Artificial Intelligence In Medicine 104 (2020) 101820

Contents lists available at ScienceDirect

Artificial Intelligence In Medicine

journal homepage: www.elsevier.com/locate/artmed

Early detection of sepsis utilizing deep learning on electronic health record event sequences

Simon Meyer Lauritsen[a,b,*], Mads Ellersgaard Kalør[a], Emil Lund Kongsgaard[a], Katrine Meyer Lauritsen[b,e], Marianne Johansson Jørgensen[d], Jeppe Lange[b,d], Bo Thiesson[a,c]

[a] Enversion A/S, Fiskerivej 12, 8000 Aarhus C, Denmark
[b] Department of Clinical Medicine, Aarhus University, Denmark
[c] Department of Engineering, Aarhus University School of Engineering, Denmark
[d] Department of Research, Horsens Regional Hospital, Denmark
[e] Department of Endocrinology and Internal Medicine, Aarhus University Hospital, Aarhus, Denmark

ARTICLE INFO

Keywords:
Sepsis
Clinical decision support systems
Machine learning
Medical informatics
Early diagnosis
Electronic health records

ABSTRACT

Background: The timeliness of detection of a sepsis incidence in progress is a crucial factor in the outcome for the patient. Machine learning models built from data in electronic health records can be used as an effective tool for improving this timeliness, but so far, the potential for clinical implementations has been largely limited to studies in intensive care units. This study will employ a richer data set that will expand the applicability of these models beyond intensive care units. Furthermore, we will circumvent several important limitations that have been found in the literature: (1) Model evaluations neglect the clinical consequences of a decision to start, or not start, an intervention for sepsis. (2) Models are evaluated shortly before sepsis onset without considering interventions already initiated. (3) Machine learning models are built on a restricted set of clinical parameters, which are not necessarily measured in all departments. (4) Model performance is limited by current knowledge of sepsis, as feature interactions and time dependencies are hard-coded into the model.

Methods: In this study, we present a model to overcome these shortcomings using a deep learning approach on a diverse multicenter data set. We used retrospective data from multiple Danish hospitals over a seven-year period. Our sepsis detection system is constructed as a combination of a convolutional neural network and a long short-term memory network. We assess model quality by standard concepts of accuracy as well as clinical usefulness, and we suggest a retrospective assessment of interventions by looking at intravenous antibiotics and blood cultures preceding the prediction time.

Results: Results show performance ranging from AUROC 0.856 (3 h before sepsis onset) to AUROC 0.756 (24 h before sepsis onset). Evaluating the clinical utility of the model, we find that a large proportion of septic patients did not receive antibiotic treatment or blood culture at the time of the sepsis prediction, and the model could, therefore, facilitate such interventions at an earlier point in time.

Conclusion: We present a deep learning system for early detection of sepsis that can learn characteristics of the key factors and interactions from the raw event sequence data itself, without relying on a labor-intensive feature extraction work. Our system outperforms baseline models, such as gradient boosting, which rely on specific data elements and therefore suffer from many missing values in our dataset.

1. Introduction

Sepsis is one of the most common causes of death globally [1]. The World Health Organization estimates that more than six million people die of sepsis annually, and many of these deaths are preventable [2]. In the United States, severe sepsis affects more than 700,000 patients each year at a cost of more than 20 billion dollars [3,4]. Early detection of sepsis has shown to improve patient outcomes, but it remains a challenging problem in medicine [5]. Even experienced physicians have difficulties in detecting sepsis early and accurately, as the early symptoms associated with sepsis may also be caused by many other clinical conditions [6].

Previous studies have shown that machine learning (ML) models trained from data in individual patient electronic health records (EHR)

* Corresponding author at: Enversion A/S, Fiskerivej 12, 8000 Aarhus C, Denmark.
E-mail address: sla@enversion.dk (S.M. Lauritsen).

https://doi.org/10.1016/j.artmed.2020.101820
Received 2 May 2019; Received in revised form 9 January 2020; Accepted 17 February 2020

may be used for the early detection of sepsis [7–12]. The ML models for sepsis detection far exceed the predictive ability of existing clinical early warning system scores, such as the National Early Warning Score (NEWS) [7,9,10,12–14]. Recently, Shimabukuro et al. demonstrated several positive effects with the use of an ML model for sepsis detection in a randomized trial, reporting an in-hospital mortality decrease of 12.4 percentage points ($p = 0.018$) and an average length of stay decrease from 13.0 to 10.3 days ($p = 0.042$) [12].

However, the current studies have limitations. First, most studies build their ML models on a limited set of clinical parameters, such as vital signs, which therefore must be collected in the given clinical setting before predictions from a model can become of any real use. Emergency departments and intensive care units (ICU) often have guidelines for frequent registration of vital signs, but this is typically not the case in many medical and surgical departments. The applicability and deployment potential of the models is therefore limited due to the comprehensive data registration requirements that are imposed on the departments in which the models are to be used. Second, during model evaluation, it is customary to report only receiver operating characteristic (ROC) curves and the derived area under the ROC curve (AUROC). This type of evaluation is chosen in spite of claims that AUROC is purely a measure of predictive ability and does not measure expected clinical usefulness, as it does not take prevalence into account [15–17]. AUROC may be misleading when applied to data sets with a high imbalance between positive and negative samples, which is often the case within the field of health science [18]. Additionally, most studies are evaluated by ROC curves at a fixed time relative to the time of sepsis onset. In a real clinical setting, the evaluation should start at the time the patient arrives at the hospital, and the algorithm should be used for inference multiple times thereafter. Finally, the clinical utility of the models is typically not investigated in relation to potential interventions. As an example, it is not reported whether sepsis treatment has, in fact, already been initiated at the time of the early detection.

In this paper, (1) we present a scalable deep learning [19] approach for early sepsis detection on the heterogeneous data set that includes hospitalizations both within and outside of the ICUs from multiple medical centers; (2) we stress the importance of clinical utility and contrast it to simple concepts of accuracy; (3) we suggest a sequence evaluation approach that provides realistic estimations of model performance; (4) we evaluate the clinical utility of the model in relation to early interventions with blood cultures and antibiotics.

2. Materials and methods

2.1. Data population and data sources

Table 1 lists the type of data sources that were used in the study. The data included health data on all citizens 18 years or older with residency in one of four Danish municipalities (Odder, Hedensted, Skanderborg, and Horsens). We used data from the secondary health sector in combination with nationwide registers for the period of 2010 to 2017.

The data from the secondary health sector contained information from the EHR, including biochemistry, medicine, microbiology, medical imaging, and the patient administration system (PAS). The data constituted raw health events as they are reported into the EHR by the healthcare professionals. Importantly, this data is of a time-stamped, sequential nature that reflects the point in time that the healthcare professionals record each event during the admission of a patient. An illustrative representation of the raw sequential data for a random patient is shown in Fig. 1. The different data sources are color coded so that events from the same data source have the same color. For example, registrations in regard to medications are colored blue, whereas blood test results are red. The size of a circle indicates how many events have been observed with the same timestamp and scales with the square root of the number of concurring events.

The EHR data was combined with data from the National Patient Register [20] and the Civil Registration System [21] containing information of a more contextual type, such as previously registered diagnoses, procedures, hospital admissions, marital status, and housing situation. These data were used to additionally include contextual covariates with information about comorbidities, age, and marital status, which had been registered preceding the current admission.

The data were extracted from the research project "CROSSTRACKS".[1]

2.2. Inclusion criteria and dataset preparation

The flowchart in Fig. 2 illustrates the stepwise construction of the data set. First, all relevant hospital contacts were identified from a set of 1,002,450 contacts. From this set, 776,219 outpatient contacts were removed, leaving 226,231 inpatient contacts to be considered. Second, 11,262 admissions with a duration of less than 3 h were removed. Finally, admissions to hospital departments with an overall prevalence of sepsis less than 2% (162.740 contacts) were removed, leaving 52,229 contacts in total in the data set. The threshold of 2% is arbitrary but reflects a trade-off between the coverage of the prediction model and the noise of errorneous predictions. With a threshold of 2%, the included departments cover 96% of all sepsis contacts in our data set. Note that the registered events – or non-registration hereof – do not in any way impose restrictions on the inclusion of a patient contact to the data set. Neither will later manipulations (see Section 2.4) of the raw EHR events affect the inclusion. For that reason, we denote this data set as the *full data* to contrast it with a second data set, called the *vital sign data*.

The vital sign data set was constructed by removing 49,103 contacts with incomplete vital sign measurements (systolic blood pressure, diastolic blood pressure, heart rate, respiratory rate, peripheral capillary oxygen saturation, and temperature) in the 3 h preceding the *label time*. For the positive sequences, the label time corresponded to the time the positive label was obtained (sepsis onset). For the negative sequences, there was no obvious label time, so we selected a pseudo-random time during the admission, excluding the first and last 3 h of the admission.

The data set was as a result of this reduced to a subset of the full data consisting of 3126 contacts in total (Fig. 2). The vital sign data was constructed to enable a direct comparison with successful models from the literature. However, note that due to the sparsity of the vital sign measurements, the clinical usefulness of models constructed from this data is severely limited compared to full data models.

For both data sets, the hospital contacts that led to sepsis were identified (see Section 2.3), and the data were split into two parts: one with sepsis-positive contacts and one with sepsis-negative contacts. Sepsis-positive contacts were further divided into training data (80%), validation data (10%), and test data (10%) and combined with sepsis-negative contacts of similar proportions. The training data were used to fit the model parameters. The validation data were used to perform an unbiased evaluation of a model fit during training, and the test data were used to provide an unbiased evaluation of the final model fit on the training data. In the training data, the sepsis-positive contacts were oversampled by a factor of ten. This entire process was repeated five times to enable five-fold cross-validation.

For each sequence, we considered at most five days of data prior to the label time. Fig. 1 shows an example with an observation window in green and the corresponding prediction window in red. The transition between the two windows is marked by the prediction time, which is not static but rather slides from the beginning of each sequence to the label time, along with the progression of the hospitalization for a patient. In this way, both windows are changing size as the prediction

[1] http://www.tvaerspor.dk/.

Table 1
Data sources.

Source system	Data type
Electronic health record (patient administration system)	Diagnoses (international classification disease – 10; ICD-10), procedures (NCSP – the NOMESCO Classification of Surgical Procedures), booking information, health content (structured notes containing physiological measurements, symptom classifications, check box data such as smoking and exercise habits)
Electronic health record (medication module)	Dates and times for prescriptions and dispensing together with information on ingredients, dose, administration routes.
Laboratory system	Microbiology and blood gas analysis
Medical imaging system	Image descriptions from computed tomography, magnetic resonance imaging, ultrasound, X-ray, positron-emission tomography
National patient register	Hospital admissions, diagnoses (ICD-10), procedures (NCSP)
Civil registration system	Patient demographics: age, address, and marital status

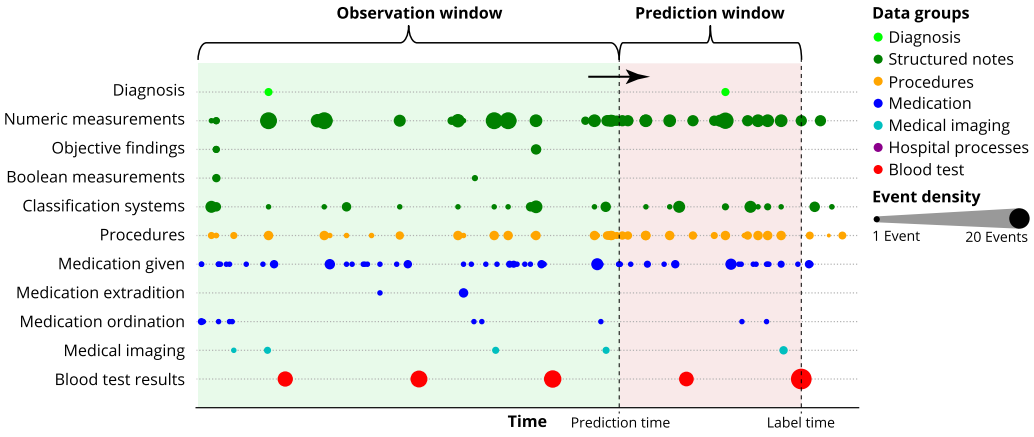

Fig. 1. A visual example of data for a randomly chosen sepsis patient. The observation window has a green background while the prediction window has a red background. The transition between the two windows is called the prediction time. The prediction time is not static, as displayed in this snapshot; instead, it is shifting from the beginning of each sequence to the label time, as the hospitalization progresses. For the sepsis-positive sequences, the label time corresponded to the time of sepsis onset. For the negative sequences, the label time was randomly chosen within the admission. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

time shifts forward in real-time.

2.3. Target definition for the early detection of sepsis

After the inclusion of hospital admissions, each admission underwent a binary classification process to denote it as either *sepsis-positive* or *sepsis-negative*. The concrete individual registration of sepsis used in our dataset occurs when an internal medicine physician based on the SIRS criteria and with a sufficient clinical suspicion of systemic infection with or without proven bacteriaemia actively register this as a sepsis event in the EHR. The classification defines whether, or not, a patient meets the gold standard for sepsis, which is based on the 2001 consensus definition of sepsis [22]. That is, the presence of two or more Systemic Inflammatory Response Syndrome (SIRS) criteria paired with a suspicion of infection. The SIRS criteria are defined as:

- heart rate >90 beats/min
- body temperature $>38^{\circ}$C or $<36^{\circ}$C
- respiratory rate >20 breaths/min or PaCO2 (alveolar carbon dioxide tension) <32 mm Hg
- white cell count $>12 \times 109$ cells/L or $<4 \times 109$ cells/L

Importantly, the fulfillment of the 2001 consensus sepsis definition is an independent EHR registration that occurs continually during admission and is reported by the treating physician. This registration is in contrast to the registration of diagnoses, which often relates to the time of discharge. Hence, in this study, the gold standard may be fulfilled and registered even though vital sign measurements have not yet been entered into the EHR. This is in great contrast to other sepsis studies that needs to estimate the sepsis onset time retrospectively. Based on this unique EHR registration in our dataset, we chose to conduct this study with the 2001 gold standard despite the introduction of the new Sepsis-3 definition introduced in 2016 by Singer et al. [23].

2.4. Data representation

In the raw data, each sample represents a given patient as a time-ordered sequence of EHR events $E = (e_1, e_2, ..., e_T)$, where e_t is an observed event ordered by $t \in 1, \ldots, T$ and T corresponds to the registered time for the patient's label assignment. Recall the visual example of the time-ordered sequence of EHR events in Fig. 1. Each event consists of three elements: a time stamp, an event category (e.g., blood pressure or medication code), and a value. The time for the registration of event e_t is denoted as $t(e_t)$, the category $c(e_t)$, and the value $v(e_t)$. For example, if the category $c(e_t)$ is blood pressure, then $v(e_t) \in \mathbb{R}^2$, as it contains both the systolic and diastolic measurements. Notice that only for the sequential neural network model, the detailed ordering of events is important.

The raw event data is transformed through a two-step vectorization of individual events. The first step will represent each event e_t by a very sparse vector e_t with an entry for all event-value types that can be observed across all patients. The size of this vector will be greater than the number of different event categories, as a category may have more than one measurement (e.g., for the blood pressure event from above, the event vector e_t will have two nonzero elements, one for each measurement in $v(e_t)$). In the second step, each vector entry is further transformed as follows. Categorical features are converted into their corresponding one-hot binary feature vector, numerical features are standard normalized, and hierarchical features, such as diagnosis codes, are represented as multi-hot vectors with an entry in each of the present levels of the diagnosis hierarchy. The resulting vectorization of a given event e_t is therefore a very sparse, but not necessarily one-hot, vector e_t of size 80,000.

A raw event vector sequence may be partitioned into intervals of time, where the raw event vectors are then aggregated within a time interval I. We will let e_I denote the interval aggregation of all e_t where $t(e_t) \in I$. Different aggregation functions are applied across different elements in the event vectors. Binary (categorical) outcomes, such as

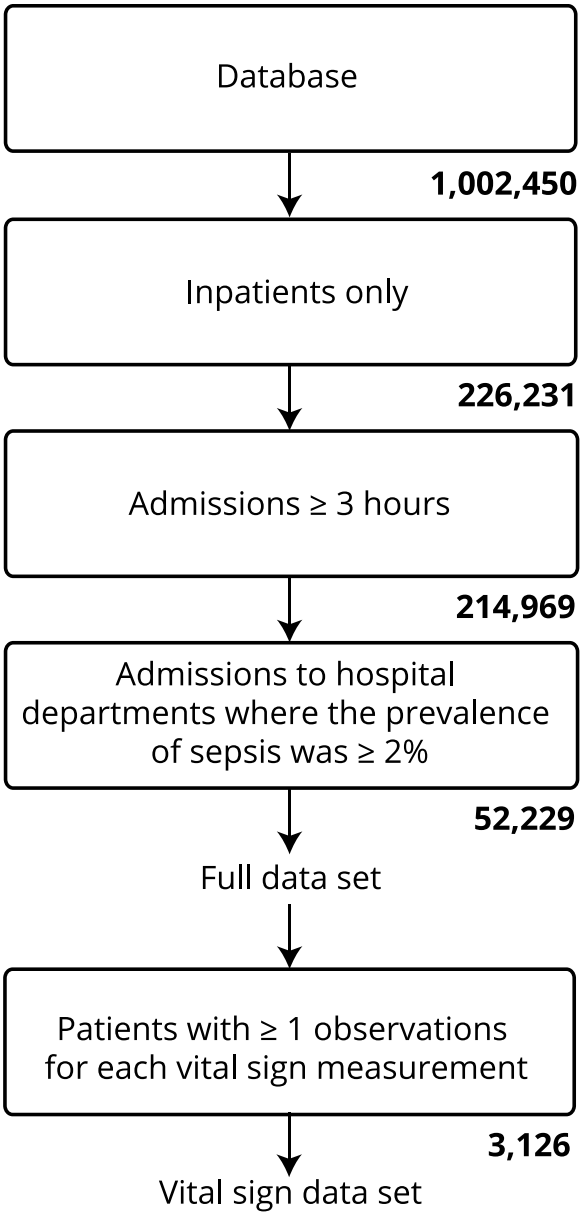

Fig. 2. Inclusion flow chart for the full and the vital sign data sets.

procedure codes, are aggregated to numerical counts, and numerical measurements, such as blood pressure, are converted to minimum, maximum, and mean values. Naturally, depending on the degree of aggregation, the ordering of the events in the entire event sequence is ignored to a greater or lesser extent.

Similar to the raw event vector, we also construct a vectorization of the context for a given patient. That is, the contextual meta data, such as demographics and the patient's comorbidities prior to the first raw event, are considered in a model. The context vector is denoted by c and is, in contrast to the raw event vectors e_t, not dependent on the sequence ordering in the model.

2.5. Data preprocessing and model design

The models in this study were built with an onset in three different approaches: (1) a classical epidemiological approach, where the model includes a small group of selected and clinically well-founded features; (2) a more data-driven approach, where all the available data is used in a slightly aggregated form to train a non-sequential neural network; and finally (3) a data-driven approach, where the available data is used in its sequenced form for the training of a sequential neural network. The first two models serve as baseline comparison models.

2.5.1. Gradient boosting

In the simplest baseline model, called GB-Vital, we replicate a well-known sepsis detection model from the literature, which has shown excellent results in a randomized study [12]. The full technical description of the model can be found in [10]. The explanatory features for this model are constructed by considering only six vital-sign events from the raw EHR event sequences: systolic blood pressure, diastolic blood pressure, heart rate, respiratory rate, peripheral capillary oxygen saturation, and temperature. The constructed features highly aggregate the sequence information, and only limited ordering information is retained. That is, for each of the six vital signs, five features are constructed to represent the average value for the current hour, the prior hour, and the hour prior to that hour, along with the trend value between two succeeding hours.

Based on these 30 features (five values from each of the six measurement channels), the GB-Vital model is constructed as a gradient boosted classifier of decision trees. As in [10], each tree in the gradient boosting model is limited to split at most six times, and no more than 1000 trees are aggregated to generate the risk prediction. The model was trained in Python using the Gradient Boosting Classifier in the Scikit-learn package.

2.5.2. Multilayer perceptron

In the more advanced baseline model, we constructed a standard multilayer feedforward neural network in the form of a multilayer perceptron (MLP). The model does not limit data to include only vital signs. Instead, features were constructed by aggregating entire event vectors in E across retrospective windows of time, including intervals of 1 h, 2 h, 4 h, 8 h, 16 h, and 32 h preceding the time of the label. Notice that with this coarse aggregation of events, the ordering of the events is basically ignored, except for the fact that the final feature vector for the model concatenates the aggregating event vectors in E across the different sized windows of time. Finally, to remove noise and reduce dimensionality, we only consider features that are present in at least 100 sample sequences of the training data, resulting in a reduction from approximately 100,000 to 5000 distinct entries in the feature vector for each retrospective timespan, or approximately 30,000 features in total.

The model structure for the MLP in this study feeds the 30,000 input units together with the 26 contextual features into two hidden layers of 200 units each, which is then followed by the binary decision. The model was trained to optimize the cross-entropy loss using the Adam optimizer [24] with mini-batches of size 50, a learning rate of 0.0001, and a dropout of 30% to prevent overfitting. Keras 2.2.2 with a TensorFlow 1.11 backend was used for the MLP experiments in this study.

2.5.3. CNN-LSTM

In this model, we considered all elements in the entire event sequence E for a patient. As with the MLP model, we only considered events that were present in at least 100 sample sequences of the training data, resulting in approximately 5000 distinct entries in the event-vector representation for an event e_t.

The event sequences are further pre-processed by (1) a temporal preserving aggregation step, (2) a gap-filling step, and (3) a context concatenation step. In the temporal preserving aggregation step, all event vectors are grouped into five-minute non-overlapping blocks $B \subseteq E$, such that the maximum time between two events in each block is five minutes. The main reason for this step is to reduce the number of inputs to the model in order to improve computational efficiency. However, the temporal aggregation also has the effect of discarding the order of events within each five-minute block, which arguably better complies with the randomness in order that healthcare professionals may introduce when entering information that close in time into the

EHR. The gap-filling step now fills the sequence with empty event vectors such that all feature vectors in the sequence are equidistant in time. In this way, there is a vector for every five minutes in the sequence, some of which are aggregations of one or more event vectors and some are empty vectors. The sequence can therefore be represented as a sparse matrix of shape $N \times K$, where N is the number of five-minute vectors in the longest sequence ($N = \left[\frac{t(e_T) - t(e_1)}{5\,\text{min}}\right]$) and K is the number of entries in the event feature vectors ($K \approx 5000$). Finally, each of the N aggregated event vectors was concatenated with the same fixed context vector such that the final sequence matrix is of shape $N \times (K + C)$, where C is the number of entries in the context vector ($C \approx 30$). Recall that the context vector contains meta data about the patient, such as age, gender, and comorbidities. The intuition for concatenating the contextual data with every event is that the importance of certain EHR registrations may be supported by such contextual information.

The classifier is structured as a convolutional neural network (CNN), followed by a recurrent layer of long short-term memory (LSTM) cells, also known as a CNN-LSTM model (or sometimes Long-term Recurrent Convolutional Network) [25]. This architecture has been shown to learn robust temporal feature representations in the convolutional layers, which makes it easier for the LSTM layer to capture temporal dependencies compared to using the raw inputs [26]. The overall architecture of the classification model is illustrated in Fig. 3. The model first projects the sparse inputs into dense 1000-dimensional vectors, reducing the dimensionality for the following convolutional layer by a factor of five. With inspiration from Conneau et al. [27], short-term temporal developments for a patient are now captured in the model by a stack of "convolutional blocks". A convolutional block consists of two one-dimensional ReLU-activated convolutional layers followed by a max-pooling layer. All convolutional layers have kernels of size 3, a stride of 1, and zero-padding is used. All max-pooling layers have a kernel size of 2 and a stride of 2, halving the temporal width of the input. To ensure that information across the convolutional blocks obeys the ordering of the input information, without contaminating the output with information from the future, all kernels are causal in the sense that they only filter input from the current time and the past.

There are five convolutional blocks in the model. The initial block has a depth of 128 for both of the convolutional layers in the block, whereas the convolutional layers in the last four blocks all have a depth of 64. After the input filters through the five convolutional blocks, the output vectors contain partly overlapping temporal information, where each vector spans 15 h and 30 min of the original input, and the temporal distance between two succeeding vectors is 2 h and 40 min. Finally, the model captures the long-term temporal development of a patient by allowing the output from the convolutional blocks to feed into an LSTM layer that incrementally builds up a representation of the temporal inputs and continually predicts an output. The LSTM layer has 64 units and is initialized with a random initial state. This layer consists of a "conventional" LSTM layer with a forget gate, as defined in [28]. Our experiments have shown that by adding the convolutional layers in front of the LSTM, we gain significant improvements in both efficiency and effectiveness compared to using a single stacked LSTM. The final prediction layer is a softmax layer.

The model was implemented in Keras 2.2.2 with a TensorFlow 1.11 backend and trained on a NVIDIA Tesla V100 GPU. Convergence was reached after approximately 90 min for experiments with this model.

2.6. Model evaluation

The models in this study produce prediction values that reflect the risk that a patient's hospital admission may result in sepsis, if not intervened upon. The predicted risk will be in the range from zero to one and should be higher for those patients at risk of later developing sepsis compared to those that are not. It is customary to evaluate the discriminative power of a binary decision model at a range of thresholds $p_\tau \in [0; 1]$ for the decision $p > p_\tau$ and then report results in the form of receiver operating characteristic (ROC) curves, precision-recall (PR) curves, area under ROC (AUROC), or mean average precision (mAP). We report these well-known measures to enable easy comparison to existing and future studies that employ evaluations of this kind.

However, while discrimination is an important statistical property, it does not properly address clinical usefulness [15–17,29–31]. For example, if a false negative decision causes greater harm than a false positive decision, a model with high sensitivity may be preferable to a model with high specificity and lower sensitivity, although the latter model might have, say, a higher AUROC. In general terms, a model is clinically useful if the use of its decisions for patients leads to a better ratio between benefits and harms than not using the model. Grounded in the utility measure from the field of decision theory, decision curve analysis (DCA) assesses the clinical usefulness of a prediction model by evaluating the so-called net benefit at varying decision thresholds for the model (see, e.g., [32,33]). Let TP and FP denote the number of, respectively, true positives and false positives that a model predicts from a sample of N cases. The net benefit, NB is now defined as

$$\text{NB} = \frac{\text{TP} - (\text{FP} \cdot \omega)}{N}, \tag{1}$$

where the weighting factor, ω, can be interpreted as the exchange ratio between the number of false positives that is acceptable in exchange for one true positive. This interpretation is important, because it is informative of how the clinician weights the harm H of a false sepsis-positive decision over the benefit B of a true decision, with the rationale being to start intervention for a patient if the expected harm compared to the benefit is above the clinician's preference of exchange ratio H/B. For example, a weighting factor of $\frac{1}{10}$ indicates that if the clinician misses one sepsis patient that could have been detected by the model, it is valued as being 10 times worse than unnecessarily classifying one

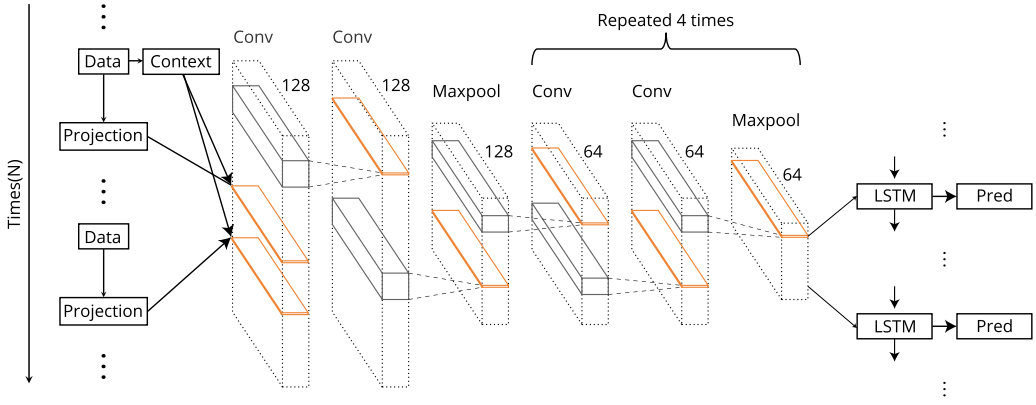

Fig. 3. The CNN-LSTM architecture for sepsis classification.

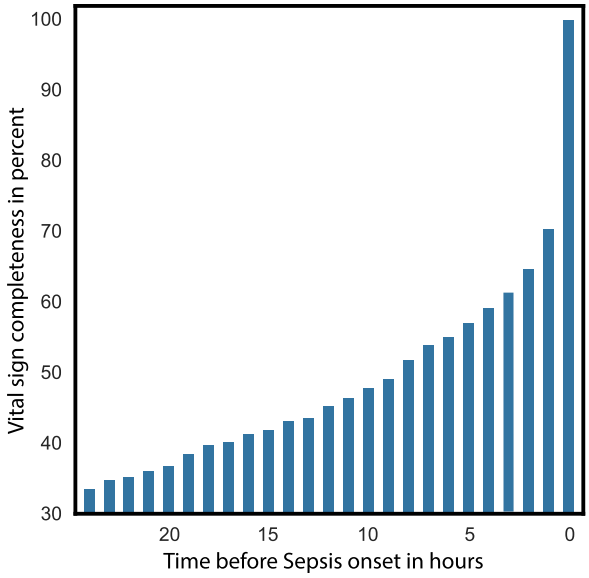

Fig. 4. Percentage of patients with two or more vital sign measurements (vital sign completeness) as a function of time before sepsis onset.

healthy person to be at risk of sepsis. Finally, by rearranging the expression for the expected utility of an intervention, one can establish the following relation between the harm/benefit exchange ratio and an operational model's decision threshold.

$$\omega = \frac{H}{B} = \frac{p_\tau}{1 - p_\tau} \tag{2}$$

The harm/benefit exchange ratio is subjective and will vary across clinicians. A decision curve in DCA illustrates the consequence of an arbitrary choice by evaluating the net benefit for the binary decision of opting into the intervention ($p > p_\tau$), or not ($p \leq p_\tau$), across a range of different decision thresholds – or equivalently, for a range of different harm-benefit exchange ratios. With different models to consider – including the extreme options of never intervening or always intervening – the clinician should favor the model with the highest net benefit at his personally determined H/B ratio. The curves also allow the clinician to alter this ratio in the context of a given patient (e.g., in accordance with the patient's preferences). See Fig. 6c for an example of DCA curves.

The attentive reader may have noticed that the above formulation of DCA exclusively focuses on the patients for whom an intervention will occur, as is the case in [32]. For the remaining patients, their hospital contact will continue as usual and not be affected by the prediction model. This is a more conservative evaluation than the formulation of DCA in [33], where harms and benefits for non-interventions are also included. However, the latter DCA formulation is more demanding on the elicitation of the weighting factor to be used in an actual clinical setting, as it now relies on the clinician's ability to state a four-way relation between harms and benefits. In the following, we will use the former definition of net benefit in our DCA reporting.

2.6.1. Sequence evaluation with a retrospective assessment of intervention potential (SERAIP)

Concerning the clinical usefulness of a prediction model, it is important to account for earlier related interventions, if any, when evaluating the effect of the model. The model will not create additional value to the clinician (and patient) if interventions are already initiated at the time of prediction. In this case, the prediction cannot lead to any new action. There are two aspects to consider when accounting for earlier interventions in the performance measure for a prediction model: (1) An intervention may have been caused by the model at an earlier timestep, or (2) an intervention may have been caused by clinical presumptions, independently of the model.

Concerning the first aspect, a model may perform very well when evaluated close to the onset of sepsis, but at the same time perform critically bad when evaluated many hours before severe sepsis symptoms occur. When the model is used in a real-time clinical setting, the effect of a positive prediction will be an intervention that cannot be withdrawn. It implies that model decisions about interventions in earlier timesteps must be carried through and accounted for when evaluating model performance in subsequent timesteps. We address the effect of past model performance by defining the sequence prediction at time t, p_t^{seq} as the maximum probability of all predictions until then. That is:

$$p_t^{\text{seq}} = \max_{0 \leq x \leq t}(p_x), \tag{3}$$

where p_x is the prediction at time x. In this way, a sepsis-positive classification will be maintained for the subsequent timesteps, as the effect of a positive prediction will be an intervention that cannot be withdrawn.

Concerning the second aspect, we suggest adding a retrospective assessment of interventions to the evaluation by looking for intravenous antibiotics and blood cultures preceding the prediction time. Here, intravenous antibiotics are identified as intravenous medications belonging to either the ATC J01 (antibacterial agents for systemic use) or ATC J02 (antimycobacterial agents) subgroups, and blood cultures are identified through the laboratory system for microbiology. We include all registrations back to 72 h before the prediction time to ensure that all registrations related to clinical presumptions on sepsis are captured. For a given timestep this assessment is performed for all true positive predictions TP, yielding a partition of TP into those with none, one of, or both the two types of interventions; intervenous antibiotics and blood culture. We are only interested in the predictions of sepsis ($p_t^{\text{seq}} > p_\tau$) that do not already have a sepsis-related intervention. That is, the partition of TP without interventions, as these indicate the non-contestable potential for early intervention. We denote this evaluation metric as *sequence evaluation with a retrospective assessment of intervention potential* (SERAIP)

3. Results

3.1. Vital sign data coverage

Fig. 4 shows, in percentage, how many patients from our multicenter dataset have had two or more vital sign registrations at each hour prior to a sepsis onset. We see a dramatic decrease in the measurements, as we move back in time. Let t denote the time of sepsis onset. Already at $t - 3$ h before the onset, only 62% of all patients have two or more vital sign measurements, and moving further back in time to $t - 10$ and $t - 24$ h, the percentages are reduced to 48% and 33%, respectively. This observation affirms the large reduction from 52,229 to 3,126 contacts that the vital sign inclusion criterion (Section 2.2) implies on the full data set.

3.2. Gradient boosting

Fig. 5a and 5b show the ROC and PR curves from evaluating the GB-Vital model on the vital sign test data. The model achieved an AUROC of 0.786 and a mAP (mean average precision) of 0.797 when evaluated 3 h before sepsis. The results from the DCA are shown in Fig. 5c. The NB of using the GB-Vital model was equal to the NB of treating all patients in the range of probability thresholds from 0% to 32%. At thresholds above 32%, the NB of using the GB-Vital model exceeded both the NB of treating none patients and the NB of treating all patients.

3.3. Multilayer perceptron

The MLP model achieved an AUROC of 0.764 and a mAP of 0.689 when evaluated 3 h before sepsis on the vital sign data set (Fig. 5a and

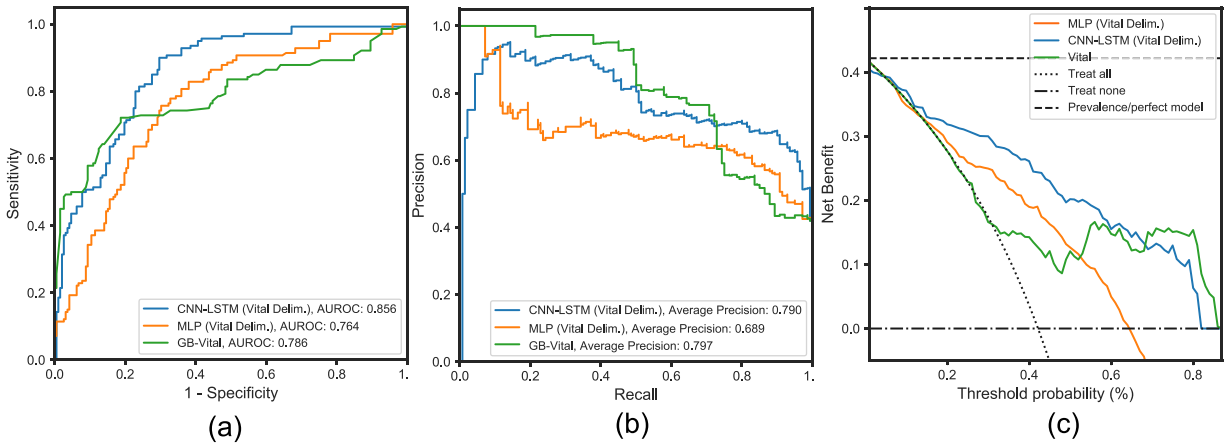

Fig. 5. Results from the vital sign test data set evaluated 3 h before sepsis onset: (a) ROC curves; (b) PR curves; (c) DCA.

b). The NB of using the MLP model was equal to the NB of treating all patients in the range of probability thresholds from 0% to 20%. At threshold values above 45%, the NB of using the MLP model exceeded both the NB of treating no patients and the NB of treating all patients.

Results from the full data set are summarized in Fig. 6. Fig. 6a and b show how AUROC and mAP change as a function of time before the labeled onset of sepsis (or not sepsis). The MLP AUROC scores on the full data set were as follows: $t - 15$ min: 0.872; $t - 3$ h: 0.871; $t - 10$ h: 0.751; and $t - 24$ h: 0.619. The highest mAP of 0.578 was achieved at 3 h. The mAP dropped on both sides of this peak (0.395 at $t - 15$ min and 0.318 at $t - 10$ h) and further decreased to 0.147 at $t - 24$ h.

The NB of using the MLP model on the full data set was slightly

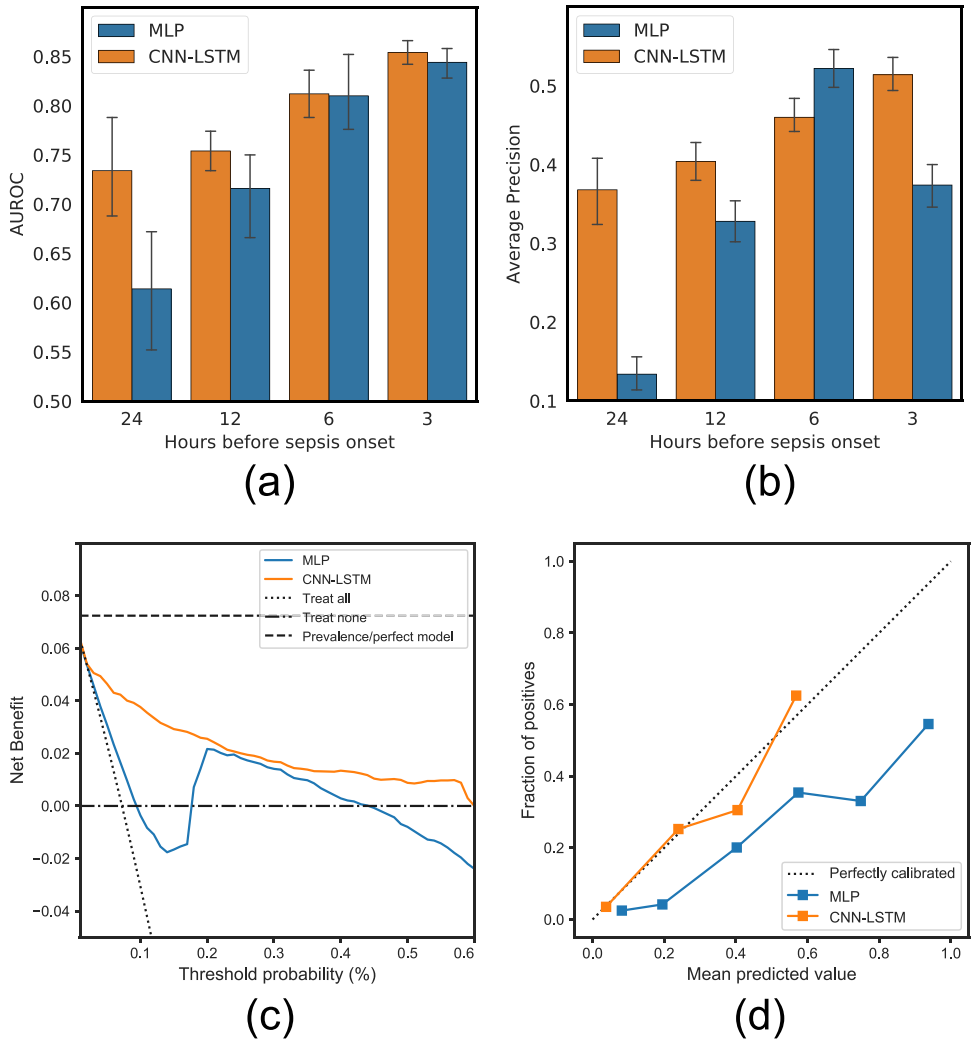

Fig. 6. Results from the full test data set: (a) AUROC at different predictions times with 95% confidence intervals. (b) mAP at different prediction times with 95 % confidence intervals c) DCA 3 h before sepsis onset. (d) Calibration curve. (For interpretation of the references to color in the text, the reader is referred to the web version of this article.)

higher than the NB of treating all patients in the range of probability thresholds from 0% to 12%. In the range from 12% to 20% and above 45%, the NB was negative. In the range from 20% to 45%, the model exceeded both the NB of treating no patients and the NB of treating all patients (Fig. 6c).

In Fig. 6, the calibration curve for the MLP model is shown (blue line). The plot provides an indication of whether future predicted probabilities agree with the observed probabilities. For example, if we predict a 35% risk of developing sepsis, the observed frequency of sepsis should be approximately 35 out of 100 patients with such a prediction. A perfectly calibrated model would have a 45 degree line along the diagonal [34,35].

3.4. CNN-LSTM

The CNN-LSTM model achieved an AUROC of 0.856 and a mAP of 0.79 when evaluated 3 h before sepsis on the vital sign test data (Fig. 5a and b). Results from the full data set showed that the CNN-LSTM model decreased from a maximum mAP of 0.531 at $t - 15$ min to 0.407 at both $t - 10$ and $t - 24$ h. The CNN-LSTM AUROC scores were as follows: $t - 15$ min: 0.879; $t - 3$ h: 0.842; $t - 10$ h: 0.792; and $t - 24$ h: 0.752. The NB of using the CNN-LSTM model on the full dataset exceeded both the NB of treating no patients and the NB of treating all patients in the range of probabilities from 5% to 60% (Fig. 6c). The calibration of the CNN-LSTM model is shown in Fig. 6d (orange line).

3.5. SERAIP

Table 2 shows the results of *SERAIP*. The columns "TP with IV antibiotics", "TP with blood culture", "TP with IV antibiotics or blood culture" and "TP with no intervention" indicate the potential for initiating interventions that the clinicians have not already thought about at the time of prediction. Using the first row of the table as an example, the model finds 17% of the positives, corresponding to 39 patients. The column "TP with IV antibiotics" shows that 8 patients out of the 39 true positives were already started with intravenous antibiotics. The column "TP with blood culture" shows that 12 out of the 39 patients had already had a blood culture, and "TP with IV antibiotics or blood culture" shows that 8 of the 12 patients for whom intravenous antibiotics had been started had also had a blood culture. Finally, the last column "TP with no intervention" shows that 27 of the 39 patients had no intervention initiated at the point of prediction.

The column "FP/FP" indicates the relationship between false positives and true positives and shows that one can expect 9.28 false alarms for every one true positive.

4. Discussion

4.1. Results

We have presented an accurate deep learning system for early sepsis detection on a multi-center data set from outside ICUs. We have compared three different approaches for early detection of sepsis: a GB-Vital model, based on vital sign features; a non-sequential MLP model with thousands of features, including those used for the GB-Vital model; and a sequential CNN-LSTM model with an equal number of features.

The GB-Vital model had reasonable performance, with an AUROC of 0.786 3 h before sepsis onset for patients with registered vital signs, but it underperforms when compared to the results of previous studies on early sepsis detection. Qingqing Mao et al. achieved an AUROC of 0.88 [10] 3 h before sepsis onset with a similar GB-Vital model. Nemati et al. reported an AUROC of 0.85 4 h before sepsis onset with a Weilbull–Cox proportional hazards model. Futoma et al. reported AUROCs of 0.86 and 0.78 3 h and 12 h before sepsis onset, respectively, with a multi-

Table 2
Results from the "sequence evaluation with retrospective assessment of intervention potential" on the full test data set. Area under the receiver operating characteristics (AUROC), false positive (FP), true negative (TN), false negative (FN), false positive (FP), intravenous (IV), Hospital (Hosp.), gastrointestinal (gastroin.), specificity (SPE) sensitivity (SEN), TP with IV antibiotics (TP anti), TP with blood culture (TP blood), TP with IV antibiotics or blood culture (TP int.), TP with no intervention (TP no int.).

Department/Hospital	Evaluated up until	SEN	SPE	FP/TP	TP	TN	FN	FP	TP anti	TP blood	TP int	TP no int.
Emergency Dept. Hosp. 1	$t - 3$ h	0.17	0.91	9.28	39	3663	197	362	8	12	12	27
	$t - 10$ h	0.13	0.91	14.77	22	3456	151	325	4	7	7	15
	$t - 24$ h	0.11	0.90	18.76	17	2899	132	319	1	2	2	15
Dept. of Oncology Hosp. 2	$t - 3$ h	0.31	0.93	7.25	4	389	9	29	3	1	3	1
	$t - 10$ h	0.40	0.93	7.25	4	372	6	29	2	1	3	1
	$t - 24$ h	0.30	0.93	9.33	3	364	7	28	2	0	2	1
Joint Emergency Dept. Hosp. 1	$t - 3$ h	0.33	0.87	3.00	4	78	8	12	0	1	1	3
	$t - 10$ h	0.09	0.10	1.50	6	1	60	9	0	1	1	5
	$t - 24$ h	0.09	0.14	1.20	5	1	50	6	0	0	0	5
Emergency Dept. Hosp. 3	$t - 3$ h	1.00	0.92	5.00	1	55	0	5	0	0	0	1
	$t - 10$ h	1.00	0.86	5.00	1	31	0	5	0	0	0	1
	$t - 24$ h	1.00	0.85	5.00	1	28	0	5	0	0	0	1
Dept. of Anaesthesiology Hosp. 1	$t - 3$ h	0.60	0.66	1.83	6	21	4	11	3	2	3	3
	$t - 10$ h	0.45	0.59	2.20	5	16	6	11	1	1	1	4
	$t - 24$ h	0.56	0.60	2.00	5	15	4	10	0	1	1	4
Dept. of Hematology Hosp. 2	$t - 3$ h	0.36	0.93	5.80	5	398	9	29	3	1	3	2
	$t - 10$ h	0.45	0.93	5.80	5	372	6	29	2	1	3	2
	$t - 24$ h	0.30	0.93	9.33	3	364	7	28	2	0	2	1
Dept. of gastroin. surgery Hosp. 2	$t - 3$ h	0.67	0.63	2.75	4	19	2	11	1	0	1	3
	$t - 10$ h	1.00	0.62	2.75	4	18	0	11	1	0	1	3
	$t - 24$ h	1.00	0.63	2.50	4	17	0	10	1	0	1	3
Dept. of Anaesthesiology Hosp. 2	$t - 3$ h	0.33	0.95	1.00	1	19	2	1	0	0	0	1
	$t - 10$ h	0.50	0.95	1.00	1	19	1	1	0	0	0	1
	$t - 24$ h	0.50	0.95	1.00	1	18	1	1	0	0	0	1

output Gaussian processes model [7]. The reason for the lower AUROC values in our GB-Vital model is likely due to the amount of missing values in our data set. All of the above studies are built solely on data from ICUs, where vital parameters are recorded frequently. Recall that in our diverse data set, only 62% of the sepsis patients had at least two vital signs measured 3 h before sepsis onset. Qingqing Mao et al. examined the direct impact of missing values in their GB-Vital model and found that AUROC decreased from 0.9 to 0.79 when increasing the percentage of missing values from 0% to 20%. Increasing the percentage of missing values even further to 60% yielded an AUROC of 0.75. The reported AUROC of 0.79 with a missing value rate of 20% is directly comparable to our GB-Vital model, which archived an AUROC of 0.786 3 h before sepsis with a similar rate of missing values. It is important to note that when Qingqing Mao et al. created a data set reminiscent of ours, our results correlate.

These numbers indicate that although the GB-Vital model performs well on ICU data, it may not be useful for the early detection of sepsis at a broader scale, where vital parameters are not recorded as frequently across various hospital departments.

In Fig. 6a, it can be seen that the MLP and CNN-LSTM models had close to equal AUROC performance at time $t - 3$, and in fact, the MLP model had a better average precision (Fig. 6b) than the CNN-LSTM model. This was probably because the DNN model was trained on data 3 h before sepsis onset. In contrast, the CNN-LSTM model appeared to be more stable when used at different times relative to sepsis onset, which may be attributed to the sequential modeling approach.

The CNN-LSTM had higher NB values in the DCA compared to the MLP model for the full range of threshold values (Fig. 6c). In addition, a slightly odd NB profile could be observed for the MLP model in the threshold range from 0.05 to 0.2, indicating that the model was not well calibrated in this area and therefore would serve poorly as a risk-estimation model. This was investigated with a calibration plot, as shown in Fig. 6d. The plot supported our presumption that the MLP model was poorly calibrated, as the observed frequency of sepsis was systematically higher than the predicted risk of developing sepsis, especially in the probability ranges of 0.05–0.2 and 0.6–0.8. The CNN-LSTM model did not seem to suffer from poor calibration.

4.2. SERAIP

The SERAIP is our attempt to create an “close to the clinic” evaluation yielding an accurate picture of how the algorithm could support the clinical work at different departments. SERAIP can be considered an extension of the real-time validation suggested by Futoma et al. [7].

We simulated real-world usage by doing a retrospective evaluation, investigating two of the most important actions that a sepsis detection model could help initiate. Intravenous antibiotics and blood culture requisitions has been analyzed in the period preceding predictions, allowing for better estimates of the clinical utility of the model. The numbers for sensitivity and specificity in Table 2 were calculated using a global probability threshold of 0.1, which was determined from inspection of ROC and DCA. Optimally, a threshold should have been chosen per department, as the patient case mix varies greatly.

Looking at the Emergency department, Hospital 1 the model had a sensitivity of 0.17 3 h before sepsis, which corresponded to finding 39 true positives, of which 27 had not received any intervention. Conversely, as many as 362 false positives (at a specificity of 0.91) must be accepted at a threshold of 0.1. The hematology department (Hospital 2) is an example of a department with a completely different patient clientele than the emergency department. Here the model had a sensitivity of 0.45 and a specificity of 0.93 10 h before sepsis onset, which corresponded to 5 true positives and 29 false positives. At the same time, no interventions were initiated for two of the five true positives. An important observation from the evaluation was that the model detected a very high proportion of sepsis patients in departments in which sepsis is not common. This was probably because septic patients differ more from the usual clientele than in, for example, emergency departments.

4.3. Limitations

4.3.1. Black box

An important improvement in relation to clinical acceptance would be to implement supporting explanation methods in the predictions, such as layer-wise relevance propagation, deep Taylor decomposition, pattern attribution, or other DL explanation approaches [36,37]. It is easy to imagine that a model that is more interpretable and supported by explanations would be more easily accepted in the clinic. Shickel et al. reached the same conclusion in a recent article reviewing the latest trends in the use of DL on EHR data [38]. They completed their review with a warning against downplaying the importance of interpretability in favor of improvements in model performance.

4.3.2. Bias and confounding

As the presented DL models (MLP and CNN-LSTM) operate in a high dimensional feature space not limited by domain specialists, it is important to consider the associated bias issues. In June 2018, Benjamin Recht and colleagues from UC Berkeley argued that many DL models may be less generalizable than we have assumed [39]. This claim was supported a month later by Zech et al., who showed that their DL models were significantly influenced by organizational and process-oriented elements [40]. Agniel et al. highlighted similar problems in a study on EHR data [41]. The authors found that data regarding the time when the blood samples were ordered were more important than the blood test results for predicting three-year survival [5]. The important message in relation to sepsis detection based on EHR data is twofold: (1) If doctors or nurses have not measured certain vital signs or ordered certain blood samples, it will not be possible for models such as GB-Vital to predict sepsis. In this case, the CNN-LSTM model could still be used to estimate whether the patient is developing sepsis. (2) On the other hand, the CNN-LSTM model will most likely contain an unfortunate bias, which may be important if process-oriented elements change, such as new IT-systems or workflows.

4.3.3. Reproducibility

We do not test our models on the MIMIC-III database as the two cohorts comprise completely different patient groups. The CROSS-TRACKS database embraces a mixed rural and urban multi center population in contrast to critical care unit at a large tertiary care hospital in the MIMIC-III database. In the CROSS-TRACKS database, patients may be hospitalized without being critically ill, unlike MIMIC-III database. If a patient in the MIMIC-III database do not have sepsis they are most likely critically ill with another condition. The in-hospital mortality is almost 15 times higher in the MIMIC-III database compared to CROSS-TRACKS and similarly, the median length of stay, and average laboratory measurements are both ten times higher among MIMIC-III patients (see Table 3). In addition, the CROSS-TRACKS database contains data about comorbidities from nationwide registers, which, to the best of our knowledge, is the not the case for MIMIC-III. Despite these differences, and not being the scope of this article, it would be very

Table 3
Comparison between the CROSS-TRACKS patient population and the MIMIC-III patient population.

	CROSS-TRACKS	MIMIC-III
Distinct patients, no.	226.320	38.597
Age, median years	55.2	65.8
Gender, male, % of total admissions	47.6	55.9
Length of stays, median hours	16	165.6
Laboratory measurements, average per admission	37	380
Hospital motality, % of total admissions	0.8	11.5

interesting to test the generalizability of sepsis prediction models between these databases in a future study. One could imagine a setup where a model is trained on the CROSS-TRACKS database and tested it on MIMIC-III and vice versa.

4.3.4. Case–control matching

In this study, we exclusively sampled our negative cases from simple naive rules, such as age and contact length. This means that our data sets potentially contain many patients that our algorithm could easily categorize as negatives. An improved sampling technique would be to match sepsis-positive contacts with "similar" sepsis-negative contacts in a case control matching approach, as suggested in [7]. In that study, the authors implemented a propensity scoring mindset that seemed to be inspired by causal inference estimation theory.

4.3.5. Dataset construction and oversampling

In Section 2.2, we described how we oversampled the positive samples by a factor of 10 and then sampled negatives until we reached a ratio of 1:5. We explored many different combinations in relation to sampling techniques and balancing. Undersampling of the negative class worsened the test performance dramatically, indicating poor sampling of the variation space. Class ratios greater than 1:5 combined with weight-adjusted loss functions also reduced test performance, as did oversampling factors greater than ten. It could make sense to try more sophisticated sampling or data augmentation techniques to achieve better training performance.

4.3.6. Gold standard

In this study we used the 2001 consensus definition of sepsis as our gold standard. In our data the fulfillment of the 2001 consensus sepsis definition is an independent EHR event, that occurs continually during admission. This is in great contrast to other sepsis studies that needs to estimate the sepsis onset time retrospectively. Based on this unique EHR registration in our dataset, we chose to conduct this study with the 2001 gold standard despite the fact that the definitions of sepsis and septic shock were revised in 2016 and the new The Third International Consensus Definitions for Sepsis and Septic Shock (Sepsis-3) was formed and published by Singer et al. In the paper by Singer et al. the 2001 consensus definition of sepsis, used in this study, was criticized for focused solely on inflammatory excess and validity of SIRS as a descriptor of sepsis pathobiology was challenged. Seymor et al. also concluded that SOFA was statistically greater than SIRS at predicting hospital mortality among ICU encounters with suspected infection. It would therefore be interesting to investigate in a future study how the model, proposed in this study, performs on the same data with new Sepsis-3 definition. [42,23]

5. Conclusion

In this multi-center retrospective study, we present a novel deep learning system for early detection of sepsis in the heterogeneous data set present outside ICUs. The system learns representations of the key factors and interactions from the raw event sequence data itself, without relying on a labor-intensive feature extraction process. Our study indicates that sequential deep learning models can be used to detect sepsis at a very early stage, and we find that our model outperforms strong baseline models, such as GB-Vital, which rely on specific data elements and therefore suffer from many missing values in our data set. We also propose a new retrospective evaluation technique for assessing the clinical utility of the model that accounts for both intravenous antibiotics and blood culture requisitions. The evaluation showed that a large proportion of sepsis patients had not initiated intravenous antibiotics or blood culture at the time of early detection, and thus the model could facilitate such interventions at an earlier point in time.

The study is conducted on retrospective observational data and the conclusions are therefore hypothesis-generating in nature. A new prospective confirmatory study is needed to the test whether the expected utility can be realized in the clinic.

Interesting directions for future work would be to add supporting explanation methods into the predictions to improve clinical acceptance, to test our models on the MIMIC-III database and finally to test the model in a prospective randomized trial.

Conflict of interest statement

The authors Simon Meyer Lauritsen, Mads Ellersgaard Kalør, Emil Lund Kongsgaard and Bo Thiesson are employed at Enversion A/S.

Acknowledgements

We acknowledge the steering committee for CROSS-TRACKS for access to data. For data acquisition, modeling, and validation, we thank the following: Julian Guldborg Birkemose, Christian Bang, Per Dahl Rasmussen, Mathias Abitz Boysen, Anne Olsvig Boilesen, Lars Mellergaard, and Jacob Høy Berthelsen. For help with deep learning modeling, we thank Mads Kristensen. We also thank the rest of the Enversion team for their support.

Funding: This work was supported by the Innovation Fund Denmark [case number 8053-00076B].

References

[1] Perner A, Lassen AT, Schierbeck J, Storgaard M, Reiter N, Benfield T. Sygdomsbyrde og definitioner af sepsis hos voksne. Ugeskrift for Laeger 2018;180(11):2–6. V09170685, ISSN: 0041-5782.

[2] Secretariat W. Improving the prevention, diagnosis and clinical management of sepsis. 2017http://apps.who.int/gb/ebwha/pdf_files/WHA70/A70_R7-en.pdf?ua=1.

[3] Murphy S, Xu J, Kochanek K. Deaths: final data for 2010. Natl Vital Stat Rep 2013;61:1–117.

[4] Angus DC, Linde-Zwirble WT, Lidicker J, Clermont G, Carcillo J, Pinsky MR. Epidemiology of severe sepsis in the united states: analysis of incidence, outcome, and associated costs of care. Crit Care Med 2001;29(7):1303–10. https://doi.org/10.1097/00003246-200107000-00002.

[5] Vincent J-L. The clinical challenge of sepsis identification and monitoring. PLOS Med 2016;13(5):e1002022. https://doi.org/10.1371/journal.pmed.1002022.

[6] Jones A, Shapiro N, Trzeciak S, Arnold R, Claremont H, Kline J. Lactate clearance vs central venous oxygen saturation as goals of early sepsis therapy: a randomized clinical trial. JAMA 2010;303:739–46. https://doi.org/10.1001/jama.2010.158.

[7] Futoma J, Hariharan S, Sendak M, Brajer N, Clement M, Bedoya A, et al. An improved multi-output gaussian process RNN with real-time validation for early sepsis detection. 2017. (arXiv preprint). arXiv:1708.05894.

[8] Horng S, Sontag DA, Halpern Y, Jernite Y, Shapiro NI, Nathanson LA. Creating an automated trigger for sepsis clinical decision support at emergency department triage using machine learning. PLOS ONE 2017;12(4):e0174708. https://doi.org/10.1371/journal.pone.0174708.

[9] Futoma J, Hariharan S, Heller K. Learning to detect sepsis with a multitask gaussian process RNN classifier. 2017. (arXiv preprint). arXiv:1706.04152.

[10] Mao Q, Jay M, Hoffman JL, Calvert J, Barton C, Shimabukuro D, et al. Multicentre validation of a sepsis prediction algorithm using only vital sign data in the emergency department general ward and ICU. Health Inform Res 2018;8(1):e017833. https://doi.org/10.1136/bmjopen-2017-017833.

[11] Nemati S, Holder A, Razmi F, Stanley MD, Clifford GD, Buchman TG. An interpretable machine learning model for accurate prediction of sepsis in the ICU. Crit Care Med 2018;46(4):547–53. https://doi.org/10.1097/ccm.0000000000002936.

[12] Shimabukuro DW, Barton CW, Feldman MD, Mataraso SJ, Das R. Effect of a machine learning-based severe sepsis prediction algorithm on patient survival and hospital length of stay: a randomised clinical trial. Crit Care 2017;4(1):e000234. https://doi.org/10.1136/bmjresp-2017-000234.

[13] Calvert J, Price D, Chettipally U, Barton C, Feldman M, Hoffman J, et al. A computational approach to early sepsis detection. Comput Biol Med 2016;74:69–73. https://doi.org/10.1016/j.compbiomed.2016.05.003.

[14] Islam MM, Nasrin T, Walther BA, Wu C-C, Yang H-C, Li Y-C. Prediction of sepsis patients using machine learning approach: a meta-analysis. Comput Methods Programs Biomed 2019;170:1–9. https://doi.org/10.1016/j.cmpb.2018.12.027.

[15] Halligan S, Altman DG, Mallett S. Disadvantages of using the area under the receiver operating characteristic curve to assess imaging tests: a discussion and proposal for an alternative approach. Eur Radiol 2015;25(4):932–9. https://doi.org/10.1007/s00330-014-3487-0.

[16] McGeechan K, Macaskill P, Irwig L, Bossuyt PM. An assessment of the relationship between clinical utility and predictive ability measures and the impact of mean risk in the population. BMC Med Res Methodol 2014;14(1). https://doi.org/10.1186/

1471-2288-14-86.
[17] Talluri R, Shete S. Using the weighted area under the net benefit curve for decision curve analysis. BMC Med Inform Decis Mak 2016;16(1):94. https://doi.org/10.1186/s12911-016-0336-x.
[18] Davis J, Goadrich M. The relationship between precision-recall and roc curves. Proceedings of the 23rd international conference on machine learning, ICML'06. New York, NY, USA: ACM; 2006. p. 233–40. https://doi.org/10.1145/1143844.1143874.
[19] LeCun Y, Bengio Y, Hinton G. Deep learning. Nature 2015;521(7553):436–44. https://doi.org/10.1038/nature14539.
[20] Schmidt M, Schmidt SAJ, Sandegaard JL, Ehrenstein V, Pedersen L, Sørensen HT. The Danish national patient registry: a review of content, data quality, and research potential. Clin Epidemiol 2015:449. https://doi.org/10.2147/clep.s91125.
[21] Pedersen CB. The Danish civil registration system. Scand J Public Health 2011;39(7_suppl):22–5. https://doi.org/10.1177/1403494810387965. pMID: 21775345.
[22] 2001SCCM/ESICM/ACCP/ATS/SIS international sepsis definitions conference. Intens Care Med 2003;29(4):530–8. https://doi.org/10.1007/s00134-003-1662-x.
[23] Singer M, Deutschman CS, Seymour CW, Shankar-Hari M, Annane D, Bauer M, et al. The third international consensus definitions for sepsis and septic shock (Sepsis-3). JAMA 2016;315(8):801–10. https://doi.org/10.1001/jama.2016.0287https://jamanetwork.com/journals/jama/articlepdf/2492881/jsc160002.pdf.
[24] Kingma DP, Ba J. Adam: a method for stochastic optimization. cite arxiv:1412.6980Comment: Published as a conference paper at the 3rd international conference for learning representations. 2014.
[25] Donahue J, Hendricks L, Rohrbach M, Venugopalan S, Guadarrama S, Saenko K, et al. Long-term recurrent convolutional networks for visual recognition and description. IEEE Tans Pattern Anal Mach Intell 2017;39:677–91. https://doi.org/10.1109/TPAMI.2016.2599174.
[26] Sainath TN, Vinyals O, Senior A, Sak H. Convolutional, long short-term memory, fully connected deep neural networks. 2015 IEEE international conference on acoustics, speech and signal processing (ICASSP) 2015. https://doi.org/10.1109/icassp.2015.7178838.
[27] Conneau A, Schwenk H, Barrault L, Lecun Y. Very deep convolutional networks for text classification. Proceedings of the 15th conference of the European chapter of the association for computational linguistics: volume 1, long papers 2017. https://doi.org/10.18653/v1/e17-1104.
[28] Graves A. Generating sequences with recurrent neural networks. 2013. (arXiv preprint). arXiv:1308.0850.
[29] ten Haaf K, Jeon J, Tammemägi MC, Han SS, Kong CY, Plevritis SK, et al. Risk prediction models for selection of lung cancer screening candidates: a retrospective validation study. PLOS Med 2017;14(4):e1002277. https://doi.org/10.1371/journal.pmed.1002277.
[30] Vickers AJ, Cronin AM. Everything you always wanted to know about evaluating prediction models (but were too afraid to ask). Urology 2010;76(6):1298–301. https://doi.org/10.1016/j.urology.2010.06.019.
[31] Steyerberg EW, Vergouwe Y. Towards better clinical prediction models: seven steps for development and an ABCD for validation. Eur Heart J 2014;35(29):1925–31. https://doi.org/10.1093/eurheartj/ehu207.
[32] Vickers AJ, Elkin EB. Decision curve analysis: a novel method for evaluating prediction models. Med Decis Mak 2006;26(6):565–74. https://doi.org/10.1177/027298906295361.
[33] Rousson V, Zumbrunn T. Decision curve analysis revisited: overall net benefit, relationships to ROC curve analysis, and application to case-control studies. BMC Med Inform Decis Mak 2011;11(1). https://doi.org/10.1186/1472-6947-11-45.
[34] H J, JD H, B B. The measurement of performance in probabilistic diagnosis. II. Trustworthiness of the exact values of the diagnostic probabilities. Methods Inf Med 1978;17:227–37.
[35] EW S, AJ V, NR C, T G, M G, N O, et al. Assessing the performance of prediction models: a framework for traditional and novel measures. Epidemiology 2010;21:128–38. https://doi.org/10.1097/EDE.0b013e3181c30fb2.
[36] Montavon G, Lapuschkin S, Binder A, Samek W, Müller K-R. Explaining nonlinear classification decisions with deep Taylor decomposition. Pattern Regconit 2017;65:211–22. https://doi.org/10.1016/j.patcog.2016.11.008.
[37] Binder A, Bach S, Montavon G, Müller K-R, Samek W. Layer-wise relevance propagation for deep neural network architectures. 2016. https://doi.org/10.1007/978-981-10-0557-2_87.
[38] Shickel B, Tighe P, Bihorac A, Rashidi P. Deep EHR: a survey of recent advances in deep learning techniques for electronic health record (EHR) analysis. 2017. (arXiv preprint). arXiv:1706.03446.
[39] Recht B, Roelofs R, Schmidt L, Shankar V. Do cifar-10 classifiers generalize to cifar-10? 2018arXiv abs/1806.00451.
[40] Zech JR, Badgeley MA, Liu M, Costa AB, Titano JJ, Oermann EK. Confounding variables can degrade generalization performance of radiological deep learning models. 2018. (arXiv preprint). arXiv:1807.00431.
[41] Agniel D, Kohane IS, Weber GM. Biases in electronic health record data due to processes within the healthcare system: retrospective observational study. BMJ 2018:k1479. https://doi.org/10.1136/bmj.k1479.
[42] Seymour CW, Liu VX, Iwashyna TJ, Brunkhorst FM, Rea TD, Scherag A, 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. https://doi.org/10.1001/jama.2016.0288https://jamanetwork.com/journals/jama/articlepdf/2492875/joi160006.pdf.