response syndrome (SIRS)

| Name | Description | MIMIC-III | eICU | HiRID | AUMC |
|---|---|---|---|---|---|
| temp | temperature | ✓ | ✓ | ✓ | ✓ |
| hr | heart rate | ✓ | ✓ | ✓ | ✓ |
| resp | respiratory rate | ✓ | ✓ | ✓ | ✓ |
| pco2 | CO2 partial pressure | ✓ | ✓ | ✓ | ✓ |
| wbc | white blood cell count | ✓ | ✓ | ✓ | ✓ |
| bnd | band form neutrophils | ✓ | ✓ | ✓ | ✓ |

**Table S3: Additional details on the hyperparameters of the deep learning models**
Additional details regarding the hyperparameter search of the investigated deep learning methods are provided: Our deep learning system, a deep self-attention model (attn), and recurrent neural networks using gated recurrent units (gru).

| Hyperparameter coarse search values | |
|---|---|
| Depth | gru: 1,2,3, attn: 2 |
| Width | 32, 64, 128, 256 |
| Learning rate | log-uniformly in range $e^{-9}$ to $e^{-7}$ |
| Dropout | 0·3, 0·4, 0·5, 0·6, 0·7 |
| Weight decay | 0·0001, 0·001, 0·01, 0·1 |

**Table S4: Patient characteristics**

Extended patient and demographic characteristics of our multi-center ICU cohort (including ethnicity and comorbidities).

| Variable | MIMIC-III | eICU | HiRID | AUMC |
|---|---|---|---|---|
| Cohort size (n) | 36,591 | 56,765 | 27,278 | 15,844 |
| Sepsis-3 prevalence (n (%)) | 9,541 (26) | 4,708 (8) | 10,170 (37) | 1,275 (8) |
| Age, years (Median (IQR)) | 65 (52-77) | 65 (53-76) | 65 (55-75) | 65 (55-75) |
| | | | | |
| Ethnicity (n (%)) | | | | |
| African American | 3,165 (9) | 5,573 (10) | - | - |
| Asian | 823 (2) | 655 (1) | - | - |
| Caucasian | 25,918 (71) | 46,216 (82) | - | - |
| Hispanic | 1,270 (3) | 952 (2) | - | - |
| Other | 5,415 (15) | 3,102 (5) | - | - |
| In-hospital mortality (n (%)) | 2,829 (8) | 3,962 (7) | 1,399 (5) | 745 (5) |
| ICU LOS, days (Median (IQR)) | 1·99 (1·15-3·63) | 1·71 (0·95-3·01) | 0·97 (0·8-1·95) | 0·97 (0·81-1·82) |
| Hospital LOS, days (Median (IQR)) | 6·43 (3·82-11·14) | 5·53 (2·99-9·89) | - | - |
| Gender, female (n (%)) | 15,944 (44) | 25,740 (45) | 9,977 (37) | 5,350 (35) |
| Gender, male (n (%)) | 20,647 (56) | 31,011 (55) | 17,301 (63) | 10,089 (65) |
| Ventilated patients (n (%)) | 16,499 (45) | 24,534 (43) | 14,021 (51) | 10,469 (66) |
| Patients on vasopressors (n (%)) | 9,669 (26) | 6,769 (12) | 7,721 (28) | 7,980 (50) |
| Patients on antibiotics (n (%)) | 21,598 (59) | 21,847 (38) | 17,152 (63) | 11,165 (70) |
| Patients with suspected infection (n (%)) | 16,349 (45) | 9,739 (17) | 15,160 (56) | 1,639 (10) |
| Initial SOFA (Median (IQR)) | 3 (1-4) | 3 (1-5) | 5 (3-8) | 6 (3-7) |
| | | | | |
| SOFA components (Median (IQR)) | | | | |
| Respiratory | 1 (0-2) | 1 (0-2) | 3 (2-4) | 2 (1-3) |
| Coagulation | 0 (0-1) | 0 (0-1) | 0 (0-1) | 0 (0-1) |
| Hepatic | 0 (0-1) | 0 (0-0) | 0 (0-1) | 0 (0-0) |
| Cardiovascular | 1 (1-1) | 1 (0-1) | 1 (1-4) | 2 (1-4) |
| CNS | 0 (0-1) | 0 (0-2) | 0 (0-1) | 0 (0-1) |
| Renal | 0 (0-1) | 0 (0-1) | 0 (0-0) | 0 (0-1) |
| | | | | |
| Admission type (n (%)) | | | | |
| Surgical | 13,836 (38) | 9,865 (19) | - | 11,905 (80) |
| Medical | 22,346 (61) | 41,674 (79) | - | 2,172 (15) |
| Other | 408 (1) | 1,346 (3) | - | 786 (5) |
| | | | | |
| Comorbidities ICD-9 (n (%)) | | | | |
| Chronic Renal Failure | 5,137 (14) | 3,950 (7) | - | - |
| Cancer | 4,687 (13) | 2,517 (4) | - | - |
| Chronic Pulmonary Disease | 7,994 (22) | 4,669 (8) | - | - |
| Liver Disease | 3,142 (9) | 1,507 (3) | - | - |
| Diabetes | 9,792 (27) | 1,919 (3) | - | - |

**Table S5: Variable units and valid ranges**

| Name | Description | Unit of meas. | Min. value | Max. value |
|---|---|---|---|---|
| hr | heart rate | bpm | 0 | 300 |
| o2sat | oxygen saturation | % | 50 | 100 |
| temp | temperature | C | 32 | 42 |
| sbp | systolic blood pressure | mmHg | 0 | 300 |
| map | mean arterial pressure | mmHg | 0 | 250 |
| dbp | diastolic blood pressure | mmHg | 0 | 200 |
| resp | respiratory rate | insp/min | 0 | 120 |
| etco2 | endtidal CO2 | mmHg | 10 | 60 |
| fio2 | fraction of inspired oxygen | % | 21 | 100 |
| be | base excess | mEq/L | -25 | 25 |
| bicar | bicarbonate | mEq/L | 5 | 50 |
| ph | pH of blood | - | 6.8 | 8 |
| pco2 | CO2 partial pressure | mmHg | 10 | 150 |
| cl | chloride | mEq/L | 80 | 130 |
| mg | magnesium | mg/dL | 0.5 | 5 |
| phos | phosphate | mg/dL | 0 | 40 |
| k | potassium | mEq/L | 0 | 10 |
| ast | aspartate aminotransferase | IU/L | 0 | - |
| bun | blood urea nitrogen | mg/dL | 0 | 200 |
| alp | alkaline phosphatase | IU/L | 0 | - |
| ca | calcium | mg/dL | 4 | 20 |
| crea | creatinine | mg/dL | 0 | 15 |
| bili_dir | bilirubin direct | mg/dL | 0 | 50 |
| glu | glucose | mg/dL | 0 | 1000 |
| lact | lactate | mmol/L | 0 | 50 |
| bili | total bilirubin | mg/dL | 0 | 100 |
| tri | troponin I | ng/mL | 0 | - |
| hct | hematocrit | % | 15 | 60 |
| hgb | hemoglobin | g/dL | 4 | 18 |
| ptt | partial thromboplastin time | sec | 0 | - |
| wbc | white blood cell count | K/uL | 0 | - |

| Name | Description | Unit of meas. | Min. value | Max. value |
|---|---|---|---|---|
| fgn | fibrinogen | mg/dL | 0 | 1500 |
| plt | platelet count | K/uL | 5 | 1200 |
| age | patient age | years | 0 | 100 |
| cai | calcium ionized | mmol/L | 0.5 | 2 |
| na | sodium | mEq/L | 110 | 165 |
| po2 | O2 partial pressure | mmHg | 40 | 600 |
| alb | albumin | g/dL | 0 | 6 |
| alt | alanine aminotransferase | IU/L | 0 | - |
| ck | creatine kinase | IU/L | 0 | - |
| ckmb | creatine kinase MB | ng/mL | 0 | - |
| crp | C-reactive protein | mg/L | 0 | - |
| tnt | troponin t | ng/mL | 0 | - |
| urine | urine output | mL | 0 | 2000 |
| basos | basophils | % | 0 | 50 |
| bnd | band form neutrophils | % | - | - |
| eos | eosinophils | % | 0 | 50 |
| esr | erythrocyte sedimentation rate | mm/hr | 0 | 200 |
| hbco | carboxyhemoglobin | - | - | - |
| inr_pt | prothrombin time/international normalized ratio | - | - | - |
| lymph | lymphocytes | % | 0 | 100 |
| mch | mean cell hemoglobin | pg | 0 | - |
| mchc | mean corpuscular hemoglobin concentration | % | 20 | 50 |
| mcv | mean corpuscular volume | fL | 50 | 150 |
| methb | methemoglobin | % | 0 | 100 |
| neut | neutrophils | % | 0 | 100 |
| pt | prothrombine time | sec | 0 | - |
| rbc | red blood cell count | m/uL | 0 | 20 |
| rdw | erythrocyte distribution width | % | 0 | 100 |
| tco2 | totcal CO2 | mEq/L | 5 | 60 |
| weight | patient weight | kg | 1 | 500 |
| height | patient height | cm | 10 | 230 |

**Table S6: Patient characteristics (development set)**

| Variable | MIMIC-III | eICU | HiRID | AUMC |
|---|---|---|---|---|
| Cohort size (n) | 32931 | 51088 | 24550 | 14259 |
| Sepsis-3 prevalence (n (%)) | 9076 (27.56) | 4243 (8.31) | 8354 (34.03) | 1215 (8.52) |
| Age, years (Median (IQR)) | 64.7 (51.53-77.13) | 65 (53-76) | 65 (55-75) | 65 (55-75) |
| In-hospital mortality (%) | 8 | 7 | 5 | 5 |
| ICU LOS, days (Median (IQR)) | 1.98 (1.15-3.59) | 1.71 (0.95-3.01) | 0.97 (0.8-1.95) | 0.97 (0.81-1.84) |
| Hospital LOS, days (Median (IQR)) | 6.4 (3.81-11.09) | 5.5 (2.98-9.86) | - | - |
| Gender, female (%) | 44 | 45 | 36 | 35 |
| Gender, male (%) | 56 | 55 | 64 | 65 |
| Ventilated patients (n (%)) | 14825 (45) | 22056 (43) | 13658 (56) | 9600 (67) |
| Patients on vasopressors (n (%)) | 8710 (26) | 6087 (12) | 6941 (28) | 7193 (50) |
| Patients on antibiotics (n (%)) | 16471 (50) | 19675 (39) | 15464 (63) | 10043 (70) |
| Patients with suspected infection (n (%)) | 12751 (39) | 8762 (17) | 13678 (56) | 1472 (10) |
| Initial SOFA (Median (IQR)) | 3 (2-5) | 3 (1-6) | 5 (3-8) | 6 (3-8) |
| Respiratory | 1 (0-2) | 1 (0-2) | 3 (1-4) | 2 (1-3) |
| Coagulation | 0 (0-1) | 0 (0-1) | 0 (0-1) | 0 (0-1) |
| Hepatic | 0 (0-1) | 0 (0-0) | 0 (0-1) | 0 (0-0) |
| Cardiovascular | 1 (1-1) | 1 (0-1) | 1 (1-4) | 2 (1-4) |
| CNS | 0 (0-1) | 0 (0-2) | 1 (0-1) | 0 (0-1) |
| Renal | 0 (0-1) | 0 (0-1) | 0 (0-0) | 0 (0-1) |
| Admission Diagnosis | | | | |
| Other | 1 | 3 | - | 5 |

| Variable | MIMIC-III | eICU | HiRID | AUMC |
|---|---|---|---|---|
| Surgical | 38 | 19 | - | 80 |
| Medical | 61 | 79 | - | 15 |
| Hispanic (%) | 3 | 2 | - | - |
| African American (%) | 9 | 10 | - | - |
| Asian (%) | 2 | 1 | - | - |
| Caucasian (%) | 71 | 82 | - | - |
| Other (%) | 15 | 5 | - | - |
| Comorbidities (ICD-9) | | | | |
| Liver Disease | 2828 (9) | 1348 (3) | - | - |
| Cancer | 4231 (13) | 2268 (4) | - | - |
| Diabetes | 8805 (27) | 1745 (3) | - | - |
| Chronic Renal Failure | 4617 (14) | 3573 (7) | - | - |
| Chronic Pulmonary Disease | 7241 (22) | 4222 (8) | - | - |

**Table S7: Patient characteristics (test set)**

| Variable | MIMIC-III | eICU | HiRID | AUMC |
|---|---|---|---|---|
| Cohort size (n) | 3660 | 5677 | 2728 | 1585 |
| Sepsis-3 prevalence (n (%)) | 1008 (27.54) | 471 (8.3) | 914 (33.5) | 129 (8.14) |
| Age, years (Median (IQR)) | 64.78 (51.8-76.89) | 65 (53-76) | 65 (55-75) | 65 (55-75) |
| In-hospital mortality (%) | 8 | 7 | 5 | 5 |
| ICU LOS, days (Median (IQR)) | 2.04 (1.17-3.91) | 1.71 (0.95-3.05) | 0.97 (0.8-1.93) | 0.97 (0.81-1.63) |
| Hospital LOS, days (Median (IQR)) | 6.75 (3.88-11.56) | 5.71 (3.01-10.15) | - | - |
| Gender, female (%) | 44 | 46 | 38 | 35 |
| Gender, male (%) | 56 | 54 | 62 | 65 |
| Ventilated patients (n (%)) | 1674 (46) | 2478 (44) | 1520 (56) | 1107 (70) |
| Patients on vasopressors (n (%)) | 959 (26) | 682 (12) | 780 (29) | 787 (50) |
| Patients on antibiotics (n (%)) | 1856 (51) | 2172 (38) | 1688 (62) | 1122 (71) |
| Patients with suspected infection (n (%)) | 1406 (38) | 977 (17) | 1482 (54) | 167 (11) |
| Initial SOFA (Median (IQR)) | 3 (2-5) | 3 (1-5) | 5 (3-8) | 6 (3-8) |
| Respiratory | 1 (0-2) | 1 (0-2) | 3 (2-4) | 2 (1-3) |
| Coagulation | 0 (0-1) | 0 (0-1) | 0 (0-1) | 1 (0-1) |
| Hepatic | 0 (0-1) | 0 (0-0) | 0 (0-1) | 0 (0-0) |
| Cardiovascular | 1 (1-1) | 1 (0-1) | 1 (1-4) | 2 (1-3) |
| CNS | 0 (0-1) | 0 (0-2) | 0 (0-1) | 0 (0-1) |
| Renal | 0 (0-1) | 0 (0-1) | 0 (0-0) | 0 (0-1) |
| Admission Diagnosis | | | | |
| Other | 1 | 2 | - | 6 |

| Variable | MIMIC-III | eICU | HiRID | AUMC |
|---|---|---|---|---|
| Surgical | 38 | 19 | - | 81 |
| Medical | 61 | 79 | - | 14 |
| Hispanic (%) | 4 | 2 | - | - |
| African American (%) | 8 | 10 | - | - |
| Asian (%) | 2 | 1 | - | - |
| Caucasian (%) | 70 | 81 | - | - |
| Other (%) | 16 | 6 | - | - |
| Comorbidities (ICD-9) | | | | |
| Liver Disease | 314 (9) | 159 (3) | - | - |
| Cancer | 456 (12) | 249 (4) | - | - |
| Diabetes | 987 (27) | 174 (3) | - | - |
| Chronic Renal Failure | 520 (14) | 377 (7) | - | - |
| Chronic Pulmonary Disease | 753 (21) | 447 (8) | - | - |

## References


1. Singer M, Deutschman CS, Seymour CW, et al. The Third International Consensus Definitions for Sepsis and Septic Shock (Sepsis-3). JAMA 2016;315(8):801–10.

2. Reyna M, Shashikumar SP, Moody B, et al. Early Prediction of Sepsis from Clinical Data: The PhysioNet/Computing in Cardiology Challenge 2019 [Internet]. 2019 Computing in Cardiology Conference (CinC). 2019;Available from: http://dx.doi.org/10.22489/cinc.2019.412

3. Seymour CW, Liu VX, Iwashyna TJ, et al. Assessment of Clinical Criteria for Sepsis: For the Third International Consensus Definitions for Sepsis and Septic Shock (Sepsis-3). JAMA 2016;315(8):762–74.

4. Thoral PJ, Peppink JM, Driessen RH, et al. Sharing ICU patient data responsibly under the Society of Critical Care Medicine/European Society of Intensive Care Medicine Joint Data Science Collaboration: The Amsterdam University Medical Centers Database (AmsterdamUMCdb) example. Crit Care Med 2021;49(6):e563–77.

5. Vaswani A, Shazeer N, Parmar N, et al. Attention is All you Need. In: Guyon I, Luxburg UV, Bengio S, et al., editors. Advances in Neural Information Processing Systems 30. Curran Associates, Inc.; 2017. p. 5998–6008.

6. Cho K, van Merrienboer B, Bahdanau D, Bengio Y. On the Properties of Neural Machine Translation: Encoder-Decoder Approaches [Internet]. arXiv [cs.CL]. 2014;Available from: http://arxiv.org/abs/1409.1259

7. Ke G, Meng Q, Finley T, et al. LightGBM: A Highly Efficient Gradient Boosting Decision Tree. In: Guyon I, Luxburg UV, Bengio S, et al., editors. Advances in Neural Information Processing Systems. Curran Associates, Inc.; 2017.

8. Tibshirani. R. Regression shrinkage and selection via the Lasso. J R Stat Soc Series B Stat Methodol 1996;58(1):267–88.

9. Lundberg SM, Lee S-I. A unified approach to interpreting model predictions. In: Proceedings of the 31st international conference on neural information processing systems. 2017. p. 4768–77.

10. Sundararajan M, Taly A, Yan Q. Axiomatic Attribution for Deep Networks. In: Precup D, Teh YW, editors. Proceedings of the 34th International Conference on Machine Learning. PMLR; 2017. p. 3319–28.